Unbalanced Gromov-Wasserstein distance

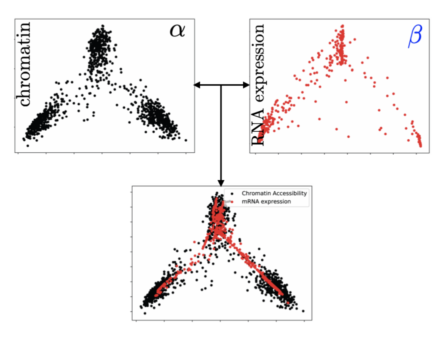
Thibault Séjourné

SMAI Minisymposium – 24th June, 2021

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

Introduction

Real world motivation: aligning genomic data[1]



Aligning RNA data

Data is:

- Heterogeneous ($\neq$ dimensions)
- imbalanced

$\Rightarrow$ Need for adaptive and robust (yet meaningful) assignments !

[1] Demetci, Pinar, et al. Gromov-Wasserstein optimal transport to align single-cell multi-omics data.

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

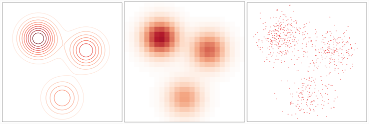
1. Background - UOT ($\bullet$)
2. Sinkhorn algorithm ($\bullet$ + $\bullet$)
3. Unbalanced Gromov-Wasserstein ($\bullet$ + $\bullet$)
4. Implementation of UGW ($\bullet$ + $\bullet$ + $\bullet$)

Unbalanced OT

Representing probabilities and measures

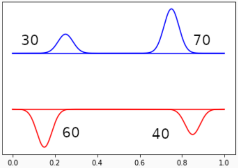
Several models for measures, most commonly pointclouds.

Measure $\alpha = \sum_i \alpha_i \delta_{x_i}$,
Mass $m(\alpha) = \sum_i \alpha_i$.



Important example: Gaussian densities in $\mathbb{R}$ with $\alpha \propto \sum_i p(x_i)\delta_{x_i}$

- β = mixture 70/30
- α = mixture 40/60

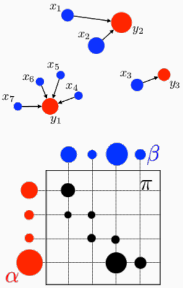


Optimal Transport (OT)

Balanced Optimal Transport Distance[2]

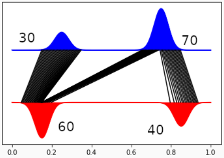
$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \geq 0} \left\{ \sum_{i,j} \mathrm{C}_{ij} \pi_{ij} \; : \; \begin{array}{l} \pi \mathbb{1} = \alpha \\ \pi^\top \mathbb{1} = \beta \end{array} \right\}.$$

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



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

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



[2] Kantorovich, L. (1942). On the transfer of masses (in Russian).

Unbalanced OT

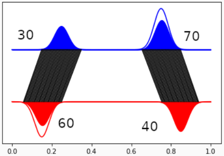
Idea: Soften the constraint $\pi 1 = \alpha$
$\rightarrow \mathrm{KL}(\pi 1|\alpha) = \sum_i \log(\frac{\pi_{1,i}}{\alpha_i})\alpha_i - m(\pi 1) + m(\alpha)$

Definition - Unbalanced OT[3]

For any **positive** measures (α, β) one defines
$\mathrm{UOT}_\rho(\alpha, \beta) = \inf_{\pi \geq 0} \mathcal{L}_{\mathrm{UOT}}(\pi)$ where

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

- **2 choices:** transport vs create/destroy
- **Other penalties:** TV, or Csiszàr div D_φ.
- **Balanced OT** $= \rho \rightarrow \infty$ or $D_\varphi = \iota_{(=)}$.



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

Entropic Optimal Transport

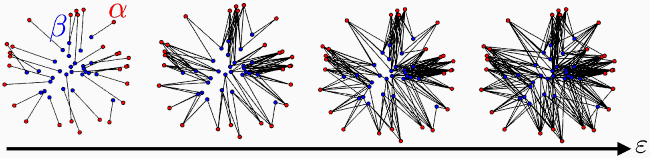
Regularization of OT

Reminder: OT is computationally expensive.

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

Entropic Unbalanced OT[4] [5]

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



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

[5] 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 $\mathrm{UOT}_{\varepsilon,\rho}$ reads

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

We consider **alternate dual ascent** to compute $\mathrm{UOT}_{\varepsilon,\rho}$:

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

Duality of regularized Balanced OT

The dual for $\mathrm{UOT}_{\varepsilon,\rho}$ reads

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

Unbalanced Sinkhorn algorithm = Alternate dual ascent

$$f_i \leftarrow \tfrac{\rho}{\varepsilon+\rho}\Big[-\varepsilon \log \sum_j e^{(g_j - C_{ij})/\varepsilon}\beta_j\Big],$$

$$g_j \leftarrow \tfrac{\rho}{\varepsilon+\rho}\Big[-\varepsilon \log \sum_i e^{(f_i - C_{ij})/\varepsilon}\alpha_i\Big].$$

Rmk: Solve dual $\Rightarrow$ Solve primal: $\pi^*_{ij} = \exp((f^*_i + g^*_j - C_{ij}/\varepsilon)\alpha_i\beta_j$.

Unbalanced Gromov-Wasserstein

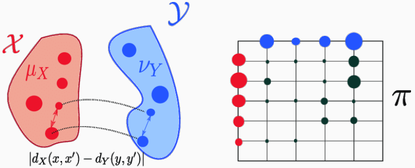
Balanced Gromov-Wasserstein distance

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

Definition - GW distance[6]

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

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



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

Isometric mm-spaces

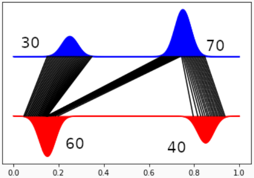
The GW distances encodes an equivalence relation of isometry.

Isometric mm-spaces

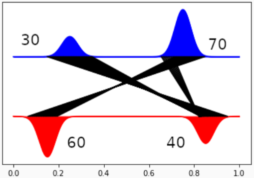
Def: $\mathcal{X} \sim \mathcal{Y} \Leftrightarrow \exists \psi : X \rightarrow Y$ bijective isometry s.t.

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

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



Balanced OT

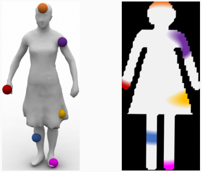


GW

Specificities of GW

Two key differences with OT

- GW is non-convex (quadratic assignment program)
- $(\mathcal{X}, \mathcal{Y})$ can differ radically in nature.[7]



[7] 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 $(\pi \otimes \pi)_{ijkl} \overset{\text{def.}}{=} \pi_{ij}\pi_{kl}$.

Definition

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

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

To be compared with

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

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

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

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

To be compared with

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

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

Unbalanced Gromov-Wasserstein

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

Definition

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

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

To be compared with

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

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

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

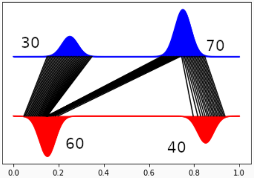
$$\mathcal{L}_{\mathrm{UGW}}(\pi) = \sum_{i,j,k,l} \left(d_{ik}^{(X)} - d_{jl}^{(Y)}\right)^2 \pi_{ij}\pi_{kl} + \rho \mathrm{KL}(\pi_1 \otimes \pi_1 | \alpha \otimes \alpha) + \rho \mathrm{KL}(\pi_2 \otimes \pi_2 | \beta \otimes \beta).$$

To be compared with

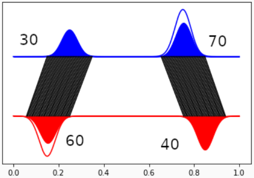
$$\mathcal{G}(\pi) = \sum_{i,j,k,l} \left(d_{ik}^{(X)} - d_{jl}^{(Y)}\right)^2 \pi_{ij}\pi_{kl},$$

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

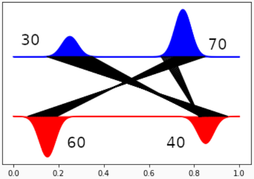
Numeric in dimension 1



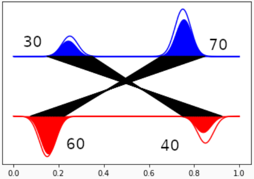
Balanced OT



UOT-KL

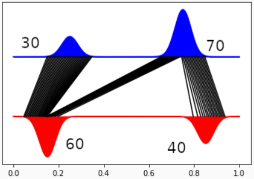


GW

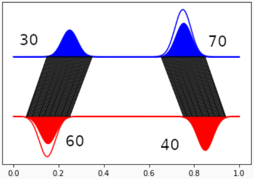


UGW

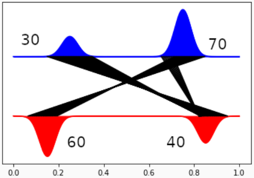
Take home message



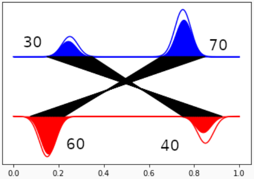
1 to 1



1 to 1 $\oplus$ reweighting



1 to 1 $\oplus$ isometry



1 to 1 $\oplus$ isom. $\oplus$ rew.

We can enrich assignments with a variety of priors.

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 of UGW and experiments

Implementing UGW

Idea: Entropic regularization + alternate minimization

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

$$\geq \inf_{\pi,\gamma\geq 0} \mathcal{F}(\pi,\gamma) + \varepsilon \mathrm{KL}(\pi\otimes\gamma, (\alpha\otimes\beta)^{\otimes 2}),$$

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

$$+ \rho \mathrm{KL}(\pi_1\otimes\gamma_1, \alpha\otimes\alpha) + \rho \mathrm{KL}(\pi_2\otimes\gamma_2, \beta\otimes\beta)$$

Reformulation of the alternate minimization

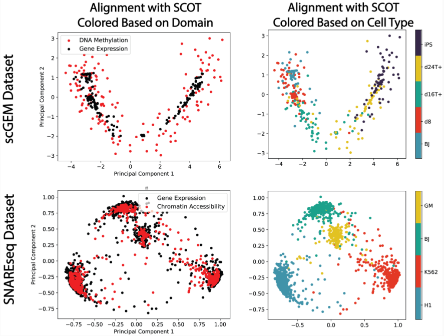
Proposition - alternate descent $\leftrightarrow$ solve UOT

For a fixed γ, $\pi \in \arg\min_\pi \mathcal{F}(\pi,\gamma) + \varepsilon \mathrm{KL}(\pi \otimes \gamma | (\alpha \otimes \beta)^{\otimes 2})$ 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 γ via a computable formula.

Back to genomic data alignment[8]



From Demetci et al.

- GW reaches state of the art performance for RNASeq data.
- UGW improves the classification performance over GW.
- See rsinghlab/SCOT on Github.

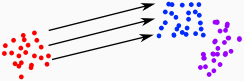
[8] Demetci, Pinar, et al. Gromov-Wasserstein optimal transport to align single-cell multi-omics data.

Application - Positive Unlabeled learning

- Domain adaptation = Propagate labels on a similar dataset.
- PU learning = supervised learning but we learn from only one class.
- **Idea:** Use a transport plan to map positive samples to unlabeled positive ones.

$\longrightarrow$





Performance results

Dataset	prior	Init (PW)	PGW	**UGW**	Dataset	prior	Init (FLB)	PGW	**UGW**
surf-C → surf-C	0.1	**89.9**	84.9	83.9	surf-C → decaf-C	0.1	85.0	85.1	**85.6**
surf-C → surf-A	0.1	81.8	82.2	**83.5**	surf-C → decaf-A	0.1	84.2	**87.1**	83.6
surf-C → surf-W	0.1	**81.9**	81.3	80.3	surf-C → decaf-W	0.1	86.2	**88.6**	86.8
surf-C → surf-D	0.1	80.0	81.4	**83.2**	surf-C → decaf-D	0.1	84.7	**91.1**	90.7
surf-C → surf-C	0.2	**79.7**	75.7	75.4	surf-C → decaf-C	0.2	74.8	75.6	**75.9**
surf-C → surf-A	0.2	65.6	66.0	**76.4**	surf-C → decaf-A	0.2	76.2	**87.9**	82.4
surf-C → surf-W	0.2	65.1	64.3	**67.3**	surf-C → decaf-W	0.2	81.5	88.4	**89.9**
decaf-C → decaf-C	0.1	**93.9**	83.0	86.8	decaf-C → surf-C	0.1	**81.7**	81.0	81.1
decaf-C → decaf-A	0.1	80.1	81.4	**85.6**	decaf-C → surf-A	0.1	80.9	81.2	**82.4**
decaf-C → decaf-W	0.1	80.1	82.7	**86.1**	decaf-C → surf-W	0.1	82.0	81.3	**83.5**
decaf-C → decaf-D	0.1	80.6	**83.8**	83.4	decaf-C → surf-D	0.1	80.0	80.8	**81.5**
decaf-C → decaf-C	0.2	**90.6**	76.7	80.5	decaf-C → surf-C	0.2	**66.6**	63.7	65.2
decaf-C → decaf-A	0.2	62.5	68.7	**74.7**	decaf-C → surf-A	0.2	62.9	62.4	**69.3**
decaf-C → decaf-W	0.2	65.7	75.9	**79.2**	decaf-C → surf-W	0.2	65.1	61.4	**83.3**

Table 1: Accuracy for all tasks. The left block are domain adaptation experiments with similar features, where both PGW and UGW are initialised with PW. The right block are domain adaptation experiments with different features, and the reported init is FLB used for UGW.

Conclusion

- 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

- $\mathrm{UOT}_{\varepsilon,\rho}$ is fast to compute but lost OT properties,
- Possible to "debias" $\mathrm{UOT}_{\varepsilon,\rho}$ to retrieve some of them.
- **Open question:** Can we debias UGW_ε? Which properties ?

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 !

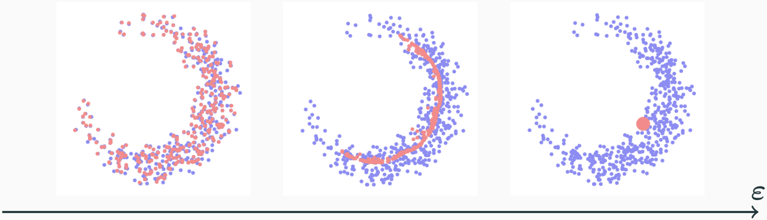
Correcting the entropic bias - Sinkhorn divergence

Entropic bias

Problem: $\mathcal{L} = \mathrm{OT}_\varepsilon$ does not retrieve β 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 ε.**

Unbalanced Sinkhorn Divergence

Definition

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

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

It extends the balanced Sinkhorn divergence[9] [10].

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

Is it positive ? Definite ? Smooth ?

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

[10] 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. $k_\varepsilon \overset{\text{def.}}{=} e^{-\frac{\mathrm{C}}{\varepsilon}}$ is a positive universal kernel, for any $\varepsilon > 0$

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

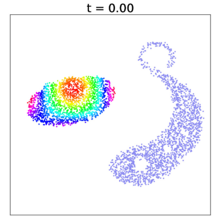
Corollary: holds for $\mathrm{C}(x, y) = \|\psi(x) - \psi(y)\|_2^2$, for ψ neural net.

Numerical insights on UOT and the Sinkhorn divergence

Numerical experiments model

Setting adapted from [Chizat '19][11].

- 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. θ



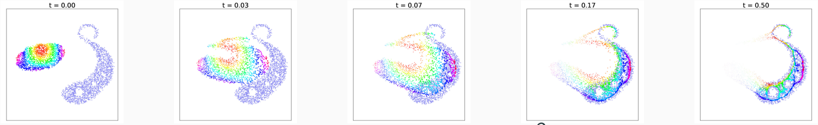
Updates of the coordinates

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

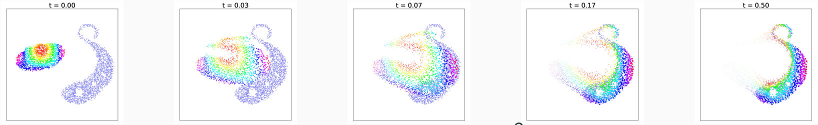
[11] Chizat, L. (2019). Sparse optimization on measures with over-parameterized gradient descent.

Numerics - Gradient descent

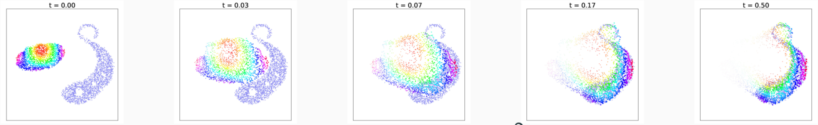
Parameters: $C(x, y) = \|x - y\|_2^2$ on $[0, 1]^2$, $\rho = 0.3$, $\eta_x = 60.0$, $\eta_r = 0.3$



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



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



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

Supplementary slides

Csiszàr divergences[12]

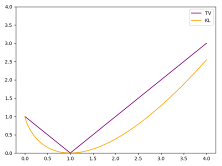
Define $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ l.s.c., convex, $\varphi(1) = 0$, $\varphi'^{\infty} = \lim_{x\to\infty} \frac{\varphi(x)}{x}$.
Write $\alpha = \sum_i \alpha_i \delta_{x_i}$ and $\beta = \sum_i \beta_i \delta_{x_i}$ (Same support (x_i))

Definition - φ-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:

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



[12] 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 γ, minimizing w.r.t. π amounts to solve a regularized UOT problem.

Algorithm

Algorithm 1 – UGW($\mathcal{X}$, $\mathcal{Y}$, ρ, ε)

Input: mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation ρ, regularization ε
Output: approximation (π, γ) minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

Initialize (π, γ) and (f, g)
while (π, γ) 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(μ, ν, $\tilde{c}$, $\tilde{\rho}$, $\tilde{\varepsilon}$)
    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 (π, γ).

Detailed algorithm

Algorithm 2 – UGW($\mathcal{X}$, $\mathcal{Y}$, ρ, ε)

Input: mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation ρ, regularization ε

Output: approximation (π, γ) minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

Initialize $\pi = \gamma = \mu \otimes \nu / \sqrt{m(\mu) m(\nu)}$, $g = 0$.

while (π, γ) has not converged **do**

    Update $\pi \leftarrow \gamma$, then $c \leftarrow c^{\varepsilon}_{\pi}$, $\tilde{\rho} \leftarrow m(\pi)\rho$, $\tilde{\varepsilon} \leftarrow m(\pi)\varepsilon$

    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 (π, γ).