# Generalizations of optimal transport: Sinkhorn divergences and Unbalanced Gromov-Wasserstein

Thibault Séjourné

LTS4 Group Meeting, EPFL – 14th October, 2020

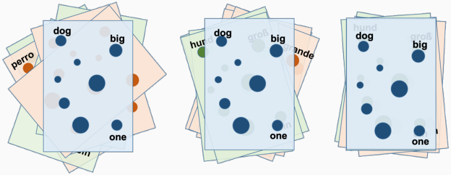
Joint work with Francois-Xavier Vialard, Gabriel Peyré, Jean Feydy and Alain Trouvé

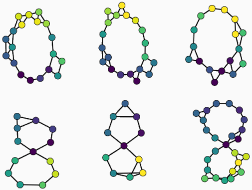
# Introduction

# Motivations

Some of machine learning challenges are

- matching point clouds of $\mathbb{R}^d$ up to isometries[1]
- graph matching[2]



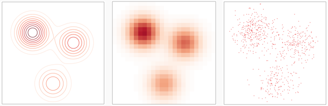



[1] Alaux, J., Grave, E., Cuturi, M., & Joulin, A. (2018). Unsupervised hyperalignment for multilingual word embeddings.

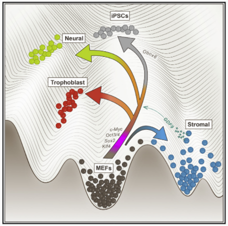
[2] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., & Courty, N. (2018). Fused Gromov-Wasserstein distance for structured objects.

# From probabilities to positive measures

Several models for measures, most commonly pointclouds $\alpha = \sum_i \alpha_i \delta_{x_i}$.



- Most often measures are normalized to mass 1 (i.e. are probabilities).
- Sometimes too restrictive:
  - $\rightarrow$ Normalizing data is unadapted.
  - $\rightarrow$ Avoid matching geometric outliers.[3]




From Schiebinger et al.

[3] Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., ... & Lee, L. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming.

# Optimal transport and its generalizations

Optimal transport displays three restrictions:

- Compares measures with same mass,
- Compares measures defined on the same space,
- Scales poorly in numerical solvers : $O(n^3 \log(n))$.

There exists extensions to overcome these issues:

- Unbalanced optimal transport,
- Gromov-Wasserstein distances,
- Entropic regularization.

# Outline of the presentation

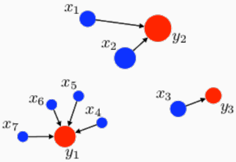
1. Background - UOT ( • )
2. Sinkhorn algorithm ( • + • )
3. Sinkhorn divergence ( • + • )
4. Unbalanced Gromov-Wasserstein ( • + • )
5. Implementation of UGW ( • + • + • )

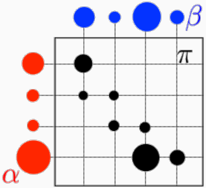
# Unbalanced OT

**Balanced Optimal Transport Distance**[4]

$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \geq 0} \left\{ \sum_{i,j} \mathrm{C}_{ij}\pi_{ij} \;:\; \begin{array}{l} \pi 1 = \alpha \\ \pi^\top 1 = \beta \end{array} \right\}.$$

Called p-Wasserstein distance for $\mathrm{C} = d^p$.




**Intuition:** Moving $\pi_{ij}$ grams from $x_i$ to $y_j$ costs $\pi_{ij} \times \mathrm{C}_{ij}$.

**Choice of C** $\rightarrow$ Choice of geometric prior.

4 Kantorovich, L. (1942). On the transfer of masses (in Russian).

# Unbalanced OT

**Idea**: Soften the constraint $\pi 1 = \alpha \rightarrow \mathrm{KL}(\pi 1 | \alpha)$

**Definition - Unbalanced OT**[5]

For any **positive** measures $(\alpha, \beta)$ one defines

$\mathrm{UOT}_\rho(\alpha, \beta) = \inf_{\pi \geq 0} \mathcal{L}_1(\pi)$ where

$$\mathcal{L}_1(\pi) \stackrel{\text{def.}}{=} \sum_{i,j} \mathsf{C}_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi 1 | \alpha) + \rho \mathrm{KL}(\pi^\top 1 | \beta).$$

- **Two dynamics:** transportation vs creation/destruction.
- **Possibility of other penalties:** $\mathrm{TV}$, or Csiszàr divergence $\mathsf{D}_\varphi$.
- **Balanced OT** is retrieved with $\rho \rightarrow \infty$ or $\mathsf{D}_\varphi = \iota_{(=)}$.

[5] Liero, M., Mielke, A., & Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures.

# Entropic Optimal Transport

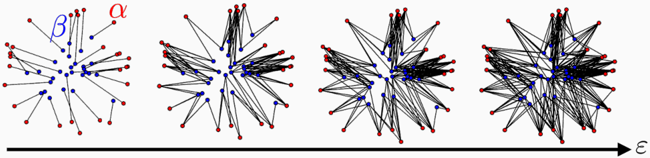
# Regularization of OT

**Reminder:** OT is computationally expensive and non-smooth

**Idea:** Add an entropic penalty $\varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta)$

**Entropic Unbalanced OT**[6] [7]

$$\mathrm{UOT}_{\varepsilon,\rho}(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \sum_{i,j} C_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi 1 | \alpha) + \rho \mathrm{KL}(\pi^\top 1, \beta) + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta)$$




[6] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport.

[7] Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F. X. (2018). Scaling algorithms for unbalanced optimal transport problems.

# Duality of regularized Balanced OT

The dual for balanced OT reads

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \sup_{f,g} \sum_i f_i \alpha_i + \sum_j g_j \beta_j - \varepsilon \sum_{i,j} (e^{\frac{f_i + g_j - \mathrm{C}_{ij}}{\varepsilon}} - 1) \alpha_i \beta_j .$$

The **alternate dual ascent** is straightforward to compute:

**Alternate dual ascent**

Given any initialization $f_0$. At time $t$ one has $(f_t, g_t)$. Then iterate until convergence:

1. Fix $f_t$ and find optimal $g$ in the dual $\rightarrow g_{t+1}$,
2. Fix $g_{t+1}$ and find optimal $f$ in the dual $\rightarrow f_{t+1}$.

# Sinkhorn algorithm for balanced OT

The dual for balanced OT reads

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \sup_{f,g} \sum_i f_i \alpha_i + \sum_j g_j \beta_j - \varepsilon \sum_{i,j} (e^{\frac{f_i + g_j - C_{ij}}{\varepsilon}} - 1)\alpha_i \beta_j.$$

## Sinkhorn algorithm in Balanced OT

Writing $K = (e^{-C_{i,j}/\varepsilon})_{ij}$, $u = (e^{f_i/\varepsilon})_i$ and $v = (e^{g_j/\varepsilon})_j$, it reads

$$u_{t+1} \leftarrow 1/K(\beta \odot v_t), \qquad v_{t+1} \leftarrow 1/K^\top(\alpha \odot u_{t+1}).$$

$\Rightarrow$ Matrix-vector operations fastly parallelizable on GPU !

## Log-stabilized Sinkhorn

$$f_i \leftarrow -\varepsilon \log \sum_j e^{(g_j - C_{ij})/\varepsilon} \beta_j, \quad g_j \leftarrow -\varepsilon \log \sum_i e^{(f_i - C_{ij})/\varepsilon} \alpha_i$$

# Unbalanced Sinkhorn algorithm

**Proposition - Unbalanced Sinkhorn algorithm**

The unbalanced Sinkhorn algorithm is the composition of its balanced counterpart with a pointwise operator $A$. It reads

$$f_i \leftarrow A\Big[-\varepsilon \log \sum_j e^{(g_j - \mathrm{C}_{ij})/\varepsilon} \beta_j\Big], \; g_j \leftarrow A\Big[-\varepsilon \log \sum_i e^{(f_i - \mathrm{C}_{ij})/\varepsilon} \alpha_i\Big].$$

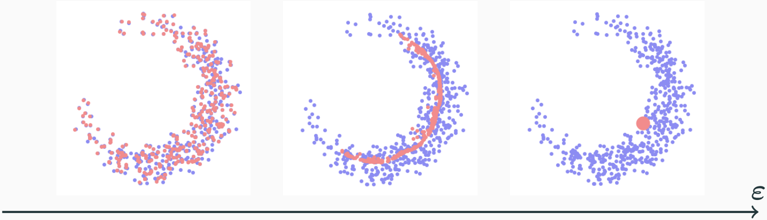
- **Thm:** Sinkhorn algorithm converges in many settings.
- $A$ is pointwise $\Rightarrow$ same complexity as Balanced Sinkhorn.
- $A$ is explicit and GPU-compatible for many settings (TV, ...)
    - Balanced OT: $A(x) = x$,
    - Kullback-Leibler ($\rho\mathrm{KL}$): $A(x) = \tau \cdot x$ with $\tau = \rho/(\varepsilon + \rho)$
- Same compositional structure with $(K, u, v)$

# Correcting the entropic bias - Sinkhorn divergence

# Entropic bias

**Problem:** $\mathcal{L} = \mathrm{OT}_\varepsilon$ does not retrieve $\beta$ for $\varepsilon > 0$.

$$\text{Not a distance:} \quad \mathrm{OT}_\varepsilon(\alpha, \alpha) > 0,$$
$$\exists \alpha \in \mathcal{M}_1^+(\mathcal{X}),\ \mathrm{OT}_\varepsilon(\alpha, \beta) < \mathrm{OT}_\varepsilon(\beta, \beta).$$




$\Rightarrow$ **One cannot crossvalidate the parameter $\varepsilon$.**

# Unbalanced Sinkhorn Divergence

**Definition**

Setting $m(\mu) = \sum_i \mu_i$, we define

$$\mathrm{S}_{\varepsilon,\rho}(\alpha, \beta) \stackrel{\text{def.}}{=} \mathrm{UOT}_{\varepsilon,\rho}(\alpha, \beta) - \frac{1}{2}\mathrm{UOT}_{\varepsilon,\rho}(\alpha, \alpha) - \frac{1}{2}\mathrm{UOT}_{\varepsilon,\rho}(\beta, \beta) + \frac{\varepsilon}{2}(m(\alpha) - m(\beta))^2.$$

It extends the balanced Sinkhorn divergence[8] [9].

**Remark**: When $\alpha = \beta$, one has $\mathrm{S}_{\varepsilon,\rho}(\alpha, \beta) = 0$.

**Is it positive ? Definite ? Smooth ?**

[8] Ramdas, A., Trillos, N. G., & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests.

[9] Genevay, A., Peyré, G., & Cuturi, M. (2018, March). Learning generative models with sinkhorn divergences.

# Main Theorem

**Theorem [S., Feydy, Vialard, Trouve, Peyre '19]**

For any Lipschitz cost C on a compact set s.t. $k_\varepsilon \stackrel{\text{def.}}{=} e^{-\frac{C}{\varepsilon}}$ is a positive universal kernel, for any $\varepsilon > 0$

- $S_{\varepsilon,\rho}$ is convex, positive, definite.
- It is (weakly) differentiable.
- One has $S_{\varepsilon,\rho}(\alpha, \beta) \to 0 \Leftrightarrow \alpha \rightharpoonup \beta$.

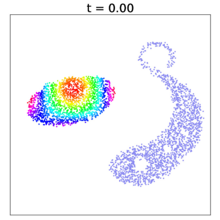
**Corollary:** holds for $C(x, y) = \|\psi(x) - \psi(y)\|_2^2$, for $\psi$ neural net.

# Numerical insights on UOT and the Sinkhorn divergence

# Numerical experiments model

Setting adapted from [Chizat '19][10].

- Position/mass parameterization $\theta = \{(x_i, r_i)_i\} \in (\mathbb{R}^d \times \mathbb{R}_+)^n$
- Model measure $\theta \mapsto \alpha(\theta) = \sum_i^n r_i^2 \delta_{x_i}$
- Minimize $\mathcal{L}(\alpha(\theta), \beta)$ w.r.t. $\theta$




## Updates of the coordinates

$$x_i^{(t+1)} = x_i^{(t)} - \eta_x \nabla_{x_i} \mathcal{L}(\alpha(\theta^{(t)}), \beta),$$
$$r_i^{(t+1)} = r_i^{(t)}.\exp\big(-2\eta_r \nabla_{r_i} \mathcal{L}(\alpha(\theta^{(t)}), \beta)\big)$$

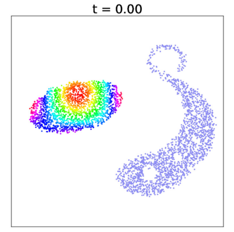
[10] Chizat, L. (2019). Sparse optimization on measures with over-parameterized gradient descent.

# Numerics 1

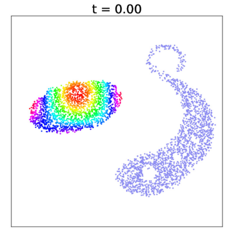
Parameters:

- $\mathrm{C}(x, y) = \|x - y\|_2^2$ on $[0,1]^2$ with $\mathrm{D}_\varphi = \rho \mathrm{KL}$
- $\rho = 0.3$, $\eta_x = 60.0$, $\eta_r = 0.3$

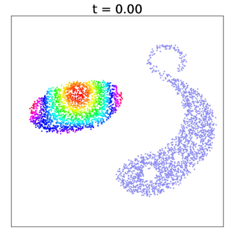
$\mathcal{L} = \mathrm{UOT}_{\varepsilon,\rho}$, $\varepsilon = 10^{-3}$




$\mathcal{L} = \mathrm{S}_{\varepsilon,\rho}$, $\varepsilon = 10^{-3}$




$\mathcal{L} = \mathrm{S}_{\varepsilon,\rho}$, $\varepsilon = 10^{-2}$



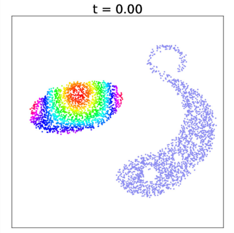

**$\mathrm{S}_{\varepsilon,\rho}$ should be prefered over $\mathrm{OT}_{\varepsilon,\rho}$, and $\varepsilon$ encodes a low pass filter effect (for statistical robustness).**

# Numerics 2

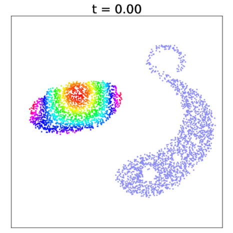
Parameters:

- $C(x,y) = \|x-y\|_2^2$ on $[0,1]^2$ with $\mathcal{L} = S_{\varepsilon,\rho}$
- $\varepsilon = 10^{-3}$, $\rho = 0.3$, $\eta_x = 60.0$, $\eta_r = 0.3$

$D_\varphi = \rho\mathrm{KL}$




$D_\varphi = \rho\mathrm{TV}$




**The choice of marginals' penalty encodes a variety of priors and behaviours, try them all !**

# Unbalanced Gromov-Wasserstein

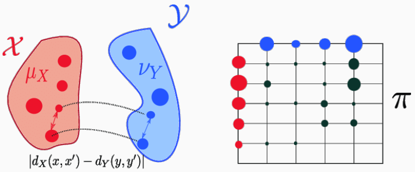
## Balanced Gromov-Wasserstein distance

**mm-space:** $\mathcal{X} = (X, d^{(X)}, \alpha)$ with $(X, d^{(X)})$ complete separable, $\alpha$ positive measure

**Definition - GW distance**[11]

Take $\mathcal{X} = (X, d^{(X)}, \alpha)$ and $\mathcal{Y} = (Y, d^{(Y)}, \beta)$ equipped with **probabilities**. One defines $GW(\mathcal{X}, \mathcal{Y}) = \inf_{\{\pi 1 = \alpha,\ \pi^\top 1 = \beta\}} \mathcal{G}(\pi)$ where

$$\mathcal{G}(\pi) \overset{\text{def.}}{=} \sum_{i,j,k,l} \left( d_{ij}^{(X)} - d_{kl}^{(Y)} \right)^2 \pi_{ik} \pi_{jl}.$$




[11] Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching.

# Specificities of GW

**Two key differences with OT**

- GW is non-convex (quadratic assignment program)
- $(\mathcal{X}, \mathcal{Y})$ can differ radically in nature.[12]

**Isometric mm-spaces**

**Def:** $\mathcal{X} \sim \mathcal{Y} \Leftrightarrow \exists \psi : X \to Y$ bijective isometry s.t.

$$d_X(x, x') = d_Y(\psi(x), \psi(x')) \quad \text{and} \quad \beta = \sum_i \alpha_i \delta_{\psi(x_i)}$$

**Prop:** With $\lambda(t) = t^q$, $GW^{\frac{1}{q}}$ distance and definite iff $\mathcal{X} \sim \mathcal{Y}$

[12] Solomon, J., Peyré, G., Kim, V. G., & Sra, S. (2016). Entropic metric alignment for correspondence problems.

# Unbalanced Gromov-Wasserstein

Define the tensor product of measures $(\pi \otimes \pi)_{ijkl} \stackrel{\text{def.}}{=} \pi_{ij}\pi_{kl}$.

**Definition**

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \sum_{i,j,k,l} \left(d_{ij}^{(X)} - d_{kl}^{(Y)}\right)^2 \pi_{ik}\pi_{jl} + \rho \mathrm{KL}(\pi_1 \otimes \pi_1, \alpha \otimes \alpha) + \rho \mathrm{KL}(\pi_2 \otimes \pi_2, \beta \otimes \beta).$$

**To be compared with**

$$\mathcal{G}(\pi) = \sum_{i,j,k,l} \left(d_{ij}^{(X)} - d_{kl}^{(Y)}\right)^2 \pi_{ik}\pi_{jl},$$

$$\mathcal{L}_1(\pi) = \sum_{i,j} \mathrm{C}_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi_1, \alpha) + \rho \mathrm{KL}(\pi_2, \beta).$$

# Unbalanced Gromov-Wasserstein

Define the tensor product of measures $(\pi \otimes \pi)_{ijkl} \stackrel{\text{def.}}{=} \pi_{ij}\pi_{kl}$.

**Definition**

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \sum_{i,j,k,l} \left( d_{ij}^{(X)} - d_{kl}^{(Y)} \right)^2 \pi_{ik}\pi_{jl} + \rho \mathrm{KL}(\pi_1 \otimes \pi_1 | \alpha \otimes \alpha)$$
$$+ \rho \mathrm{KL}(\pi_2 \otimes \pi_2 | \beta \otimes \beta).$$

**To be compared with**

$$\mathcal{G}(\pi) = \sum_{i,j,k,l} \left( d_{ij}^{(X)} - d_{kl}^{(Y)} \right)^2 \pi_{ik}\pi_{jl},$$

$$\mathcal{L}_1(\pi) = \sum_{i,j} \mathrm{C}_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi_1 | \alpha) + \rho \mathrm{KL}(\pi_2 | \beta).$$

# Unbalanced Gromov-Wasserstein

Define the tensor product of measures $(\pi \otimes \pi)_{ijkl} \stackrel{\text{def.}}{=} \pi_{ij}\pi_{kl}$.

**Definition**

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \sum_{i,j,k,l} \left( d_{ij}^{(X)} - d_{kl}^{(Y)} \right)^2 \pi_{ik}\pi_{jl} + \rho \mathrm{KL}(\pi_1 \otimes \pi_1 | \alpha \otimes \alpha) + \rho \mathrm{KL}(\pi_2 \otimes \pi_2 | \beta \otimes \beta).$$

**To be compared with**

$$\mathcal{G}(\pi) = \sum_{i,j,k,l} \left( d_{ij}^{(X)} - d_{kl}^{(Y)} \right)^2 \pi_{ik}\pi_{jl},$$

$$\mathcal{L}_1(\pi) = \sum_{i,j} \mathrm{C}_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi_1 | \alpha) + \rho \mathrm{KL}(\pi_2 | \beta).$$

# Unbalanced Gromov-Wasserstein

Define the tensor product of measures $(\pi \otimes \pi)_{ijkl} \stackrel{\text{def.}}{=} \pi_{ij}\pi_{kl}$.

**Definition**

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \sum_{i,j,k,l} \left(d_{ij}^{(X)} - d_{kl}^{(Y)}\right)^2 \pi_{ik}\pi_{jl} + \rho \mathrm{KL}(\pi_1 \otimes \pi_1 | \alpha \otimes \alpha) + \rho \mathrm{KL}(\pi_2 \otimes \pi_2 | \beta \otimes \beta).$$

**To be compared with**

$$\mathcal{G}(\pi) = \sum_{i,j,k,l} \left(d_{ij}^{(X)} - d_{kl}^{(Y)}\right)^2 \pi_{ik}\pi_{jl},$$

$$\mathcal{L}_1(\pi) = \sum_{i,j} \mathsf{C}_{ij}\pi_{ij} + \rho \mathrm{KL}(\pi_1 | \alpha) + \rho \mathrm{KL}(\pi_2 | \beta).$$

# Theoretical results and conic formulation

- UOT is not convenient to prove the triangle inequality.
- Need to use another formulation called "conic" (COT)

$\rightarrow$ COT = OT on a lifted space $\mathfrak{C} = X \times \mathbb{R}_+$

- **Thm 1:** UOT is definite.
- **Thm 2:** COT is a distance between positive measures.
- **Thm 3:** One has UOT $=$ COT.

**Theorem [S., Vialard, Peyré]**

1. UGW is definite up to isometries.
2. There exists a conic formulation CGW which is a distance between mm-spaces up to isometry.
3. One has UGW $\geq$ CGW.

# Implementation and numerics

# Implementing UGW

**Idea:** Entropic regularization + alternate minimization

$$\begin{aligned}
\mathrm{UGW}_\varepsilon(\mathcal{X},\mathcal{Y}) &\overset{\text{def.}}{=} \inf_{\pi\geq 0} \mathcal{L}_2(\pi) + \varepsilon \mathrm{KL}(\pi\otimes\pi, (\alpha\otimes\beta)^{\otimes 2}) \\
&\geq \inf_{\pi,\gamma\geq 0} \mathcal{F}(\pi,\gamma) + \varepsilon \mathrm{KL}(\pi\otimes\gamma, (\alpha\otimes\beta)^{\otimes 2}),
\end{aligned}$$

$$\begin{aligned}
\text{where } \mathcal{F}(\pi,\gamma) &\overset{\text{def.}}{=} \sum_{i,j,k,l} \left(d^{(X)}_{ij} - d^{(Y)}_{kl}\right)^2 \pi_{ik}\gamma_{jl} \\
&+ \rho \mathrm{KL}(\pi_1\otimes\gamma_1, \alpha\otimes\alpha) + \rho \mathrm{KL}(\pi_2\otimes\gamma_2, \beta\otimes\beta)
\end{aligned}$$

- Alternate descent = sequence of $\mathrm{UOT}_{\varepsilon,\rho}$ problems (Sinkhorn).
- Invariance: $(\pi,\gamma)$ optimal $\Rightarrow \forall s>0$, $(s\pi, \frac{1}{s}\gamma)$ optimal.
- Numerically we see at optimality $\pi^* = \gamma^*$ (if $m(\pi) = m(\gamma)$).

$\Rightarrow$ **Computes a local minimizer of $\mathrm{UGW}_\varepsilon$.**

# Reformulation of the alternate minimization

Define $m(\gamma) = \sum_{i,j} \gamma_{ij}$ with $\gamma_{1,i} = \sum_j \gamma_{ij}$ and $\gamma_{2,j} = \sum_i \gamma_{ij}$

**Proposition - alternate descent $\leftrightarrow$ solve UOT**

For a fixed $\gamma$, $\pi \in \arg\min_{\pi} \mathcal{F}(\pi, \gamma) + \varepsilon \mathrm{KL}(\pi \otimes \gamma | (\alpha \otimes \beta)^{\otimes 2})$ is the solution of

$$\min_{\pi} \sum_{i,j} c_{ij}^{\varepsilon,\gamma} \pi_{ij} + \rho m(\gamma) \mathrm{KL}(\pi_1 | \alpha) + \rho m(\gamma) \mathrm{KL}(\pi_2 | \beta) + \varepsilon m(\gamma) \mathrm{KL}(\pi | \alpha \otimes \beta), \qquad \text{where}$$

$$c_{ij}^{\varepsilon,\gamma} \stackrel{\text{def.}}{=} \sum_{k,l} \left( d_{ik}^{(X)} - d_{jl}^{(Y)} \right)^2 \gamma_{kl} + \rho \sum_i \log\left(\frac{\gamma_{1,i}}{\alpha_i}\right) \gamma_{1,i} + \rho \sum_j \log\left(\frac{\gamma_{2,j}}{\beta_j}\right) \gamma_{2,j} + \varepsilon \sum_{i,j} \log\left(\frac{\gamma_{ij}}{\alpha_i \beta_j}\right) \gamma_{ij}.$$

# Reformulation of the alternate minimization

Define $m(\gamma) = \sum_{i,j} \gamma_{ij}$ with $\gamma_{1,i} = \sum_j \gamma_{ij}$ and $\gamma_{2,j} = \sum_i \gamma_{ij}$

**Proposition - alternate descent $\leftrightarrow$ solve UOT**

For a fixed $\gamma$, $\pi \in \arg\min_\pi \mathcal{F}(\pi, \gamma) + \varepsilon \mathrm{KL}(\pi \otimes \gamma | (\alpha \otimes \beta)^{\otimes 2})$ is the solution of

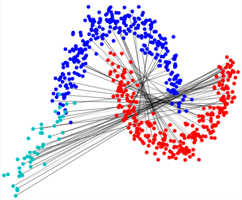
$$\min_\pi \sum_{i,j} \tilde{c}_{ij} \pi_{ij} + \tilde{\rho} \mathrm{KL}(\pi_1 | \alpha) + \tilde{\rho} \mathrm{KL}(\pi_2 | \beta) + \tilde{\varepsilon} \mathrm{KL}(\pi | \alpha \otimes \beta),$$

where $(\tilde{c}, \tilde{\rho}, \tilde{\varepsilon})$ depend on the fixed measure $\gamma$ via a computable formula.

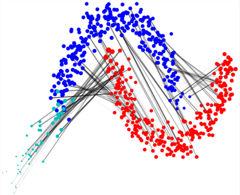
# Numerical experiment - outlier discardance

**Setup:** $(\mathcal{X}, \mathcal{Y})$ = two moons + outlier with Euclidean distance, $(\alpha, \beta)$ uniform probabilities

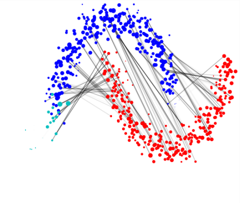
GW



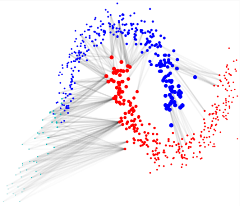
UGW $\rho = 10^0$



UGW $\rho = 10^{-1}$



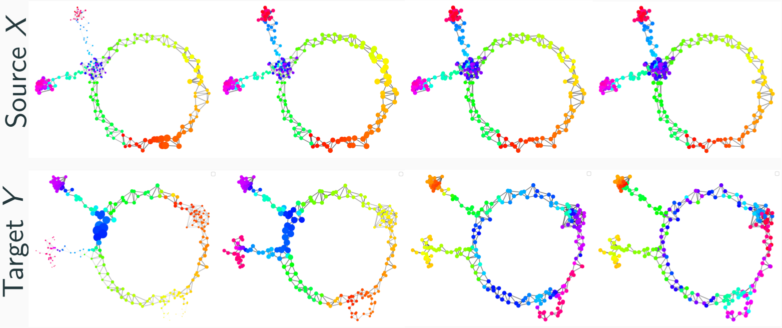
UOT $\rho = 10^{-2}$



**UGW performs lazy matchings to avoid outliers.**

# Numerical experiment - graph matching

**Setup:** $(\mathcal{X}, \mathcal{Y})$ = graphs with $n$ nodes and geodesic distance, $(\alpha, \beta)$ uniform probabilities




$\rho = 0.1$ $\quad$ $\rho = 1$ $\quad$ $\rho = 10$ $\quad$ GW $(\rho = \infty)$

**UGW encodes partial and/or geometrically consistent matchings.**

# Perspective - debiasing GW

$\mathrm{UGW}_\varepsilon$ also suffer from **entropic bias**. Adapting from regularized UOT, a potential divergence candidate is

$$\begin{aligned}\mathrm{SUGW}(\mathcal{X},\mathcal{Y}) \overset{\text{def.}}{=} \mathrm{UGW}_\varepsilon(\mathcal{X},\mathcal{Y}) - \tfrac{1}{2}\,\mathrm{UGW}_\varepsilon(\mathcal{X},\mathcal{X}) - \tfrac{1}{2}\,\mathrm{UGW}_\varepsilon(\mathcal{Y},\mathcal{Y}) \\ + \tfrac{\varepsilon}{2}(m(\alpha)^2 - m(\beta)^2)^2.\end{aligned}$$

**Open question:** Does it debias $\mathrm{UGW}_\varepsilon$? Is it positive ? Definite ?

# Conclusion

- Beware of entropic regularization: favor $\mathrm{S}_{\varepsilon,\rho}$ over $\mathrm{UOT}_{\varepsilon,\rho}$
- Flexibility of UOT models through $(\mathrm{C}, \rho, \varepsilon) + \mathrm{KL} \rightsquigarrow \mathrm{D}_\varphi$

- Blending of UOT with GW distances
- Computations on GPUs $\rightarrow$ UGW
- Theoretical aspects $\rightarrow$ CGW distance

## Implementations - github repositories

- thibsej/unbalanced-ot-functionals
- jeanfeydy/geomloss
- thibsej/unbalanced_gromov_wasserstein

## References


- Feydy, J., Séjourné, T., Vialard, F. X., Amari, S. I., Trouvé, A., & Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences.
- Séjourné, T., Feydy, J., Vialard, F. X., Trouvé, A., & Peyré, G. (2019). Sinkhorn Divergences for Unbalanced Optimal Transport.
- Séjourné, T., Vialard, F. X., & Peyré, G. (2020). The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation.


**Thank you !**

# Supplementary slides

# Csiszàr divergences[13]

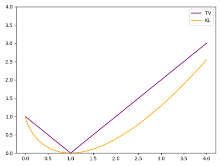
Define $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ l.s.c., convex, $\varphi(1) = 0$, $\varphi'^{\infty} = \lim_{x\to\infty} \frac{\varphi(x)}{x}$.
Write $\alpha = \sum_i \alpha_i \delta_{x_i}$ and $\beta = \sum_i \beta_i \delta_{x_i}$ (Same support $(x_i)$)

## Definition - $\varphi$-divergence

$$\mathrm{D}_\varphi(\alpha, \beta) = \sum_{\beta_i \neq 0} \varphi\left(\frac{\alpha_i}{\beta_i}\right)\beta_i + \varphi'^{\infty} \sum_{\beta_i = 0} \alpha_i.$$

**Examples:**

- $\mathrm{KL}(\alpha, \beta) = \sum_i (\log(\frac{\alpha_i}{\beta_i})\alpha_i - \alpha_i + \beta_i)$: $\varphi(x) = x \log x - x + 1$,
- $\mathrm{TV}(\alpha, \beta) = \sum_i |\alpha_i - \beta_i|$: $\varphi(x) = |x - 1|$.




[13] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation.

# Alternate UGW = sequence of Sinkhorn updates

- Focus on $\lambda(t) = t^2$ for improved time and memory complexity
- Focus on $\mathrm{D}_\varphi = \mathrm{KL}$ which verifies

$$\begin{aligned}\mathrm{KL}(\mu \otimes \nu, \alpha \otimes \beta) = m(\nu)\mathrm{KL}(\mu, \alpha) + m(\mu)\mathrm{KL}(\nu, \beta) \\ +(m(\mu) - m(\alpha))(m(\nu) - m(\beta)).\end{aligned}$$

$\Rightarrow$ Given $\gamma$, minimizing w.r.t. $\pi$ amounts to solve a regularized UOT problem.

# Algorithm

**Algorithm 1 – UGW($\mathcal{X}$, $\mathcal{Y}$, $\rho$, $\varepsilon$)**

**Input:** mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation $\rho$, regularization $\varepsilon$
**Output:** approximation $(\pi, \gamma)$ minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

Initialize $(\pi, \gamma)$ and $(f, g)$
**while** $(\pi, \gamma)$ has not converged **do**
    Update $\gamma \leftarrow \pi$ and compute the cost $\tilde{c} \leftarrow c^{\varepsilon,\gamma}$
    Update parameters $(\tilde{\rho}, \tilde{\varepsilon}) \leftarrow (m(\pi)\rho, m(\pi)\varepsilon)$
    Compute $(f, g)$ that solves UOT($\mu$, $\nu$, $\tilde{c}$, $\tilde{\rho}$, $\tilde{\varepsilon}$)
    Update $\gamma_{ij} \leftarrow \exp\left[(f_i + g_j - \tilde{c}_{ij})/\tilde{\varepsilon}\right] \alpha_i \beta_j$
    Rescale $\gamma \leftarrow \sqrt{m(\pi)/m(\gamma)}\gamma$
Return $(\pi, \gamma)$.

# Detailed algorithm

**Algorithm 2 – UGW($\mathcal{X}$, $\mathcal{Y}$, $\rho$, $\varepsilon$)**

**Input:** mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation $\rho$, regularization $\varepsilon$
**Output:** approximation $(\pi, \gamma)$ minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

Initialize $\pi = \gamma = \mu \otimes \nu / \sqrt{m(\mu) m(\nu)}$, $g = 0$.
**while** $(\pi, \gamma)$ has not converged **do**
  Update $\pi \leftarrow \gamma$, then $c \leftarrow c_\pi^\varepsilon$, $\tilde{\rho} \leftarrow m(\pi)\rho$, $\tilde{\varepsilon} \leftarrow m(\pi)\varepsilon$
  **while** $(f, g)$ has not converged **do**
    $\forall x,\ f(x) \leftarrow -\frac{\tilde{\varepsilon}\tilde{\rho}}{\tilde{\varepsilon}+\tilde{\rho}} \log \left( \int e^{(g(y)-c(x,y))/\tilde{\varepsilon}} \mathrm{d}\nu(y) \right)$
    $\forall y,\ g(y) \leftarrow -\frac{\tilde{\varepsilon}\tilde{\rho}}{\tilde{\varepsilon}+\tilde{\rho}} \log \left( \int e^{(f(x)-c(x,y))/\tilde{\varepsilon}} \mathrm{d}\mu(x) \right)$
  Update $\gamma(x,y) \leftarrow \exp \left[ (f(x) + g(y) - c(x,y))/\tilde{\varepsilon} \right] \mu(x)\nu(y)$
  Rescale $\gamma \leftarrow \sqrt{m(\pi)/m(\gamma)}\gamma$
Return $(\pi, \gamma)$.