

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○○
○○○○○○○○

Conclusion

Bridging the gap between Optimal Transport and MMD with Sinkhorn Divergences

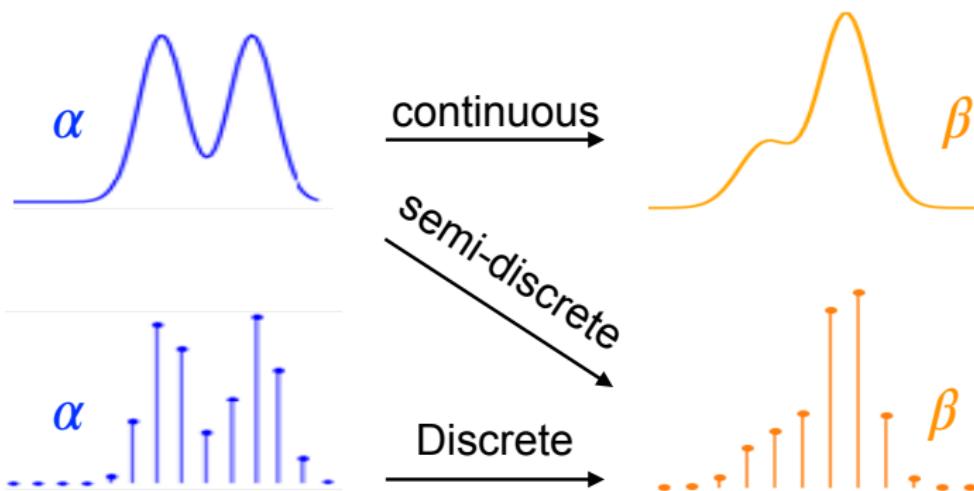
Aude Genevay

MIT CSAIL

UCL CSML Seminar - Feb. 2020

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

Comparing Probability Measures



Discrete Setting (Quantization)

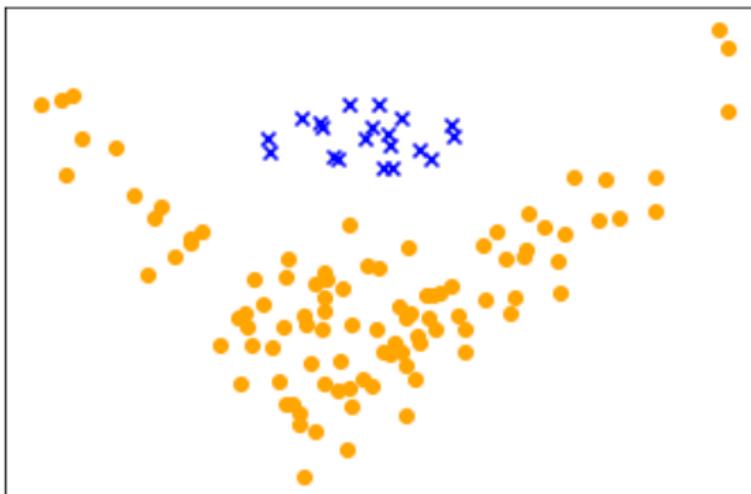


Figure 1 – $\min_{(x_1, \dots, x_k)} \mathcal{D}\left(\frac{1}{k} \sum_{i=1}^k \delta_{x_i}, \frac{1}{n} \sum_{j=1}^n \delta_{y_j}\right)$

Discrete Setting (Quantization)

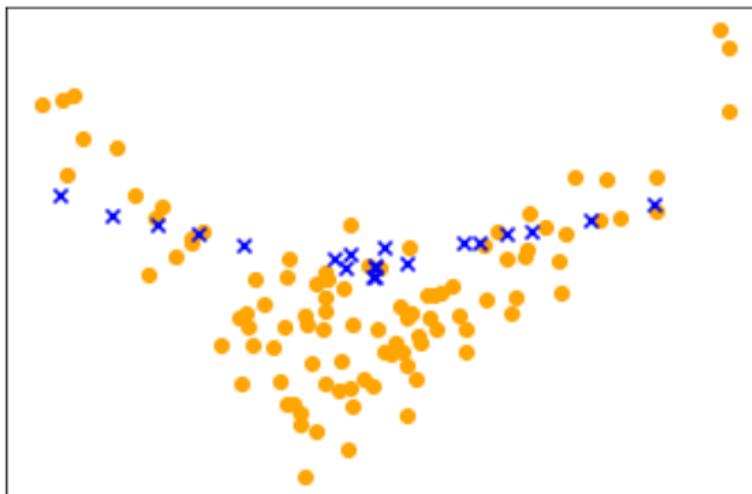


Figure 1 – $\min_{(x_1, \dots, x_k)} \mathcal{D}\left(\frac{1}{k} \sum_{i=1}^k \delta x_i, \frac{1}{n} \sum_{j=1}^n \delta y_j\right)$

Discrete Setting (Quantization)

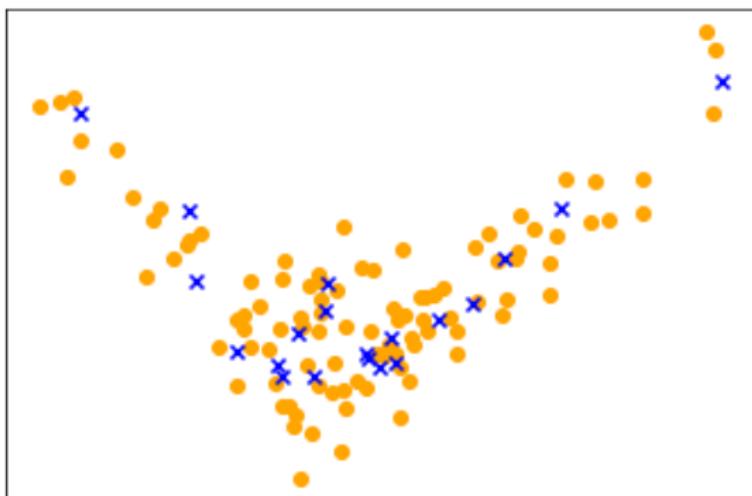


Figure 1 – $\min_{(x_1, \dots, x_k)} \mathcal{D}\left(\frac{1}{k} \sum_{i=1}^k \delta x_i, \frac{1}{n} \sum_{j=1}^n \delta y_j\right)$

Discrete Setting (Quantization)

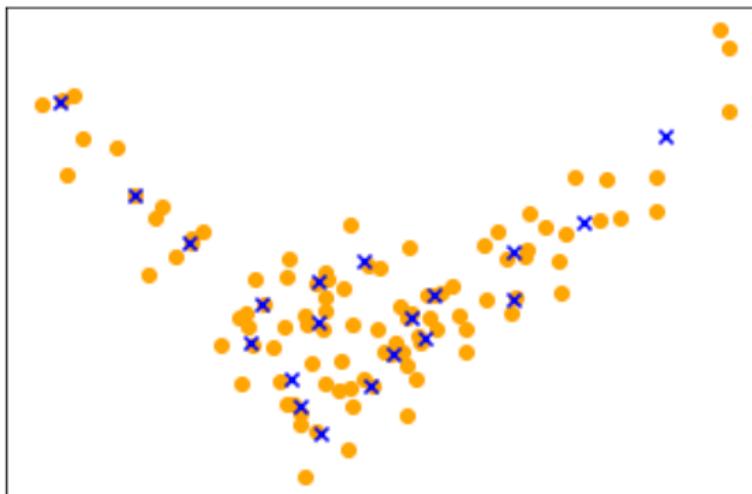


Figure 1 – $\min_{(x_1, \dots, x_k)} \mathcal{D}\left(\frac{1}{k} \sum_{i=1}^k \delta x_i, \frac{1}{n} \sum_{j=1}^n \delta y_j\right)$

Semi-discrete Setting (Density Fitting)

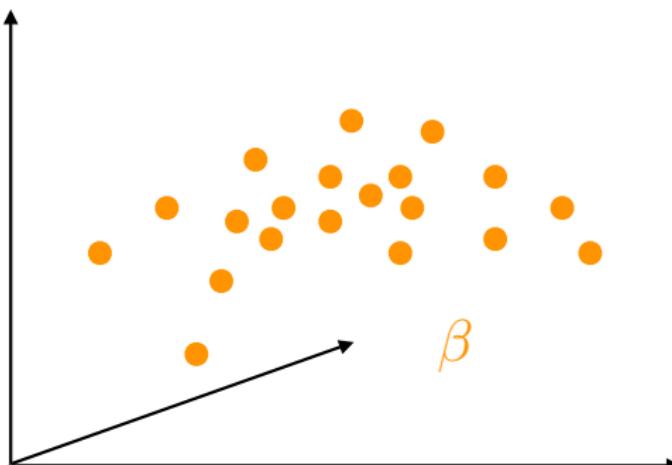


Figure 2 – $\min_{\theta} \mathcal{D}(\alpha_{\theta}, \beta)$

Semi-discrete Setting (Density Fitting)

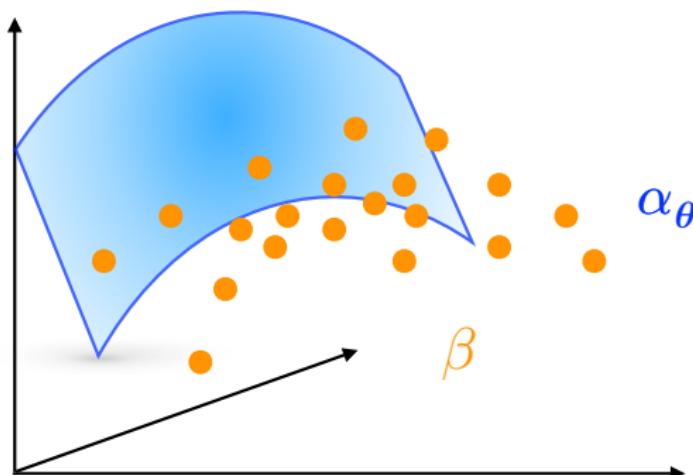


Figure 2 – $\min_{\theta} \mathcal{D}(\alpha_{\theta}, \beta)$

Semi-discrete Setting (Density Fitting)

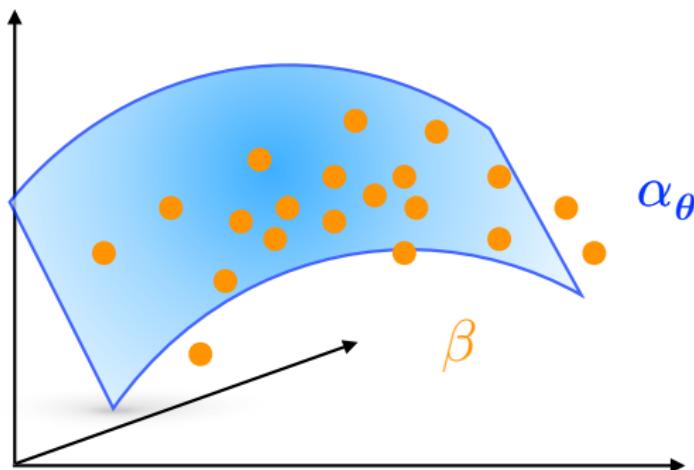


Figure 2 – $\min_{\theta} \mathcal{D}(\alpha_{\theta}, \beta)$

Semi-discrete Setting (Density Fitting)

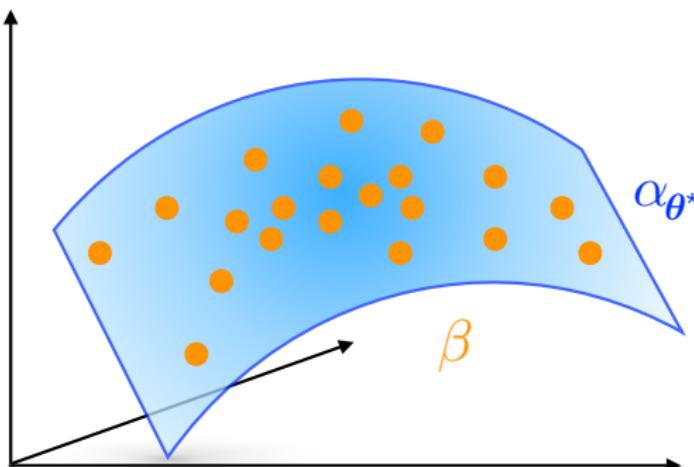


Figure 2 – $\min_{\theta} \mathcal{D}(\alpha_{\theta}, \beta)$

Distances

Entropic Regularization

oooo
o
ooo

Sinkhorn Divergences

oooooo
oooooooo

Conclusion

- 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 :

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

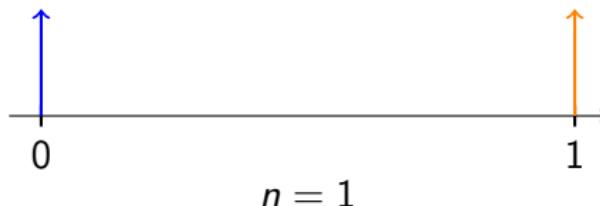
Example (Kullback Leibler Divergence)

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

Weak Convergence of measures

Example

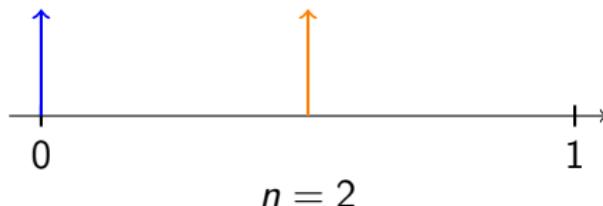
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

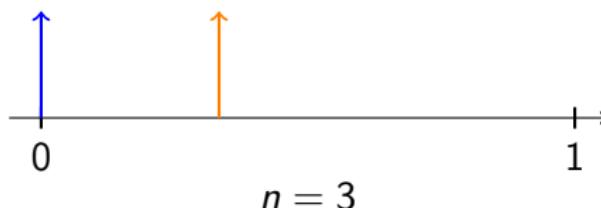
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

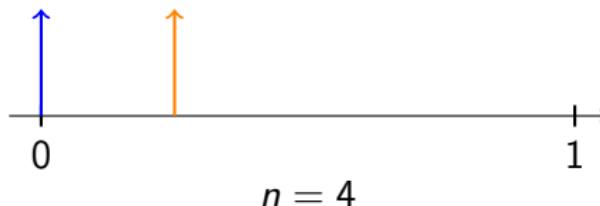
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

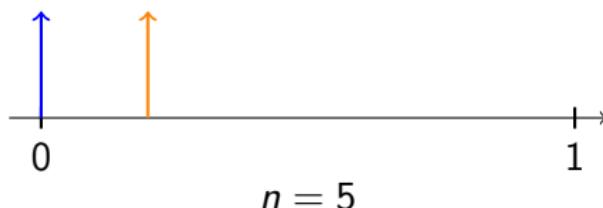
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

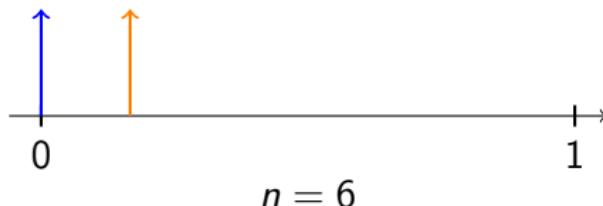
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

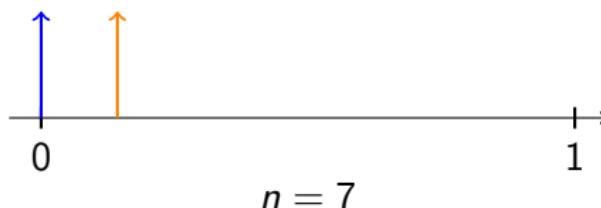
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

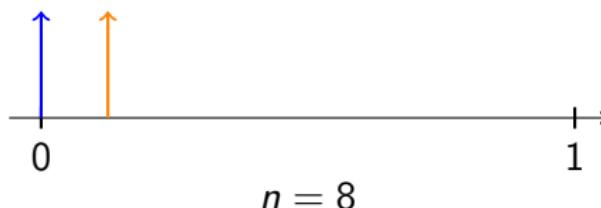
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

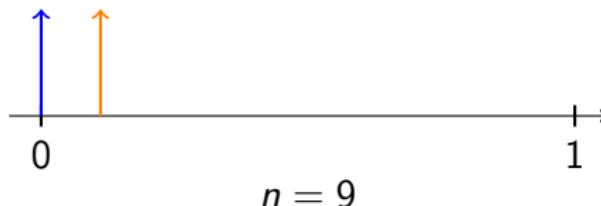
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

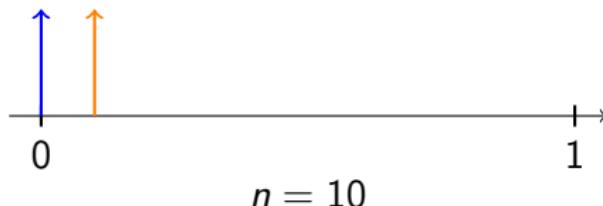
On \mathbb{R} , $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{KL}(\alpha_n | \alpha) = +\infty$.



Weak Convergence of measures

Example

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



Definition (Weak Convergence)

α_n weakly converges to α , (denoted $\alpha_n \rightharpoonup \alpha$)

$\Leftrightarrow \int f(x) d\alpha_n(x) \rightarrow \int f(x) d\alpha(x) \forall f \in \mathcal{C}_b(\mathcal{X})$.

Let \mathcal{D} distance between measures , \mathcal{D} metrises weak convergence IFF $(\mathcal{D}(\alpha_n, \alpha) \rightarrow 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha)$.

Maximum Mean Discrepancies (Gretton '06)

Definition (RKHS)

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

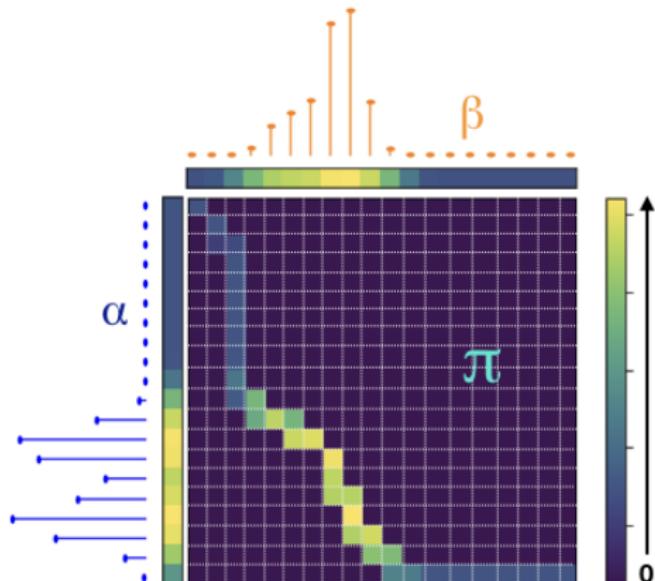
- ① $\forall x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H},$
- ② $\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}.$

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

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

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

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

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

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○
○○○○○○○○

Conclusion

Optimal Transport vs. MMD

MMD**OT**

sample complexity

$$\left(\frac{1}{\sqrt{n}}\right)$$

$$O(n^{-1/d})$$

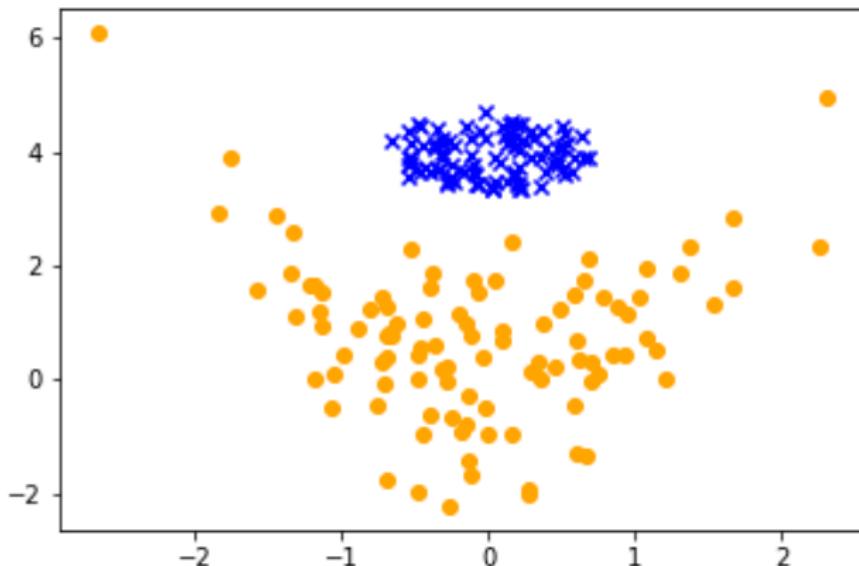
(curse of dimension)

computation

$$O(n^2)$$

$$O(n^3 \log(n))$$

Simple example



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

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○
○○○○○○○○

Conclusion

Discrete gradient flow of *MMD*

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○
○○○○○○○○

Conclusion

Discrete gradient flow of OT

Distances

Entropic Regularization

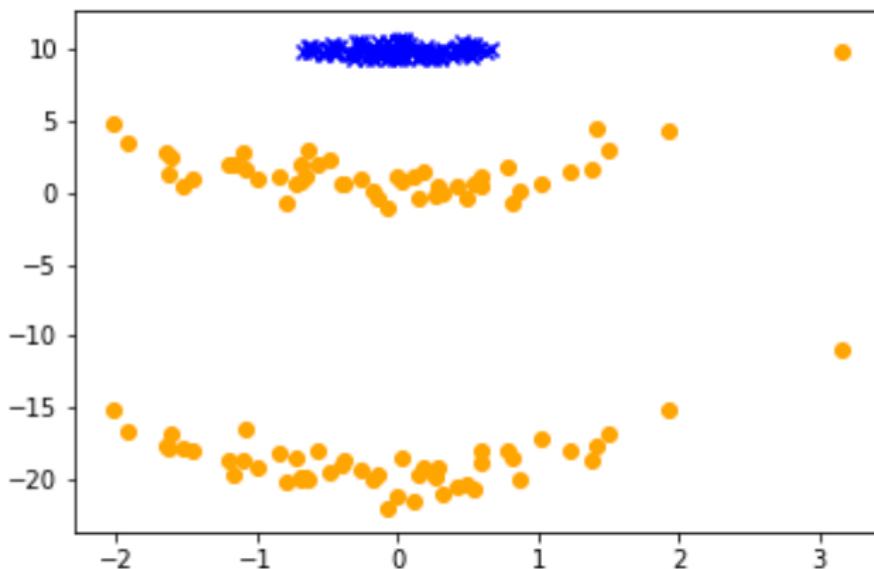
oooo
o
ooo

Sinkhorn Divergences

oooooo
oooooooo

Conclusion

Another example



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

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○
○○○○○○○○

Conclusion

Discrete gradient flow of *MMD*

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○
○○○○○○○○

Conclusion

Discrete gradient flow of OT

Distances

Entropic Regularization

○○○○
○○○
○○○

Sinkhorn Divergences

○○○○○
○○○○○○○○

Conclusion

Optimal Transport vs. MMD

MMD

OT

sample complexity

$$(\frac{1}{\sqrt{n}})$$

$$O(n^{-1/d})$$

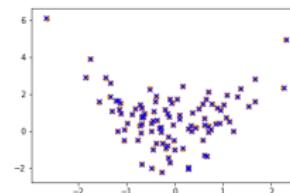
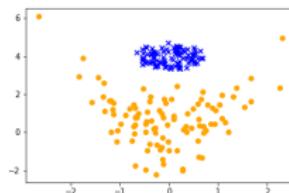
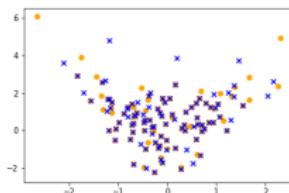
(curse of dimension)

computation

$$O(n^2)$$

$$O(n^3 \log(n))$$

better gradients!



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

① Notions of Distance between Measures

② Entropic Regularization of Optimal Transport

The basics

A magic regularizing tool !

Sample Complexity

③ Sinkhorn Divergences : Interpolation between OT and MMD

④ Conclusion

The basics

Entropic Regularization (Cuturi '13)

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

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

The basics

Entropic Regularization (Cuturi '13)

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

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

where

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

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

The basics



Entropic Regularization

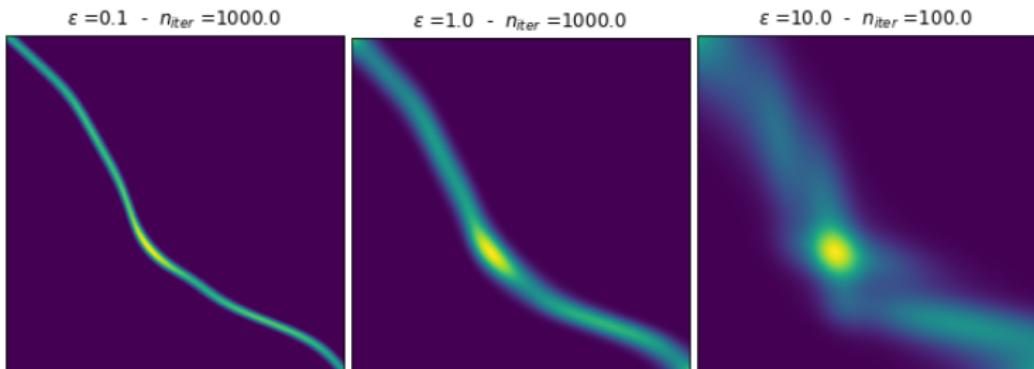


Figure 3 – Influence of the regularization parameter ε on the transport plan π .

Intuition : the entropic penalty ‘smoothes’ the problem and avoids over fitting (think of ridge regression for least squares)

The basics

Dual Formulation

Contrary to standard OT, no constraint on the dual problem :

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

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

The basics

Dual Formulation

Contrary to standard OT, no constraint on the dual problem :

$$\begin{aligned}
 W_{c,\varepsilon}(\alpha, \beta) &= \max_{\substack{\mathbf{u} \in \mathcal{C}(\mathcal{X}) \\ \mathbf{v} \in \mathcal{C}(\mathcal{Y})}} \int_{\mathcal{X}} \mathbf{u}(x) d\alpha(x) + \int_{\mathcal{Y}} \mathbf{v}(y) d\beta(y) \\
 &\quad - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} e^{\frac{\mathbf{u}(x) + \mathbf{v}(y) - c(x,y)}{\varepsilon}} d\alpha(x) d\beta(y) + \varepsilon. \\
 &= \max_{\substack{\mathbf{u} \in \mathcal{C}(\mathcal{X}) \\ \mathbf{v} \in \mathcal{C}(\mathcal{Y})}} \mathbb{E}_{\alpha \otimes \beta} \left[f_{\varepsilon}^{XY}(\mathbf{u}, \mathbf{v}) \right] + \varepsilon, \tag{D_\varepsilon}
 \end{aligned}$$

with $f_{\varepsilon}^{XY}(\mathbf{u}, \mathbf{v}) \stackrel{\text{def.}}{=} \mathbf{u}(x) + \mathbf{v}(y) - \varepsilon e^{\frac{\mathbf{u}(x) + \mathbf{v}(y) - c(x,y)}{\varepsilon}}$



The basics

Sinkhorn's Algorithm

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

Sinkhorn's Algorithm

Iterative algorithm : alternate between optimizing over \mathbf{u} with fixed \mathbf{v} and optimizing over \mathbf{v} with fixed \mathbf{u} .

Sinkhorn's Algorithm

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

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

Complexity of each iteration : $O(n^2)$,

Linear convergence, constant degrades when $\varepsilon \rightarrow 0$.



A magic regularizing tool !



Differentiable approximation of OT

Bonus : Sinkhorn procedure is fully differentiable with auto-diff tools (e.g TensorFlow) \Rightarrow yields a differentiable approximation of OT !

Some applications :

- Differentiable sorting (Cuturi et al '19)
- Differentiable (or 'soft') assignments
- Differentiable clustering (G. et al '19)
- Learning with a regularized Wasserstein loss
(\rightarrow more on that later...)



Sample Complexity

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.*

→ Bad sample complexity implies bad generalization (over-fitting).

Known cases :

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

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



Sample Complexity

'Sample Complexity' of W_ε .

Theorem (G., Chizat, Bach, Cuturi, Peyré '19) (Mena, Weed '19)

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

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

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



Sample Complexity

'Sample Complexity' of W_ε .

We get the following asymptotic behavior

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

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

→ A large enough regularization breaks the curse of dimension.

- ① Notions of Distance between Measures
- ② Entropic Regularization of Optimal Transport
- ③ Sinkhorn Divergences : Interpolation between OT and MMD
 - Definition and properties
 - Learning with Sinkhorn Divergences
- ④ Conclusion

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○○○
○○○○○○○○

Conclusion

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,

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

Distances

Entropic Regularization

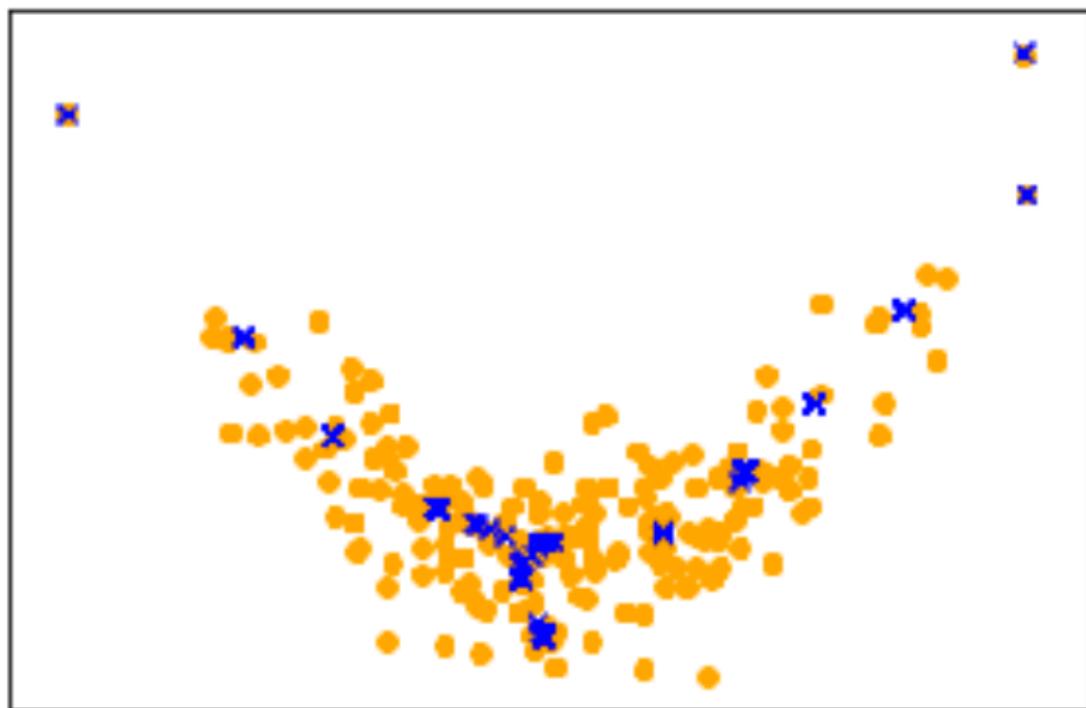
oooo
o
ooo

Sinkhorn Divergences

oooooo
oooooooo

Conclusion

The effect of entropy



Definition and properties

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

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

Definition and properties

Interpolation Property

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

Sinkhorn Divergences have the following asymptotic behavior :

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

$$\text{when } \varepsilon \rightarrow +\infty, \quad SD_{c,\varepsilon}(\alpha, \beta) \rightarrow \frac{1}{2} MMD_{-c}^2(\alpha, \beta). \quad (2)$$

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.

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○●○○
○○○○○○○○

Conclusion

Definition and properties

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

Distances

Entropic Regularization

○○○○
○
○○○

Sinkhorn Divergences

○○○●○
○○○○○○○○

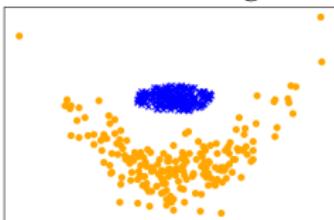
Conclusion

Definition and properties

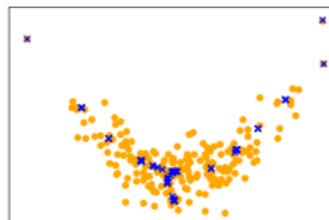
Discrete gradient flow of SD_ε , $\varepsilon = 1$

Summary

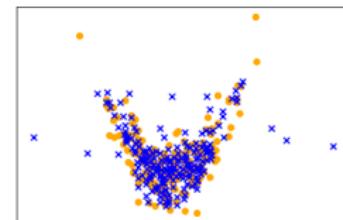
Initial Setting



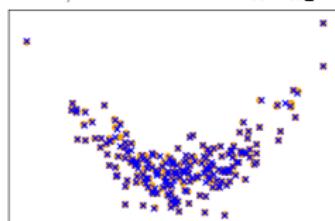
$W_{c,\varepsilon} - \varepsilon = 1, c = \|\cdot\|_2^{1.5}$



$ED_p - p = 1.5$



$SD_{c,\varepsilon} - \varepsilon = 1, c = \|\cdot\|_2^{1.5}$



$SD_{c,\varepsilon} - \varepsilon = 10^2, c = \|\cdot\|_2^{1.5}$

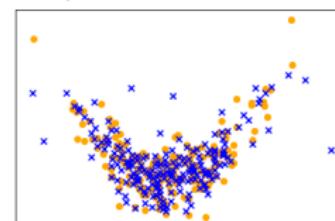
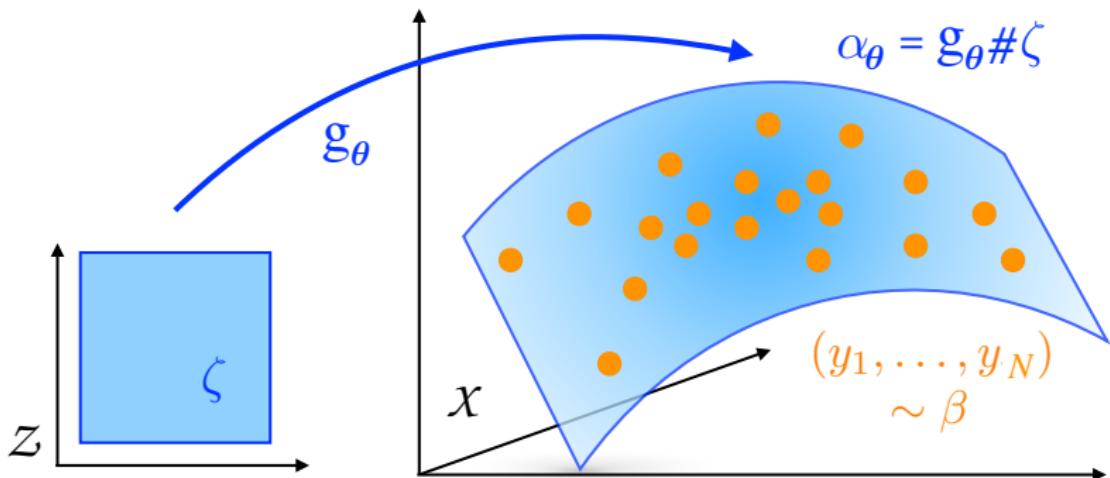


Figure 4 – Goal : Recover the positions of the Diracs with gradient descent. Orange circles : target distribution β , blue crosses : parametric model after convergence α_{θ^*} . Upper right : initial setting α_{θ_0} .

Generative Models



Problem Formulation

- β the **unknown** measure of the data :
finite number of samples $(y_1, \dots, y_N) \sim \beta$
- α_θ the parametric model of the form $\alpha_\theta \stackrel{\text{def.}}{=} g_\theta \# \zeta$:
to sample $x \sim \alpha_\theta$, draw $z \sim \zeta$ and take $x = g_\theta(z)$.

We are looking for the optimal parameter θ^* defined by

$$\theta^* \in \operatorname{argmin}_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta)$$

NB : α_θ and β are only known via their samples.



The Optimization Procedure

We want to solve by gradient descent

$$\min_{\theta} SD_{c,\varepsilon}(\alpha_\theta, \beta)$$

At each descent step k instead of approximating $\nabla_{\theta} SD_{c,\varepsilon}(\alpha_\theta, \beta)$:

- we approximate $SD_{c,\varepsilon}(\alpha_{\theta(k)}, \beta)$ by $SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta(k)}, \hat{\beta})$ via
 - minibatches : draw n samples from $\alpha_{\theta(k)}$ and m in the dataset (distributed according to β),
 - L Sinkhorn iterations : we compute an approximation of the SD between both samples with a fixed number of iterations
- we compute the gradient $\nabla_{\theta} SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta(k)}, \hat{\beta})$ by backpropagation (with automatic differentiation library)
- we do an update $\theta^{(k+1)} = \theta^{(k)} - C_k \nabla_{\theta} SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta(k)}, \hat{\beta})$

Learning

Computing the Gradient in Practice

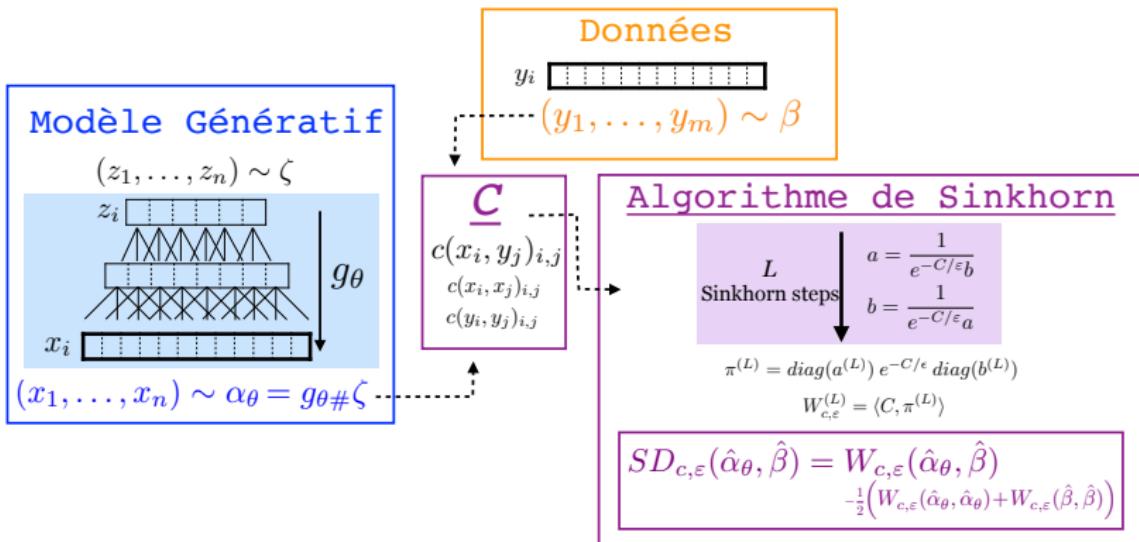
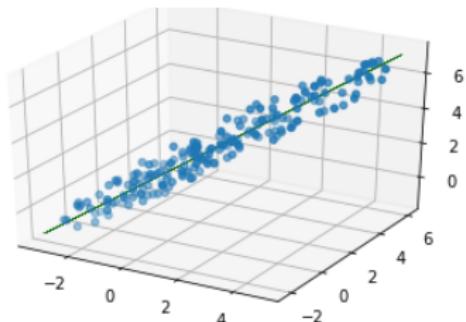


Figure 5 – Scheme of the approximation of the Sinkhorn Divergence from samples (here, $g_\theta : z \mapsto x$ is represented as a 2-layer NN).

Empirical Results

$$W_{c,\varepsilon} - \varepsilon = 1, c = \|\cdot\|_2^2$$



$$SD_{c,\varepsilon} - \varepsilon = 1, c = \|\cdot\|_2^2$$

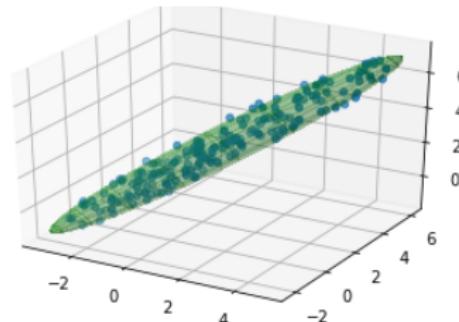
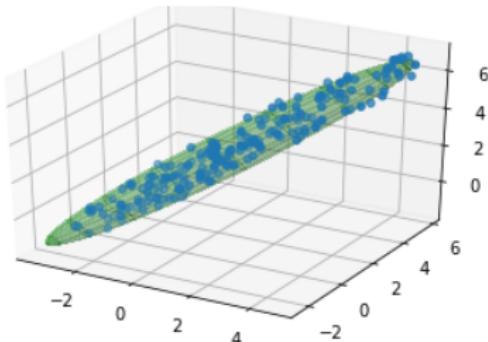


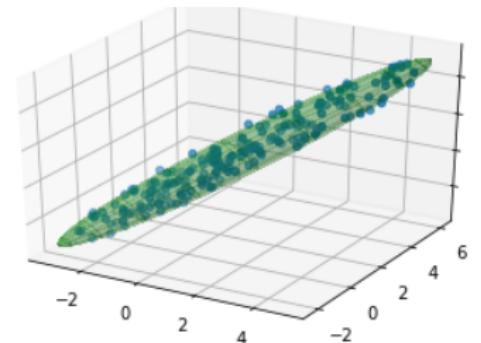
Figure 6 – Influence of the ‘debiasing’ of the Sinkhorn Divergence (SD_ε) compared to regularized OT (W_ε). Data are generated uniformly inside an ellipse, we want to infer the parameters A, ω (covariance and center).

Empirical Results

$$ED_p - p = 1.5$$



$$SD_{c,\varepsilon} - \varepsilon = 1, c = \|\cdot\|_2^2$$



ED_p		
1.5,-		
3.12	1.74	2.08
2.25	2.83	2.09
2.30	1.74	3.07
(0.63 , 1.75 , 2.75)		

ground truth		
3	2	2
2	3	2
2	2	3
(1,2,3)		

$SD_{c,\varepsilon}$		
2, 1		
2.90	1.96	2.13
2.02	3.03	2.10
2.06	1.95	3.03
(0.94 , 1.96 , 2.90)		

Figure 7 – Comparison of the Sinkhorn Divergence ($SD_{c,\varepsilon}$) and Energy Distance (ED_p) on the ellipse fitting task (we retained best parameters for each).

Learning

Learning the cost function

In high dimension (e.g. images), the Euclidean distance is not relevant → choosing the cost c is a complex problem.

Idea : the cost should yield high values for the Sinkhorn Divergence when $\alpha_\theta \neq \beta$ to differentiate between synthetic samples (from α_θ) and 'real' data (from β). (Li and al '18)

We learn a parametric cost of the form :

$$c_\varphi(x, y) \stackrel{\text{def.}}{=} \|f_\varphi(x) - f_\varphi(y)\|^p \quad \text{where} \quad f_\varphi : \mathcal{X} \rightarrow \mathbb{R}^{d'},$$

The optimization problem becomes a min-max on (θ, φ)

$$\min_{\theta} \max_{\varphi} SD_{c_\varphi, \varepsilon}(\alpha_\theta, \beta)$$

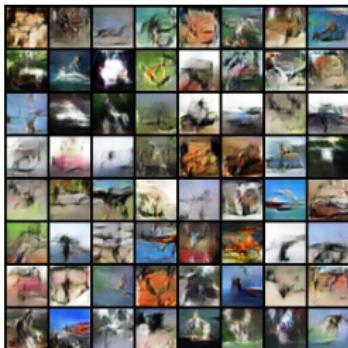
→ GAN-type problem, cost c acts as a discriminator.

Learning

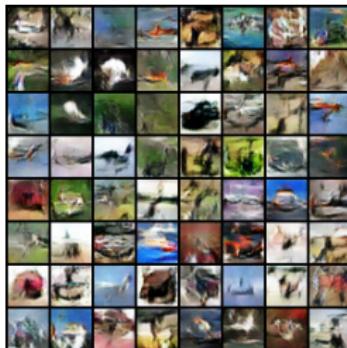
○○○○
○
○○○

○○○○○
○○○○○○●

Empirical Results - CIFAR10



(a) MMD

(b) $\varepsilon = 100$ (c) $\varepsilon = 1$

MMD (Gaussian)

 $\varepsilon = 100$ $\varepsilon = 10$ $\varepsilon = 1$ 4.56 ± 0.07 4.81 ± 0.05 4.79 ± 0.13 4.43 ± 0.07

Table 1 – Inception Scores on CIFAR10 (same setting as MMD-GAN paper (Li et al. '18)).

Distances

Entropic Regularization

oooo
o
ooo

Sinkhorn Divergences

oooooo
oooooooo

Conclusion

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

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