Sinkhorn Divergences for Unbalanced Optimal Transport

Thibault Séjourné

SMAI MODE – 7th September, 2020

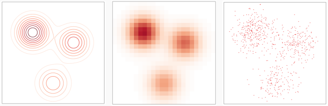
Joint work with Jean Feydy, Francois-Xavier Vialard, Alain Trouvé and Gabriel Peyré

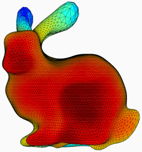
Outline

Introduction

Machine Learning setting with probabilities

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





Shape registration

Supervised Learning

Unsupervised Learning

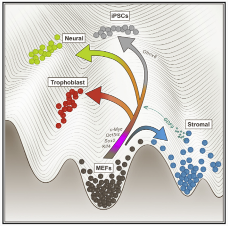
- Then we optimize via GD & backpropagation a loss $\mathcal{L}$

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

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

From probabilities to positive measures

- Most often measures are normalized to mass 1 (i.e. are probabilities).
- Sometimes too restrictive when:
  - → Normalizing data is unadapted.
  - → Dampening the geometric prior's importance is necessary (outliers).[1]
  - → Introducing mass variation dynamics is relevant.[2]



From Schiebinger et al.

[1] 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.

[2] Chizat, L., & Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport.

Prerequisites of Loss functions

We require that the loss verifies at least the following axioms for any $(\alpha, \beta) \in \mathcal{M}_+(\mathcal{X})$:

- **Positivity:** $\mathcal{L}(\alpha, \beta) \geq 0$.
- **Definiteness:** $\mathcal{L}(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$.
- **Convexity**.
- **Metrizing weak* convergence (convergence in law):**
$$\mathcal{L}(\alpha, \beta) \to 0 \Leftrightarrow \alpha \rightharpoonup \beta,$$
where $\alpha \rightharpoonup \beta \Leftrightarrow \forall \mathrm{f} \in \mathcal{C}(\mathcal{X}),\ \int_{\mathcal{X}} \mathrm{fd}\alpha \to \int_{\mathcal{X}} \mathrm{fd}\beta$.
- **Differentiability** (for backpropagation).

Csiszàr divergences

Csiszàr divergences

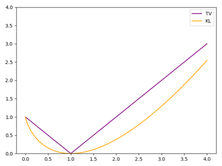
Definitions[3]

- **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:** $\forall(\alpha, \beta)$, $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^{\top}$
- **φ-divergence:** $\mathrm{D}_{\varphi}(\alpha, \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta})\mathrm{d}\beta + \varphi'^{\infty} \int_{\mathcal{X}} \mathrm{d}\alpha^{\top}$

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

Examples:

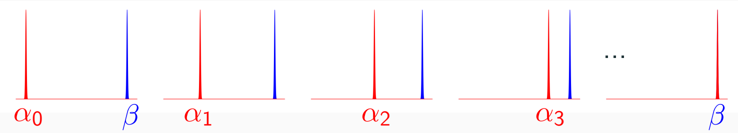
- KL: $\varphi(x) = x \log x - x + 1$, $\varphi'^{\infty} = +\infty$,
- TV: $\varphi(x) = |x - 1|$ and $\varphi'^{\infty} = 1$.



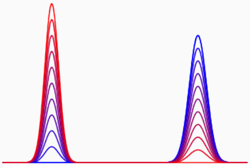
[3] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation.

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



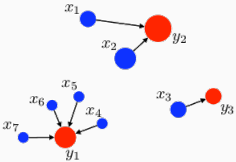
Optimal Transport

Optimal Transport (OT)

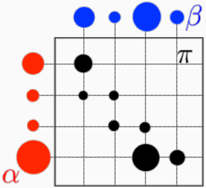
Balanced Optimal Transport Distance[4]

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

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



Discrete: $\int_{\mathcal{X}} \mathrm{C} d\pi = \sum_{i,j} \pi_{ij} \mathrm{C}_{ij}$
Intuition: Moving π_{ij} grams from x_i to y_j costs $\pi_{ij} \times \mathrm{C}_{ij} = \pi_{ij} \times \mathrm{C}(x_i, y_j)$.



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

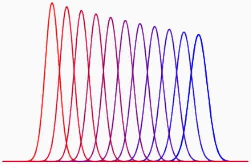
[4] Kantorovich, L. (1942). On the transfer of masses (in Russian).

Properties of OT

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

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

- Metric on $\mathcal{X}$ $\rightarrow$ metric on $\mathcal{M}_+^1(\mathcal{X})$



☺ Metrizes convergence in law

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

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

Unbalanced Optimal Transport

Unbalanced optimal transport

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

Definition - Unbalanced OT[5]

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

$$\mathrm{OT}_\rho(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \int_{\mathcal{X}} \mathrm{C} \mathrm{d}\pi + \rho \mathrm{D}_\varphi(\pi 1, \alpha) + \rho \mathrm{D}_\varphi(\pi^\top 1, \beta).$$

- **Intuition**: Hybridizing vertical and horizontal geometries
- **Transport radius ρ:** $\mathrm{OT}_\rho \xrightarrow[\rho \to +\infty]{} \mathrm{OT}_b$.
- **Choice of D_φ:** prior on the mass variation dynamics
- Balanced OT is retrieved with $\mathrm{D}_\varphi = \iota_{(=)}$

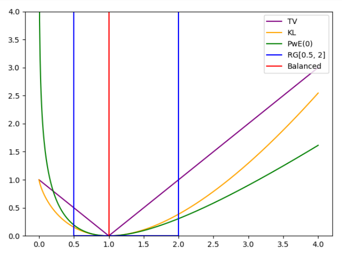
[5] Liero, M., Mielke, A., & Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures.

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**: $\varphi(x) = \frac{1}{p(p-1)}(x^p - p(x-1) - 1)$, $p \in \mathbb{R}$.

$\rightarrow$ Includes Hellinger and Berg entropies

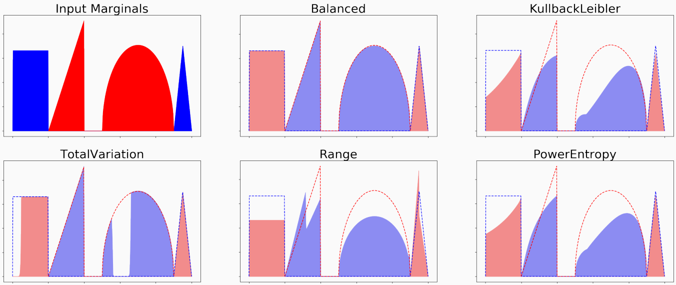
- **Range**: $\varphi(x) = \iota_{[a,b]}(x)$ $(a \leq 1 \leq b)$, i.e $a\alpha \leq \pi_1 \leq b\alpha$.



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 $\pi^\top 1 \approx \beta$
- Input marginals are dashed.



Entropic Optimal Transport

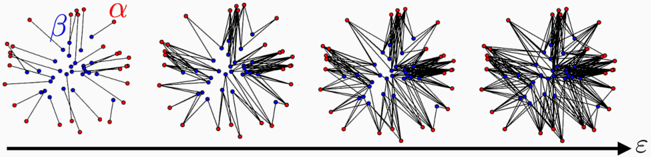
Regularization of OT

Reminder: OT is computationally expensive and non-smooth

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

Entropic Unbalanced OT[6] [7]

$$\mathrm{OT}_{\varepsilon,\rho}(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \int_{\mathcal{X}} \mathrm{C} d\pi + \rho \mathrm{D}_{\varphi}(\pi 1, \alpha) + \rho \mathrm{D}_{\varphi}(\pi^\top 1, \beta) + \varepsilon \mathrm{KL}(\pi, \alpha \otimes \beta)$$



[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 OT

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

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

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

Alternate dual ascent

Given any initialization $f_0 \in \mathcal{C}(\mathcal{X})$. 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 $\to f_{t+1}$.

Unbalanced Sinkhorn algorithm

Proposition - Unbalanced Sinkhorn algorithm

Define the following operators

- (Softmin / LogSumExp) $\text{Smin}^{\varepsilon}_{\alpha}(\mathrm{f}) \stackrel{\text{def.}}{=} -\varepsilon \log \left(\int_{\mathcal{X}} e^{-\mathrm{f}/\varepsilon} \mathrm{d}\alpha \right)$
- (Anisotropic Prox) $\text{aprox}(p) = \arg\min_{q \in \mathbb{R}} \varepsilon e^{(p-q)/\varepsilon} + \varphi^*(q)$

The optimality condition defines the Sinkhorn algorithm

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

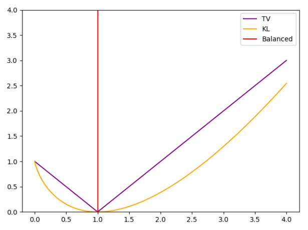
$$\mathrm{f}_{t+1}(x) = -\,\text{aprox}(\,-\text{Smin}^{\varepsilon}_{\beta}\,(\mathrm{C}(x,.) - \mathrm{g}_{t+1})\,).$$

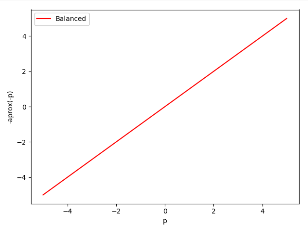
Theorem [S., Feydy, Vialard, Trouve, Peyre '19]

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

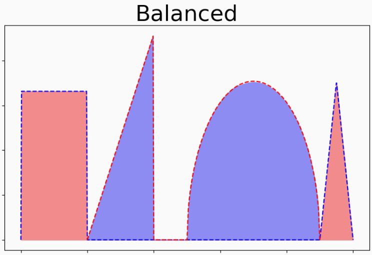
Examples of Anisotropic prox - Balanced

Entropy and Aprox



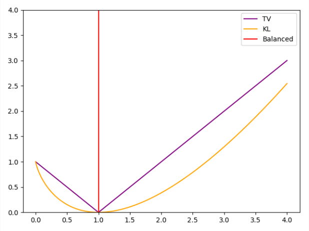


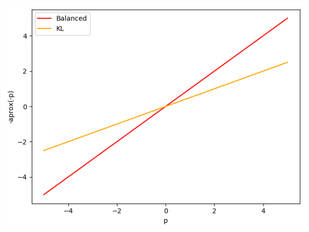
$$D_\varphi = \iota_{(=)}$$
$$\varphi(x) = \iota_{\{1\}}(x)$$
$$\text{aprox}(x) = x$$



Examples of Anisotropic prox - KL

Entropy and Aprox

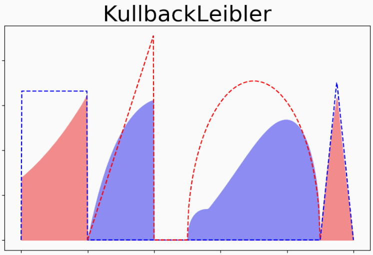




$$\mathrm{D}_\varphi = \rho \mathrm{KL}$$

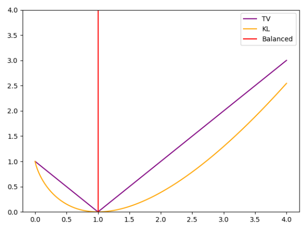
$$\varphi(x) = \rho(x \log x - x + 1)$$

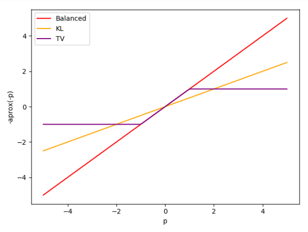
$$\mathrm{aprox}(x) = \frac{\rho}{\rho+\varepsilon} x$$



Examples of Anisotropic prox - TV

Entropy and Aprox

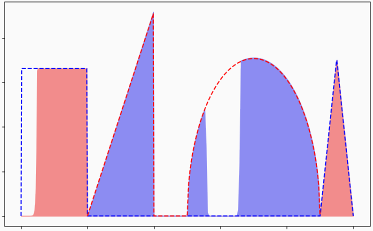




$$\mathrm{D}_\varphi = \rho \mathrm{TV}$$

$$\varphi(x) = \rho|x - 1|$$

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



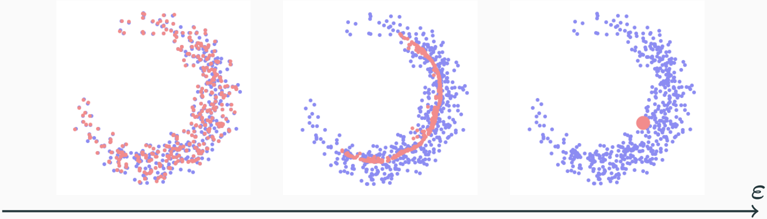
Correcting the entropic bias - Sinkhorn divergence

Entropic bias

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

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

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



Unbalanced Sinkhorn Divergence

Definition

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

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

It extends the balanced case[8] [9].

Remark: entropic bias + mass bias induced by $\varepsilon\mathrm{KL}(\pi|\alpha \otimes \beta)$.

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

- For any entropy, $S_{\varepsilon,\rho}$ is convex, positive, definite.
- For φ^* differentiable and strictly convex, it is (weakly) differentiable.
- For $D_\varphi = \rho \mathrm{KL}$ one has $S_{\varepsilon,\rho}(\alpha, \beta) \to 0 \Leftrightarrow \alpha \rightharpoonup \beta$.

Numerical highlights

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 $\theta \mapsto \alpha(\theta) = \sum_i^n r_i^2 \delta_{x_i}$
- Flow $\partial_t \theta(t) = -\nabla_\theta \mathrm{S}_{\varepsilon,\rho}(\alpha(\theta), \beta)$

Updates of the coordinates

$$x_i^{(t+1)} = x_i^{(t)} - \eta_x \nabla_{x_i} \mathrm{S}_{\varepsilon,\rho}(\alpha(\theta^{(t)}), \beta), \tag{1}$$

$$r_i^{(t+1)} = r_i^{(t)} . \exp\big(- 2\eta_x \nabla_{r_i} \mathrm{S}_{\varepsilon,\rho}(\alpha(\theta^{(t)}), \beta)\big) \tag{2}$$

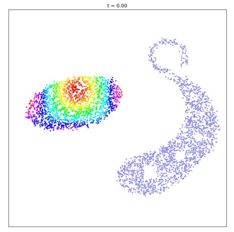
[10] Chizat, L. (2019). Sparse optimization on measures with over-parameterized gradient descent.

Numerics 1

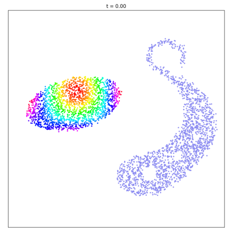
Parameters:

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

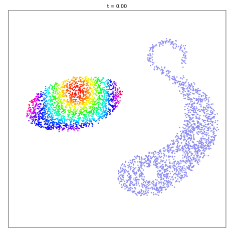
$\mathcal{L} = \mathrm{OT}_{\varepsilon,\rho}$, $\varepsilon = 10^{-3}$



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



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

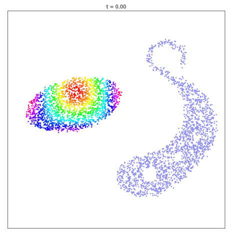


Numerics 2

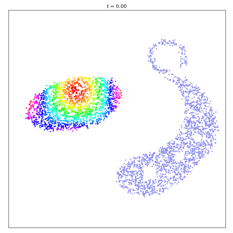
Parameters:

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

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



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



Conclusion

- Family of parametric losses with appealing properties (convexity, differentiability, positivity...)
- Algorithm with linear convergence
- Several parameters to crossvalidate (ε,ρ,φ)
- Improvement of the statistical complexity (Not detailed here)
- Implementations available:

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

It remains to experiment new ML applications!