From Entropy-Regularized OT to Sinkhorn Divergences

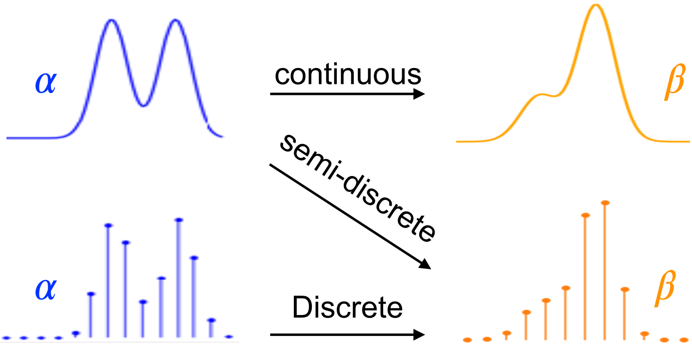
Aude Genevay

MIT CSAIL

OTML Worskshop - NeurIPS 2019

Joint work with Francis Bach, Lénaïc Chizat , Marco Cuturi, Gabriel Peyré

Comparing Probability Measures



Discrete Setting (Quantization)

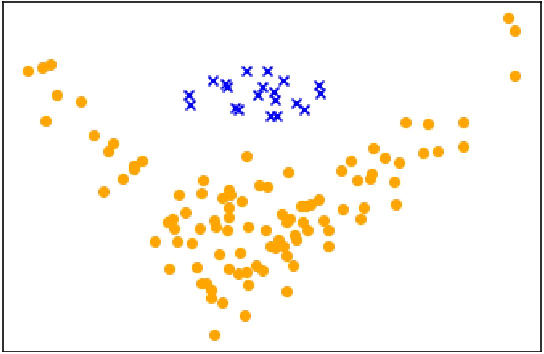


Figure 1 – $\min_{(x_1,\dots,x_k)} \mathcal{D}\left(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j\right)$

Discrete Setting (Quantization)

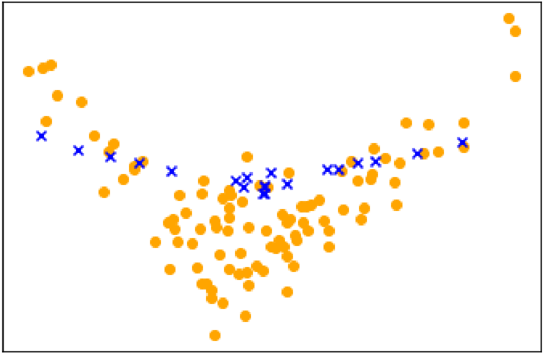


Figure 1 – $\min_{(x_1,\dots,x_k)} \mathcal{D}(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j)$

Discrete Setting (Quantization)

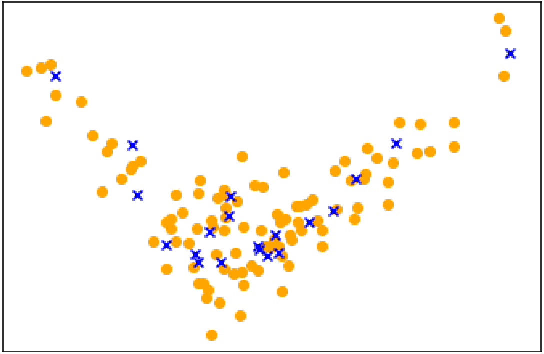


Figure 1 – $\min_{(x_1,\ldots,x_k)} \mathcal{D}(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j)$

Discrete Setting (Quantization)

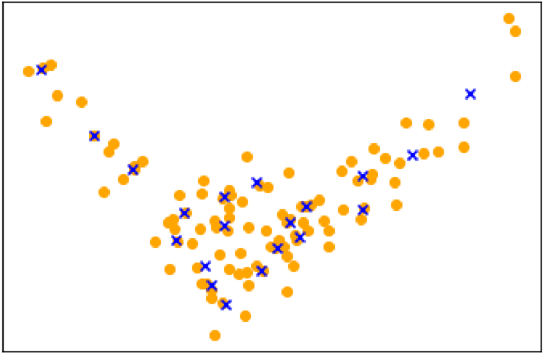


Figure 1 – $\min_{(x_1,\dots,x_k)} \mathcal{D}\left(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j\right)$

Semi-discrete Setting (Density Fitting)

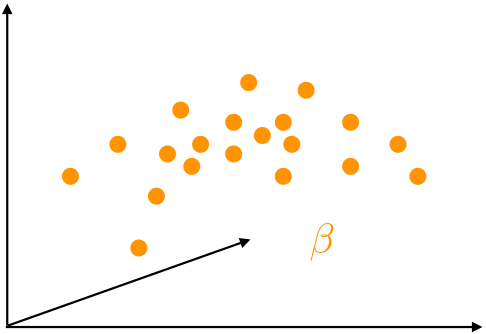


Figure 2 – $\min_\theta \mathcal{D}(\alpha_\theta, \beta)$

Semi-discrete Setting (Density Fitting)

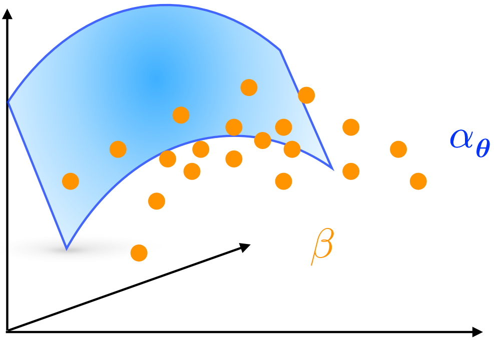


Figure 2 – $\min_\theta \mathcal{D}(\alpha_\theta, \beta)$

Semi-discrete Setting (Density Fitting)

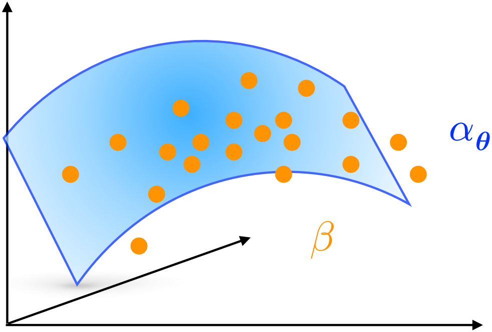


Figure 2 – $\min_\theta \mathcal{D}(\alpha_\theta, \beta)$

Semi-discrete Setting (Density Fitting)

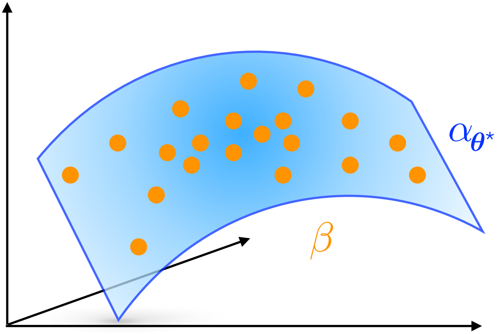


Figure 2 – $\min_\theta \mathcal{D}(\alpha_\theta, \beta)$

1. Notions of Distance between Measures
2. Entropic Regularization of Optimal Transport
3. Sinkhorn Divergences : Interpolation between OT and MMD
4. Conclusion

φ-divergences (Czisar '63)

Definition (φ-divergence)

Let φ convex l.s.c. function such that $\varphi(1) = 0$, the φ-divergence D_φ between two measures α and β is defined by :

$$D_\varphi(\alpha|\beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi\Big(\frac{\mathrm{d}\alpha(x)}{\mathrm{d}\beta(x)}\Big)\mathrm{d}\beta(x).$$

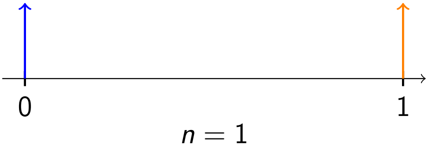
Example (Kullback Leibler Divergence)

$$D_{KL}(\alpha|\beta) = \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\right)\mathrm{d}\alpha(x) \quad \leftrightarrow \quad \varphi(x) = x\log(x)$$

Weak Convergence of measures

Example

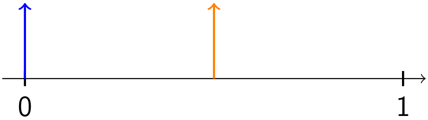
On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.



Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.

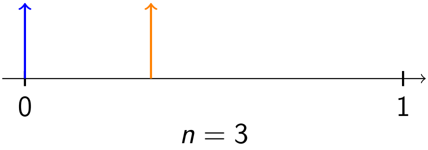


$n = 2$

Weak Convergence of measures

Example

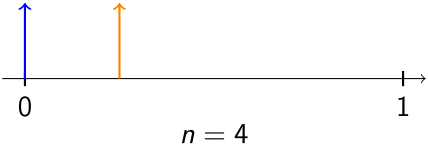
On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.



Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.



Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.

$n = 5$

Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.

$n = 6$

Weak Convergence of measures

Example

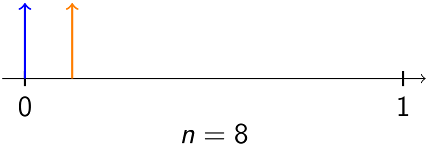
On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.

$n = 7$

Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.



Weak Convergence of measures

Example

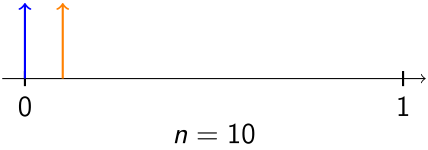
On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.

$n = 9$

Weak Convergence of measures

Example

On $\mathbb{R}$, $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n}$: $D_{KL}(\alpha_n|\alpha) = +\infty$.



Definition (Weak Convergence)

α_n **weakly converges** to α, (denoted $\alpha_n \rightharpoonup \alpha$)
$\Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{D}$ distance between measures , $\mathcal{D}$ **metrises weak convergence** IFF$\left(\mathcal{D}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Max. Mean Discrepancies (Gretton '06)

Definition (RKHS)

Let $\mathcal{H}$ a Hilbert space with kernel k, then $\mathcal{H}$ is a Reproduicing Kernel Hilbert Space (RKHS) IFF :

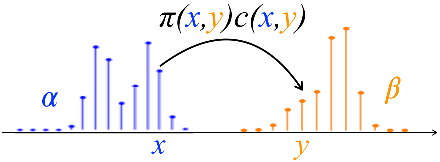
1. $\forall x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H},$
2. $\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}.$

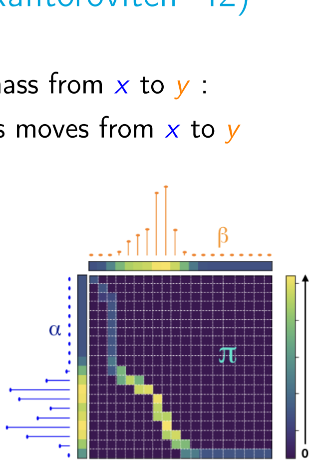
Let $\mathcal{H}$ a RKHS avec kernel k, the distance **MMD** between two probability measures α and β is defined by :

$$
\begin{aligned}
MMD_k^2(\alpha, \beta) \quad &\stackrel{\text{def.}}{=} \quad \left(\sup_{\{f \mid \|f\|_{\mathcal{H}} \leqslant 1\}} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))| \right)^2 \\
&= \quad \mathbb{E}_{\alpha \otimes \alpha}[k(X, X')] + \mathbb{E}_{\beta \otimes \beta}[k(Y, Y')] \\
&\qquad -2\mathbb{E}_{\alpha \otimes \beta}[k(X, Y)].
\end{aligned}
$$

Optimal Transport (Monge 1781, Kantorovitch '42)

- $c(x, y)$: cost of moving a unit of mass from x to y :
- $\pi(x, y)$ (coupling) : how much mass moves from x to y





The Wasserstein Distance

Minimal cost of moving ALL the mass from α to β ?

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_c(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \qquad (\mathcal{P})$$

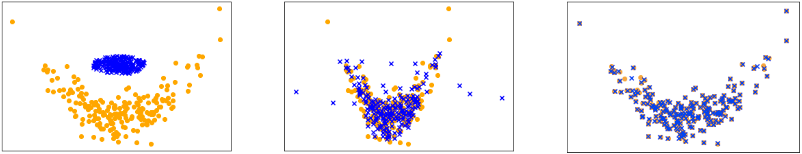
For $c(x, y) = \|x - y\|_2^p$, $W_c(\alpha, \beta)^{1/p}$ is the **p-Wasserstein distance**.

Optimal Transport vs. MMD

	MMD	**OT**
sample complexity	$(\frac{1}{\sqrt{n}})$	$O(n^{-1/d})$ (curse of dimension)
computation	$O(n^2)$	$O(n^3 \log(n))$

Optimal Transport vs. MMD

	MMD	**OT**
sample complexity	$(\frac{1}{\sqrt{n}})$	$O(n^{-1/d})$ (curse of dimension)
computation	$O(n^2)$	$O(n^3 \log(n))$
		better gradients !



Initial Setting — MMD_k - k = - $\|\cdot\|_2^{1.5}$ — W_c - c = $\|\cdot\|_2^{1.5}$

$\underset{(x_1,\dots,x_k)}{min} \mathcal{D}(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j)$ after 200 steps of grad. descent.

Entropic Regularization (Cuturi '13)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_c \ (\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y) \qquad (\mathcal{P})$$

Entropic Regularization (Cuturi '13)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_{c,\varepsilon}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi|\alpha \otimes \beta), \quad (\mathcal{P}_\varepsilon)$$

where

$$H(\pi|\alpha \otimes \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\pi(x,y).$$

relative entropy of the transport plan π with respect to the product measure $\alpha \otimes \beta$.

Entropic Regularization

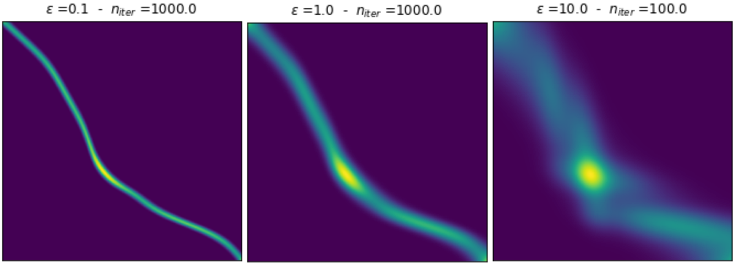


Figure 3 – Influence of the regularization parameter ε on the transport plan π.

The entropic penalty smoothes the coupling matrix, yielding fuzzy assignments.

Dual Formulation

Convex dual of standard OT : constrained dual problem

$$W_c\ (\alpha, \beta) = \max_{\substack{u \in \mathcal{C}(\mathcal{X}) \\ v \in \mathcal{C}(\mathcal{Y})}} \int_{\mathcal{X}} u(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y) \qquad (\mathcal{D})$$

such that $\quad \{u(x) + v(y) \leqslant c(x, y)\, \forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}\}$

Dual Formulation

Convex dual of regularized OT : unconstrained dual problem

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{\substack{u \in \mathcal{C}(\mathcal{X}) \\ v \in \mathcal{C}(\mathcal{Y})}} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon. \tag{$\mathcal{D}_\varepsilon$}$$

Iterative algorithm : alternate between optimizing over u with fixed v and optimizing over v with fixed u.

Sinkhorn's Algorithm

When $\alpha = \sum_{i=1}^{n} \boldsymbol{\alpha}\delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \boldsymbol{\beta}\delta_{y_j}$

Sinkhorn's Algorithm

Let $\mathsf{K}_{ij} = e^{-\frac{c(x_i,y_j)}{\varepsilon}}, \mathbf{a} = e^{\frac{\mathbf{u}}{\varepsilon}}, \mathbf{b} = e^{\frac{\mathbf{v}}{\varepsilon}}$.

$$\mathbf{a}^{(\ell+1)} = \frac{1}{\mathsf{K}(\mathbf{b}^{(\ell)} \odot \boldsymbol{\beta})} \quad ; \quad \mathbf{b}^{(\ell+1)} = \frac{1}{\mathsf{K}^T(\mathbf{a}^{(\ell+1)} \odot \boldsymbol{\alpha})}$$

Complexity of each iteration : $O(n^2)$ (matrix vector multiplications)
Linear convergence, constant degrades when $\varepsilon \to 0$.
Bonus : Fully differentiable with auto-diff tools (e.g TensorFlow)
$\Rightarrow$ differentiable approximation of OT ! (Salimans et al., G.P.C. '18)

The 'sample complexity'

Informal Definition

*Given a distance between measures , its **sample complexity** corresponds to the error made when approximating this distance with samples of the measures.*

Known cases :

- OT : $\mathbb{E}|W(\alpha, \beta) - W(\hat{\alpha}_n, \hat{\beta}_n)| = O(n^{-1/d})$
  $\Rightarrow$ curse of dimension (Dudley '84, Weed and Bach '18)
- MMD : $\mathbb{E}|MMD(\alpha, \beta) - MMD(\hat{\alpha}_n, \hat{\beta}_n)| = O(\frac{1}{\sqrt{n}})$
  $\Rightarrow$ independent of dimension (Gretton '06)

What about $\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)|$?

'Sample Complexity' of W_ε.

Theorem (G., C., B., C., P. '19) (Mena, Weed '19)

Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ bounded , and $c \in \mathcal{C}^\infty$ L-Lipschitz. Then

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{1}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}\right) \qquad \text{when } \varepsilon \to 0$$

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{1}{\sqrt{n}}\right) \qquad \text{when } \varepsilon \to +\infty.$$

where constants depend on $|\mathcal{X}|$, $|\mathcal{Y}|$, d, and $\left\|c^{(k)}\right\|_\infty$ pour $k = 0 \dots \lfloor d/2 \rfloor + 1$.

$\rightarrow$ A large enough regularization breaks the curse of dimension.

Discrete gradient flow of W_ε, $\varepsilon = 1$

The effect of entropy

Entropic Transport is Maximum Likelihood under Gaussian noise (Rigollet Weed '18)

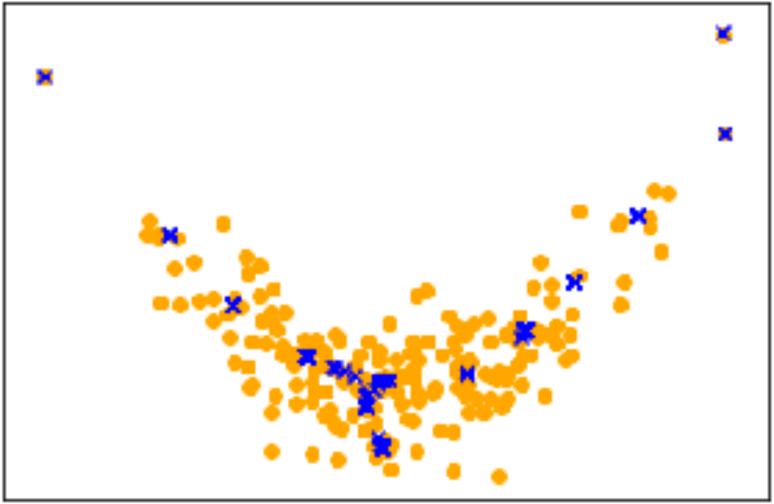
Consider a sample $(x_1, \dots, x_n) \sim X$ from the model

$$X = Y + \zeta \quad \text{where } Y \sim \alpha_\theta,\ \zeta \sim \mathcal{N}(0, \varepsilon)$$

. Then,

$$\hat{\theta}^{MLE} = min_\theta W_\varepsilon(\alpha_\theta, \frac{1}{n}\sum_{i=1}^{n} \delta x_i)$$

The effect of entropy



Sinkhorn Divergences

'Issue' of regularized Wass. Distance : $W_{c,\varepsilon}(\alpha, \alpha) \neq 0$
Proposed Solution : introduce corrective terms to 'debias' regularized Wasserstein distance.

Definition (Sinkhorn Divergences)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$SD_{c,\varepsilon}(\alpha, \beta) \overset{\text{def.}}{=} W_{c,\varepsilon}(\alpha, \beta) - \frac{1}{2} W_{c,\varepsilon}(\alpha, \alpha) - \frac{1}{2} W_{c,\varepsilon}(\beta, \beta),$$

Interpolation Property

Theorem (G., Peyré, Cuturi '18), (Ramdas and al. '17)

Sinkhorn Divergences have the following asymptotic behavior :

$$\text{when } \varepsilon \to 0, \quad SD_{c,\varepsilon}(\alpha, \beta) \to W_c(\alpha, \beta), \tag{1}$$

$$\text{when } \varepsilon \to +\infty, \quad SD_{c,\varepsilon}(\alpha, \beta) \to \frac{1}{2} MMD^2_{-c}(\alpha, \beta). \tag{2}$$

Remark : To get an MMD, $-c$ must be positive definite. For $c = \| \cdot \|_2^p$ with $0 < p < 2$, the MMD is called Energy Distance.

Discrete gradient flow of SD_ε, $\varepsilon = 1$

Discrete gradient flow of *MMD*

Summary

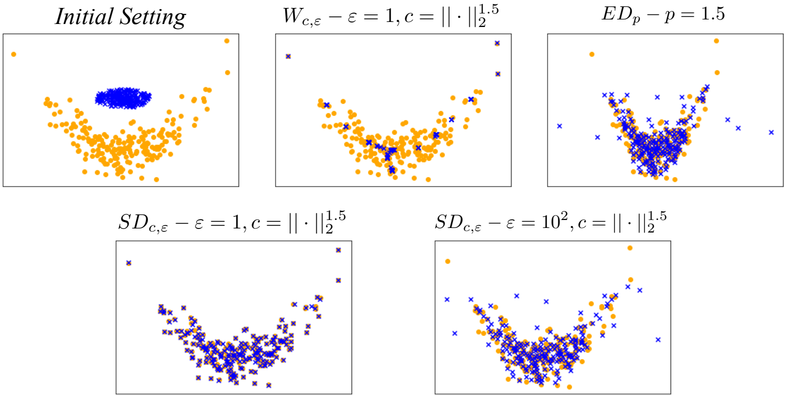


Figure 4 – $\min_{(x_1,\ldots,x_k)} \mathcal{D}(\frac{1}{k}\sum_{i=1}^k \delta x_i, \frac{1}{n}\sum_{i=1}^n \delta y_j)$.

Take Home Message

Sinkhorn Divergences are a great notion of distance between measures !

- 'debias' regularized Wasserstein Distance
- interpolate between OT (small ε) and MMD (large ε) and get the best of both worlds :
    - inherit geometric properties from OT
    - break curse of dimension for ε large enough
- fast algorithms for implementation in ML tasks