Optimal Transport for Transfer Learning Across Spaces

Quang Huy Tran
Under the supervision of Nicolas Courty, Karim Lounici and Rémi Flamary

[1]Institut de Recherche en Informatique et Systèmes Aléatoires - IRISA
Université Bretagne Sud

[2]Centre de Mathématiques Appliquées - CMAP
Ecole Polytechnique

PhD Defense, 16 May 2024, Palaiseau

Table of Contents

1. Introduction
2. Unbalanced CO-Optimal Transport
3. Fused Unbalanced Gromov-Wasserstein
4. Augmented Gromov-Wasserstein
5. Conclusion and Perspective

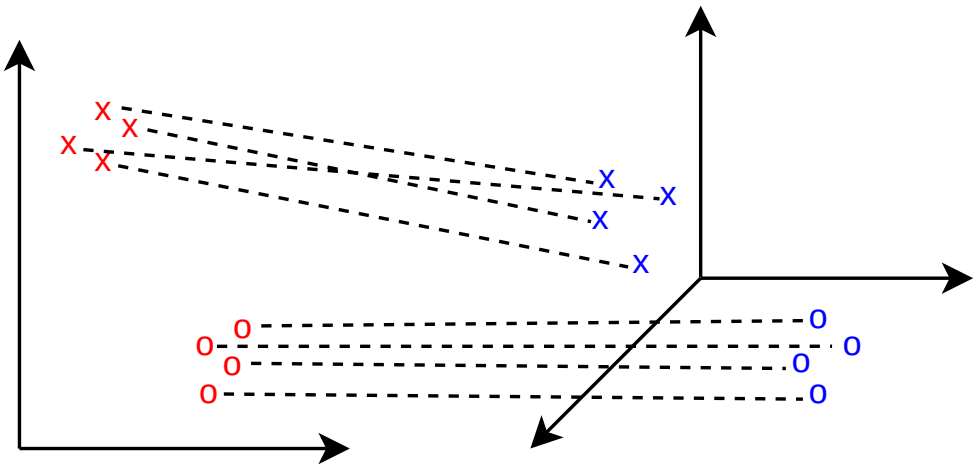
Across-space comparison

Gromov-Wasserstein distance (Mémoli 2007; Mémoli 2011)

Given two metric-measure spaces $\mathcal{X}_1 = (X_1, \mu_1, d_1)$ and $\mathcal{X}_2 = (X_2, \mu_2, d_2)$,

$$\mathrm{GW}_p^p(\mathcal{X}_1, \mathcal{X}_2) = \inf_{\pi \in U(\mu_1, \mu_2)} \iint \left| d_1(x_1, x_1') - d_2(x_2, x_2') \right|^p \mathrm{d}\pi(x_1, x_2) \mathrm{d}\pi(x_1', x_2')$$

1. Metric properties + Isometries.
2. Not the only way to compare incomparable spaces $\Rightarrow$ Gromov-Hausdorff distance.
3. Add something more

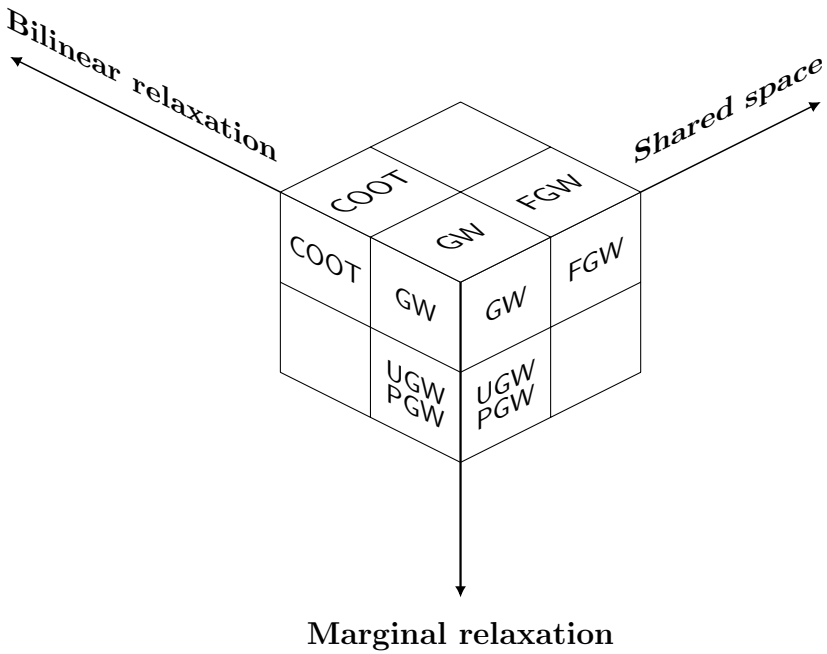


Extensions

1. Measure networks (Chowdhury and Mémoli 2019).
2. Sliced-GW (Vayer, Flamary, et al. 2019),
3. Low-rank GW (Scetbon, Peyré, and Cuturi 2021).

- Our focus

1. Bilinear relaxation.
2. Marginal relaxation.
3. Shared space.



Unbalanced Gromov-Wasserstein

Definition (Unbalanced GW (Séjourné, Vialard, and Peyré 2021))

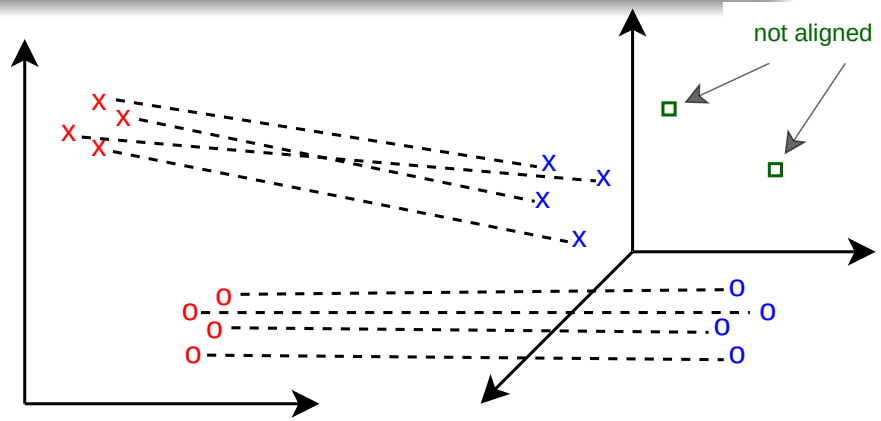
Given two compact metric-measure spaces $\mathcal{X}_1 = (X_1, \mu_1, d_1)$, $\mathcal{X}_2 = (X_2, \mu_2, d_2)$ and a Csiszár divergence D_φ,

$$\mathrm{UGW}_p^p(\mathcal{X}_1, \mathcal{X}_2) = \inf_{\pi \in \mathcal{M}^+(X_1 \times X_2)} L_{\mathrm{UGW}}(\pi),$$

where

$$L_{\mathrm{UGW}}(\pi) = \iint |d_1(x_1, x_2) - d_2(y_1, y_2)|^p \, \mathrm{d}\pi(x_1, y_1)\mathrm{d}\pi(x_2, y_2) + \rho_1 D_\varphi(\pi_{\#1} \otimes \pi_{\#1} | \mu_1 \otimes \mu_1) + \rho_2 D_\varphi(\pi_{\#2} \otimes \pi_{\#2} | \mu_2 \otimes \mu_2).$$

- In practice: D_φ = Kullback-Leibler div.
- Characterizing isometries.
- Practical robustness to outliers.
- Approx. with/without entropic reg.



Fused Gromov-Wasserstein

Definition (Fused GW (Vayer, Chapel, et al. 2019))

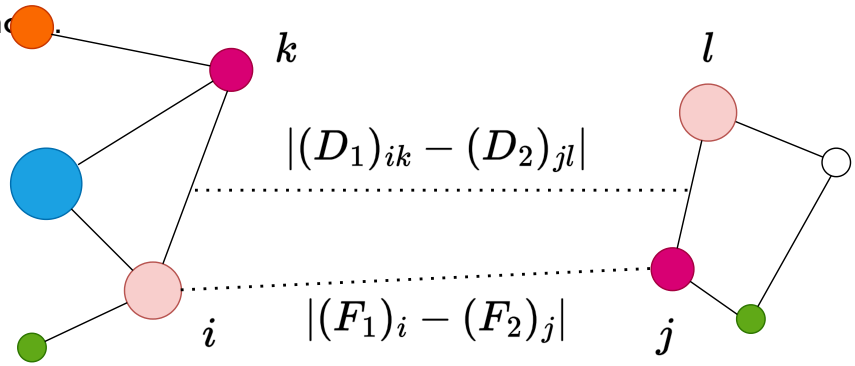
Consider two attributed graphs $\mathcal{X}_1 = (D_1, F_1, \mu_1)$ and $\mathcal{X}_2 = (D_2, F_2, \mu_2)$, where $D_k \in \mathbb{R}^{n_k \times n_k}$, $F_k \in \mathbb{R}^{n_k \times d}$ and $\mu_k \in \Delta_{n_k}$. For $\alpha \in [0, 1]$,

$$\mathrm{FGW}_p^p(\mathcal{X}_1, \mathcal{X}_2) = \inf_{\pi \in U(\mu_1, \mu_2)} L_{FGW}(\pi),$$

where

$$L_{FGW}(\pi) = (1-\alpha) \sum_{i,j,k,l} |(D_1)_{ik} - (D_2)_{jl}|^p \pi_{ij} \pi_{kl} + \alpha \sum_{ij} ||(F_1)_i - (F_2)_j||^q \pi_{ij}.$$

- Interpolation between GW and Wasserstein distance.
- Feature-preserving isometries.



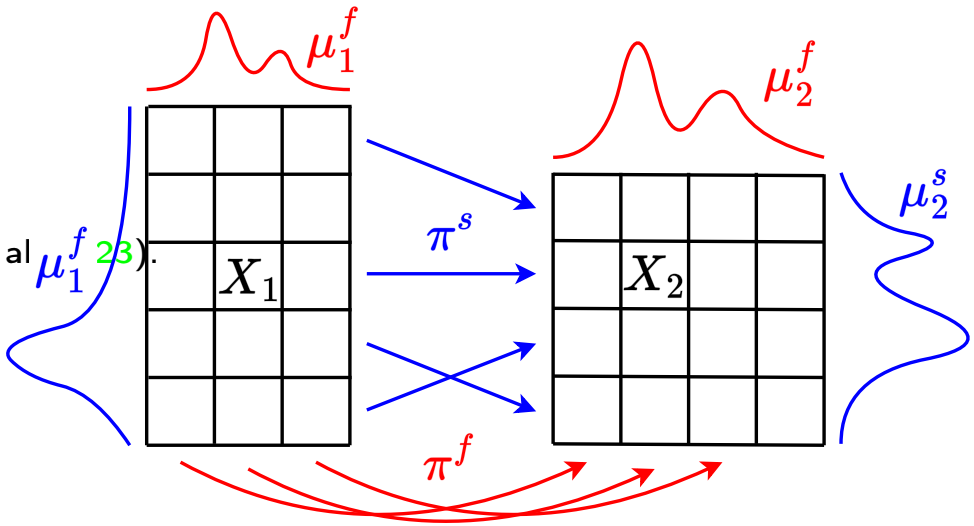
Co-Optimal Transport

Definition (Discrete COOT (Redko et al. 2020))

Given two weighted matrices $\mathcal{X}_1 = (X_1, \mu_1^s, \mu_1^f)$ and $\mathcal{X}_2 = (X_2, \mu_2^s, \mu_2^f)$, where $X_k \in \mathbb{R}^{n_k \times d_k}$ and histograms $\mu_k^s \in \Delta_{n_k}$, $\mu_k^f \in \Delta_{d_k}$

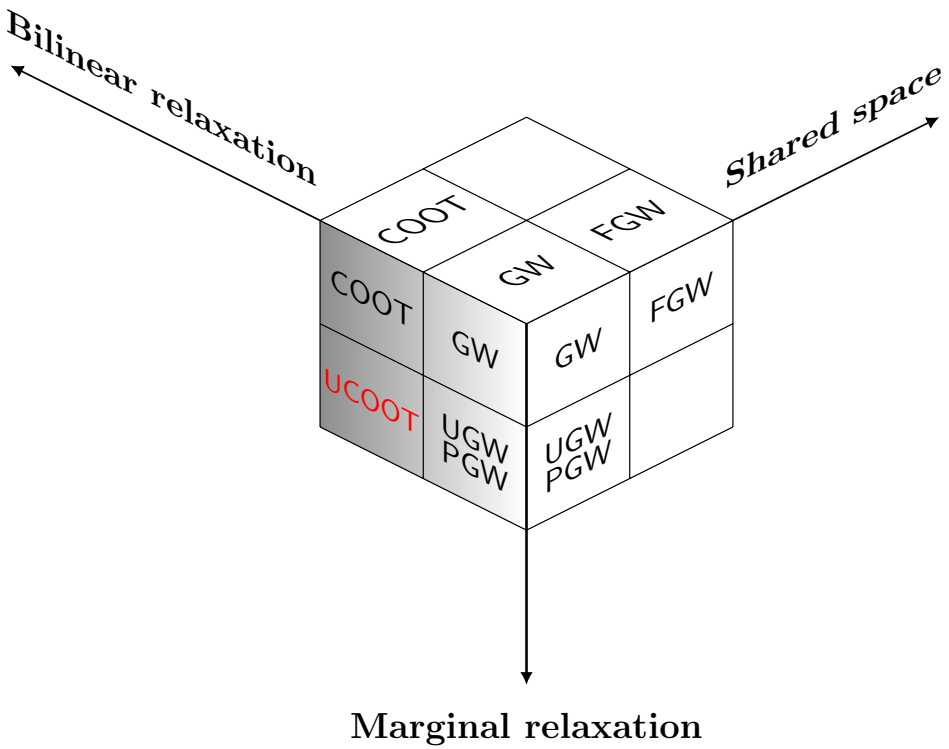
$$\text{COOT}_p^p(\mathcal{X}_1, \mathcal{X}_2) = \min_{\substack{\pi^s \in U(\mu_1^s, \mu_2^s) \\ \pi^f \in U(\mu_1^f, \mu_2^f)}} \sum_{i,j,k,l} |(X_1)_{ik} - (X_2)_{jl}|^p \pi^s_{ij} \pi^f_{kl}$$

- Comparing arbitrary-size matrices.
- Meaningful feature coupling π^f.
- Metric properties.
- Connection with GW distance.
- Continuous extension (Chowdhury, Needham, et al 23).

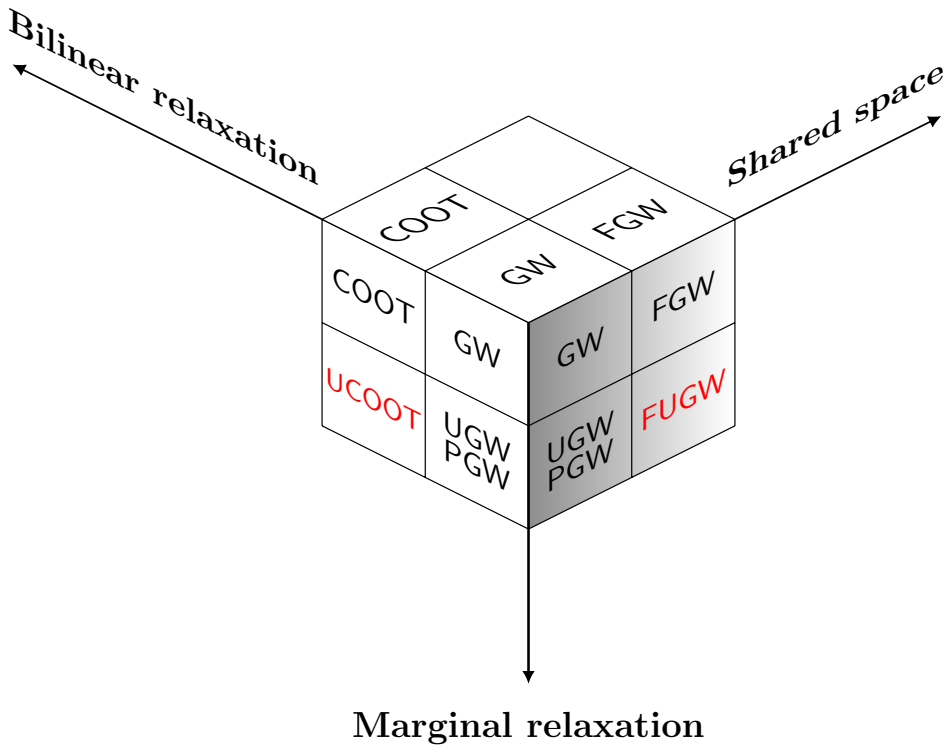


Summary of contributions

Publication: Unbalanced Co-Optimal Transport. **QHT**, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci and Ritambhara Singh. *AAAI Conference on Artificial Intelligence*, 2023.



Summary of contributions



Summary of contributions

Publication:

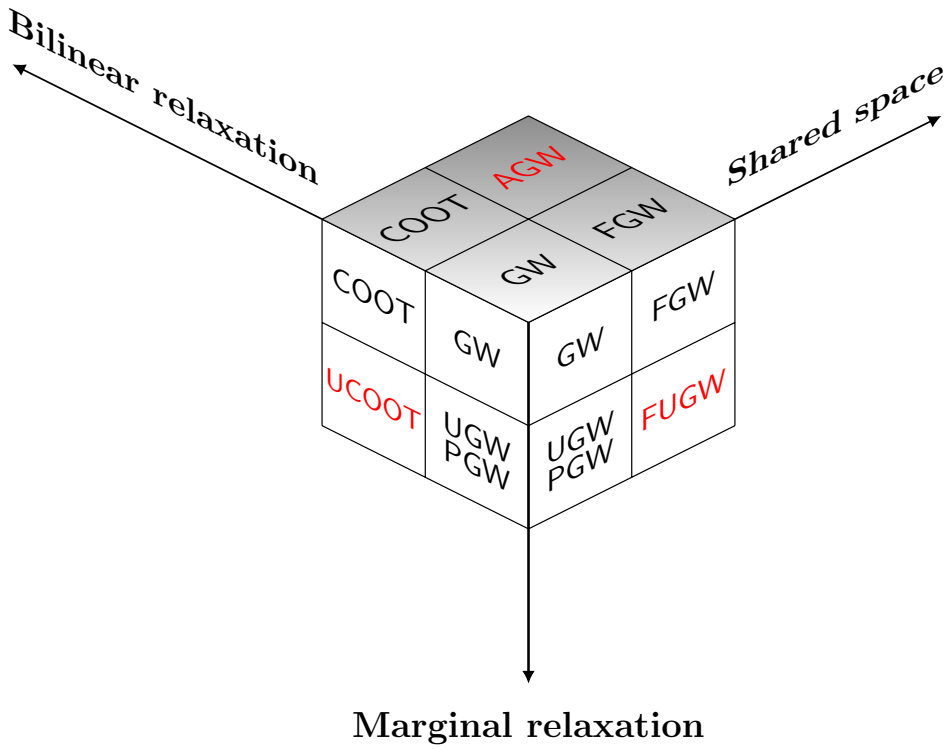
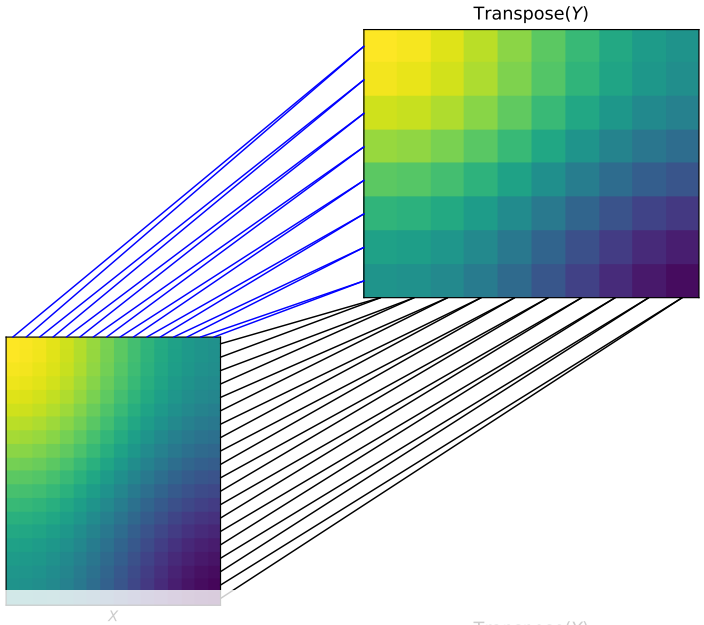
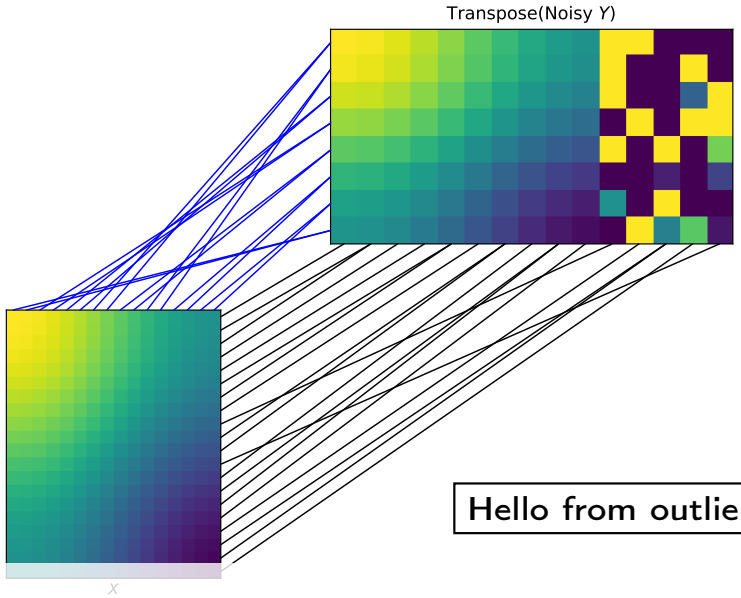


Table of Contents

1. Introduction
2. Unbalanced CO-Optimal Transport
3. Fused Unbalanced Gromov-Wasserstein
4. Augmented Gromov-Wasserstein
5. Conclusion and Perspective

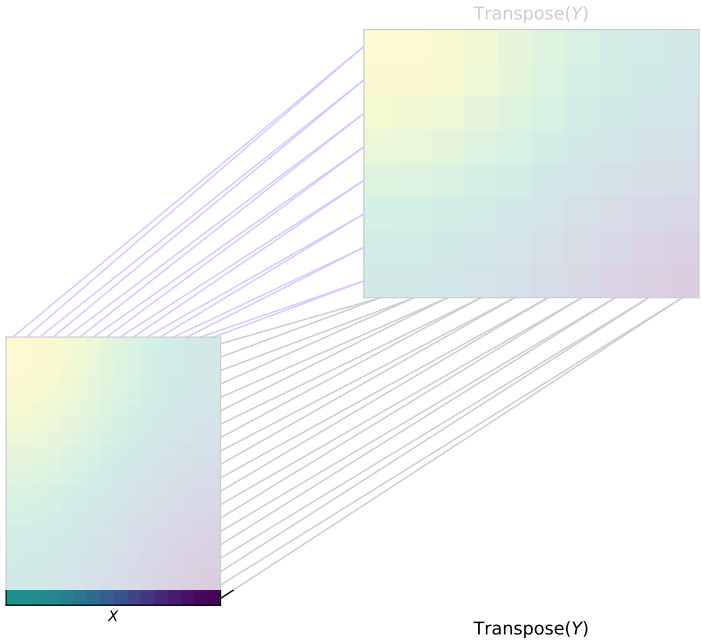
Motivation (1)

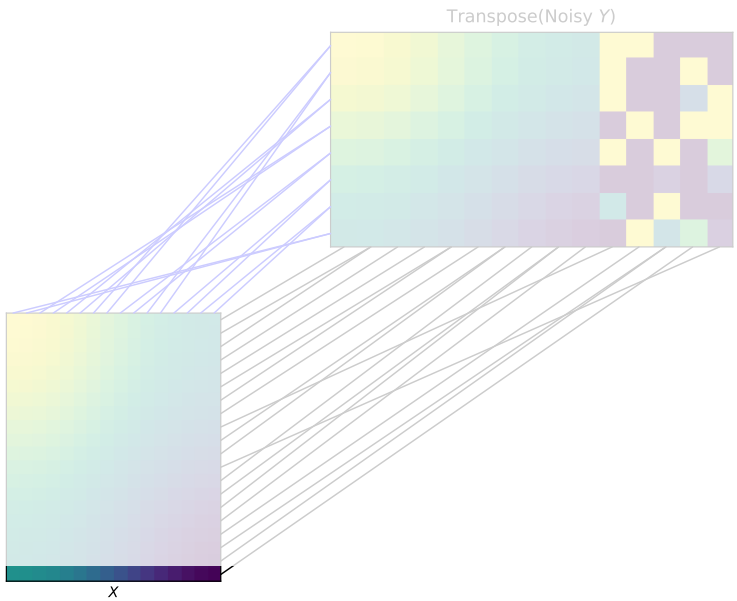


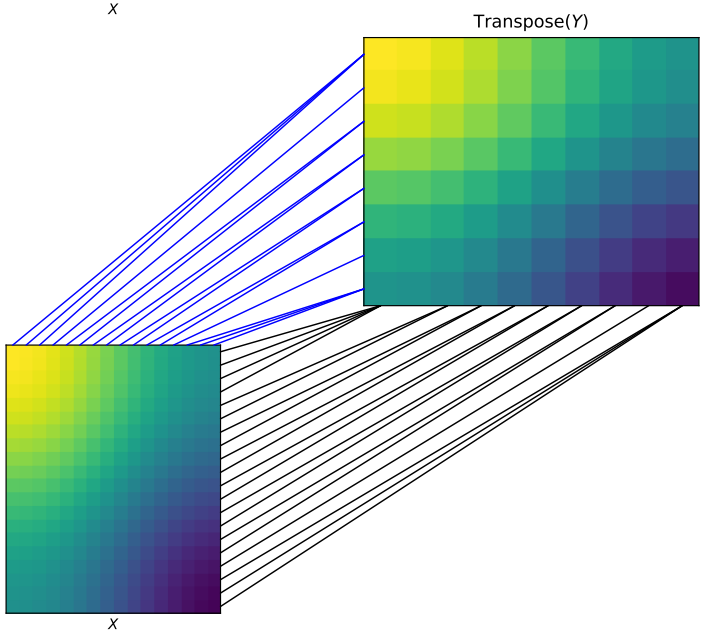


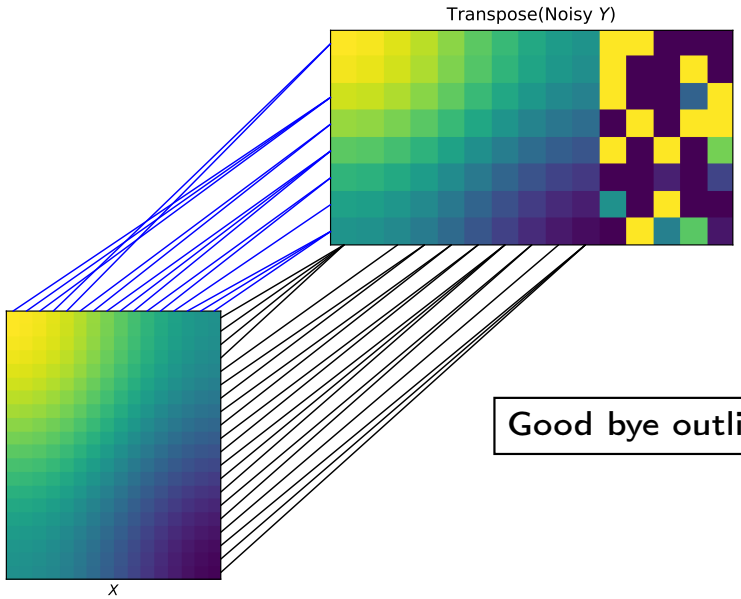
Hello from outliers

Motivation (2)









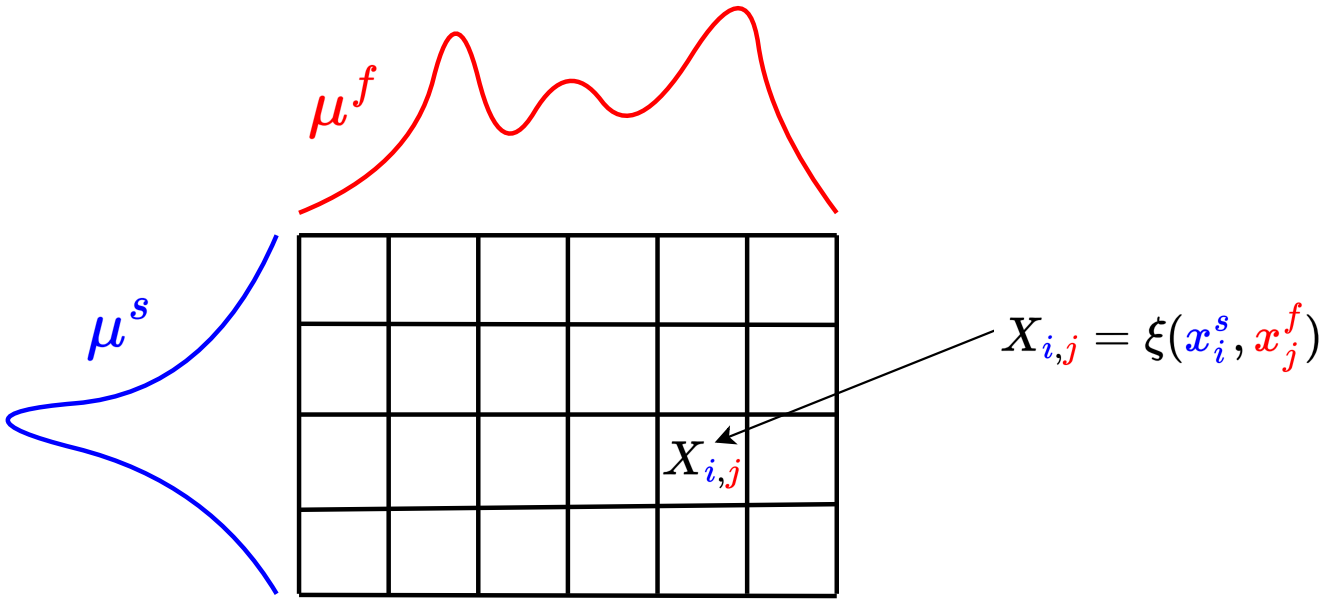
Good bye outliers

Unbalanced CO-Optimal Transport (1)

Definition (Sample - feature space)

Let (X^s, μ^s) and (X^f, μ^f) be compact measure spaces, where $\mu^s \in \mathcal{M}^+(X^s)$ and $\mu^f \in \mathcal{M}^+(X^f)$. Let ξ be a scalar integrable function in $L^p(X^s \times X^f, \mu^s \otimes \mu^f)$, for some $p \geq 1$. We call

- The triplet $\mathcal{X} = \Big((X^s, \mu^s), (X^f, \mu^f), \xi\Big)$ a *sample - feature space*.
- The function ξ an *interaction*.



Unbalanced CO-Optimal Transport (2)

Definition (UCOOT)

Given $\lambda_1, \lambda_2 > 0$ and two s.f. spaces $\mathcal{X}_1 = \left((X_1^s, \mu_1^s), (X_1^f, \mu_1^f), \xi_1\right)$ and $\mathcal{X}_2 = \left((X_2^s, \mu_2^s), (X_2^f, \mu_2^f), \xi_2\right)$ is defined by:

$$\mathrm{UCOOT}_\lambda(\mathcal{X}_1, \mathcal{X}_2) := \inf_{\substack{\pi^s \in \mathcal{M}^+(X_1^s \times X_2^s) \\ \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f)}} F(\pi^s, \pi^f).$$

where

$$F(\pi^s, \pi^f) = \underbrace{\iint |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p \mathrm{d}\pi^s(x_1^s, x_2^s) \mathrm{d}\pi^f(x_1^f, x_2^f)}_{\text{transport cost of sample-feature pairs}}$$
$$+ \underbrace{\sum_{k=1}^{2} \lambda_k \mathrm{KL}(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu_k^s \otimes \mu_k^f)}_{\text{mass destruction / creation penalty}}.$$

Why quadratic divergence?

1. Homogeneity.
2. Existence of solution.
3. Bilinear relation between minimum and minimizer.

$$\text{UCOOT}_{\lambda}(\mathcal{X}_{\mathbf{1}}, \mathcal{X}_{\mathbf{2}}) = \sum_{k=\mathbf{1}}^{\mathbf{2}} \lambda_k m(\mu_k^s) m(\mu_k^f) - (\lambda_{\mathbf{1}} + \lambda_{\mathbf{2}}) m(\pi_*^s) m(\pi_*^f).$$

4. Restriction to equal-mass solutions: $m(\pi_*^s) = m(\pi_*^f)$.
5. Robustness to outlier.

Provable robustness of UCOOT (1)

Definition

Consider two **clean** s.f. spaces $\mathcal{X}_k = \Big((X_k^s, \mu_k^s), (X_k^f, \mu_k^f), \xi_k\Big)$, for $k = 1, 2$.

1. **Noisy s.f. space:** $\widetilde{\mathcal{X}_1} = \Big((X_1^s \cup O^s, \widetilde{\mu}_1^s), (X_1^f \cup O^f, \widetilde{\mu}_1^f), \xi_1\Big)$, where
   - **Noisy distribution on sample space:** $\widetilde{\mu}_1^s = \alpha_s \mu_1^s + (1 - \alpha_s)\varepsilon^s$, where $\alpha_s \in [0, 1], \varepsilon^s \in \mathcal{M}^+(O^s)$.
   - **Noisy distribution on feature space:** $\widetilde{\mu}_1^f = \alpha_f \mu_1^f + (1 - \alpha_f)\varepsilon^f$, where $\alpha_f \in [0, 1], \varepsilon^f \in \mathcal{M}^+(O^f)$.
2. **Minimal cost:** $\Delta_0 := \min_{\substack{x_1^s \in O^s, x_1^f \in O^f \\ x_2^s \in X_2^s, x_2^f \in X_2^f}} \left|\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)\right|^p$.
3. **Maximal cost:** $\Delta_\infty := \max_{\substack{x_1^s \in X_1^s \cup O^s, x_1^f \in X_1^f \cup O^f \\ x_2^s \in X_2^s, x_2^f \in X_2^f}} \left|\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)\right|^p$.

$\Longrightarrow$ Both costs can explode if the outliers are too impactful.

Provable robustness of UCOOT (2)

Proposition

1. *COOT is sensitive to outliers*

$$COOT(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \geq (1 - \alpha_s)(1 - \alpha_f)\Delta_0.$$

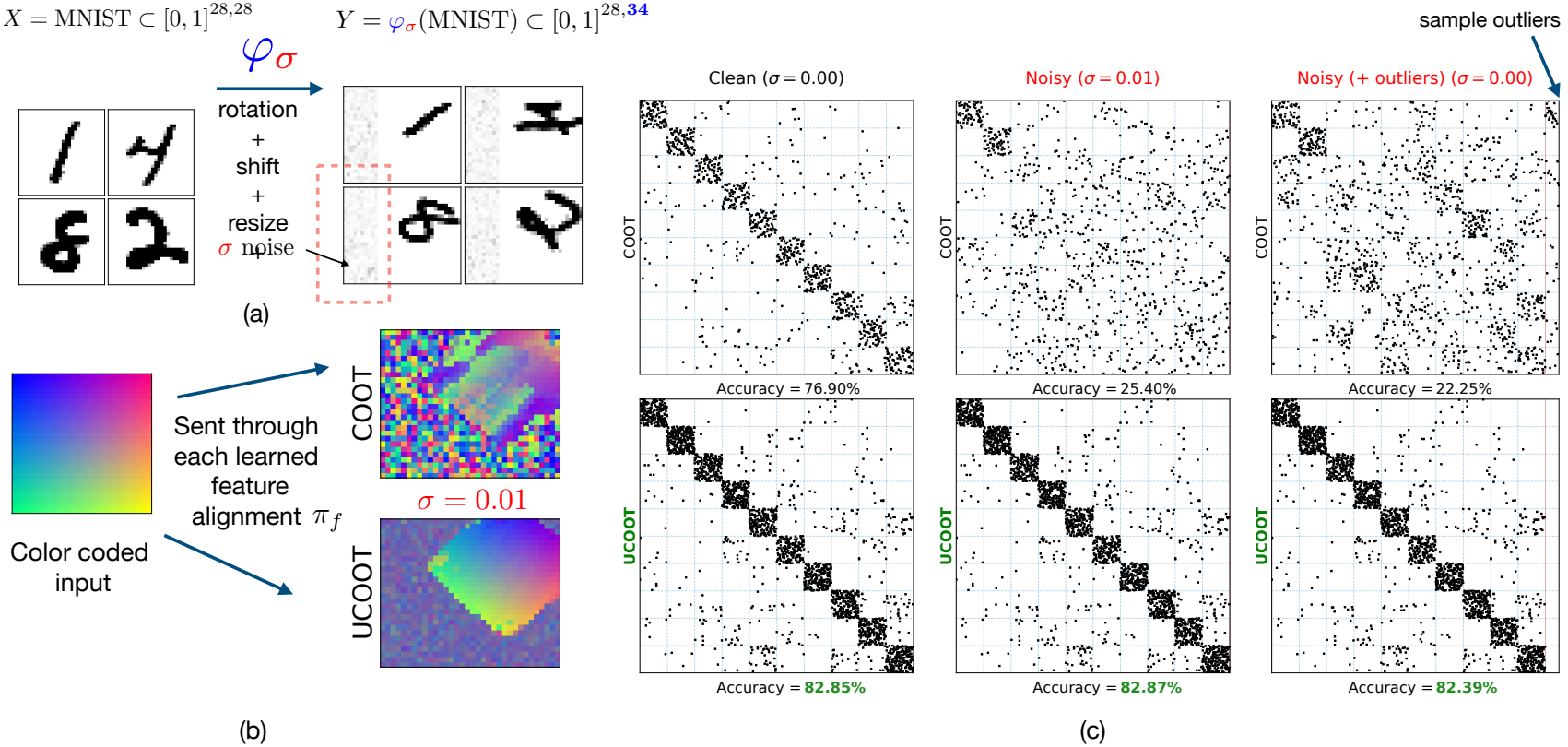
2. *Denote*

- $\delta = 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f)$.
- $M = m(\pi^s) = m(\pi^f)$: *mass of OT plans between clean data.*
- $K = M + \frac{1}{M} UCOOT_\lambda(\mathcal{X}_1, \mathcal{X}_2) + \delta$.

Then

$$\underbrace{UCOOT_\lambda(\widetilde{\mathcal{X}_1}, \mathcal{X}_2)}_{\text{"noisy divergence"}} \leq \alpha_s\alpha_f \underbrace{UCOOT_\lambda(\mathcal{X}_1, \mathcal{X}_2)}_{\text{"clean divergence"}} + \underbrace{\delta M \left[1 - \exp\left(-\frac{\Delta_\infty(1 + M) + K}{\delta M}\right)\right]}_{\text{saturates quickly if } \Delta_\infty \to \infty}.$$

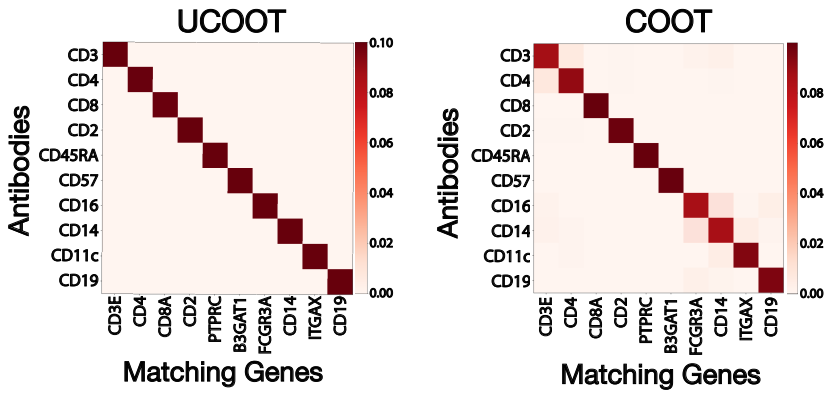
Illustration on MNIST images



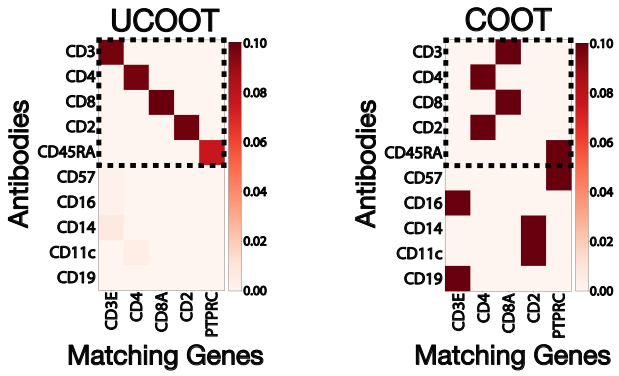
Example illustrating and interpreting the feature alignment π^f learned by UCOOT and its robustness to outliers.

Illustration on MNIST images

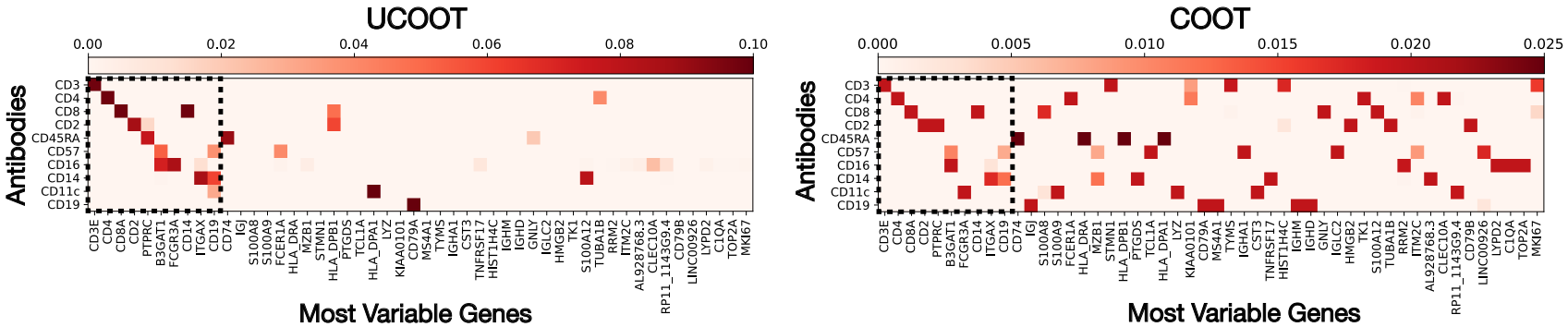
(a) Balanced scenario: aligning matching features



(b) Unbalanced scenario: aligning a subset of the matching features



(c) Unbalanced scenario: aligning antibodies with the top 50 most variable genes (including matching features)



Example illustrating and interpreting the feature alignment π^f learned by UCOOT and its robustness to outliers.

Table of Contents

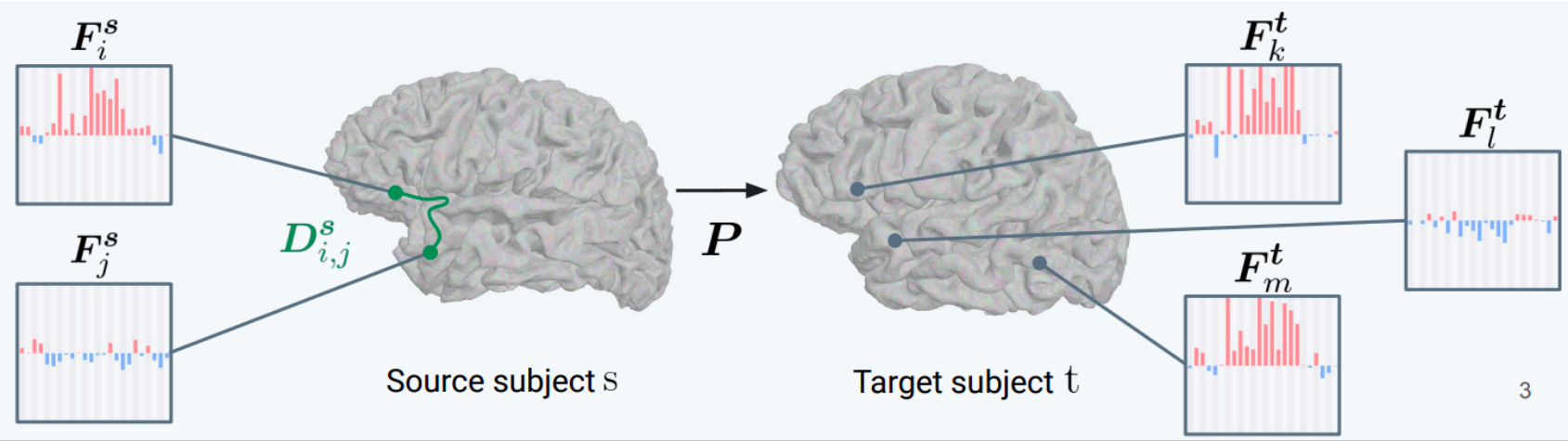
1. Introduction
2. Unbalanced CO-Optimal Transport
3. Fused Unbalanced Gromov-Wasserstein
4. Augmented Gromov-Wasserstein
5. Conclusion and Perspective

Motivation

Formulation

Definition

$$\inf_{P \in \mathbb{R}_{\geq \mathbf{0}}^{m \times n}} \quad (1-\alpha) \sum_{i,j,k,l} |D_{ik}^s - D_{jl}^t|^{\mathbf{2}} P_{ij} P_{kl} + \alpha \sum_{i,j} ||F_i^s - F_j^t||_{\mathbf{2}}^{\mathbf{2}} P_{ij}$$
$$+ \lambda \left[\mathrm{KL}(P_{\#\mathbf{1}} \otimes P_{\#\mathbf{1}} | w^s \otimes w^s) + \mathrm{KL}(P_{\#\mathbf{2}} \otimes P_{\#\mathbf{2}} | w^t \otimes w^t) \right]$$



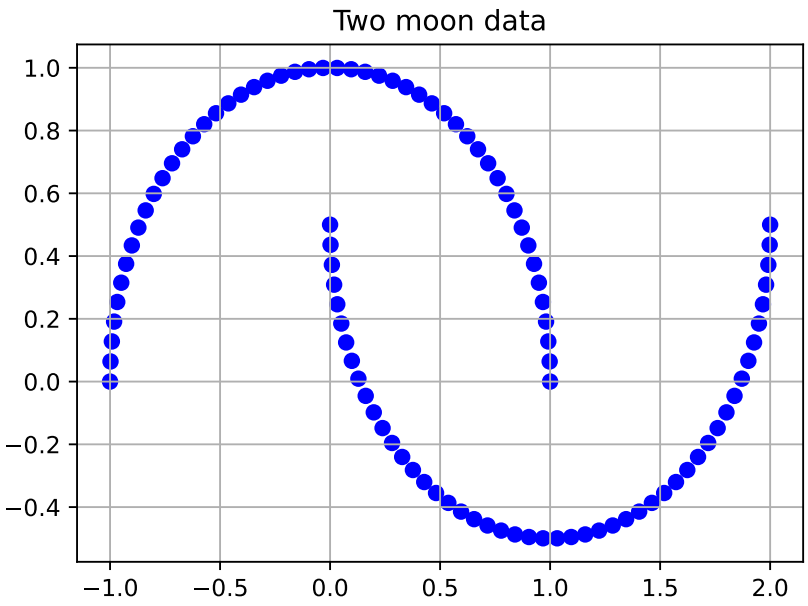
Neuroscientific interpretation

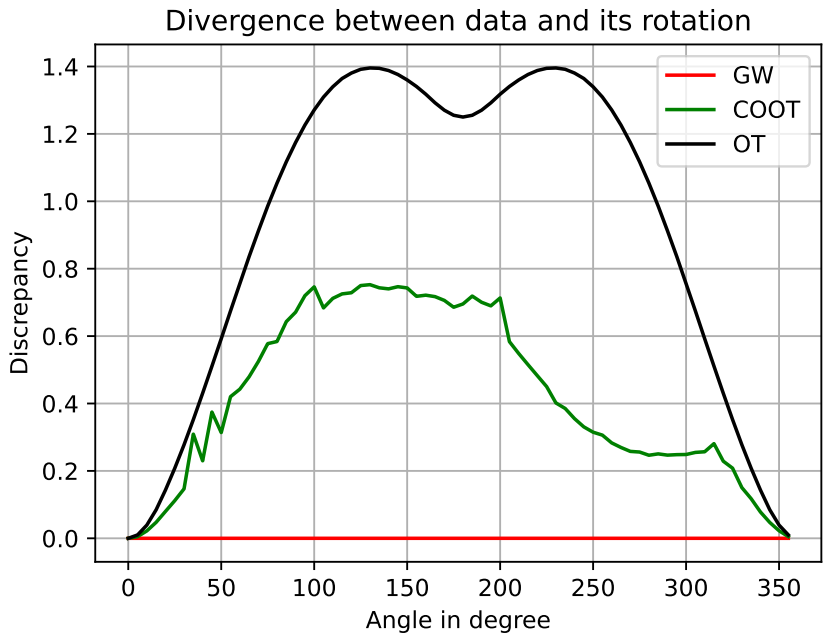
Table of Contents

1. Introduction
2. Unbalanced CO-Optimal Transport
3. Fused Unbalanced Gromov-Wasserstein
4. Augmented Gromov-Wasserstein
5. Conclusion and Perspective

Motivation

1. Isometries are useful for across-space comparison.
2. All isometries are created equal, but some are more equal than others.
3. GW induces all isometries but none for COOT.





Formulation

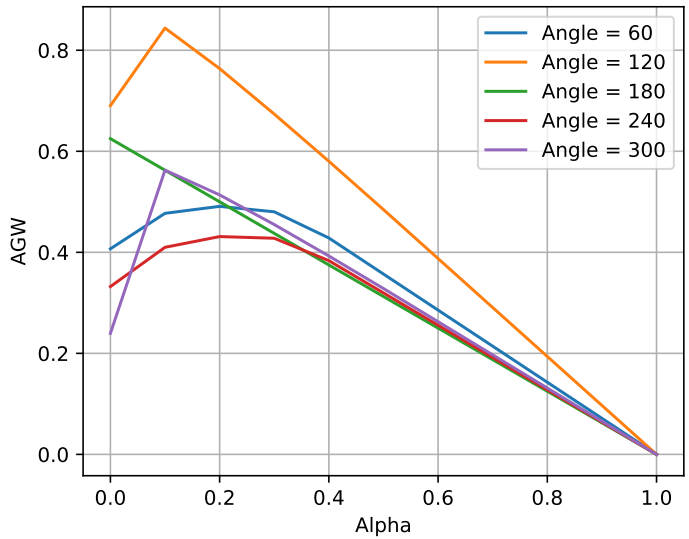
Definition

Given two weighted matrices $\mathcal{X}_1 = (X_1, \mu_1^s, \mu_1^f)$ and $\mathcal{X}_2 = (X_2, \mu_2^s, \mu_2^f)$, for $0 \leq \alpha \leq 1$,

$$\mathrm{AGW}_\alpha(\mathcal{X}_1, \mathcal{X}_2) := \min_{\substack{\pi^s \in U(\mu_1^s, \mu_2^s) \\ \pi^f \in U(\mu_1^f, \mu_2^f)}} L_\alpha(\pi^s, \pi^f),$$

where $L_\alpha(\pi^s, \pi^f) = \alpha \langle L(D_X, D_Y) \otimes \pi^s, \pi^s \rangle + (1-\alpha) \langle L(X, Y) \otimes \pi^s, \pi^f \rangle$.

- Interpolating between GW and COOT.
- Satisfying relaxed triangle inequality.



Isometries

- Only finitely many isometries such that $\mathrm{AGW}_\alpha(\mathcal{X}_1, \mathcal{X}_2) = 0$.

Assumption

Given an input matrix $X \in \mathbb{R}^{n \times d}$, assume

(A1) *$n \geq d$: Low-dimensional setting.*

(A2) *X is full rank: Easily met by preprocessing.*

(A3) *X has d distinct singular values. Fact: set of Hermitian matrices with repeated eigenvalues has zero Lebesgue measure.*

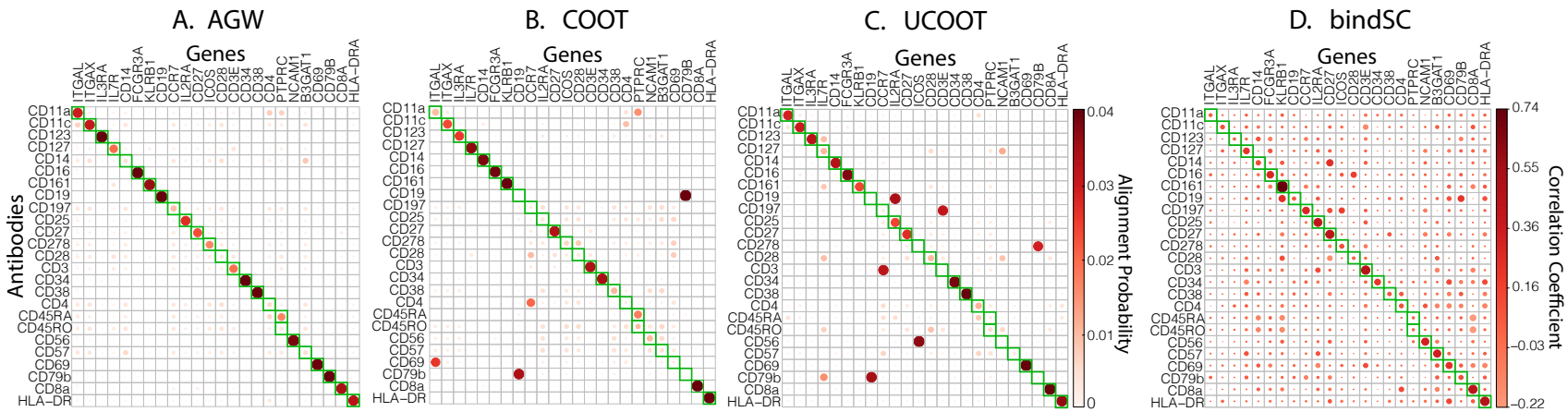
Proposition

Given two weighted matrices $\mathcal{X}_1 = (X_1, \mu_1^s, \mu_1^f)$ and $\mathcal{X}_2 = (X_2, \mu_2^s, \mu_2^f)$,

($\Rightarrow$) *If $\mu_1^s = \mu_2^s$ and X_2 is obtained by permuting columns of X_1 via the permutation σ_c (so $\mu_2^f = (\sigma_c)_\# \mu_1^f$), then $AGW_\alpha(\mathcal{X}_1, \mathcal{X}_2) = 0$.*

($\Leftarrow$) *Suppose X_1 satisfies A1-A3. For any $0 < \alpha < 1$, if $AGW_\alpha(\mathcal{X}_1, \mathcal{X}_2) = 0$, then there exist a symmetric orthogonal matrix $O \in \mathcal{O}_d$ and a permutation matrix $P \in \mathcal{P}_d$ such that $X_2 = X_1 OP$.*

Experiments



Neuroscientific interpretation

Table of Contents

1. Introduction
2. Unbalanced CO-Optimal Transport
3. Fused Unbalanced Gromov-Wasserstein
4. Augmented Gromov-Wasserstein
5. Conclusion and Perspective

Conclusion and Perspective

- Conclusion

1.

- Perspectives

1. In general, statisical aspects of unbalanced across-space divergences are still little understood.
2. Practical consideration: UGW, FUGW, UCOOT rely on BCD $\Rightarrow$ More efficient UOT solvers.

Thank you for your attention

References I

- Chowdhury, Samir and Facundo Mémoli (2019). "The Gromov-Wasserstein distance between networks and stable network invariants". In: *Information and Inference: A Journal of the IMA* 8 (4), pp. 757–787 (cit. on p. 5).
- Chowdhury, Samir, Tom Needham, et al. (2023). "Hypergraph Co-Optimal Transport: Metric and Categorical Properties". In: *Journal of Applied and Computational Topology* 40 (cit. on p. 8).
- Mémoli, Facundo (2007). "On the use of Gromov-Hausdorff Distances for Shape Comparison". In: *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association (cit. on p. 4).
- – (2011). "Gromov-Wasserstein distances and the metric approach to object matching". In: *Foundations of Computational Mathematics*, pp. 1–71 (cit. on p. 4).
- Redko, Ievgen et al. (2020). "CO-Optimal Transport". In: *Advances in Neural Information Processing Systems* 33 (cit. on p. 8).

References II

- Scetbon, Meyer, Gabriel Peyré, and Marco Cuturi (2021). "Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs". In: *arXiv preprint arXiv:2106.01128* (cit. on p. 5).
- Séjourné, Thibault, François-Xavier Vialard, and Gabriel Peyré (2021). "The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation". In: *Advances in Neural Information Processing Systems* 34 (cit. on p. 6).
- Vayer, Titouan, Laetita Chapel, et al. (2019). "Optimal Transport for structured data with application on graphs". In: *International Conference on Machine Learning* 97 (cit. on p. 7).
- Vayer, Titouan, Rémi Flamary, et al. (2019). "Sliced Gromov-Wasserstein". In: *Advances in Neural Information Processing Systems* 32, pp. 14726–14736 (cit. on p. 5).

Why bilinear relation.