Régularisation Entropique du Transport Optimal pour l'Apprentissage Statistique

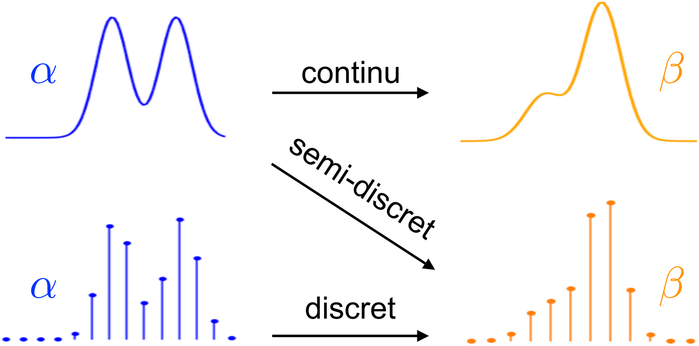
Soutenance de Thèse d'Aude Genevay

DMA - Ecole Normale Supérieure - CEREMADE - Université Paris Dauphine

13 Mars 2019

Travail effectué sous la direction de Gabriel Peyré

Comparer des Mesures de Probabilité



Cadre Discret

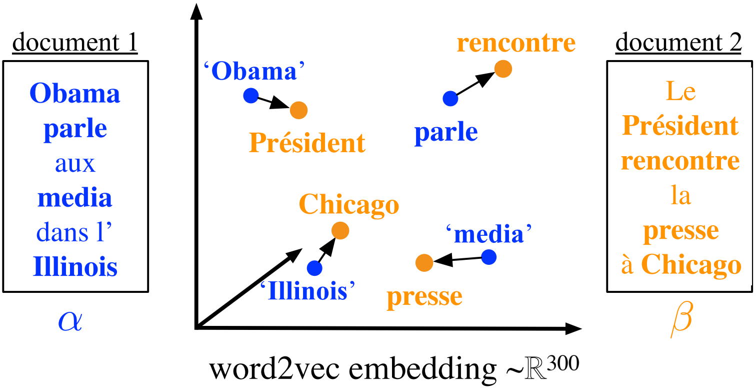
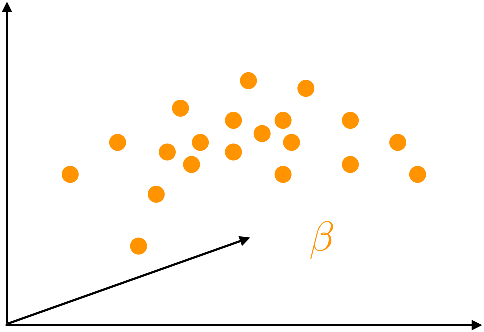
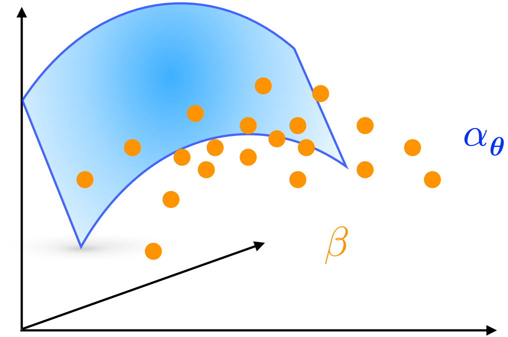


Figure 1 – Exemple de représentation de données sous forme de nuage de point (extrait de Kusner '15)

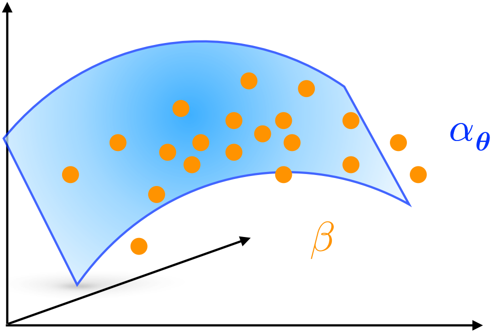
Cadre Semi-discret



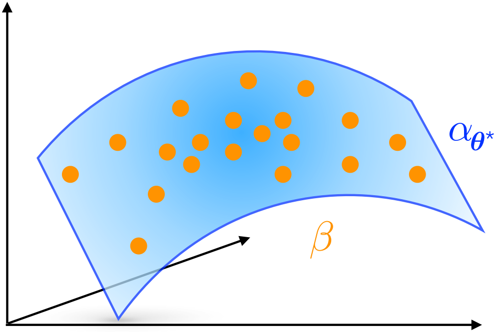
Cadre Semi-discret



Cadre Semi-discret



Cadre Semi-discret



1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

φ-divergences (Czisar '63)

Définition (φ-divergence)

Soit φ une fonction convexe semi-continue inférieurement telle que $\varphi(1) = 0$, la φ-divergence D_φ entre deux mesures α et β est définie par :

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

Exemple (Divergence de Kullback Leibler)

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

Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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

$n = 1$

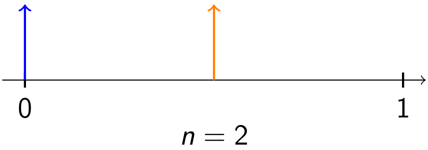
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\, \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x) \; \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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

$n = 3$

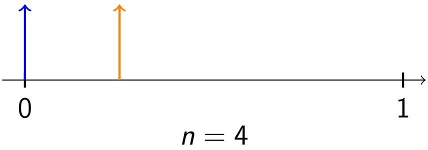
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



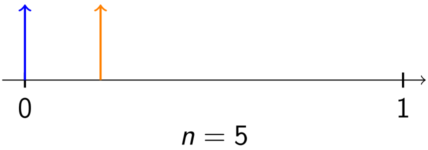
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



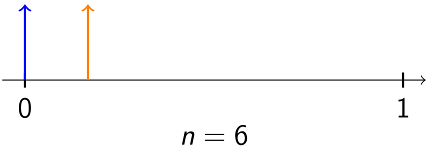
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



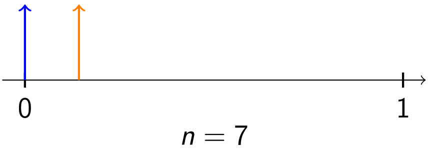
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\, \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



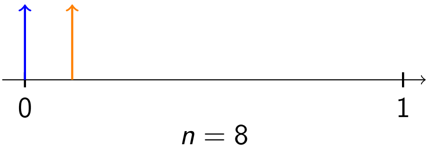
Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\, \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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



Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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

$n = 9$

Convergence Faible de Mesures

Définition (Convergence Faible)

Soit $(\alpha_n)_n \in \mathcal{M}_+^1(\mathcal{X})^{\mathbb{N}}$, $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.
La suite α_n **converge faiblement** vers α, i.e.
$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha(x)\ \forall f \in \mathcal{C}_b(\mathcal{X})$.
Soit $\mathcal{L}$ une distance entre mesures, $\mathcal{L}$ **métrise la convergence faible** SSI $\left(\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha\right)$.

Exemple

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

$n = 10$

Maximum Mean Discrepancies (Gretton '06)

Definition (RKHS)

Soit $\mathcal{H}$ un espace de Hilbert avec noyau k, alors $\mathcal{H}$ est à Noyau Reproduisant (RKHS) si et seulement si :

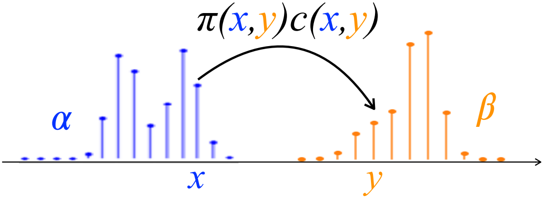
1. $\forall x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H}$,
2. $\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$.

Soit $\mathcal{H}$ un RKHS avec noyau k, la distance **MMD** entre deux mesures de probabilité α et β est définie par :

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

Le Transport Optimal (Monge 1781, Kantorovitch '42)

- Coût de déplacer une unité de masse de x vers y : $c(x, y)$



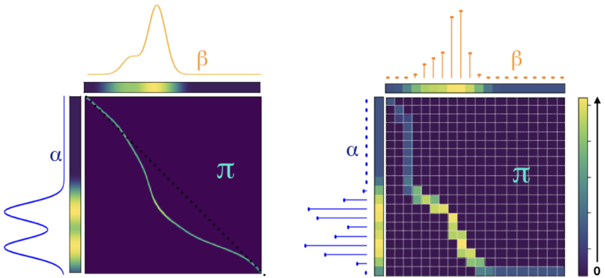
- Quel est le couplage π qui minimise le coût total de bouger TOUTE la masse de α vers β ?

La distance de Wasserstein

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

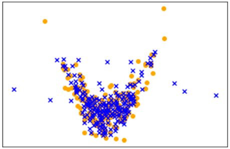
$$W_c(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \qquad (\mathcal{P})$$

Pour $c(x, y) = \|x - y\|_2^p$, $W_c(\alpha, \beta)^{1/p}$ est la distance de Wasserstein.

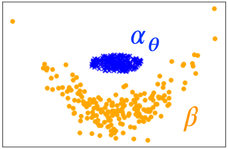


Transport Optimal vs. MMD

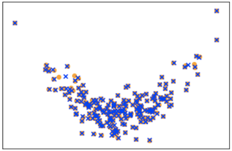
MMD	Transport Optimal
estimation en $O(n^2)$	estimation en $O(n^3 \log(n))$
estimation robuste par échantillions	malédiction de la dimension
capte mal les phénomènes loin des zones denses	s'adapte à la géometrie du problème via la fonction de coût c



MMD_k - $k = - || \cdot ||_2^{1.5}$



Configuration initiale



W_c - $c = || \cdot ||_2^{1.5}$

1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

La Régularisation Entropique (Cuturi '13)

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

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

La Régularisation Entropique (Cuturi '13)

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

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

La Régularisation Entropique (Cuturi '13)

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

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

où

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

entropie relative du plan de transport π par rapport à la mesure produit $\alpha \otimes \beta$.

La Régularisation Entropique

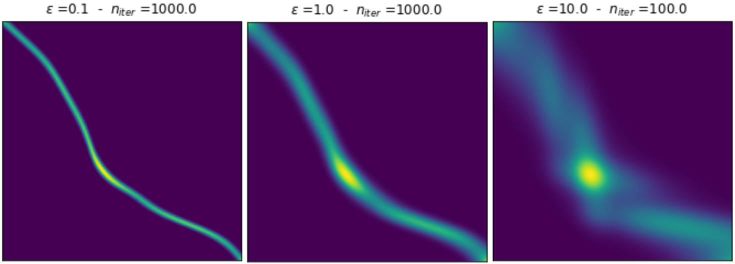


Figure 2 – Influence du paramètre de régularisation ε sur le plan de transport π.

Intuition : La pénalisation entropique permet de 'lisser' le problème et d'empêcher l'overfitting / sur-apprentissage (comme la régularisation ridge sur les moindres carrés)

Formulation Duale

Contrairement au transport classique, pas de contrainte sur le dual :

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

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

Formulation Duale

Contrairement au transport classique, pas de contrainte sur le dual :

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

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

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

L'Algorithme de Sinkhorn

Conditions de premier ordre pour $(\mathcal{D}_\varepsilon)$, concave en (u, v) :

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

$\rightarrow$ (u, v) vérifient une équation de point fixe.

L'Algorithme de Sinkhorn

Conditions de premier ordre pour $(\mathcal{D}_\varepsilon)$, concave en (u, v) :

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

$\rightarrow (u, v)$ vérifient une équation de point fixe.

Algorithme de Sinkhorn

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

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

Complexité de chaque iteration : $O(n^2)$,
Convergence linéaire, constante se dégrade quand $\varepsilon \rightarrow 0$.

Extensions

- Autres regularisations : $D_\varphi(\pi|\alpha \otimes \beta)$ avec φ convexe de domaine $\mathbb{R}^+$.
  $\rightarrow$ formulation duale sous forme d'espérance
- Transport ‘unbalanced’ (mesures de masse quelconque) avec régularisation convexe $\rightarrow$ formulation duale sous forme d'espérance

1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

Les Divergences de Sinkhorn

Problème du transport entropique : $W_{c,\varepsilon}(\alpha, \alpha) \neq 0$
Solution proposée : introduction de termes correctifs pour 'débiaiser' le transport régularisé

Définition (Divergences de Sinkhorn)

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

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

Propriété d'Interpolation

Théorème (G., Peyré, Cuturi '18), (Ramdas et al. '17)

Les Divergences de Sinkhorn ont le comportement limite suivant :

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

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

Remarque : Pour avoir un MMD, $-c$ doit induire un noyau défini positif. Pour $c = \| \cdot \|_2^p$ avec $0 < p < 2$, le MMD associé s'appelle l'Energy Distance.

Illustration Numérique

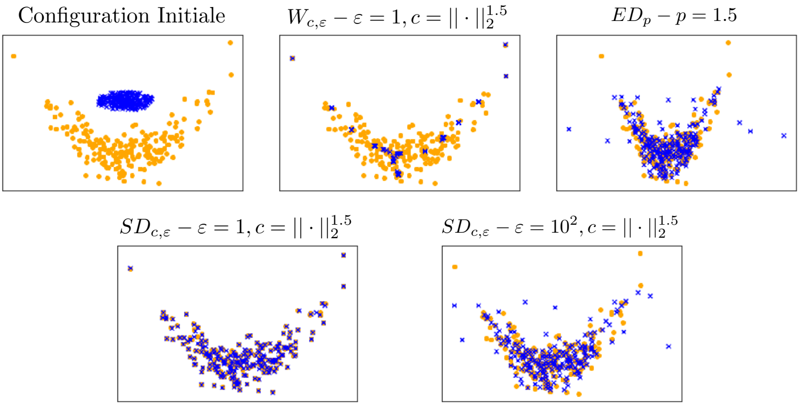


Figure 3 – But : Retrouver les positions des Diracs par descente de gradient. Cercles oranges : distribution cible β, croix bleues modèle appris α_{θ^*}. En haut à droite : distribution initiale α_{θ_0}.

La 'sample complexity'

Définition informelle

Etant donnée une distance entre mesures, sa ***sample complexity*** *correspond à l'erreur d'approximation lorsque l'on évalue cette distance à l'aide d'échantillons des mesures.*

→ Mauvaise sample complexity implique mauvaise généralisation (sur-apprentissage) car on colle trop au bruit des données.

Cas connus :

- TO : $\mathbb{E}|W(\alpha, \beta) - W(\hat{\alpha}_n, \hat{\beta}_n)| = O(n^{-1/d})$
  $\Rightarrow$ fléau de la dimension (Dudley '84, Weed et Bach '18)
- MMD : $\mathbb{E}|MMD(\alpha, \beta) - MMD(\hat{\alpha}_n, \hat{\beta}_n)| = O(\frac{1}{\sqrt{n}})$
  $\Rightarrow$ indépendant de la dimension (Gretton '06)

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

Propriétés des Potentiels Duaux

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

Soient $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ bornés , et $c \in \mathcal{C}^\infty$. Alors les paires de potentiels duaux optimales (u, v) sont uniformément bornées dans le Sobolev $\mathbf{H}^{\lfloor d/2 \rfloor + 1}(\mathbb{R}^d)$ et leur norme vérifie

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

avec des constantes dépendant de $|\mathcal{X}|$ (ou $|\mathcal{Y}|$ pour v), d, et $\left\|c^{(k)}\right\|_\infty$ pour $k = 0, \ldots, \lfloor d/2 \rfloor + 1$.

$\mathbf{H}^{\lfloor d/2 \rfloor + 1}(\mathbb{R}^d)$ est un RKHS $\rightarrow$ le dual $(\mathcal{D}_\varepsilon)$ est la maximisation d'une espérance dans une boule d'un RKHS.

'Sample Complexity' des Div. de Sinkhorn

Theorème (Bartlett-Mendelson '02)

Soit $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{X})$, ℓ une fonction B-Lipschitz et $\mathcal{H}$ un RKHS avec noyau k borné sur $\mathcal{X}$ par K. Alors

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

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

Soient $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ bornés , et $c \in \mathcal{C}^\infty$ L-Lipschitz. Alors

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

où $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$ et les constantes dépendent de $|\mathcal{X}|$, $|\mathcal{Y}|$, d, et $\left\|c^{(k)}\right\|_\infty$ pour $k = 0 \dots \lfloor d/2 \rfloor + 1$.

'Sample Complexity' des Div. de Sinkhorn

En particulier, on obtient le comportement asymptotique suivant

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

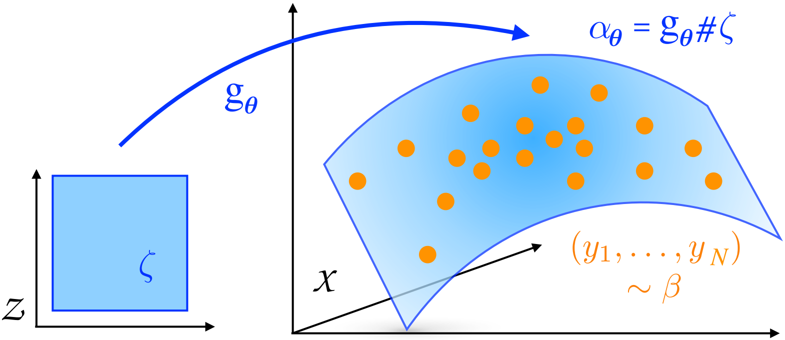
$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = O\left(\frac{1}{\sqrt{n}}\right) \qquad \text{quand } \varepsilon \to +\infty.$$

$\rightarrow$ On retrouve la propriété d'interpolation,

$\rightarrow$ Une régularisation assez grande casse le fléau de la dimension.

1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

Les Modèles Génératifs



Formulation du Problème

- β la mesure **inconnue** des données :
  nombre fini de points $(y_1, \ldots, y_N) \sim \beta$
- α_θ le modèle paramétrique de la forme $\alpha_\theta \stackrel{\text{def.}}{=} g_{\theta\#}\zeta$:
  pour obtenir $x \sim \alpha_\theta$, on tire $z \sim \zeta$ et on prend $x = g_\theta(z)$.

On cherche le paramètre optimal θ^* défini par

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

NB : α_θ et β ne sont connues QUE via leurs échantillons.

La Procédure d'Optimisation

On veut résoudre par descente de gradient

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

A chaque pas de descente k lieu d'approximer $\nabla_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta)$:

- on approxime $SD_{c,\varepsilon}(\alpha_{\theta^{(k)}}, \beta)$ par $SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta^{(k)}}, \hat{\beta})$ via
    - minibatches : on tire n échantillons selon $\alpha_{\theta^{(k)}}$ et m dans le jeu de données (distribuées selon β),
    - L iterations de Sinkhorn : on calcule une approximation de la distance de transport entre les deux échantillons avec un nombre fixé d'itérations
- on calcule le gradient $\nabla_\theta SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta^{(k)}}, \hat{\beta})$ par backpropagation
- on effectue un update $\theta^{(k+1)} = \theta^{(k)} - C_k \nabla_\theta SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta^{(k)}}, \hat{\beta})$

Le Calcul du Gradient en Pratique

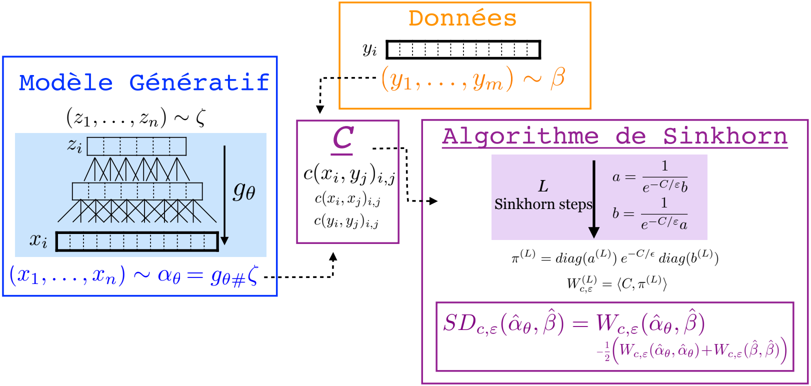


Figure 4 – Schéma d'approximation de la Divergence de Sinkhorn à partir d'échantillons (ici, $g_\theta : z \mapsto x$ est représenté sous forme d'un réseau de neurones à 2 couches).

Résultats Numériques

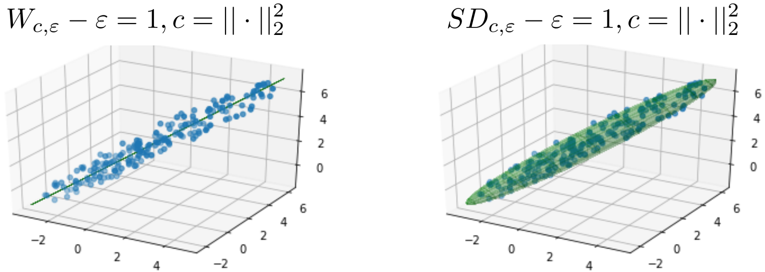


Figure 5 – Influence de la ‘normalisation’ de la Divergence de Sinkhorn (SD_ε) par rapport au transport régularisé (W_ε). Les données sont générées uniformément à l’intérieur d’une ellipse, dont on souhaite retrouver les paramètres A, ω (covariance et centre).

Résultats Numériques - MNIST

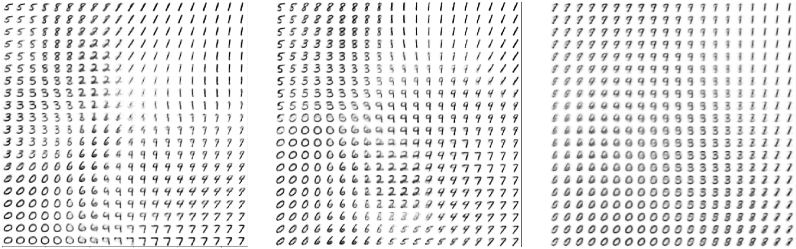


Figure 6 – Influence des hyperparametres sur les chiffres générés. gauche : $\varepsilon = 1,\ m = 200, L = 10$; milieu : $\varepsilon = 10^{-1},\ m = 200, L = 100$; droite : $\varepsilon = 10^{-1}, m = 10, L = 300$

Apprendre la fonction de coût

En grande dimension (e.g. pour des images), la distance euclidienne n'est pas pertinente $\rightarrow$ le choix du coût c est un problème complexe.

Idée : le coût doit induire de grandes valeurs pour la Divergence de Sinkhorn lorsque $\alpha_\theta \neq \beta$ pour bien différencier les échantillons synthétiques (selon α_θ) des 'vraies' données (selon β). (Li et al '18)

On apprend un coût paramétrique de la forme :

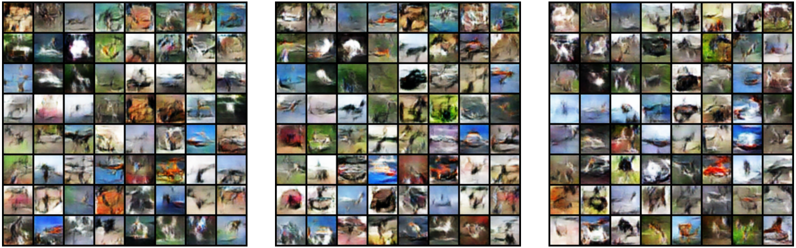
$$c_\varphi(x, y) \stackrel{\text{def.}}{=} \|f_\varphi(x) - f_\varphi(y)\|^p \quad \text{where} \quad f_\varphi : \mathcal{X} \rightarrow \mathbb{R}^{d'},$$

Le problème d'optimisation devient un min-max sur (θ, φ)

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

$\rightarrow$ problème de type GAN, coût c joue le rôle du discriminateur.

Résultats Numériques - CIFAR10



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

Figure 7 – Points générés par α_{θ^*} entrainé sur 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 sur CIFAR10 (expériences réalisées dans le même cadre que le papier MMD-GAN (Li et al. '18)).

1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

Motivations

- Sinkhorn algorithme purement discret : nécessite d'échantillonner les mesures au préalable
- Méthode 'batch' : chaque iteration coute $O(n^2)$

Idée : exploiter la formulation du TO régularisé comme max d'espérance avec des méthodes d'**optimisation stochastique**.

- nécessite seulement de pouvoir générer des points selon les mesures $\rightarrow$ pas de biais de discrétisation
- méthodes 'en ligne' : chaque itération coûte $O(n)$

Formulation Semi-Duale

Si l'une des mesures est discrète, e.g.

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

En exploitant la condition de premier ordre du dual (relation entre v et u), on obtient la formulation *semi-duale* :

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

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

Cas Semi-Discret : SGD

On cherche à résoudre

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

par montée de gradient sur $G_\varepsilon(\mathbf{v})$.

Problème : On ne sait pas calculer le gradient (α n'est pas connue)

Idée : A chaque itération, on tire $x^{(k)} \sim \alpha$ et $\nabla g_\varepsilon^{x^{(k)}}$ sert d'approximation pour ∇G_ε.

Cas Semi-Discret : SGD

Les itérées de SGD sont de la forme :

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

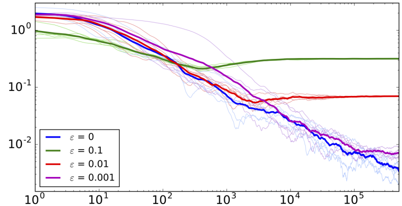
Proposition (Convergence de SGD)

Soit $\mathbf{v}_\varepsilon^*$ un minimiseur du semi-dual et $\bar{\mathbf{v}}^{(k)} \stackrel{\text{def.}}{=} \frac{1}{k} \sum_{i=1}^k \mathbf{v}^{(k)}$ la moyenne des itérées de SGD. Alors

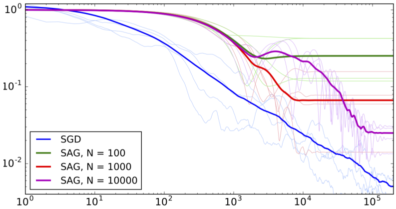
$$|G_\varepsilon(\mathbf{v}_\varepsilon^*) - G_\varepsilon(\bar{\mathbf{v}}^{(k)})| = O(1/\sqrt{k}).$$

Complexité de chaque itération $O(n)$.

Cas Semi-Discret : SGD - Application



(a) convergence de SGD pour différentes régularisations ε



(b) comparaison de SGD (bleu) contre un algorithme discret

Cas Discret : SAG

Deux mesures sont discrètes : $\alpha = \sum_{j=1}^{m} \boldsymbol{\alpha}_j \delta_{x_j}$; $\beta = \sum_{i=1}^{n} \boldsymbol{\beta}_i \delta_{y_i}$.
Le semi dual devient un problème de maximization de m fonctions :

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^n} \frac{1}{m} \sum_{j=1}^{m} \left[g_\varepsilon^{x_j}(\mathbf{v}) \right] \qquad (\mathcal{S}_\varepsilon)$$

On résout le problème avec l'algorithme **Stochastic Averaged Gradients** (SAG)
$\rightarrow$ même idée que SGD mais approximation du gradient différente.

Proposition (Convergence de SAG)

Soit $\mathbf{v}_\varepsilon^*$ un minimiseur du problème semi-dual. Alors $\mathbf{v}^{(k)}$ vérifie

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

Complexité de chaque iteration $O(n)$.

Cas Discret : SAG - Application

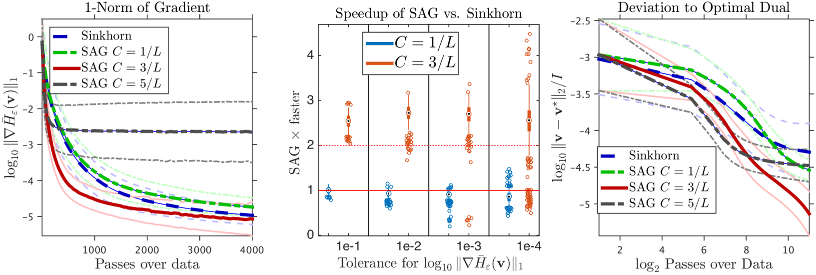


Figure 8 – Calcul de 595 ‘word mover’s distances’ 2 à 2 pour 35 documents, représentés comme des histogrammes avec $n = 20,000$.

Cas Continu : Formulation Duale

Idée : Remplacer les potentiels(u, v) dans le dual par leur expansion dans un RKHS bien choisi

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

Le problème devient

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

avec

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

Cas Continu : Kernel-SGD

Soit $\mathcal{H}$ un RKHS avec noyau κ. Les itérées de Kernel-SGD s'écrivent :

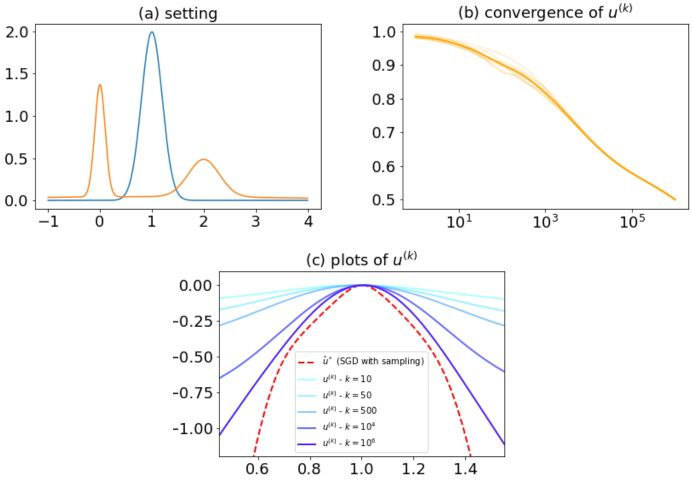
$$\begin{cases} u^{(k)} & \overset{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, x_i) \\ v^{(k)} & \overset{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, y_i)), \end{cases} \quad \text{avec} \quad \begin{cases} (x_i)_{i=1\ldots k} \sim \alpha \\ (y_i)_{i=1\ldots k} \sim \beta \end{cases}$$

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

Proposition (Convergence de Kernel-SGD)

Si α et β sont à support borné dans $\mathbb{R}^d$, alors pour κ le noyau de Matern ou un noyau universel (e.g. Gaussien) les itérées $(u^{(k)}, v^{(k)})$ convergent vers une solution du dual $(\mathcal{D}_\varepsilon)$.

Cas Continu : Kernel-SGD - Illustration



Cas Continu : Kernel-SGD - Accélération

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

Problème : l'itération k a un coût $O(k)$

Idée : remplacer le noyau κ par une approximation de la forme

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

$\rightarrow$ Le coût de chaque itération est alors fixe $O(p)$.

Exemples : Décomposition de Cholesky, Random Fourier Features (RFF)

Cas Continu : Kernel-SGD - Accélération

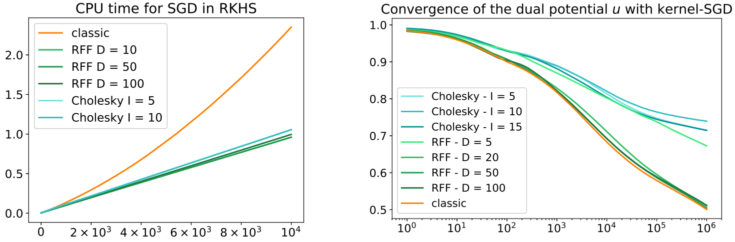


Figure 9 – Effets de la procédure d'accélération le temps de calcul et la précision

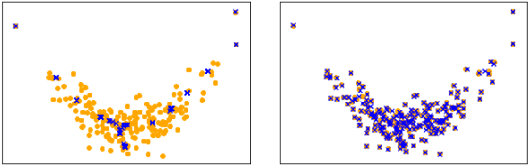
→ Pour 10^6 itérations, kernel-SGD prend 6 heures
→ L'accélération RFF avec $D = 20$ prend 3 minutes, et obtient la même précison !

1. Notions de Distance entre Mesures
2. Régularisation Entropique du Transport Optimal
3. Les Divergences de Sinkhorn : Interpolation entre TO et MMD
4. Apprentissage Non-Supervisé avec les Divergences de Sinkhorn
5. Optimisation Stochastique pour le Transport Régularisé
6. Conclusion

Contributions Principales

- Divergences de Sinkhorn :
  - Débiaisage du transport régularisé,

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



Contributions Principales

- Divergences de Sinkhorn :
  - Débiaisage du transport régularisé,
  - Interpolation entre TO et MMD,
  - Application aux modèles génératifs (type GAN) grâce à la différentiation automatique,

Contributions Principales

- Divergences de Sinkhorn :
  - Débiaisage du transport régularisé,
  - Interpolation entre TO et MMD,
  - Application aux modèles génératifs (type GAN) grâce à la différentiation automatique,
- Sample complexity du transport régularisé
  $\rightarrow$ **une régularisation suffisante casse le fléau de la dimension,**

Contributions Principales

- Divergences de Sinkhorn :
  - Débiaisage du transport régularisé,
  - Interpolation entre TO et MMD,
  - Application aux modèles génératifs (type GAN) grâce à la différentiation automatique,
- Sample complexity du transport régularisé
  $\rightarrow$ **une régularisation suffisante casse le fléau de la dimension,**
- Méthodes d'optimisation en ligne pour le transport régularisé sous toutes ses formes : discret / semi-discret / continu

En bref

Les Divergences de Sinkhorn présentent de bonnes propriétés pour les applications en apprentissage statistique, comme illustré sur les modèles génératifs :

- propriétés géométriques héritées du transport
- meilleure sample complexity grâce à la régularisation
- algorithmes rapides pour l'utilisation dans les problèmes de ML.

Perspectives

- Barycentres de Divergences de Sinkhorn
  $\rightarrow$ effet du débiaisage sur le barycentre ?
- Evaluation des modèles génératifs en utilisant les DS comme métrique sur les modèles appris
- Peut-on casser le fléau de la dimension pour l'estimation du Transport Optimal (non régularisé) ?