Régularisation entropique décroissante pour le transport optimal semi-discret

Ferdinand Genans [1] & Antoine Godichon-Baggioni [2] & Olivier Wintenberger [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, France, fgenans@lpsm.paris*
[2] *Laboratoire de Probabilités, Statistique et Modélisation, France, antoine.godichon_ baggioni@sorbonne-universite.fr*
[3] *Laboratoire de Probabilités, Statistique et Modélisation, France, olivier.wintenberger@sorbonne-universite.fr*

Résumé. Le transport optimal est une méthode de comparaison des distributions de probabilité, utilisée dans diverses disciplines, y compris l'économie, l'apprentissage automatique et la biologie. Néanmoins, résoudre les problèmes de transport optimal est coûteux en calcul, ce qui a incité à introduire le transport optimal entropique. Cette dernière approche incorpore un terme de régularisation entropique pour faciliter des calculs plus abordables et efficaces. La sélection du paramètre de régularisation ε devient alors une préoccupation pratique. Une régularisation plus faible est préférée pour la précision mais est souvent associée à une convergence plus lente, établissant un compromis entre le taux de convergence et la précision. Dans le cadre du transport optimal semi-discret, nous introduisons un algorithme de Descente de Gradient Stochastique, incorporant un schéma de régularisation décroissante.

Mots-clés. Transport Optimal, Optimisation Stochastique, Descente de Gradient, Régularisation Entropique.

Abstract. Optimal transport is a method for comparing probability distributions, used in various disciplines, including economics, machine learning, and biology. Nonetheless, solving optimal transport problems is computationally expensive, which has led to the introduction of entropic optimal transport. This latter approach incorporates an entropic regularization term to facilitate more affordable and efficient calculations. The selection of the regularization parameter ε then becomes a practical concern. Weaker regularization is preferred for accuracy but is often associated with slower convergence, establishing a trade-off between convergence rate and accuracy. In the context of semi-discrete optimal transport, we introduce a Stochastic Gradient Descent algorithm, incorporating a decreasing regularization scheme.

Keywords. Optimal transport, Stochastic Optimization, Gradient Descent, Entropic Regularization.

1 Introduction

1.1 Transport Optimal

La théorie du transport optimal (TO) remonte aux travaux de Monge (1781) et a été généralisée plus tard par Kantorovich (1942). Ce cadre offre une méthode puissante pour comparer et transporter des mesures de probabilité et a démontré son efficacité dans divers domaines comme l'apprentissage automatique (Courty et al., 2014; Genevay et al., 2018; Bigot et al., 2017), la biologie (Schiebinger et al., 2019), et l'économie (Galichon, 2018).

Étant donné des mesures de probabilité source et cible $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ et une fonction $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ représentant le coût pour déplacer la masse, la formulation de Kantorovich du transport optimal est

$$\mathrm{TO}_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \mathrm{d}\pi(x, y). \tag{1}$$

Habituellement, nous utilisons $c(x, y) = \|x - y\|^p$ pour $p \geq 1$ et dans ce cas, $\mathrm{TO}_c^{\frac{1}{p}}$ représente une distance entre mesures de probabilité, appelée la distance de Wasserstein p (Villani, 2009; Santambrogio, 2015). Dans ce travail, nous nous concentrerons sur le coût $c(x, y) = \frac{1}{2}\|x - y\|^2$.

Distance de Wasserstein 2 et théorème de Brenier.

Soit $\mu \in \mathcal{P}(\mathbb{R}^d)$ une mesure source ayant une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ et une mesure cible $\nu \in \mathcal{P}(\mathbb{R}^d)$. De plus, supposons que μ et ν ont des moments d'ordre deux. Dans ce contexte, le théorème de Brenier (Brenier, 1991) stipule qu'il existe une carte unique $T_{\mu,\nu}$, que nous appelons la carte de Brenier, telle que

$$T^*_{\mu,\nu} := \operatorname*{argmin}_{T \in \mathcal{T}(\mu,\nu)} \int \frac{1}{2}\|x - T(x)\|^2 \mathrm{d}\mu(x), \tag{2}$$

où $\mathcal{T}(\mu, \nu) := \left\{T : \mathbb{R}^d \times \mathbb{R}^d, \mu(T^{-1}(B)) = \nu(B) \text{ pour tous } B \in \mathcal{B}(\mathbb{R}^d)\right\}$ est l'ensemble des cartes transportant μ sur ν.

Cette carte résout également la formulation de Kantorovich du transport optimal dans le sens où la carte $(\mathrm{Id}, T_{\mu,\nu})_{\#}\mu$ résout (1) pour le coût $c(x, y) = \frac{1}{2}\|x - y\|^2$. De plus, $T^*_{\mu,\nu}$ est le gradient d'une fonction convexe et nous avons que

$$T^*_{\mu,\nu}(x) = x - \nabla f^*(x), \tag{3}$$

où f^* est un potentiel de Kantorovich, c'est-à-dire, (f^*, f^{*c}) maximise le problème dual de Kantorovich

$$f^* \in \arg\max_{f \in C(\mathbb{R}^d)} \int_{\mathbb{R}^d} f \mathrm{d}\mu + \int_{\mathbb{R}^d} f^c \mathrm{d}\nu, \tag{4}$$

où nous

notons f^c la c-transformation de f donnée pour tout $y \in \mathbb{R}^d$ par

$$f^c(y) := \inf_{x \in \mathbb{R}^d} \left[\frac{1}{2}\|x-y\|^2 - f(x) \right].$$

Couplage de transport optimal entropique et formulation duale.

L'ajout d'un terme entropique au transport optimal a été présenté dans les travaux de Cuturi (2013) pour le cas discret et est devenu populaire grâce à son calcul efficace via l'algorithme de Sinkhorn. Avec un $\varepsilon > 0$ fixé, le Transport Optimal Entropique (TOE) reformule Kantorovich comme

$$\mathrm{TO}_c^\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \mid \mu \otimes \nu), \tag{5}$$

où

$$\mathrm{KL}(\pi \mid \mu \otimes \nu) \overset{\text{def.}}{=} \int_{\mathbb{R}^d \times \mathbb{R}^d} \ln\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)} \right) \mathrm{d}\pi(x,y)$$

est l'entropie relative entre π et $\mu \otimes \nu$. Comme mentionné dans Nutz and Wiesel (2022), le TOE a une solution unique $\gamma_{\mu,\nu}^{\varepsilon,*} \in \Pi(\mu,\nu)$ et a la densité suivante :

$$\frac{\mathrm{d}\gamma_{\mu,\nu}^{\varepsilon,*}}{\mathrm{d}(\mu \otimes \nu)}(x,y) = \exp\left(\frac{f_\varepsilon^*(x) + \mathbf{g}_\varepsilon^*(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon} \right), \tag{6}$$

pour deux fonctions mesurables $f_\varepsilon^* : \mathbb{R}^d \to \mathbb{R}$ et $\mathbf{g}_\varepsilon^* : \mathbb{R}^d \to \mathbb{R}$, parfois appelées les potentiels de Schrödinger. Ces potentiels sont uniques à une constante près ajoutée à une fonction et soustraite à l'autre (Nutz and Wiesel, 2022) et solution du problème dual du TOE

$$\sup_{f \in L^1(\mu), g \in L^1(\nu)} \int f \mathrm{d}\mu + \int g \mathrm{d}\nu - \varepsilon \int \exp\left(\frac{f(x) + g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon} \right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) + \varepsilon. \tag{7}$$

En notant (f^*, f^{*c}), résolvant (4), Carlier et al. (2017) fournit la convergence de $(f_\varepsilon^*, g_\varepsilon^*)$ vers (f^*, f^{*c}) lorsque $\varepsilon \to 0$.

2 Cadre: transport optimal semi-discret

Dans ce travail, nous nous intéressons au TO semi-discret, cas où la mesure source μ est continue, tandis que la mesure cible ν est discrète. Nous considérons les hypothèses suivantes, présentes dans Delalande (2022) et Divol et al. (2022).

Hypothèses. ***(A)*** *La mesure source μ est une mesure continue avec* $\mathrm{Supp}(\mu) \subset B(0,R)$ *et une densité p satisfaisant* $0 < p_{\min} \leq p \leq p_{\max} < \infty$ *pour certaines constantes* $p_{\min}, p_{\max}$ *et* R.

(B) *La mesure cible ν est une mesure discrète de la forme*

$$\nu = \sum_{j=1}^{M} b_j \delta_{\mathbf{y_j}}$$

avec $\mathbf{b} = (b_1, \ldots, b_M)$ *ses poids de probabilité et* $(\mathbf{y_1}, \ldots, \mathbf{y_M}) \in B(0, R)^M$ *son support.*

Formulation semi-duale.

Supposant que nous pouvons échantillonner à partir de la mesure source μ, Genevay et al. (2016) a suggéré d'utiliser une formulation semi-duale pour résoudre le problème du TOE semi-discret. Cela aboutit à un problème d'optimisation convexe lisse consistant à trouver un minimiseur de la fonction H_ε définie pour tout $\mathbf{g} = (g_1, \ldots, g_M)$ par

$$H_\varepsilon(\mathbf{g}) \stackrel{\text{def.}}{=} -\int_{\mathbb{R}^d} \mathbf{g}^{c,\varepsilon}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) - \sum_{j=1}^{M} g_j b_j, \tag{8}$$

où pour $\mathbf{x} \in \mathbb{R}^d$, $g^{c,\varepsilon}(\mathbf{x})$, souvent désigné comme la e (c, ε)-transformée (Peyré et al., 2019) est défini par

$$\mathbf{g}^{c,\varepsilon}(x) := -\varepsilon \ln \left(\sum_{j=1}^{M} \exp\left(\frac{g_j - \frac{1}{2}\|x - y_j\|^2}{\varepsilon} \right) b_j \right).$$

Notons que la formulation (8) permet de ne pas avoir un biais de discrétisation, comme on aurait eu en échantillonnant directement μ pour appliquer l'algorithme de Sinkhorn Cuturi (2013). Nous pouvons calculer le coût du TOE avec (3) puisque nous avons $\min_{\mathbf{g} \in \mathbb{R}^M} H_\varepsilon(\mathbf{g}) = \mathrm{OT}_c^\varepsilon(\mu, \nu)$ (Genevay et al., 2016). L'un des intérêts de cette formule est que la fonctionnelle H_ε est différentiable et son gradient est défini pour tout $\mathbf{g} \in \mathbb{R}^M, \mathbf{x} \in \mathbb{R}^d$ et toute coordonnée j par

$$\nabla_{\mathbf{g}} H_\varepsilon(\mathbf{g}, \mathbf{x})_j = -b_j + \int_{\mathbb{R}^d} \chi_j^\varepsilon(\mathbf{x}, \mathbf{g}) \mathrm{d}\mu(\mathbf{x}),$$

où pour $\mathbf{x} \in \mathbb{R}^d, \mathbf{g} \in \mathbb{R}^M$, nous avons

$$\chi_j^\varepsilon(\mathbf{x}, \mathbf{g}) = \frac{\exp\left(\frac{-c(\mathbf{x}, \mathbf{y}_j) + g_j}{\varepsilon}\right) b_j}{\sum_{k=1}^{M} \exp\left(\frac{-c(\mathbf{x}, \mathbf{y}_k) + g_k}{\varepsilon}\right) b_k}.$$

Résoudre le semi-discret avec l'optimisation stochastique et en ligne.

Étant donné une mesure source arbitraire dont nous pouvons échantillonner, la formulation semi-duale est particulièrement pertinente pour l'estimation du potentiel de Brenier. En effet, cette formulation nous permet de ne pas discrétiser la mesure source et ainsi de surmonter un biais de discrétisation qui aurait été obligatoire pour utiliser la Programmation Linéaire

ou les algorithmes de Sinkhorn. On peut considérer le problème de minimisation donné par (8) comme un problème d'Optimisation Stochastique. En effet, la fonctionnelle H_ε peut être réécrite comme

$$H_\varepsilon(\mathbf{g}) = \mathbb{E}\left[h_\varepsilon(X, \mathbf{g})\right] = \int_{\mathbb{R}^d} h_\varepsilon(\mathbf{x}, \mathbf{g}) \mathrm{d}\mu(\mathbf{x}),$$

où X représente une variable aléatoire tirée de la distribution source μ, et pour tout $(\mathbf{x}, \mathbf{g}) \in \mathbb{R}^d \times \mathbb{R}^M$

$$h_\varepsilon(\mathbf{x}, \mathbf{g}) = -\mathbf{g}^{c,\varepsilon}(\mathbf{x}) - \sum_{j=1}^{M} g_j b_j.$$

Pour tout $\mathbf{g} \in \mathbb{R}^M$, étant donné $\mathbf{x} \sim \mu$, un estimateur non biaisé du gradient est alors, pour toute coordonnée $1 \leq j \leq M$

$$\nabla_{\mathbf{g}} h_\varepsilon(\mathbf{g}, \mathbf{x})_j = -b_j + \chi_j^\varepsilon(\mathbf{x}, \mathbf{g}).$$

Pour une régularisation fixe $\varepsilon > 0$, les méthodes stochastiques de premier ordre sont principalement employées pour ce cadre. Spécifiquement, dans Genevay et al. (2016), le Descente de Gradient Stochastique Moyennée (ASGD) est utilisée. Ayant une initialisation $\mathbf{g}_1 \in \mathbb{R}^M$, et fixant $\gamma_t = \gamma_0/t^b$ avec $b \in (0, 1)$, ASGD consiste à chaque itération à considérer un bloc de $n \geq 1$ données i.i.d $\mathbf{x}_t = (\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,n}) \sim \mu^{\otimes n}$ et ensuite faire la mise à jour

$$\mathbf{g}_{t+1} = \mathbf{g}_t - \gamma_t \nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_t), \tag{9}$$

$$\bar{\mathbf{g}}_{t+1} = \frac{1}{t+2}\mathbf{g}_{t+1} + \frac{t+1}{t+2}\bar{\mathbf{g}}_t \tag{10}$$

où $\nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_t) := \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_{t,i})$.. Nous parlons de moyennisation puisque $\bar{\mathbf{g}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{g}_t$. A noter que les algorithmes ASGD sont largement étudiés (Polyak and Juditsky, 1992; Pelletier, 2000; Bach and Moulines, 2013) et on peut se référer Bercu and Bigot (2021) pour le cas spécifique du TOE.

3 Régularisation entropique décroissante

Comme nous l'avons vu, la régularisation entropique permet d'obtenir une fonction objectif différentiable et lisse, facilitant ainsi l'utilisation des méthodes d'optimisation basées sur le gradient. Cependant, le choix du coefficient de régularisation ε reste critique. Plus ε est faible, plus l'algorithme se rapproche du problème non régularisé. Sa convergence, cependant, ralentit en fonction de ε, établissant un équilibre entre la vitesse de convergence et la précision de l'approximation. Pour approximer le TO non régularisé, notamment le Transport de Brenier, l'utilisation d'une régularisation décroissante au fil du temps, notée ε_t, semble naturel dans un cadre stochastique ou en ligne. En se basant sur le travail récent de Delalande (2022) sur la différence d'optimiseur entre deux régularisations différentes lorsque nos mesures vérifient les hypothèses (A) et (B), nous introduisons dans cette section un algorithme de Descente de Gradient Stochastique avec régularisation décroissante.

Étape de projection.

Dans la descente de gradient stochastique et en ligne, l'incorporation d'une étape de projection avec une connaissance préalable de l'espace des solutions peut conduire à une convergence plus rapide. Ceci est particulièrement vrai lorsque l'étape de projection est presque sans coût computationnel. Nous introduisons ici deux ensembles de projection différents qui peuvent être utilisés pour les problèmes OT et EOT, en tirant parti de la régularité des solutions dans (8). À cette fin, remarquons que pour toute fonction de coût bornée dans $\mathrm{Supp}(\mu) \times \mathrm{Supp}(\nu)$ et pour toute régularisation $\varepsilon > 0$, il existe une solution optimale $\mathbf{g}_\varepsilon^* = (g_1^*, \dots, g_M^*)$ de la formulation semi-duale telle que pour tous $j = 1, \dots, M$ nous avons $0 \leq g_j^* \leq \|c\|_\infty$, où nous notons $\|c\|_\infty$ le supremum de la fonction de coût sur $\mathrm{Supp}(\mu) \times \mathrm{Supp}(\nu)$ (voir Nutz and Wiesel (2022) par exemple). Nous pouvons donc également ajouter une étape de projection sur l'ensemble convexe compact suivant

$$C_\infty := [0, \|c\|_\infty]^M.$$

Cette projection est presque sans coût puisqu'elle consiste uniquement à clipper chaque coordonnée de notre vecteur. Cependant, pour certains coûts, nous pouvons trouver un meilleur ensemble de projection, en utilisant la régularité de la fonction de coût. Cet ensemble de projection est donné par le lemme suivant.

Lemme 1. *Si pour tous $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathbb{R}^d$, il existe des constantes K, β telles que*

$$|c(\mathbf{x}, \mathbf{y}) - c(\mathbf{x}, \mathbf{y}')| \leq K\|\mathbf{y} - \mathbf{y}'\|^\beta,$$

alors, pour tous $k = 1, \dots, M$, il existe une solution unique $\mathbf{g}^*$ à (3) dans l'ensemble convexe compact suivant

$$C_k := \{\mathbf{g} \in \mathbb{R}^M;\, g_k = 0 \text{ et } |g_j| \leq K\|\mathbf{y_k} - \mathbf{y_j}\|^\beta,\ j = 1, \dots, M\}\ .$$

Algorithme ASGD projeté avec régularisation décroissante.

Tirant parti de notre étape de projection, nous proposons de changer l'étape usuelle de descente de gradient avec régularisation fixe, par une régularisation décroissante ε_t avec $\varepsilon_t \to 0$, ce qui conduit à l'algorithme de gradient projeté suivant

$$\mathbf{g}_t = \Pi_C \left(\mathbf{g}_{t-1} - \frac{\gamma_1}{t^b} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{x}_t, \mathbf{g}_{t-1}) \right).$$

Ce mécanisme de régularisation offre l'avantage d'une forte régularisation en début d'algorithme, qui s'affine progressivement pour améliorer la précision de notre estimateur au fil du temps, visant ainsi à obtenir une estimation précise et non biaisée du potentiel de Brenier. La version complète de l'algorithme est décrite ci-dessous.

Algorithm 1 Decreasing Regularization Projected ASGD (DRPASGD)

Parameters: $(\gamma_1, \gamma, a, b, n)$
Initialize $\mathbf{g}_0 \in C$ and $\bar{\mathbf{g}}_0 = \mathbf{g}_0$
for $t = 0$ to $T-1$ **do**
  $\varepsilon_t = 1/t^a$
  $\gamma_t = \gamma_1/t^b$
  $x_t \sim \mu^{\otimes n}$
  $\mathbf{g}_{t+1} = \mathrm{Proj}_C\left(\mathbf{g}_t - \gamma_{t+1}\frac{1}{n}\sum_{k=1}^{n} \nabla_{\mathbf{g}} h_{\varepsilon_t}(x_{t,i}, \mathbf{g}_t)\right)$
  $\bar{\mathbf{g}}_{t+1} = \dfrac{1}{t+2}\mathbf{g}_{t+1} + \dfrac{t+1}{t+2}\bar{\mathbf{g}}_t$
end for
return $\bar{\mathbf{g}}_T$

3.1 Illustrations numériques

Convergence de l'estimateur.

Nous considérons une expérience synthétique pour l'illustration de la convergence de DRPASGD. Soit $\mu = \mathrm{Unif}\left([0,1]^{10}\right)$, et fixons $M = 100$. Nous générons aléatoirement $y_1, \ldots, Y_{100} \in [0,1]^{10}$, ainsi qu'un potentiel $\mathbf{g}^* \in [0,1]^{100}$ aléatoire, et considérons le transport optimal

$$T_0(x) = \mathrm{argmin}_{j \in [\![1,M]\!]} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{y}_j\|2 - g_j^* \right\}.$$

Nous définissons $\nu = (T_0)_\sharp P$, comme cela $\mathbf{g}^*$ est bien un potentiel optimal, et approximons les poids $\mathbf{b}$ par Monte-Carlo avec 10^6 échantillons de μ.

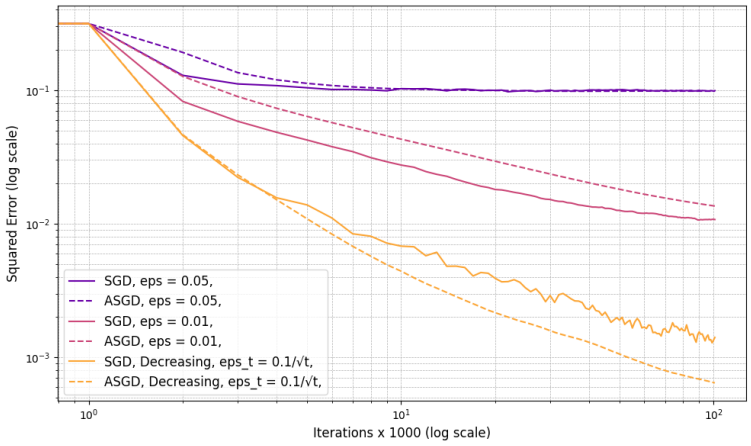


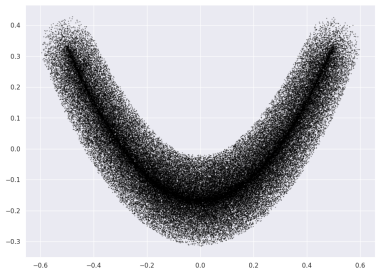
Figure 1: Évolution moyenne de l'erreur $\|\mathbf{g}_t - \mathbf{g}^*\|_2^2$ pour chaque estimateur, $t \in [0, 10^5]$, sur 20 expériences par configuration.

Nous illustrons dans la Figure 1 l'utilité de DRPASGD, en fixant $\gamma_t = \frac{5}{t^{3/4}}$ pour ASGD et DRPASGD, pour les deux algorithmes, nous utilisons la projection sur $[0,1]^{100}$ et choisissons

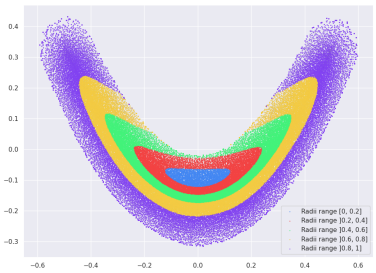
une régularisation décroissante $\varepsilon_t = 0.1/\sqrt{t}$. On constate que DRPASGD bénéficie d'une accélération tout en ne s'arrêtant pas à un optimum biaisé contrairement aux régularisations fixes.

Quantiles de Monge-Kantorovich.

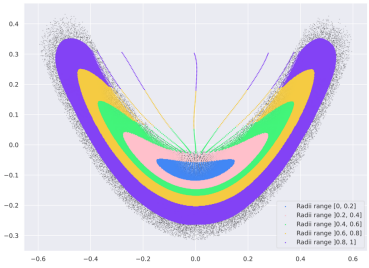
Nous illustrons ici notre méthode pour déterminer les régions quantiles de Monge-Kantorovich Chernozhukov et al. (2017), qui définissent une généralisation en dimension $d > 1$ de la notion habituelle de quantiles, en s'appuyant sur le théorème de Brenier Brenier (1991). Dans ce cas, la mesure source μ est la mesure uniforme sur la boule unité euclidienne. Nous fixons ici $d = 2$ et prenons comme mesure discrète cible,discrétisation d'une mesure en forme de "banane" avec $M = 10^5$ points, exemple que l'on peut retrouver dans Chernozhukov et al. (2017) et Bercu et al. (2023). Nous illustrons le résultat obtenu par DRPASGD et le comparons avec une régularisation fixe de $\varepsilon = 0.002$. En commençant au centre, chaque région de couleur correspond à la région quantile $[0.2k, 0.2(k+1)]$ où $k \in [0.4]$, c'est à dire, là où les points de $\mathcal{B}(0, 0.2(k+1)) \setminus \mathcal{B}(0, 0.2k)$ sont envoyés.



(a) Mesure discrète cible, $M = 10^5$.



(b) MK-régions, DRPASGD, $\varepsilon_t = 0.1/\sqrt{t}$.



(c) MK-régions, transport entropique, $\varepsilon = 0.002$.

Figure 2: MK-Quantiles pour notre distribution "banane", après $T = 10^6$ itérations.

Comme on le voit sur la figure 2, notre algorithme permet d'avoir des régions quantiles recouvrant toute la distribution, tandis qu'avec une régularisation fixe, le dernier contour ne contient pas toute la distribution.

Bibliographie

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

B. Bercu and J. Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. 2021.

B. Bercu, J. Bigot, and G. Thurin. Stochastic optimal transport in banach spaces for regularized estimation of multivariate quantiles. *arXiv preprint arXiv:2302.00982*, 2023.

J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic pca in the wasserstein space by convex pca. 2017.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. 2017.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances In Neural Information Processing Systems*, 26, 2013.

A. Delalande. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *International Conference On Artificial Intelligence And Statistics*, pages 1619–1642, 2022.

V. Divol, J. Niles-Weed, and A.-A. Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

A. Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances In Neural Information Processing Systems*, volume 29, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

M. Nutz and J. Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

M. Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM Journal on Control and Optimization*, 39(1):49–72, 2000.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

F. Santambrogio. Optimal transport for applied mathematicians. 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.