Geometric Losses for Machine Learning - Sinkhorn Divergences for Unbalanced Optimal Transport

Thibault Séjourné

PGMO Seminar – 4th December, 2019

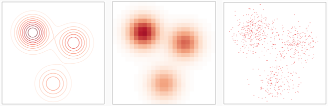
Joint work with Jean Feydy, Francois-Xavier Vialard, Alain Trouvé and Gabriel Peyré

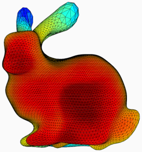
Outline

Introduction

Machine Learning setting

- Given an empirical measure β,
- And a model α_θ parametrized by θ.





Shape registration

Supervised Learning

Unsupervised Learning

- Then we optimize via GD & backpropagation a loss $\mathcal{L}$

$$\theta^* \in \arg\min_\theta \mathcal{L}(\alpha_\theta, \beta).$$

Which loss $\mathcal{L}$ should we use to introduce a geometric prior w.r.t. the data and compare weighted point clouds ?

From discrete to continuous setting (and back)

Definitions

- Continuous function: $f \in \mathcal{C}(\mathcal{X})$
- Positive measure: Linear form $\alpha \in \mathcal{M}_+(\mathcal{X})$
- Dual product: $\langle \alpha, f \rangle = \int_{\mathcal{X}} f(x) d\alpha(x) = \mathbb{E}_\alpha[f]$.

Discretizing measures

When $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i}$ one implements α on a computer with $(\alpha_i) \in \mathbb{R}^n$ and $(x_i) \in \mathbb{R}^{n \times d}$. Then functions are vectors $(f_i) = (f(x_i))_i \in \mathbb{R}^n$ and $\langle \alpha, f \rangle = \sum \alpha_i f_i$.

Small Take home message

Some algorithms are better understood using a continuous formalism.

Prerequisites of Loss functions

We require that the loss verifies at least the following axioms:

- Positivity: $\forall(\alpha, \beta)$, $\mathcal{L}(\alpha, \beta) \geq 0$.
- Definiteness: $\forall(\alpha, \beta)$, $\mathcal{L}(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$.
- Metrizing weak* convergence (convergence in law):

$$\forall(\alpha, \beta),\ \mathcal{L}(\alpha, \beta) \to 0 \Leftrightarrow \alpha \rightharpoonup \beta,$$

  where $\alpha \rightharpoonup \beta \Leftrightarrow \forall \mathrm{f} \in \mathcal{C}(\mathcal{X}),\ \langle \alpha, \mathrm{f} \rangle \to \langle \beta, \mathrm{f} \rangle$.
- Differentiability (for backpropagation).

Csiszar divergences

Csiszàr divergences

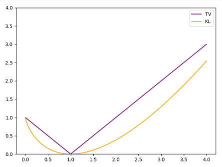
Definitions [Csiszàr'67]

- Entropy φ: nonnegative, l.s.c., convex on $\mathbb{R}_+$ s.t. $\varphi(1) = 0$
- Recession constant: $\varphi'^{\infty} = \lim_{x \to \infty} \varphi(x)/x$
- Lebesgue decomposition: $\forall(\alpha, \beta)$, $\alpha = \frac{d\alpha}{d\beta}\beta + \alpha^{\top}$
- φ-divergence: $D_\varphi(\alpha, \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi(\frac{d\alpha}{d\beta})d\beta + \varphi'^{\infty} \int_{\mathcal{X}} d\alpha^{\top}$

$\rightarrow$ Discretized: $D_\varphi(\alpha, \beta) = \sum_{\beta_i \neq 0} \varphi(\frac{\alpha_i}{\beta_i})\beta_i + \varphi'^{\infty} \sum_{\beta_i = 0} \alpha_i$

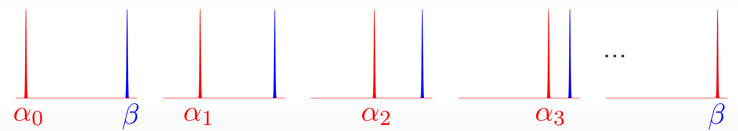
Examples:

- KL: $\varphi(x) = x \log x - x + 1$, $\varphi'^{\infty} = +\infty$,
- TV: $\varphi(x) = |x - 1|$ and $\varphi'^{\infty} = 1$.

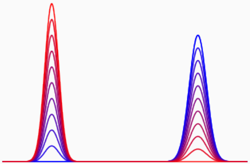


Properties of Csiszàr divergences

Consider the sequence $\alpha_n = \delta_{1/n}$ and $\beta = \delta_0$. One has $\alpha_n \rightharpoonup \beta$, but $\mathrm{KL}(\alpha_n|\beta) = \infty$ and $\mathrm{TV}(\alpha_n|\beta) = 2$.



- ☺ Simple and cheap to compute
- ☹ Ignores the geometry and do not metrize convergence in law



Optimal Transport

Optimal Transport Distance

$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \geq 0} \left\{ \langle \pi, \mathrm{C} \rangle \ : \ \begin{array}{l} \pi 1 = \alpha \\ \pi^\top 1 = \beta \end{array} \right\}.$$

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

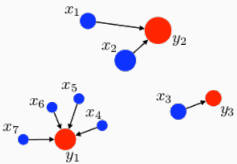
Discrete: $\langle \pi, \mathrm{C} \rangle = \sum_{\mathrm{i,j}} \pi_{\mathrm{ij}} \mathrm{C}_{\mathrm{ij}}$

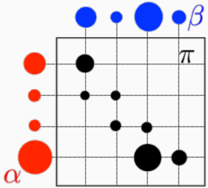
Intuition: Moving π_{ij} grams from $\mathrm{x_i}$ to $\mathrm{y_j}$ costs $\pi_{\mathrm{ij}} \times \mathrm{C}_{\mathrm{ij}} = \pi_{\mathrm{ij}} \times \mathrm{C}(\mathrm{x_i}, \mathrm{y_j})$.

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

$\Rightarrow$ Learn it !

[Kantorovich'42]



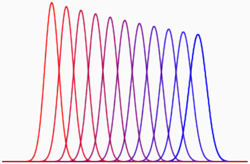


Properties of OT

- One has $\mathrm{OT}(\delta_{\mathrm{x}}, \delta_{\mathrm{y}}) = \mathrm{C}(\mathrm{x}, \mathrm{y})$

$\Rightarrow$ $\mathrm{OT}(\delta_{1/\mathrm{n}}, \delta_0) \xrightarrow{\mathrm{n}\to\infty} 0$

- Metric on $\mathcal{X}$ $\to$ metric on $\mathcal{M}_+(\mathcal{X})$



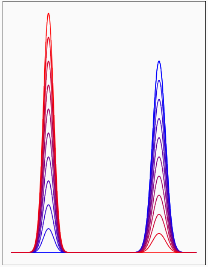
☺ Metrizes convergence in law

☹ Computation complexity $\mathcal{O}(\mathrm{n}^3 \log \mathrm{n})$, not differentiable

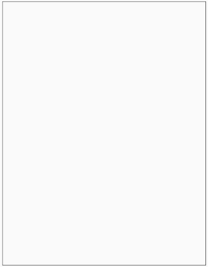
☹ Only compares probabilities, i.e. normalized weighted point clouds

Unbalanced Optimal Transport

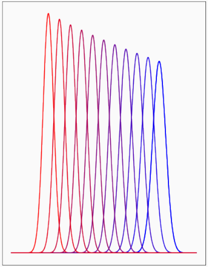
Hybridizing Vertical and Horizontal Geometries



Vertical



In between ?



Horizontal

Unbalanced optimal transport

Hybridizing $\Rightarrow$ Soften the hard constraint $\pi 1 = \alpha \rightarrow \rho \mathrm{D}_\varphi(\pi 1 | \alpha)$.

Allows for creation/destruction of mass locally.

Definition - Unbalanced OT [Liero, Mielke, Savaré '18]

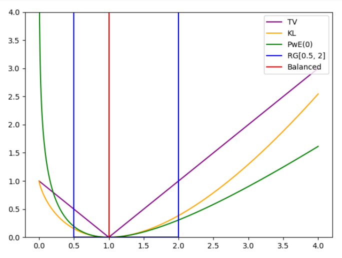
For any φ-divergence D_φ and any measures (α, β) one defines:

$$\mathrm{OT}_\rho(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \langle \pi, \mathrm{C} \rangle + \rho \mathrm{D}_\varphi(\pi_1, \alpha) + \rho \mathrm{D}_\varphi(\pi_2, \beta).$$

- Add a parameter ρ: Transport radius. ($\mathrm{OT}_\rho \xrightarrow[\rho \to +\infty]{} \mathrm{OT}$).
- Choice of D_φ: prior on the mass variation dynamics
- Balanced OT is retrieved with $\mathrm{D}_\varphi = \iota_{(=)}$

Examples of entropies

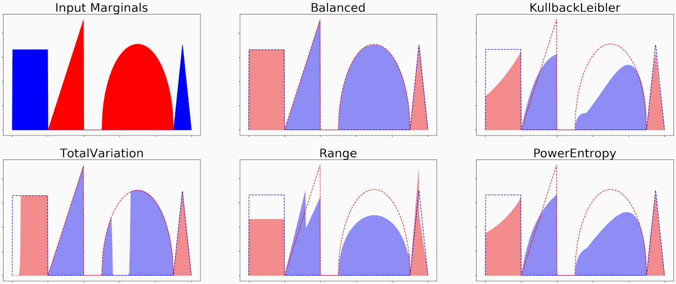
- Balanced: $\varphi(x) = \iota_{\{1\}}(x)$ with $D_\varphi(\pi_1, \alpha) = \iota_{(=)}(\pi_1, \alpha)$.
- TV: $\varphi(x) = |x - 1|$
- KL: $\varphi(x) = x \log x - x + 1$
- Power entropy: $\varphi(x) = \frac{1}{p(p-1)}(x^p - p(x-1) - 1)$, $p \in \mathbb{R}$.
- Range: $\varphi(x) = \iota_{[a,b]}(x)$ $(a \leq 1 \leq b)$, i.e $a\alpha \leq \pi_1 \leq b\alpha$.



Numerical examples

Reminder: Local mass creation and destruction is allowed

- Shows how α is matched onto β and vice versa through π.
- Plots $\pi_1 \approx \alpha$ and $\pi_2 \approx \beta$
- Input marginals are dashed.



Entropic Optimal Transport

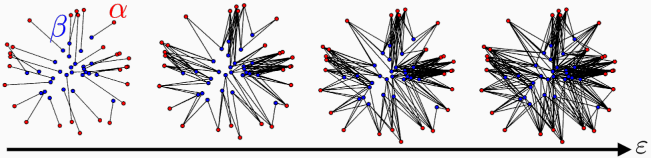
Regularization of OT

Reminder: OT is computationally expensive and non-smooth

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

Entropic Unbalanced OT [Cuturi'13, Chizat'18]

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \inf_{\pi \geq 0} \langle \pi, \mathrm{C} \rangle + \rho \mathrm{D}_\varphi(\pi_1, \alpha) + \rho \mathrm{D}_\varphi(\pi_2, \beta) \\ + \varepsilon \mathrm{KL}(\pi, \alpha \otimes \beta)$$



Duality of regularized OT

Writing $\varphi^*(x) = \sup_{y \geq 0} xy - \varphi(y)$, the dual reads

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \sup_{f,g \in \mathcal{C}(\mathcal{X})} \langle \alpha, -(\rho\varphi)^*(-f) \rangle + \langle \beta, -(\rho\varphi)^*(-g) \rangle - \varepsilon \langle \alpha \otimes \beta, e^{\frac{f(x)+g(y)-C(x,y)}{\varepsilon}} - 1 \rangle$$

The alternate dual ascent is straightforward to compute:

Alternate dual ascent

given any initialization $f_0 \mathcal{C}(\mathcal{X})$. At time t one has (f_t, g_t). Then

- Fix f_t and find optimal g in the dual $\to g_{t+1}$,
- Fix g_{t+1} and find optimal f in the dual $\to f_{t+1}$,
- Iterate until convergence.

Unbalanced Sinkhorn algorithm

Proposition - Unbalanced Sinkhorn algorithm

Define the following operators

- (Softmin / LogSumExp) $\mathrm{Smin}_{\alpha}^{\varepsilon}(\mathrm{f}) \stackrel{\text{def.}}{=} -\varepsilon \log\langle \alpha, e^{-\mathrm{f}/\varepsilon}\rangle$
- (Anisotropic Prox) $\mathsf{aprox}(\mathrm{p}) = \arg\min_{\mathrm{q}\in\mathbb{R}} \varepsilon e^{(\mathrm{p}-\mathrm{q})/\varepsilon} + \varphi^*(\mathrm{q})$

The optimality condition defines the Sinkhorn algorithm

$$\mathrm{g}_{t+1}(\mathrm{y}) = -\,\mathsf{aprox}(\,-\mathrm{Smin}_{\alpha}^{\varepsilon}\,(\mathrm{C}(.,\mathrm{y}) - \mathrm{f}_t)\,)$$

$$\mathrm{f}_{t+1}(\mathrm{x}) = -\,\mathsf{aprox}(\,-\mathrm{Smin}_{\beta}^{\varepsilon}\,(\mathrm{C}(\mathrm{x},.) - \mathrm{g}_{t+1})\,).$$

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

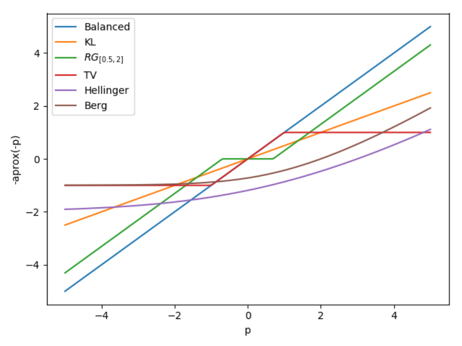
The Sinkhorn algorithm converges towards the optimal (f,g) of $\mathrm{OT}_\varepsilon(\alpha,\beta)$ when φ^* is strictly convex, but also for TV, Range and Balanced OT.

Examples of Anisotropic prox

- Balanced Sinkhorn = Softmin
- Unbalanced Sinkhorn = $\mathsf{aprox} \circ$Softmin

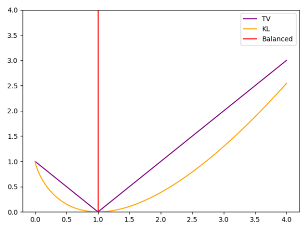
$\Rightarrow$ Unbalanced Sinkhorn = Readjusting Balanced Sinkhorn with the operator aprox.

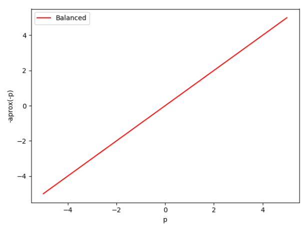
- Sinkhorn algorithm is a (weakly) contractive algorithm



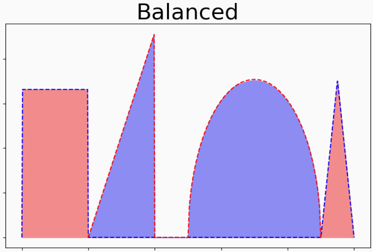
Examples of Anisotropic prox - Balanced

Entropy and Aprox



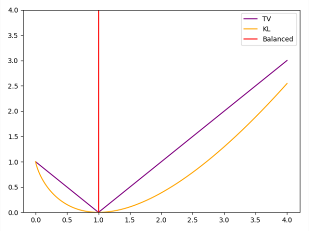


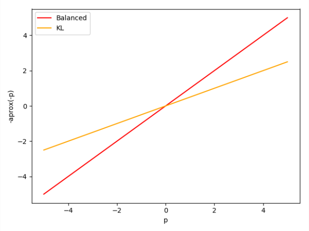
$$D_\varphi = \iota_{(=)}$$
$$\varphi(x) = \iota_{\{1\}}(x)$$
$$\mathrm{aprox}(x) = x$$



Examples of Anisotropic prox - KL

Entropy and Aprox

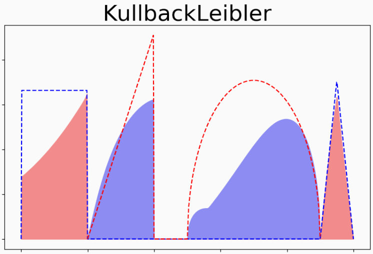




$$D_\varphi = \rho \mathrm{KL}$$

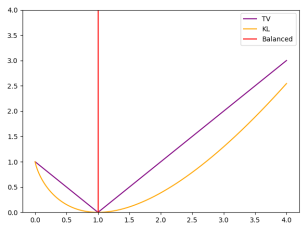
$$\varphi(\mathrm{x}) = \rho(\mathrm{x}\log \mathrm{x} - \mathrm{x} + 1)$$

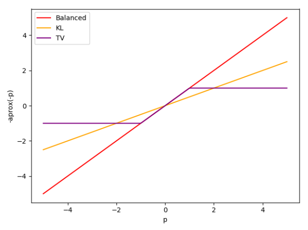
$$\mathsf{aprox}(\mathrm{x}) = \frac{\rho}{\rho+\varepsilon}\,\mathrm{x}$$



Examples of Anisotropic prox - TV

Entropy and Aprox

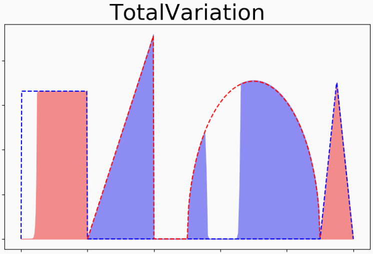




$$\mathrm{D}_\varphi = \rho \mathrm{TV}$$

$$\varphi(\mathrm{x}) = \rho|\mathrm{x} - 1|$$

$\mathsf{aprox}(\mathrm{x}) = \mathrm{x}$ if $\mathrm{x} \in [-\rho, \rho]$, ρ if $\mathrm{x} \geq \rho$ and $-\rho$ if $\mathrm{x} \leq -\rho$



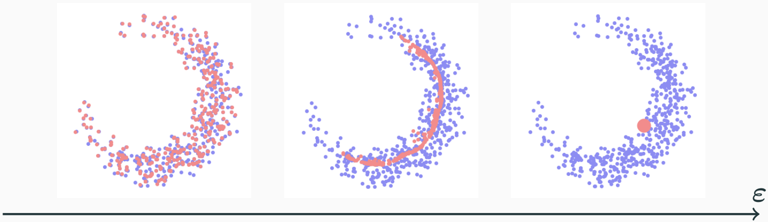
Correcting the entropic bias - Sinkhorn divergence

Entropic bias

Problem: OT_ε does not metrize weak* convergence for $\varepsilon > 0$. ☹

$$\exists \alpha \in \mathcal{M}_1^+(\mathcal{X}), \mathrm{OT}_\varepsilon(\alpha, \beta) < \mathrm{OT}_\varepsilon(\beta, \beta).$$

$$\mathrm{OT}_0(\alpha, \beta) \xleftarrow{0 \leftarrow \varepsilon} \mathrm{OT}_\varepsilon(\alpha, \beta) \xrightarrow{\varepsilon \to \infty} \alpha^\top \mathrm{C} \beta.$$



Main Theorem

Unbalanced Sinkhorn Divergence

Setting $\mathrm{m}(\mu)$ to be the total mass of the measure μ, we define

$$\begin{aligned} \mathrm{S}_{\varepsilon,\rho}(\alpha, \beta) \overset{\text{def.}}{=} & \mathrm{OT}_{\varepsilon,\rho}(\alpha, \beta) - \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\rho}(\alpha, \alpha) - \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\rho}(\beta, \beta) \\ & + \tfrac{\varepsilon}{2}(\mathrm{m}(\alpha) - \mathrm{m}(\beta))^2. \end{aligned}$$

It extends the balanced case from [Ramdas '15][Genevay '18].

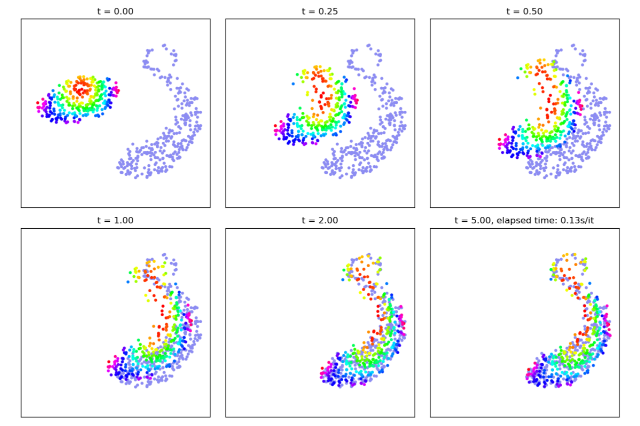
Theorem [S., Feydy, Vialard, Trouve, Peyre '19]

For any Lipschitz cost C s.t. $\mathrm{k}_\varepsilon \overset{\text{def.}}{=} \mathrm{e}^{-\frac{\mathrm{C}}{\varepsilon}}$ is a positive universal kernel, for any $\varepsilon > 0$ and strictly convex φ^*

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

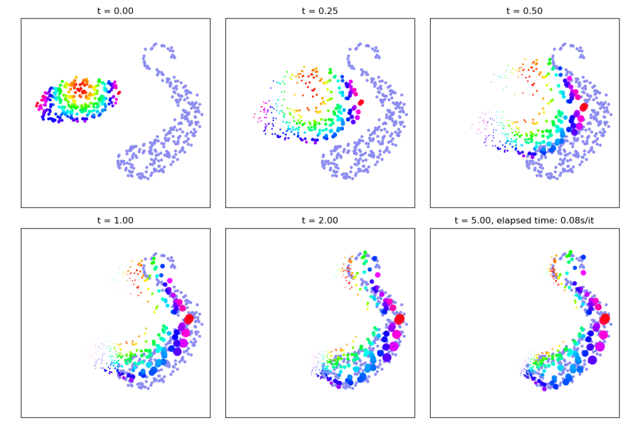
Numerics - Balanced Gradient Flow

- Model: $\alpha_\theta = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ with $\theta = (x_i)$
- Loss: S_ε with balanced OT and $\varepsilon = 0.01$



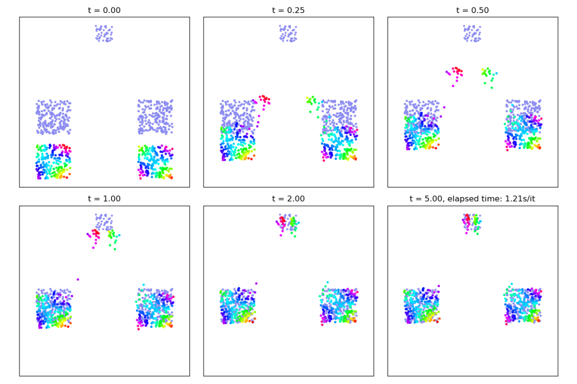
Numerics - Unbalanced Gradient Flow

- Model: $\alpha_\theta = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ with $\theta = (x_i, \alpha_i)$
- Loss: S_ε with KL UOT and $(\varepsilon, \rho) = (0.01, 0.3)$



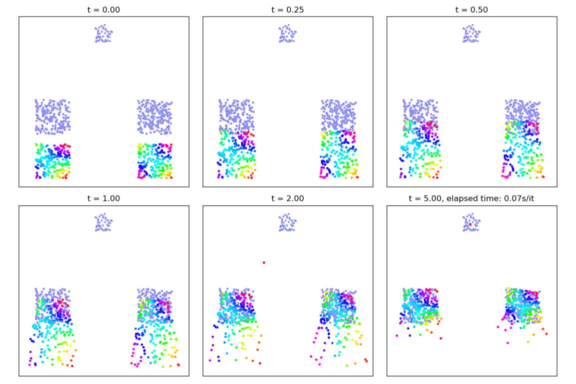
Numerics - Avoiding overfitting

- Model: $\alpha_\theta = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ with $\theta = (x_i)$
- Loss: S_ε with balanced OT and $\varepsilon = 0.01$



Numerics - Avoiding overfitting

- Model: $\alpha_\theta = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ with $\theta = (x_i)$
- Loss: S_ε with KL UOT and $(\varepsilon, \rho) = (0.01, 0.3)$



Unbalanced OT allows to avoid overfitting of outliers !

Implementation

- Sinkhorn divergences can be fastly computed via GPU-friendly routines

+ Efficient optimization heuristics (annealing + subsampling)

- Available losses (Balanced + KL) on Jean Feydy's repository:

  http://www.kernel-operations.io/geomloss/
- Two modes:
  - Keops backend for huge measures without overflow ($\sim$1 million points)
  - Mini-batch mode for machine learning.

- Implementation of other unbalanced divergences at: https://github.com/thibsej/unbalanced-ot-functionals

Conclusion

- Family of parametric losses with appealing properties (convexity, differentiability, positivity...)
- Algorithm with linear convergence
- Consistent behavior which allows to crossvalidate w.r.t. ε
- Improvement of the statistical complexity (Not detailed here)

It remains to experiment new ML applications!

http://www.kernel-operations.io/geomloss/
https://github.com/thibsej/unbalanced-ot-functionals