Sinkhorn Divergences : Interpolating between Optimal Transport and MMD

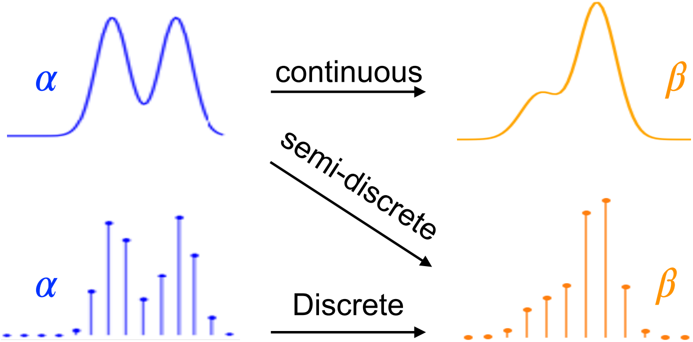
Aude Genevay

DMA - Ecole Normale Supérieure - CEREMADE - Université Paris Dauphine

NYU - April 2019

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

Comparing Probability Measures



Discrete Setting

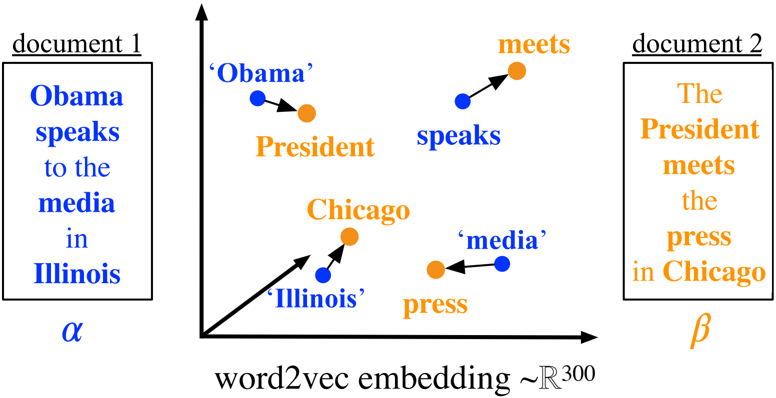
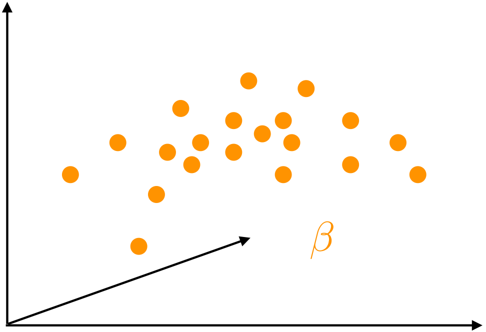
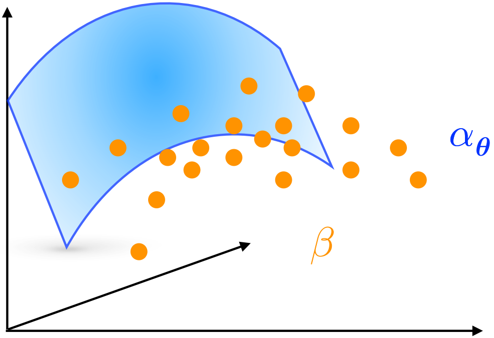


Figure 1 – Exemple of data representation as a point cloud (from Kusner '15)

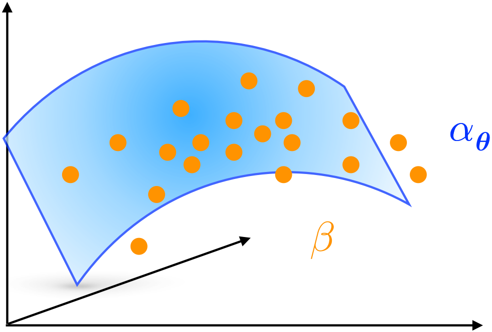
Semi-discrete Setting



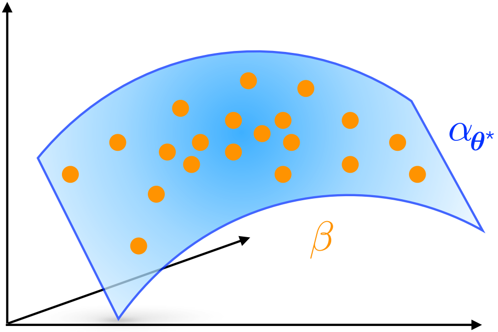
Semi-discrete Setting



Semi-discrete Setting



Semi-discrete Setting



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

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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

$n = 1$

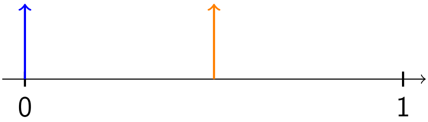
Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Example

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



$n = 2$

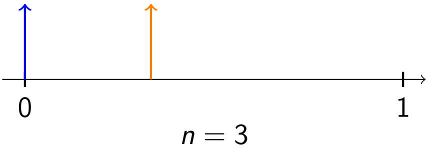
Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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



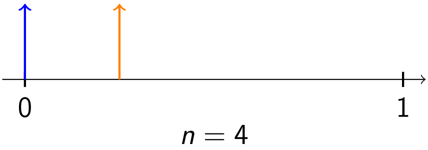
Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x) \; \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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



Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

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

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

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

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

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

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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

$n = 8$

Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\, \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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

$n = 9$

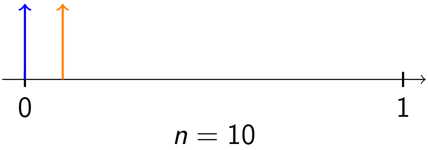
Weak Convergence of measures

Definition (Weak Convergence)

Let $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
The sequence α_n **weakly converges** to α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Let $\mathcal{L}$ a distance between measures , $\mathcal{L}$ **metrises weak convergence** IFF$\Big(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\Big)$.

Example

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



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 :

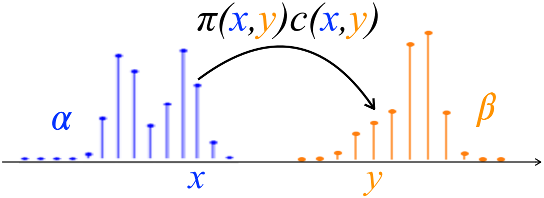
1. $\forall x \in \mathcal{X}, \quad k(x,\cdot) \in \mathcal{H}$,
2. $\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$.

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

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

Optimal Transport (Monge 1781, Kantorovitch '42)

- Cost of moving a unit of mass from x to y : $c(x, y)$



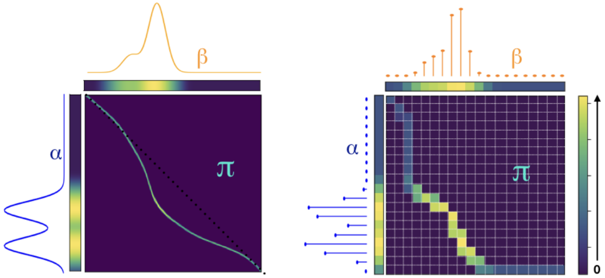
- What is the coupling π that minimizes the total cost of moving ALL the mass from α to β ?

The Wasserstein Distance

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

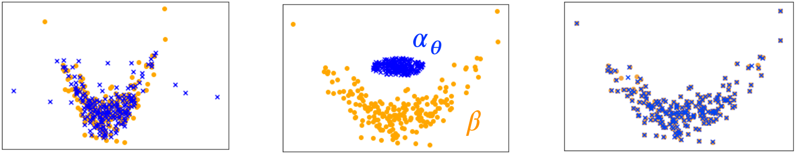
$$W_c(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \qquad (\mathcal{P})$$

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



Transport Optimal vs. MMD

MMD	**Optimal Transport**
estimation robust to sampling	curse of dimension
computed in $O(n^2)$	computed in $O(n^3 \log(n))$
inefficient outside of dense areas	recovers full support of measures



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

Figure 2 – Goal : fit the the discrete measure β with α_θ, where θ encodes the positions of the Diracs. Method : minimize $MMD(\alpha_\theta, \beta)$ or $W_c(\alpha_\theta, \beta)$ with gradient 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 $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

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

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

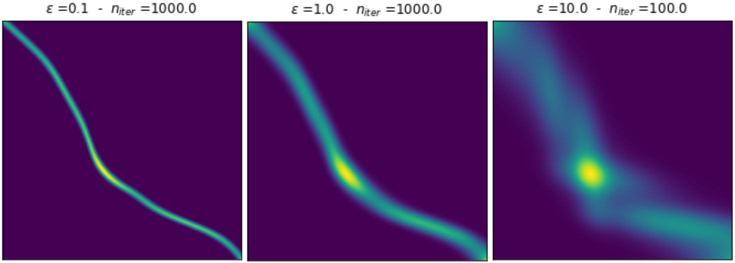


Figure 3 – Influence of the regularization parameter ε on the transport plan π.

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

Dual Formulation

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

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

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

Dual Formulation

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

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

$$- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon.$$

$$= \max_{\substack{u\in\mathcal{C}(\mathcal{X})\\ v\in\mathcal{C}(\mathcal{Y})}} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon, \qquad (\mathcal{D}_\varepsilon)$$

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

Sinkhorn's Algorithm

First order conditions for $(\mathcal{D}_\varepsilon)$, concave in (u, v) :

$$e^{u(x)/\varepsilon} = \frac{1}{\int_{\mathcal{Y}} e^{\frac{v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)} \quad ; \quad e^{v(y)/\varepsilon} = \frac{1}{\int_{\mathcal{X}} e^{\frac{u(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)}$$

$\to (u, v)$ solve a fixed point equation.

Sinkhorn's Algorithm

First order conditions for $(\mathcal{D}_\varepsilon)$, concave in (u, v) :

$$e^{u_i/\varepsilon} = \frac{1}{\sum_{j=1}^m e^{\frac{v_j - c_{ij}}{\varepsilon}} \beta_j} \quad ; \quad e^{v_j/\varepsilon} = \frac{1}{\sum_{i=1}^n e^{\frac{u_i - c_{ij}}{\varepsilon}} \alpha_i}$$

$\to (u, v)$ solve a fixed point equation.

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 $\varepsilon \to 0$.

Sinkhorn Divergences

Issue of entropic transport : $W_{c,\varepsilon}(\alpha, \alpha) \neq 0$
Proposed Solution : introduce corrective terms to 'debias' entropic transport

Definition (Sinkhorn Divergences)

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

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

Interpolation Property

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

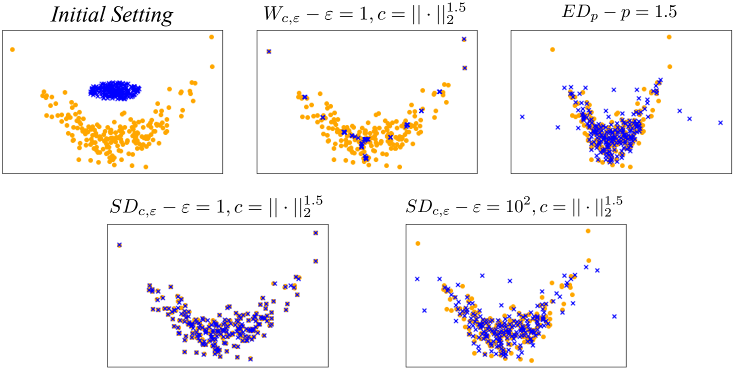
Sinkhorn Divergences have the following asymptotic behavior :

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

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

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.

Empirical Illustration



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.*

$\rightarrow$ Bad sample complexity implies bad generalization (over-fitting).

Known cases :

- OT : $\mathbb{E}|W(\alpha, \beta) - W(\hat{\alpha}_n, \hat{\beta}_n)| = O(n^{-1/d})$
  $\Rightarrow$ 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 $\mathbb{E}|SD_\varepsilon(\alpha, \beta) - SD_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)|$?

Properties of Dual Potentials

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

Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ bounded , and $c \in \mathcal{C}^\infty$. Then the optimal pairs of dual potentials (u, v) are uniformly bounded in the Sobolev $\mathbf{H}^{\lfloor d/2 \rfloor + 1}(\mathbb{R}^d)$ and their norm verifies :

$$\|u\|_{\mathbf{H}^{\lfloor d/2 \rfloor+1}} = O\left(1 + \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\right) \text{ et } \|v\|_{\mathbf{H}^{\lfloor d/2 \rfloor+1}} = O\left(1 + \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\right),$$

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

$\mathbf{H}^{\lfloor d/2 \rfloor + 1}(\mathbb{R}^d)$ is a RKHS $\rightarrow$ the dual $(\mathcal{D}_\varepsilon)$ est the maximization of an expectation in a RKHS ball.

‘Sample Complexity’ of Sinkhorn Div.

Theorem (Bartlett-Mendelson ’02)

Let $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{X})$, ℓ a B-Lipschitz function and $\mathcal{H}$ a RKHS with kernel k bounded on $\mathcal{X}$ by K. Then

$$\mathbb{E}_{\mathbb{P}}\left[\sup_{\{g \mid \|g\|_{\mathcal{H}} \leqslant \lambda\}} \mathbb{E}_{\mathbb{P}}\ell(g,X) - \frac{1}{n}\sum_{i=1}^{n}\ell(g,X_i)\right] \leqslant 2B\frac{\lambda K}{\sqrt{n}}.$$

Theorem (G., Chizat, Bach, Cuturi, Peyré ’19)

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

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

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

'Sample Complexity' of Sinkhorn Div.

We get the following asymptotic behavior

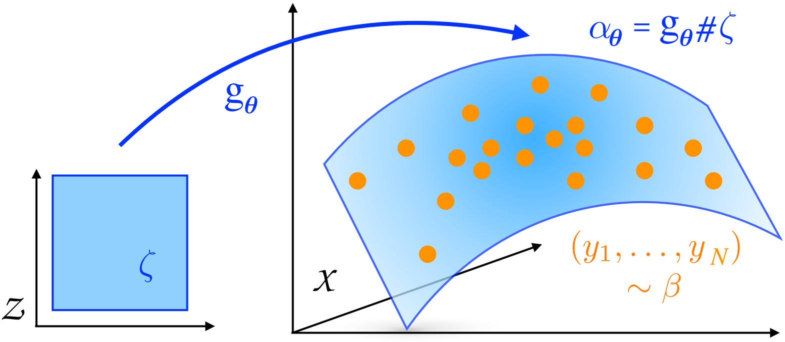
$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}\right) \qquad \text{quand } \varepsilon \to 0$$

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

$\rightarrow$ We recover the interpolation property,

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

Generative Models



Problem Formulation

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

We are looking for the optimal parameterθ^* defined by

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

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 bewteen 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 $\theta^{(k+1)} = \theta^{(k)} - C_k \nabla_\theta SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta^{(k)}}, \hat{\beta})$

Computing the Gradient in Practice

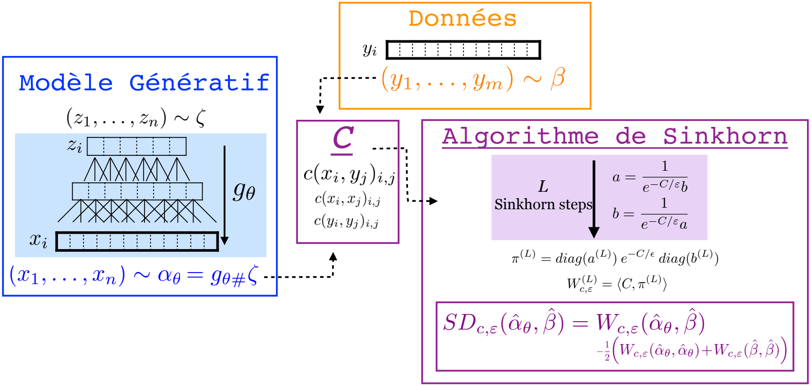


Figure 5 – Scheme of the approximation of the Sinkhorn Divergence from samples (here, $g_\theta : z \mapsto x$ is represented as a 2-layer NN).

Empirical Results

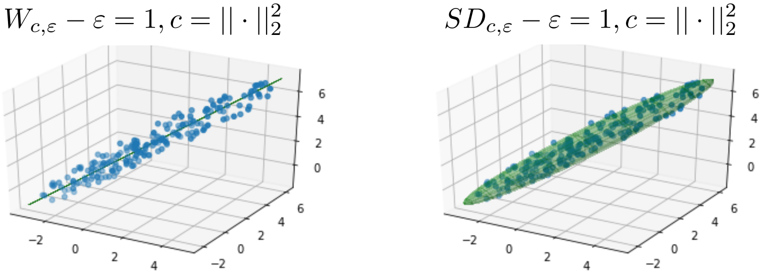
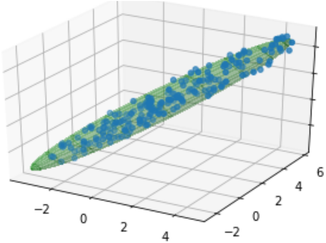


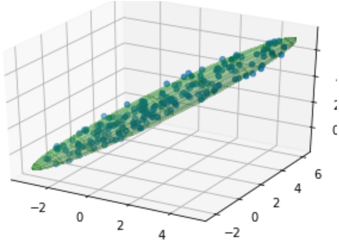
Figure 6 – Influence of the ‘debiasing’ of the Sinkhorn Divergence (SD_ε) compared to regularized OT (W_ε). Data are generated uniformly inside an ellipse, we want to infer the parametersLes données sont générées A, ω (covariance and center).

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 the cost function

In high dimension (e.g. images), the euclidean distance is not relevant $\rightarrow$ 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 differenciate between synthetic samples (from α_θ) and 'real' data (from β). (Li and al '18)

We learn a parametric cost of the form :

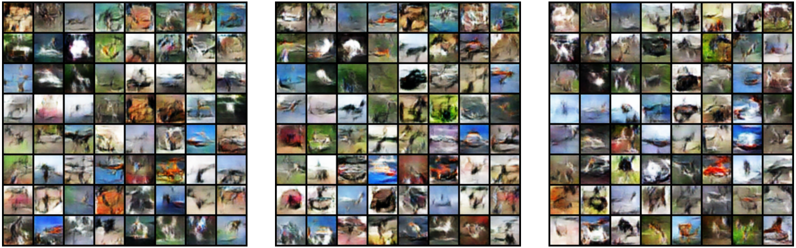
$$c_\varphi(x, y) \stackrel{\text{def.}}{=} \|f_\varphi(x) - f_\varphi(y)\|^p \quad \text{where} \quad f_\varphi : \mathcal{X} \rightarrow \mathbb{R}^{d'},$$

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

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

$\rightarrow$ GAN-type problem, cost c acts as a discriminator.

Empirical Results - CIFAR10



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

Figure 8 – Images generated by α_{θ^*} trained on CIFAR 10

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)).

Motivations

- Sinkhorn purely discrete algorithm : requires sampling from the measures **beforehand**
- 'Batch' method :each iteration costs $O(n^2)$

Ides : exploit OT formulation as max of an expectation by using **stochastic optimization**.

- Only requires being able to sample from the measures $\rightarrow$ no discretization bias
- 'Online' method : each iteration costs $O(n)$

Semi-Dual Formulation

When one mesure is discrete, e.g.

$$\beta \overset{\text{def.}}{=} \sum_{i=1}^{n} \boldsymbol{\beta}_i \delta y_i \quad \rightarrow \quad v = (\mathbf{v}_i)_{i=1}^{n} \overset{\text{def.}}{=} (\mathbf{v}_(x_i), \ldots, \mathbf{v}(x_n) \in \mathbb{R}^n.$$

Using first order condition on dual problem (relation between v and u), we get the *semi-dual* formulation :

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^n} \mathbb{E}_\alpha \left[g_\varepsilon^X(\mathbf{v}) \right] \qquad (\mathcal{S}_\varepsilon)$$

where $$g_\varepsilon^x(\mathbf{v}) = \sum_{j=1}^{m} \mathbf{v}_i \boldsymbol{\beta}_i + \begin{cases} -\varepsilon \log \left(\sum_{i=1}^{n} \exp(\frac{\mathbf{v}_i - c(x, y_i)}{\varepsilon}) \right) \boldsymbol{\beta}_i) & \text{si } \varepsilon > 0, \\ \min_j \left(c(x, y_i) - \mathbf{v}_i \right) & \text{si} \varepsilon = 0. \end{cases}$$

Semi-Discrete Case : SGD

We want to solve

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^n} \mathbb{E}_\alpha \left[g_\varepsilon^X(\mathbf{v}) \right] \stackrel{\text{def.}}{=} G_\varepsilon(\mathbf{v}) \qquad (\mathcal{S}_\varepsilon)$$

by gradient ascent on $G_\varepsilon(\mathbf{v})$.

Problem : We can't compute the gradient (α is not known)

Idea : At each iteration, we draw $x^{(k)} \sim \alpha$ and $\nabla g_\varepsilon^{x^{(k)}}$ is a proxy for ∇G_ε.

Semi-Discrete Case : SGD

The iterates of SGD are :

$$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \frac{C}{\sqrt{k}} \nabla_{v} g_{\varepsilon}^{x^{(k)}}(\mathbf{v}^{(k+1)}) \quad \text{where} \quad x^{(k)} \sim \alpha. \tag{3}$$

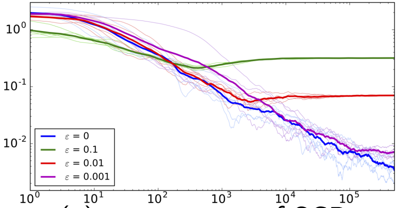
Proposition (Convergence of SGD)

Let $\mathbf{v}_{\varepsilon}^{*}$ a minimizer of the semi-dual and $\bar{\mathbf{v}}^{(k)} \stackrel{\text{def.}}{=} \frac{1}{k}\sum_{i=1}^{k} \mathbf{v}^{(k)}$ the average of the SGD iterates. Then

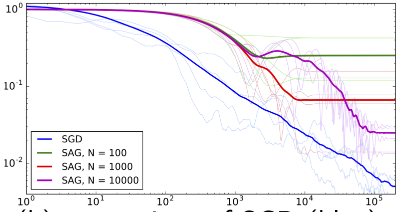
$$|G_{\varepsilon}(\mathbf{v}_{\varepsilon}^{*}) - G_{\varepsilon}(\bar{\mathbf{v}}^{(k)})| = O(1/\sqrt{k}).$$

Complexity of each iteration $O(n)$.

Semi-Discrete Case : SGD - Application



(a) convergence of SGD
for different levels of regularization ε



(b) comparison of SGD (blue)
against a discrete algorithm

Continuous Case : Dual Formulation

Idea : Replace dual potentials(u, v) by their expansion in a well chosen RKHS

$$u(x) \leftarrow \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}} \qquad v(y) \leftarrow \langle v, \kappa(\cdot, y)\rangle_{\mathcal{H}}$$

The dual problem becomes

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}(\mathcal{X})} \mathbb{E}_{\alpha \otimes \beta}\left[f_\varepsilon^{XY}(u, v)\right] + \varepsilon, \qquad (\mathcal{D}_\varepsilon)$$

with

$$f_\varepsilon^{xy}(u, v) \stackrel{\text{def.}}{=} \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}} + \langle v, \kappa(\cdot, y)\rangle_{\mathcal{H}} - \varepsilon \exp\left(\frac{\langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}} + \langle v, \kappa(\cdot, y)\rangle_{\mathcal{H}} - c(x, y)}{\varepsilon}\right)$$

Continuous Case : Kernel-SGD

Let $\mathcal{H}$ a RKHS with kernel κ. The iterates of Kernel-SGD read :

$$\begin{cases} u^{(k)} & \stackrel{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, x_i) \\ v^{(k)} & \stackrel{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, y_i)), \end{cases} \quad \text{with} \quad \begin{cases} (x_i)_{i=1\ldots k} \sim \alpha \\ (y_i)_{i=1\ldots k} \sim \beta \end{cases}$$

$$\text{et} \quad w^{(i)} \stackrel{\text{def.}}{=} \frac{C}{\sqrt{i}} \left(1 - \exp\left(\frac{u^{(i-1)}(x_i) + v^{(i-1)}(y_i) - c(x_i, y_i)}{\varepsilon} \right) \right),$$

Proposition (Convergence of Kernel-SGD)

If α and β have bounded supports in $\mathbb{R}^d$,then for κ the Matern kernel or a universal Kernel (e.g. Gaussian) the iterates $(u^{(k)}, v^{(k)})$ converge to a solution of the dual $(\mathcal{D}_\varepsilon)$.

Continuous Case : Kernel-SGD - Illustration

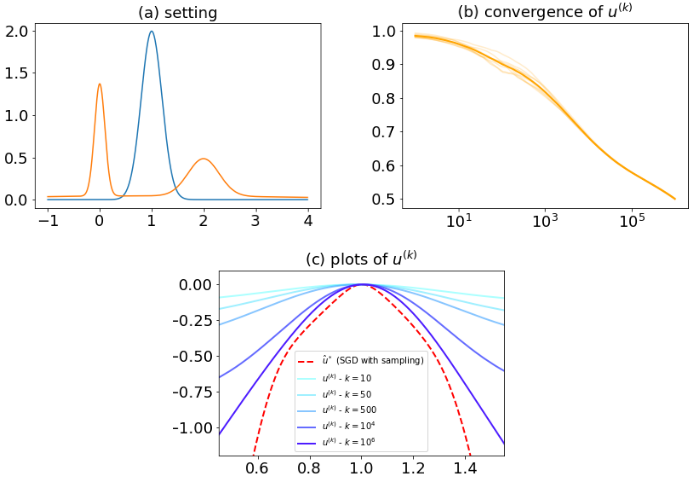


Figure 9 – Illustration of the convergence of kernel-SGD on a simple case in 1D

Continuous Case : Kernel-SGD - Acceleration

At iteration k, need to compute

$$\begin{cases} u^{(k-1)}(x_k) = \sum_{i=1}^{k-1} w^{(i)} \kappa(x_k, x_i) \\ v^{(k-1)}(y_k) = \sum_{i=1}^{k-1} w^{(i)} \kappa(y_k, y_i) \end{cases}$$

Problem : itération k costs $O(k)$

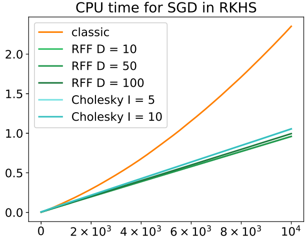
Idea : replace kernel κ by an approximation of the form

$$\hat{\kappa}(x, x') = \langle \varphi(x), \varphi(x') \rangle \quad \text{où} \quad \varphi : \mathcal{X} \to \mathbb{R}^p.$$

$\to$ The cost of each iteration is then fixed as $O(p)$.

Examples : Cholesky Decomposition, Random Fourier Features (RFF)

Continuous Case : Kernel-SGD - Acceleration



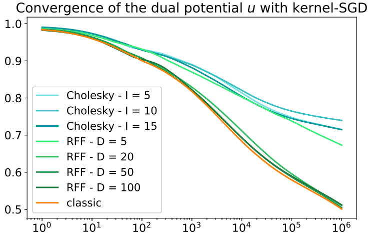


Figure 10 – Effects of the acceleration procedure on CPU time and precision

$\rightarrow$ For 10^6 iterations, kernel-SGD takes 6 hours
$\rightarrow$ The accelerated version with RFF and $D = 20$ takes 3 minutes, and we get the same level of precison !

Take Home Message

Sinkhorn Divergences 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