Sinkhorn Divergences for Unbalanced Optimal Transport

Thibault Séjourné

Mokameeting – 6th April, 2020

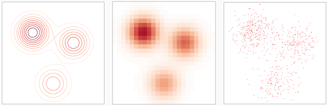
Joint work with Jean Feydy, Francois-Xavier Vialard, Alain Trouvé and Gabriel Peyré

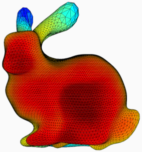
Outline

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 compare probability measures ?

Desired properties of $\mathcal{L}$:

- Positive, definite and convex
- Metrizes the weak* convergence
  $\alpha_\mathrm{n} \rightharpoonup \alpha \Leftrightarrow \forall \mathrm{f} \in \mathcal{C}(\mathcal{X}), \int \mathrm{fd}\alpha_\mathrm{n} \to \int \mathrm{fd}\alpha.$
- Differentiable

Possible losses between measures:

- Csiszar divergences (KL, TV, Hellinger, etc...)
- Maximum mean discrepancies / kernel norms
- Optimal transport distances

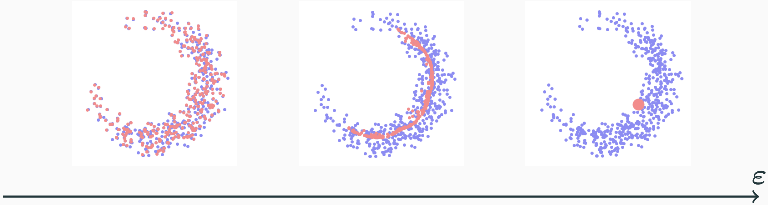
Entropic bias

OT issues: non-smooth + complexity + curse of dimensionality

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \langle \pi, \mathrm{C} \rangle + \varepsilon \mathrm{KL}(\pi, \alpha \otimes \beta)$$

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

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



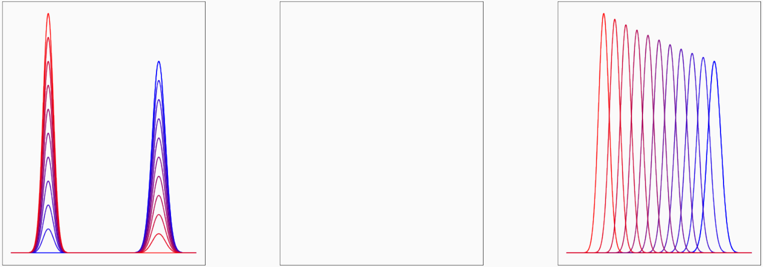
$\Rightarrow$ Debias: $\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta, \beta)$

Unbalanced OT

Goal of Unbalanced OT

Mitigate between vertical and horizontal geometries on $\mathcal{X}$.

+ avoids normalizing data and geometric outliers.



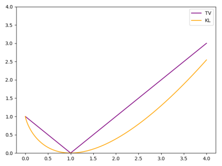
Vertical (Csiszar) | In between ? | Horizontal (OT)

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: $D_{\varphi}(\alpha, \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi(\frac{d\alpha}{d\beta}) d\beta + \varphi'^{\infty} \int_{\mathcal{X}} d\alpha^{\top}$

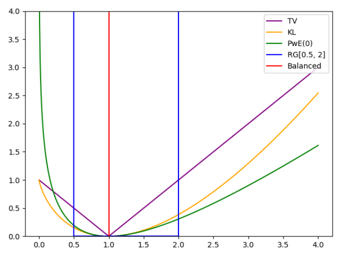
Examples:

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



Examples of entropies

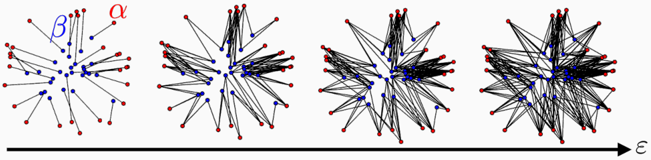
- Balanced: $\varphi(x) = \iota_{\{1\}}(x)$ with $D_\varphi(\pi_1, \alpha) = \iota_{(=)}(\pi_1, \alpha)$.
- TV: $\varphi(x) = |x - 1|$
- KL: $\varphi(x) = x \log x - x + 1$
- Power entropy: $\varphi(x) = \frac{1}{p(p-1)}(x^p - p(x-1) - 1)$, $p \in \mathbb{R}$.
- Range: $\varphi(x) = \iota_{[a,b]}(x)$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



Entropically regularized Unbalanced OT

Entropic Unbalanced OT [Chizat'18]

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

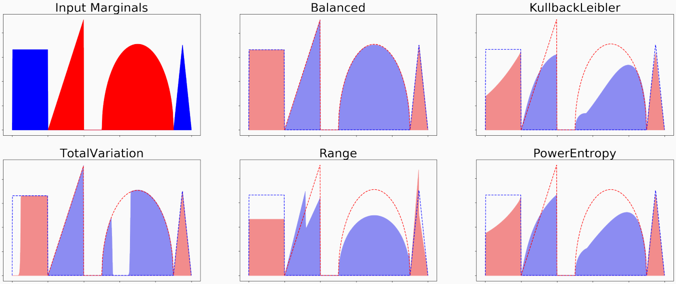


choice of C and D_φ = priors on geometry and mass dynamics

Numerical examples ($\varepsilon = 0$)

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.



Sinkhorn divergence

Unbalanced Sinkhorn Divergence

Definition

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

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

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

Impact of KL: entropic bias + mass bias ($\mathrm{m}(\pi) \to \mathrm{m}(\alpha \otimes \beta)$).

Bound on the Sinkhorn divergence

Proposition

Assuming φ^* strictly convex, denoting $\mathrm{k}_\varepsilon \overset{\text{def.}}{=} \mathrm{e}^{-\frac{\mathrm{C}}{\varepsilon}}$ and f_α and g_β the optimal symmetric potentials of $\mathrm{OT}_\varepsilon(\alpha, \alpha)$ and $\mathrm{OT}_\varepsilon(\beta, \beta)$ respectively, one has

$$\mathrm{S}_{\varepsilon,\rho}(\alpha, \beta) \geq \frac{\varepsilon}{2} \|\alpha \mathrm{e}^{\frac{\mathrm{f}_\alpha}{\varepsilon}} - \beta \mathrm{e}^{\frac{\mathrm{g}_\beta}{\varepsilon}}\|^2_{\mathrm{k}_\varepsilon}.$$

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

For any Lipschitz cost C s.t. $\mathrm{k}_\varepsilon \stackrel{\text{def.}}{=} \mathrm{e}^{-\frac{\mathrm{C}}{\varepsilon}}$ 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.
- $\mathrm{S}_{\varepsilon,\rho}(\alpha, \beta) \to 0 \Leftrightarrow \alpha \rightharpoonup \beta$.

Berg's Theorem: $\mathrm{e}^{-\mathrm{C}/\varepsilon}$ positive kernel $\Leftrightarrow -\mathrm{C}$ positive kernel.

Sinkhorn algorithm

Duality of regularized OT

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

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

Proposition - Unbalanced Sinkhorn algorithm

Define the following operators

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

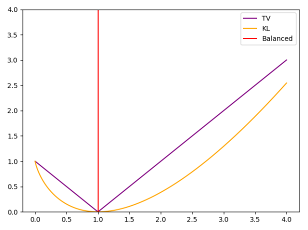
The optimality condition defines the Sinkhorn algorithm

$$g_{t+1}(y) = -\,\mathrm{aprox}(\,-\mathrm{Smin}_{\alpha}^{\varepsilon}\,(C(.,y) - f_t)\,)$$

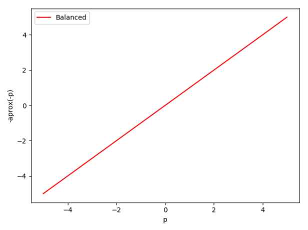
$$f_{t+1}(x) = -\,\mathrm{aprox}(\,-\mathrm{Smin}_{\beta}^{\varepsilon}\,(C(x,.) - g_{t+1})\,).$$

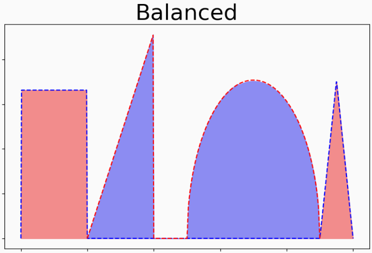
Examples of Anisotropic prox - Balanced

Entropy and Aprox



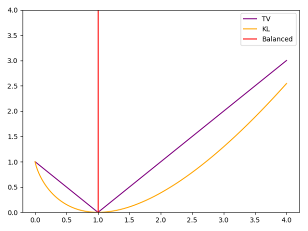
$$D_\varphi = \iota_{(=)}$$
$$\varphi(x) = \iota_{\{1\}}(x)$$
$$\mathsf{aprox}(x) = x$$

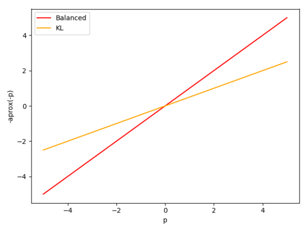




Examples of Anisotropic prox - KL

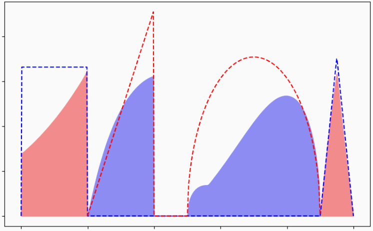
Entropy and Aprox



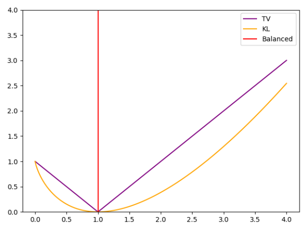


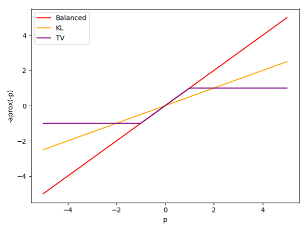
$$D_\varphi = \rho \mathrm{KL}$$
$$\varphi(x) = \rho(x \log x - x + 1)$$
$$\mathrm{aprox}(x) = \frac{\rho}{\rho+\varepsilon} x$$

KullbackLeibler



Entropy and Aprox

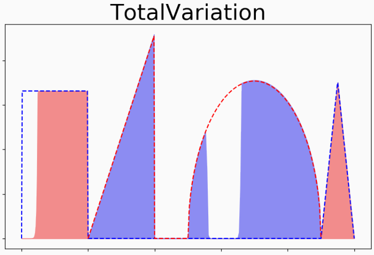




$$D_\varphi = \rho \mathrm{TV}$$

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

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



Stability of Softmin and Aprox

Proposition

One has for any $(\mathrm{f}, \mathrm{g}) \in \mathcal{C}(\mathcal{X})$

$$\left|\mathrm{Smin}_{\alpha}^{\varepsilon}(\mathrm{f}) - \mathrm{Smin}_{\alpha}^{\varepsilon}(\mathrm{g})\right| \leq \|\mathrm{f} - \mathrm{g}\|_{\infty}. \tag{1}$$

$$\left\|\mathsf{aprox}_{\varphi^*}^{\varepsilon}(\mathrm{f}) - \mathsf{aprox}_{\varphi^*}^{\varepsilon}(\mathrm{g})\right\|_{\infty} \leq \|\mathrm{f} - \mathrm{g}\|_{\infty}. \tag{2}$$

$\Rightarrow$ The algorithm is numerically stable. If there exists a fixed point and compactness, the algorithm then converges linearly towards it.

Convergence of Sinkhorn algorithm

Assume either:

- φ^* strictly convex and $\partial\varphi^*$ goes to zero or $+\infty$ as x goes to 0 or $+\infty$.
- The entropy corresponds to Balanced, TV or Range.

Theorem - Existence and convergence

Assume C is Lipschitz on a compact space $\mathcal{X}$. Then:

1. The space of dual potentials can be restricted to a relatively compact set, thus there is existence of dual maximizers in $\mathcal{C}(\mathcal{X})$.
2. The Sinkhorn algorithm converges towards fixed points which are dual maximizers.

Numerical illustrations - Gradient flows

Numerical experiments model

Setting adapted from [Chizat '19].

- Position/mass parameterization $x = \{(x_i, r_i)_i\} \in (\mathbb{R}^d \times \mathbb{R})^n$
- Model measure $x \mapsto \alpha(x) = \sum_i^n r_i^2 \delta_{x_i}$
- Cone metric $\langle (x_1, r_1), (x_2, r_2) \rangle_{(x,r)} = \frac{\eta_x}{r^2} \langle x_1, x_2 \rangle_x + \eta_r r_1 r_2$
- Flow $\nabla x(t) = -\nabla_x S_{\varepsilon,\rho}(\alpha(x), \beta)$

Updates of the coordinates

$$x_i^{(t+1)} = x_i^{(t)} - \eta_x \nabla_{x_i} S_{\varepsilon,\rho}(\alpha^{(t)}, \beta), \quad (3)$$

$$r_i^{(t+1)} = r_i^{(t)} . \exp\left(-2\eta_x \nabla_{r_i} S_{\varepsilon,\rho}(\alpha^{(t)}, \beta)\right) \quad (4)$$

Plots - debiasing effect

OT_ε-KL

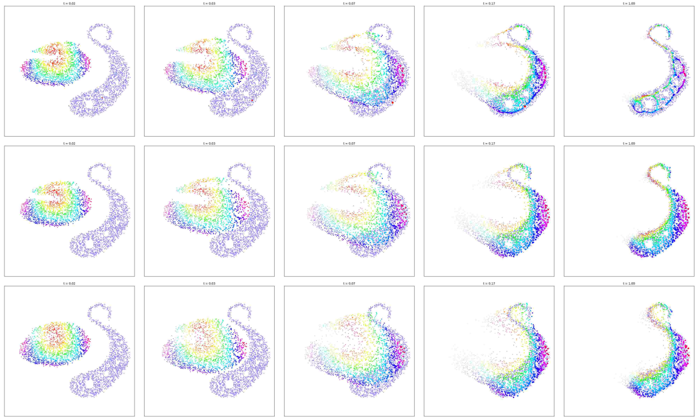
$(10^{-3}, 0.3)$

$\mathrm{S}_{\varepsilon,\rho}$-KL

$(10^{-3}, 0.3)$

$\mathrm{S}_{\varepsilon,\rho}$-KL

$(10^{-2}, 0.3)$



Plots - With other entropies

$S_{\varepsilon,\rho}$-Bal

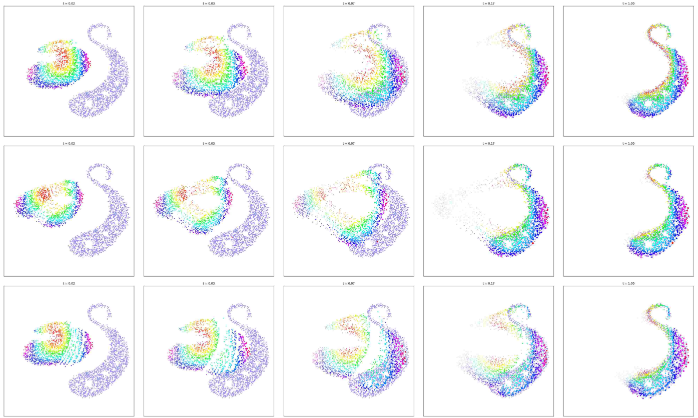
(10^{-3})

$S_{\varepsilon,\rho}$-TV

$(10^{-3}, 0.1)$

$S_{\varepsilon,\rho}$-RG

$(10^{-3}, [0.7, 1.3])$



Conclusion

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

http://www.kernel-operations.io/geomloss/

https://github.com/thibsej/unbalanced-ot-functionals