Approximation de mesures en grande dimension par des méthodes d'optimisation

Christophe Vauthier [a]
Quentin Mérigot [b]
Anna Korba [c]

Université Paris-Saclay

[a]Doctorant au LMO, Université Paris-Saclay
[b]Directeur de la thèse, LMO, Université Paris-Saclay
[c]Co-directrice, CREST, ENSAE

1. Introduction

Nous nous intéressons ici au problème de *quantisation uniforme*, consistant à approximer une mesure $\pi \in \mathcal{P}(\mathbb{R}^d)$ par une mesure discrète supportée sur un nombre fini de points. Autrement dit, étant donné une divergence D entre mesures de probabilités sur $\mathbb{R}^d$, nous voulons résoudre le problème suivant :

$$\min_{X \in (\mathbb{R}^d)^N} D\left(\mu_X := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}, \pi\right)$$

Ce problème est important en apprentissage statistique, avec des applications notamment dans l'étude des modèles génératifs tels que les réseaux génératifs adversariaux (GANs) ou les auto-encodeurs variationnels : typiquement, on dispose d'une mesure $\pi_r \in \mathcal{P}(\mathbb{R}^d)$, dans un espace de grande dimension, qu'on veut approcher à l'aide d'un réseau de neurones T_θ paramété par $\theta \in \Theta$ et qui engendre des éléments selon une distribution π_θ. Or, bien que les distributions π_r et π_θ soient souvent intractables, ce qui empêche de résoudre directement $\min_{\theta\in\Theta} D(\pi_\theta, \pi_r)$, on peut généralement facilement les échantillonner, ce qui permet de se ramener à l'étude d'un problème de quantisation.

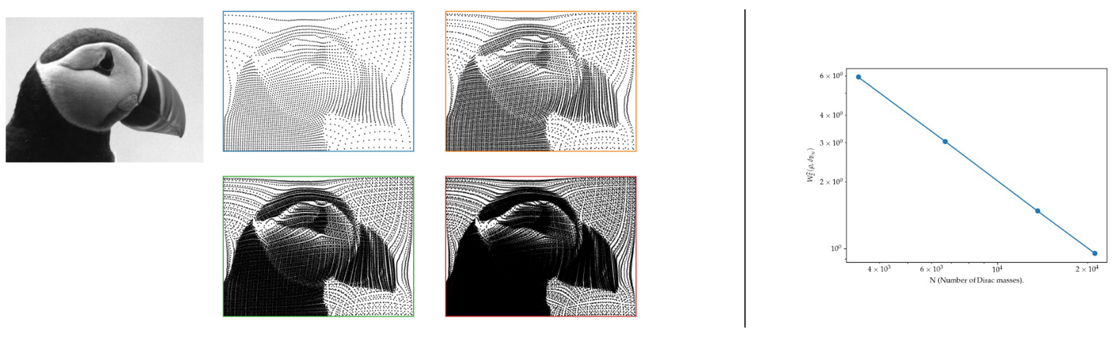

Figure 1: *Quantisation optimale d'une densité ρ correspondant à une image en niveaux de gris (Wikimedia Commons, CC BY-SA 3.0). À gauche : nuages de points obtenus après une étape de l'algorithme de Lloyd, en partant d'une grille régulière de taille $N \in \{3750, 7350, 15000, 43350\}$. À droite : erreur de quantisation $\mathrm{W}_2^2(\rho, \delta_{B_N})\ N^{-1.00}$ [2]*

2. Arrière-plan et notations

Soit $p \in [1, +\infty)$ et $\mathcal{P}_p(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int_{\mathbb{R}^d} \|x\|^p d\mu(x) < +\infty\right\}$ l'ensemble des mesures de probabilité sur $\mathbb{R}^d$ ayant un moment d'ordre p fini. La distance de Wasserstein d'ordre p entre $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ est définie comme

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y), \tag{1}$$

où $\|\cdot\|$ désigne la norme euclidienne, et $\Pi(\mu, \nu)$ l'ensemble des mesures de probabilité sur $\mathbb{R}^d \times \mathbb{R}^d$ dont les marginales par rapport aux première et seconde variables sont données respectivement par μ et ν (qu'on appelle *plans de transport* entre μ et ν).
La distance de Sliced-Wasserstein définit une métrique alternative pratique en tirant parti du fait qu'on sait calculer efficacement W_p^p pour les distributions univariées. Elle est définie par

$$\mathrm{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(P_{\theta\#}\mu, P_{\theta\#}\nu) d\theta, \quad \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d) \tag{2}$$

où $\mathbb{S}^{d-1}$ est la sphère unité de dimension $d-1$, $d\theta$ la distribution uniforme sur $\mathbb{S}^{d-1}$, et P_θ la projection $\mathbb{R}^d \to \mathbb{R}\theta$. Là où les méthodes standard utilisées pour résoudre le programme linéaire dans (1) ont une complexité computationnelle dans le pire cas en $\mathcal{O}(n^3 \log(n))$, et tendent à avoir un coût super-cubique en pratique, (2) se calcule en $O(Ln\log(n))$ par une méthode de Monte-Carlo (où L est le nombre de directions)

3. Erreur de quantisation minimale

On définit l'*erreur de quantisation minimale* de la distance de Slice-Wasserstein pour une mesure $\pi \in \mathcal{P}(\mathbb{R}^d)$ fixée par

$$e_N(\pi) := \min_{X_1,\dots,X_N \in \mathbb{R}^d} \mathrm{SW}_p\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}, \pi\right)$$

Quelles bornes supérieures et inférieures a-t-on sur cette erreur ? On sait que, sous certaines conditions sur π, $e_N(\pi)$ est au plus de l'ordre de $O(\frac{1}{\sqrt{N}})$ [3] et [1]. En revanche, aucune borne inférieure raisonnable pour l'erreur de quantisation n'a été prouvée à ce jour.

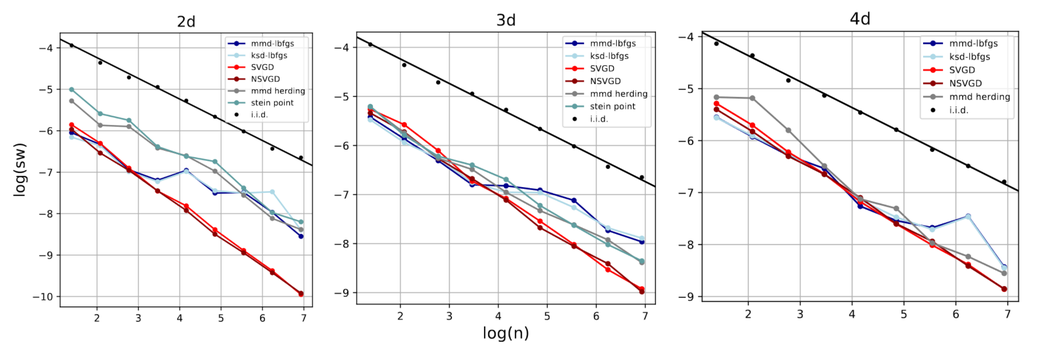

Figure 2: *Erreur de quantisation pour SW_1 obtenue avec divers algorithmes*
Avec $\pi \sim \mathcal{N}(0, \frac{1}{d} I_d)$, et $d \in \{2, 3, 4\}$. L'erreur de quantisation est de l'ordre de $N^{-\frac{1}{2}-\frac{1}{2d}}$, ce qui suggère qu'il est possibe d'améliorer les bornes théoriques connues de l'erreur de quantisation minimale [4]

4. Descente de gradient et stabilité

Pour calculer un nuage de points $X \in \mathbb{R}^{d\times N}$ avec une petite erreur, on veut minimiser par descente de gradient la fonction

$$F_N : X \in \mathbb{R}^{d\times N} \mapsto \frac{1}{2}\mathrm{SW}_2^2\left(\mu_X := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i},\ \pi\right)$$

Cependant, F_N n'est pas convexe (en fait, $\mu \mapsto \mathrm{SW}_p^p(\mu, \pi)$ est convexe, mais cette convexité est perdue lorsqu'on se restreint à des nuages de particules $X \mapsto \mu_X$), et il peut donc exister plus d'un minimum local. Par conséquent, il importe de comprendre le comportement des points critiques de F : en existe-t-il de haute énergie, vers lesquels la descente de gradient pourrait converger ? Ces questions se posaient déjà pour W_2^2, avec $G_N : X \mapsto \mathrm{W}_2^2(\mu_X, \pi)$.

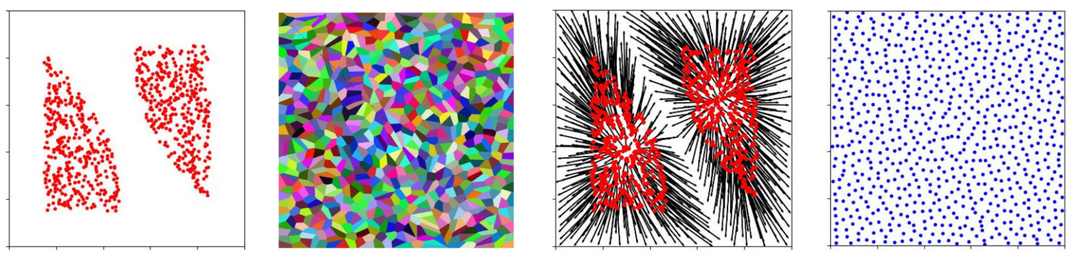

Figure 3: *Optimisation d'un nuage de points pour G_N avec l'algorithme de Lloyd*
De gauche à droite : (i) Un nuage de points Y_0 sur la densité uniforme π sur le carré $[0,1]\times[0,1]$ (ii) Cellules du carré envoyées sur chaque point de Y_0 par le transport optimal de π vers μ_{Y_0} (iii) Le gradient $\nabla G_N(Y_0)$ est $\frac{1}{N}(Y_0 - B_N(Y_0))$ où $B_N(Y_0)$ sont les barycentres des cellules associées à Y_0 (iv) Si les points Y_0 sont "assez dispersés", $G_N(B_N(Y_0))$ est de l'ordre de $N^{-1/d}$ (donc proche de l'optimum) [2]. L'algorithme de Lloyd consiste en fait à faire un pas de descente de gradient pour G_N avec un pas de N

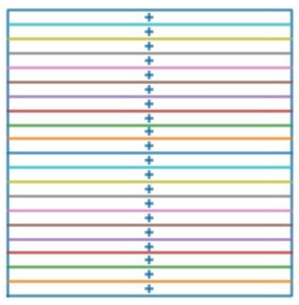

Figure 4: *Point critique de haute énergie pour G_N*

π est la mesure uniforme sur le carré $[0,1]\times[0,1]$. Le point critique Y_N est défini par

$$Y_N = \left(\left(\frac{1}{2}, \frac{1}{2N}\right), \left(\frac{1}{2}, \frac{3}{2N}\right), ..., \left(\frac{1}{2}, \frac{2N-1}{2N}\right)\right)$$

On a en fait $\lim_{N\to+\infty} \frac{G_N(Y_N)}{\min G_N} = +\infty$ [2]

References

[1] T. Manole, S. Balakrishnan, and L. Wasserman. Minimax Confidence Intervals for the Sliced Wasserstein Distance. *Electronic Journal of Statistics*, 16(1), Jan. 2022. arXiv:1909.07862 [math, stat].

[2] Q. Mérigot, F. Santambrogio, and C. Sarrazin. Non-asymptotic convergence bounds for wasserstein approximation using point clouds. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[3] K. Nadjahi. *Sliced-Wasserstein distance for large-scale machine learning : theory, methodology and extensions*. phdthesis, Institut Polytechnique de Paris, Nov. 2021.

[4] L. Xu, A. Korba, and D. Slepcev. Accurate Quantization of Measures via Interacting Particle-based Optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 24576–24595. PMLR, June 2022.