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

Christophe Vauthier

Octobre 2023

Contents

1 Le problème de quantisation uniforme **1**

2 Introduction au transport optimal **2**
2.1 Les problèmes de Monge et de Kantorovich 2
2.2 Solution de cas particuliers 4
2.3 Distances de Wasserstein 5
2.4 Distances de Sliced-Wasserstein 5

1 Le problème de quantisation uniforme

Nous nous intéressons ici au problème, dit 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 :

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

Il s'agit d'un problème important en apprentissage statistique. Il a des applications notamment dans l'étude des modèles génératifs tels que les réseaux génératifs adversariaux (GANs) [Arjovsky et al. (2017)] ou les auto-encodeurs variationnels [Tolstikhin et al. (2018)] : 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.

Un choix de divergence D courant est la *divergence de Kullback-Leibler* KL (aussi appelée *entropie relative*), définie par

$$KL(\mu, \nu) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x) d\rho(x)$$

si μ et ν sont toutes deux absolument continues par rapport à une certaine mesure ρ, et de densités respectives f et g ; sinon

$$KL(\mu, \nu) = \infty$$

Cette divergence présente cependant l'inconvénient d'être infinie lorsque les mesures comparées sont supportées sur des variétés de petite dimension disjointes, ce qui peut arriver dans le cas des distributions qu'on peut se retrouver à comparer en pratique (par exemple c'est probablement le cas pour les π_r et π_θ discutés ci-dessus). C'est ce qui justifie de se tourner vers d'autres divergences, capables de mieux prendre en compte les propriétés géométriques des distributions, telles que les *distances de Wasserstein*, issues de la théorie du transport optimal, et que la section suivante aura pour but de définir.

2 Introduction au transport optimal

2.1 Les problèmes de Monge et de Kantorovich

Le premier à avoir formulé le problème connu aujourd'hui sous le nom de **transport optimal** fut Gaspard Monge, dans un mémoire soumis en 1781 à l'Académie des Sciences. Il y étudiait le problème des *déblais et des remblais*, qui consistait, comme son nom l'indique, à transporter des terres d'un endroit à un autre en minimisant le prix total du transport. La formulation moderne de ce problème est la suivante : soient $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ deux mesures de probabilités dans des espaces mesurables X et Y (par exemple $X = Y \subseteq \mathbb{R}^d$), et une fonction de coût $c : X \times Y \mapsto \mathbb{R}$ représentant le coût du déplacement d'un point $x \in X$ à un point $y \in Y$, on cherche à résoudre le *problème de Monge* donné par

$$\text{(MP)} \quad \inf\left\{ M(T) := \int_X c(x, T(x)) d\mu(x),\ T_{\#}\mu = \nu \right\}$$

Ici T parcourt l'ensemble des applications mesurables $T : X \mapsto Y$ telles que le *poussé en avant* de μ par T, c'est-à-dire la mesure de probabilités sur Y notée $T_{\#}\mu$ et définie par $A \mapsto \mu(T^{-1}(A))$, est égal à ν. Une telle application T est appelée *application de transport*.

En raison de la contrainte sur T (qui, en particulier, n'a pas de propriété de convexité), ce problème est longtemps resté difficile à attaquer, jusqu'à la percée réalisée par le mathématicien Leonid Kantorovich dans les années 1940. Celui-ci a introduit une généralisation du problème de Monge, appelée *problème*

de Kantorovich :

$$\text{(KP)} \quad \inf \left\{ K(\gamma) := \int_{X\times Y} c(x,y) d\gamma(x,y),\ \gamma \in \Pi(\mu,\nu) \right\}$$

avec

$$\Pi(\mu,\nu) := \{\gamma \in \mathcal{P}(X\times Y) \mid \pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu\}$$

où π_1 et π_2 sont les projections de $X\times Y$ sur X et Y respectivement. En d'autres termes, au lieu de considérer une application de transport T, on considère un *plan de transport* γ, qui est une mesure de probabilités sur $X \times Y$ dont les mesures marginales sont μ et ν (concrètement, cela signifie qu'au lieu d'envoyer toute la terre d'un point $x \in X$ vers un point $y = T(x)$, on s'autorise à scinder la masse de terre en x et à la répartir sur Y). Le problème de Kantorovich est bien une généralisation de celui de Monge : en effet, si T est une application de transport, $\gamma_T = (Id, T)_{\#}\mu$ est un plan de transport tel que $M(T) = K(\gamma_T)$. Ainsi on a bien $\min(\text{KP}) \leq \min(\text{MP})$.

Le grand avantage du problème de Kantorivich par rapport à celui de Monge est que l'espace $\Pi(\mu,\nu)$ des plans de transport est beaucoup plus facile à manipuler que celui des applications de transport. En effet, il est convexe et non vide (car $\mu\otimes\nu \in \Pi(\mu,\nu)$). De plus, on peut formuler le *problème dual* de celui de Kantorovich : on commence par remarquer que si $\gamma \in \mathcal{M}_+(X\times Y)$ est une mesure (positive) sur $X\times Y$, alors

$$\sup_{\varphi\in C_b(X)} \int_X \varphi d\mu - \int_X \varphi d\pi_{1\#}\gamma = \begin{cases} 0 & \text{si} \quad \pi_{1\#}\gamma = \mu \\ +\infty & \text{sinon} \end{cases}$$

où $C_b(X)$ est l'espace des applications continues bornées sur X. Par conséquent

$$\begin{aligned} \inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c d\gamma &= \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{X\times Y} c d\gamma + \sup_{\varphi,\psi} \int_X \varphi d(\mu - \pi_{1\#}\gamma) + \int_Y \psi d(\nu - \pi_{2\#}\gamma) \\ &= \inf_{\gamma} \sup_{\varphi,\psi} \int_{X\times Y} (c - \varphi - \psi) d\gamma + \int_X \varphi d\mu + \int_Y \psi d\nu \end{aligned}$$

L'interversion du supremum et de l'infimum donne alors le problème dual (DP)

$$\text{(DP)} \quad \sup_{\varphi,\psi} \int_X \varphi d\mu + \int_Y \psi d\nu + \inf_{\gamma} \int_{X\times Y} (c - \varphi - \psi) d\gamma$$

or

$$\inf_{\gamma} \int_{X\times Y} (c - \varphi - \psi) d\gamma = \begin{cases} 0 & \text{si} \quad \varphi + \psi \leq c \\ -\infty & \text{sinon} \end{cases}$$

et le problème dual peut donc être mis sous la forme

$$\text{(DP)} \quad \sup \left\{ \int_X \varphi d\mu + \int_Y \psi d\nu \mid \varphi \in C_b(X), \psi \in C_b(Y), \varphi + \psi \leq c \right\}$$

Il n'est pas difficile de vérifier que $\sup(\text{DP}) \leq \inf(\text{KP})$. De plus, si on définit pour $\varphi : X \mapsto \mathbb{R} \cup \{-\infty\}$ (non égale à $-\infty$ partout) sa c-transformée $\varphi^c : Y \mapsto \mathbb{R} \cup \{-\infty\}, y \mapsto \inf_{x \in X} c(x,y) - \varphi(x)$, alors on voit que si X et Y sont compacts, on a

$$(\text{DP}) \quad \sup \left\{ \int_X \varphi d\mu + \int_Y \varphi^c d\nu \mid \varphi \in C(X) \right\}$$

où cette fois φ parcourt l'espace $C(X)$ des fonctions continues sur X. On appelle en outre *potentiel de Kantorovich* toute fonction φ qui réalise le sup.

2.2 Solution de cas particuliers

Armés de ces définitions, nous pouvons énoncer quelques résultats (pour leur démonstration, voir [Santambrogio (2015)], Théorèmes 1.17 et 1.22) :

Théorème 2.1. *Soient μ et ν des mesures de probabilités sur un compact $\Omega \subseteq \mathbb{R}^d$, alors il existe un plan de transport optimal γ pour toute fonction de coût $c(x,y) = h(x-y)$ où h est strictement convexe. De plus, si μ est absolument continue et $\partial\Omega$ est négligeable, alors γ est unique et de la forme $(Id, T)_{\#}\mu$ avec T une application de transport optimale. En outre il existe un potentiel de Kantorovich φ, tel que pour presque tout $x \in \Omega$,*

$$T(x) = x - (\nabla h)^{-1}(\nabla \varphi(x))$$

Théorème 2.2. *Soient μ et ν des mesures de probabilités sur $\mathbb{R}^d$ de moment d'ordre 2 fini, et $c(x,y) = \frac{1}{2}|x-y|^2$. On suppose de plus que μ ne donne pas de masse à des surfaces C^2 de dimension $d-1$. Alors le résultat du théorème précédent s'applique, et l'application de transport optimale T est de la forme $T = \nabla u$ avec u une fonction convexe.*

Un cas particulier dans lequel le problème se résout de façon très simple est le cas en une dimension (pour un traitement plus détaillé, voir [Santambrogio (2015)], chapitre 2.1). En effet, l'application de transport se construit dans ce cas très simplement à partir des fonctions de distribution cumulatives des mesures. Pour rappel, si μ est une mesure de probabilités sur $\mathbb{R}$, sa fonction de distribution cumulative (FDC) $F_\mu : \mathbb{R} \mapsto [0;1]$ est définie par $F_\mu(a) := \mu((-\infty; a])$ pour tout $a \in \mathbb{R}$. On peut alors définir son *pseudo-inverse* par

$$F_\mu^{[-1]}(x) := \inf\{t \in \mathbb{R} \mid F_\mu(t) \geq x\}, \; x \in [0;1]$$

On a alors le résultat suivant :

Théorème 2.3. *Soient $\mu, \nu \in \mathcal{P}(\mathbb{R})$, et une fonction de coût de la forme $c(x,y) = h(x-y)$ avec h strictement convexe, telle que $\min$ (KP) $< +\infty$. Alors le problème de Kantorovich a une unique solution donnée par $\gamma := (F_\mu^{[-1]}, F_\nu^{[-1]})_{\#}\mathcal{L}^1_{|[0;1]}$; où $\mathcal{L}^1_{|[0;1]}$ est la mesure de Lebesgue sur $\mathbb{R}$ restreinte au segment $[0;1]$. De plus si μ est sans atome, γ est induite par l'application de transport $T = F_\nu^{[-1]} \circ F_\mu$ qui est optimale.*

Donnons quelques exemples : si $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ et $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$ sont des mesures discrètes supportées sur le même nombre de points, avec $x_1 < ... < x_n$ et $y_1 < ... < y_n$, alors $T(x_i) = y_i$ est une application de transport optimale entre μ et ν pour tout coût de la forme $|x-y|^p, p \geq 1$. Si à la place on considère μ une mesure à densité, avec $\mu((-\infty; a_i]) = \frac{i}{n}$, alors l'application T qui envoie $(-\infty; a_1]$ sur y_1, $(a_1; a_2]$ sur y_2... est une application de transport optimale entre μ et ν.

2.3 Distances de Wasserstein

Soit $p \geq 1$ et $\mathcal{P}_p(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} |x|^d d\mu(x) < +\infty\}$ l'ensemble des mesures de probabilités de moment d'ordre p fini. On définit alors pour tout $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ la *distance de Wasserstein* d'ordre p entre μ et ν par

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int |x-y|^p d\gamma(x,y) \right)^{\frac{1}{p}}$$

On peut montrer qu'il s'agit bien d'une distance sur $\mathcal{P}_p(\mathbb{R}^d)$. De plus, lorsqu'on la restreint à $\mathcal{P}(\Omega)$ avec $\Omega \subseteq \mathbb{R}^d$ compact, alors la topologie induite par W_p est la topologie faible (en d'autres termes $W_p(\mu_n, \mu) \to 0$ ssi μ_n converge faiblement vers μ, c'est-à-dire $\langle \mu_n, \varphi \rangle \to \langle \mu, \varphi \rangle$ pour tout $\varphi \in C_b(\Omega)$).

La figure 2.3 permet d'illustrer certaines propriétés des distances de Wasserstein, en les comparant avec des distances usuelles comme la distance L^p (on considère des mesures absolument continues de densité respectives f et g). Là où la distance L^p ne tient compte que des distances "verticales" $|f(x) - g(x)|$, de sorte que $\|f(.+h) - g\|_p^p = \|f\|_p^p + \|g\|_p^p$ lorsque h est assez grand, la distance de Wasserstein $W_p^p(f(.+h), g)$ est de l'ordre de $|h|$ pour h assez grand. Nous voyons ainsi que les distances de Wasserstein permettent de tenir compte des distances "horizontales" entre les densités. C'est cette propriété, à laquelle nous avions fait allusion dans l'introduction, qui fait de ces distances de Wasserstein des substituts intéressants à la divergence de Kullback-Leibler.

2.4 Distances de Sliced-Wasserstein

Les distances de Wasserstein présentent cependant, en pratique, au moins deux inconvénients.

Le premier est la malédiction de la dimension (*curse of dimensionality* en langue anglaise). L'erreur minimale qu'on peut obtenir en approximant une mesure avec avec N points est de l'ordre de $N^{-\frac{1}{d}}$ ([Graf and Luschgy (2000)], [Kloeckner (2012)]) : en d'autres termes, pour $\pi \in \mathcal{P}(\mathbb{R}^d)$ une mesure à densité fixée, il existe une constante $C > 0$ telle que pour tout $N > 0$,

$$\min_{X_1,...,X_N \in \mathbb{R}^d} W_p\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}, \pi\right) > CN^{-\frac{1}{d}}$$

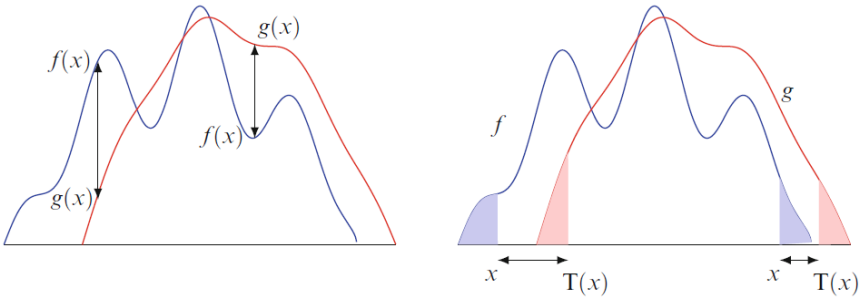

Figure 1: Distance L^p (à gauche) et distance de Wasserstein (à droite) entre deux densités f et g
Sur l'image de droite, la distance de transport T est calculée à partir des fonctions de distribution cumulatives, comme décrit précédemment, de sorte que les aires en bleu et en rouge sont égales.
Figure issue de [Santambrogio (2015)], figure 5.1.

Ainsi, plus on travaille sur un espace de grande dimension, plus il faut de points pour approximer précisément une mesure.

Le second est la complexité algorithmique du calcul de W_2 : les algorithmes pour calculer la distance de transport optimal entre deux mesures discrètes supportées sur n points μ_n et ν_n ont une complexité de l'ordre de $O(n^3 \log(n))$ (voir [Peyré and Cuturi (2019)], Chapitre 3).

Ces deux problèmes sont les raisons de l'intérêt porté aux distances dites de Sliced-Wasserstein. Elles tirent parti de la simplicité du calcul du transport optimal en une dimension : en effet, pour calculer la distance de transport optimal entre deux mesures discrètes en une dimension $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ et $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, il suffit de trier les n points de chaque distribution et de considérer leurs distances deux à deux, ce qui se fait en temps $O(n \log(n))$. Ainsi, pour comparer deux mesures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, on peut, pour chaque $\theta \in \mathbb{S}^{d-1}$ de la sphère unité, calculer la distance de Wasserstein entre les projections $W_p^p(P_{\theta\#}\mu, P_{\theta\#}\nu)$ (où $P_\theta(x) = \langle \theta, x \rangle$), et les intégrer sur $\mathbb{S}^{d-1}$ pour définir ainsi la *distance de Sliced-Wasserstein*

$$SW_p^p(\mu, \nu) := \int_{\mathbb{S}^{d-1}} W_p^p(P_{\theta\#}\mu, P_{\theta\#}\nu) d\theta, \ \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$$

où $d\theta$ est la mesure de Hausdorff sur $\mathbb{S}^{d-1}$, normalisée à $\int d\theta = 1$. Dans le cas de mesures à support discret, la distance $SW_p(\mu, \nu)$ peut être approximée par une méthode de Monte-Carlo en faisant la moyenne de $W_p^p(P_{\theta_i\#}\mu, P_{\theta_i\#}\nu)$ pour $\theta_1, ..., \theta_L \in \mathbb{S}^{d-1}$ L directions tirées (uniformément) au hasard, de sorte que le calcul se fasse en temps $O(Ldn + Ln\log(n))$, ce qui est (pour L assez petit) une amélioration substantielle par rapport à la distance de Wasserstein.

On peut vérifier que la distance SW_p est bien une distance (en particulier, que si $SW_p(\mu, \nu) = 0$, alors on a bien $\mu = \nu$). De plus, on a $SW_p \leq W_p$, et, sur un domaine compact $\Omega \subseteq \mathbb{R}^d$, SW_p induit la même topologie que W_p [Bonnotte (2013), chapitre 5]. Cependant, bien que les applications pratiques de la distance de Sliced-Wasserstein semblent prometteuses [Nadjahi (2021)], son comportement et ses propriétés théoriques restent mal connus.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arjovsky17a.html`.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, 2013. URL `http://www.theses.fr/2013PA112297`. Thèse de doctorat dirigée par Ambrosio, Luigi et Santambrogio, Filippo Mathématiques Paris 11 2013.

Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2000. ISBN 978-3-540-67394-1 978-3-540-45577-6. doi: 10.1007/BFb0103945. URL `http://link.springer.com/10.1007/BFb0103945`.

Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):343–359, April 2012. ISSN 1292-8119, 1262-3377. doi: 10.1051/cocv/2010100. URL `http://arxiv.org/abs/1003.1035`. arXiv:1003.1035 [math].

Kimia Nadjahi. *Sliced-Wasserstein distance for large-scale machine learning : theory, methodology and extensions*. phdthesis, Institut Polytechnique de Paris, November 2021. URL `https://theses.hal.science/tel-03533097`.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–206, 01 2019. doi: 10.1561/2200000073.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2. URL `https://link.springer.com/10.1007/978-3-319-20828-2`.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.