Sujet de stage de Master 2 Recherche
Institut de Mathématiques de Bordeaux - Université de Bordeaux
Toulouse School of Economics - Université de Toulouse

Algorithmes stochastiques pour le transport optimal de mesures de probabilités. Applications en modèles génératifs d'images (GAN) pour l'IA.

Encadrants : Bernard BERCU, Jérémie BIGOT & Sébastien GADAT
bernard.bercu@u-bordeaux.fr jeremie.bigot@u-bordeaux.fr
sebastien.gadat@tse-fr.eu

Contexte du sujet de stage

Dans de nombreux problèmes d'apprentissage statistique, les données se présentent sous une forme plus complexe que de simples vecteurs réels. C'est par exemple le cas lorsqu'on utilise des représentations de ces données sous la forme de mesures de probabilités. C'est en particulier nécessaire lorsqu'on étudie des exemples spécifiques en analyse de textes (nuages de mots pour l'étude du langage en *Natural Language Processing*), en vision par ordinateur ou traitement du signal. L'utilisation de la notion de distance de Wasserstein associée au problème de transport optimal entre des mesures de probabilités est un outil privilégié pour la comparaison de ce type de données. Pour une présentation de nombreux exemples des applications du transport optimal en apprentissage automatique (machine learning), on pourra consulter le livre récent (et tutoriels associés) de Cuturi & Peyré sur les aspects numériques du transport optimal à l'URL :

https://optimaltransport.github.io/

Un ouvrage de référence très complet sur les aspects mathématiques du transport optimal est également le livre [20] de Cédric Villani.

L'utilisation d'outils issus de la théorie du transport optimal permet ainsi d'atteindre les performances de l'état de l'art pour des applications variées, en particulier dans les modèles génératifs (d'images) à l'aide d'approches basés sur les GAN (Generative Adversarial Networks) [14]. Ces modèles génératifs (dont Yann LeCun a jugé qu'il s'agissait "de l'approche la plus intéressante en Machine Learning de ces 10 dernières années") ont des applications très vastes en I.A., dont on pourra trouver quelques exemples d'application ici :

https://en.wikipedia.org/wiki/Generative_adversarial_network

Au delà de son intérêt statistique incontournable, la pierre angulaire et le facteur limitant dans l'utilisation de distances de Wasserstein pour l'apprentissage statistique est le coût de calcul numérique (et plus précisément son approximation) du transport optimal entre deux mesures de probabilités. Le sujet de ce stage porte sur l'étude des propriétés asymptotiques (convergence presque sûre et théorème central limite) d'estimateurs d'une distance de transport (possiblement régularisée) basés sur des algorithmes stochastiques qui ont été récemment introduits dans [13].

Plus précisément, soient $\mathcal{X}$ et $\mathcal{Y}$ deux espaces métriques, et notons $\mathcal{M}_+^1(\mathcal{X})$ et $\mathcal{M}_+^1(\mathcal{Y})$ les espaces des mesures de probabilités à support sur $\mathcal{X}$ et $\mathcal{Y}$. Pour deux mesures $\mu \in \mathcal{M}_+^1(\mathcal{X})$ et $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, on note $\Pi(\mu, \nu)$ l'espace des mesures de probabilités à support sur l'espace produit $\mathcal{X} \times \mathcal{Y}$ dont les marginales sont μ et ν. Le problème du transport optimal (et sa version régularisée [8]) entre deux mesures $\mu \in \mathcal{M}_+^1(\mathcal{X})$ et $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ est alors définie par le problème de minimisation convexe suivant (formulation de Kantorovich) :

$$W_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu), \tag{1.1}$$

où $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$ est la fonction de coût du déplacement d'une masse de la position $x \in \mathcal{X}$ à $y \in \mathcal{Y}$, $\varepsilon \geq 0$ est un (possible) paramètre de regularisation, et KL est une divergence de type Kullback-Leibler entre un plan de transport π et la mesure produit $\xi = \mu\otimes\nu$ sur $\mathcal{X}\times\mathcal{Y}$ définie par

$$\mathrm{KL}(\pi|\xi) = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\log\Big(\frac{d\pi}{d\xi}(x,y)\Big) - 1\Big) d\pi(x,y).$$

Dans le cas particulier où $\mathcal{X} = \mathcal{Y}$ est un espace muni d'une métrique d et $c = d^p$ avec $p \geq 1$, la valeur $W_0^{1/p}(\mu,\nu)$ est classiquement appelée la p-distance Wasserstein distance sur l'ensemble des mesures de probabilités $\mathcal{M}_+^1(\mathcal{X})$ (avec une condition supplémentaire de moment).

La formulation (1.1) est un problème d'optimisation convexe délicat dont la minimisation n'admet généralement pas de forme explicite. Une formulation semi-duale de ce problème a été récemment introduite dans [13] qui permet d'écrire l'expression de $W_\varepsilon(\mu,\nu)$ comme le problème de maximisation suivant :

$$W_\varepsilon(\mu,\nu) = \max_{v\in\mathcal{C}(\mathcal{Y})} \mathbb{E}[h_\varepsilon(X,v)] = \max_{v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} h_\varepsilon(x,v)d\mu(x), \tag{1.2}$$

où

- $\mathcal{C}(\mathcal{Y})$ représente l'espace des fonctions continues sur $\mathcal{Y}$,
- $X \in \mathcal{X}$ est une variable aléatoire de loi μ,
- $\mathbb{E}$ correspond à l'espérance de la variable aléatoire $h_\varepsilon(X,v)$,
- $h_\varepsilon : \mathcal{X}\times\mathcal{C}(\mathcal{Y}) \to \mathbb{R}$ est une fonction connue et simple à calculer qui dépend de la loi ν et de la fonction de coût.

La formulation semi-duale (1.2) permet alors d'envisager un algorithme stochastique qui vise à approcher un maximiseur $v^* \in \mathcal{C}(\mathcal{Y})$, et donc d'estimer ainsi $W_\varepsilon(\mu,\nu) = \mathbb{E}[h_\varepsilon(X,v^*)]$.

Algorithmes stochastiques en Machine Learning

Les algorithmes stochastiques ont connu un regain d'intérêt ces 10 dernières années après le développement de certaines applications en Deep Learning (algorithme stochastique de rétro-propagation du gradient pour l'apprentissage de réseaux de neurones). Ce stage cherchera en particulier à appliquer certains progrès dans le domaine des algorithmes d'optimisation stochastique à l'optimisation semi-duale obtenue dans (1.2).

Le principe de l'optimisation stochastique est de maximiser des fonctions objectifs s'exprimant sous la forme d'espérance (comme $W_\epsilon(\mu,\nu)$) en utilisant une suite de variables aléatoires i.i.d. $X_1,\ldots,X_n$. Cette suite de variables aléatoires permet alors de considérer un algorithme d'optimisation stochastique qui construit une suite *on-line* de vecteurs aléatoires $V_1,\ldots,V_n$ qui converge vers un minimiseur de la fonction objectif v^*. Dans le cas particulier où ν est une measure discrete (somme de diracs) à support sur J points de $\mathcal{Y}$, alors (1.2) revient à un problème d'optimisation sur l'espace $\mathbb{R}^J$.

Des algorithmes de descente de gradient stochastique (de type Robbins-Monro [10]) ont été récemment proposés dans [4, 13] pour résoudre le problème d'optimisation (1.2). Le but de ce stage est d'étudier les possibilités de développer des algorithmes stochastiques du second ordre de type Newton (i.e. exploitant la connaissance ou l'estimation de la Hessienne de la fonction objectif au point courant) pour résoudre ce problème en s'appuyant par exemple sur l'approche proposée dans [5] pour la régression logistique.

Principaux objectifs du stage

Les objectifs du stage seront donc multiples, à la fois théoriques et appliqués. Il s'agira dans un premier temps d'étudier la convergence d'un algorithme stochastique de Newton pour construire une suite de vecteurs aléatoires $V_1, \ldots, V_n$ qui approxime v^*. On s'intéressera en particulier :

- à l'étude de la convergence presque sûre lorsque $n \to \infty$ de $(V_n)_{n\geq 1}$ et à la recherche d'un théorème central limite identifiant la vitesse de convergence de $(V_n)_{n\geq 1}$ ainsi que sa variance asymptotique,
- au problème de la construction d'un estimateur $\hat{W}_{\varepsilon,n}$ de $W_\varepsilon(\mu,\nu)$ pour lequel on pourra également étudier la convergence en loi (TCL).

Les principales étapes du stage pourront être les suivantes.

(1) Dans un premier temps, on étudiera le cas régularisé ($\epsilon > 0$) en distingant les situations suivantes :
   - le cas où ν est une mesure discrète (transport discret ou semi-discret) pour lequel la formulation semi-duale (1.2) devient un problème d'optimisation stochastique bien connu sur un espace Euclidien de dimension finie (voir par exemple [10, 19]) et où des résultats non-asymptotiques peuvent même être établis [11],
   - cas où ν est une mesure absolument continue (transport continu) pour lequel la formulation semi-duale (1.2) devient un problème d'optimisation stochastique sur un espace de Hilbert, et on fera le lien avec des travaux récents [2, 7] sur ce thème.

(2) Dans un second temps, on s'intéressa à la situation potentiellement plus délicate du cas non-régularisé ($\epsilon = 0$) où l'absence de forte convexité induit des difficultés théoriques significatives (voir [11]).

(3) Des exemples numériques pourront également être étudiés à l'aide de données réelles ou synthétiques afin d'illustrer les résultats théoriques obtenus.

Le stage est de nature à la fois théorique et numérique. Les principales notions abordées feront appel à des outils de probabilité, statistique mathématique et d'optimisation avec des applications possibles en traitement de données en grande dimension. Il nécessite une bonne formation en mathématiques appliquées, ainsi que la maîtrise du langage de programmation Python pour le calcul scientifique et l'analyse de données.

2. Poursuite en thèse

En fonction de son déroulement, le stage pourra déboucher sur une thèse autour de la thématique "**Propriétés statistiques des modèles génératifs d'images (GAN) basés sur le transport optimal, et applications en IA.**" s'inscrivant dans le cadre du projet ANR 2019-2023 "**Masdol: Mathematics of Stochastic and Deterministic Optimization for Deep Learning**".

Une version des modèles génératifs adversariaux à base de réseaux de neurones (GAN - Generative adversarial networks) [14] basée sur le transport optimal [1] peut se formaliser de la façon suivante. On suppose que l'on observe une suite de variables aléatoires i.i.d. $Y_1, \ldots, Y_J$ à valeurs dans $\mathcal{Y}$ de loi inconnue ν (par exemple un ensemble d'images 2D), et l'on souhaite pouvoir construire un générateur aléatoire de données dont les valeurs "sont proches" d'une variable aléatoire qui serait simulée selon la loi ν.

Plus précisément, il est supposé que l'on choisit une classe de générateurs $\mathcal{G} = (g_\theta)_{\theta\in\Theta}$ où pour tout $\theta \in \Theta \subset \mathbb{R}^p$, $g_\theta : \mathbb{R}^d \to \mathcal{Y}$ est une fonction paramétrique où d est typiquement petit par rapport à la dimension de $\mathcal{Y}$ (par exemple espace des images). Classiquement,

g_θ est construit à partir d'un réseau de neurones profond dont le nombre p de paramètres (dimension du vecteur θ) peut être très grand de sorte à donner beaucoup de flexibilité au générateur.

Etant donné une variable latente Z de loi connue ζ (par exemple la loi uniforme sur $[0,1]^d$) et un paramètre $\theta \in \Theta$, on génère de façon simple un nouvel élément à valeur dans $\mathcal{Y}$ en considérant $X_\theta = g_\theta(Z)$ qui est une variable aléatoire de la loi μ_θ (mesure image de la loi de Z par l'application g_θ). Afin de générer de nouveaux éléments dont les valeurs "sont proches" de celles observées dans l'échantillon $Y_1, \dots, Y_J$, un modèle génératif basé sur le transport optimal (possiblement régularisé) consiste à résoudre le problème variationnel suivant

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} W_\varepsilon(\mu_\theta, \hat{\nu}_J) \quad \text{où} \quad \hat{\nu}_J = \frac{1}{J}\sum_{j=1}^{J} \delta_{Y_j}, \tag{2.1}$$

c'est à dire trouver un paramètre $\hat{\theta}$ telle que la loi $\mu_{\hat{\theta}}$ soit proche de celle de la mesure empirique $\hat{\nu}_J$.

Dans ce contexte, il est à noter que les mesures μ_θ et $\hat{\nu}_J$ sont à support dans le même espace $\mathcal{X} = \mathcal{Y}$. La génération de nouvelles données (par exemple des images), se fait donc ensuite simplement en tirant aléatoirement une variable Z et en considérant $X_{\hat{\theta}} = g_{\hat{\theta}}(Z)$. Ce type d'approche peut conduire à la génération d'images de visages qui semblent très réalistes, voir par exemple cet article :

https://en.wikipedia.org/wiki/StyleGAN

Le problème variationnel (2.1) peut également se ré-écrire comme un problème d'optimisation minimax

$$\min_{\theta\in\Theta} \max_{v\in\mathbb{R}^J} \int_{\mathbb{R}^d} h_\varepsilon(g_\theta(z), v) d\zeta(z) \tag{2.2}$$

qui est un problème très relié à la théorie des jeux à somme nulle [17]. La fonction $h_\varepsilon : \mathcal{X} \times \mathbb{R}^J \to \mathbb{R}$ est définie par (ici $\mathcal{X} = \mathcal{Y}$)

$$h_\varepsilon(x,v) = \begin{cases} \frac{1}{J}\sum_{j=1}^J v_j + \min_{1\le j\le J}\{c(x,y_j) - v_j\} & \text{pour} \quad \varepsilon = 0, \\ \frac{1}{J}\sum_{j=1}^J v_j - \varepsilon \log\Big(\frac{1}{J}\sum_{j=1}^J \exp\Big(\frac{v_j - c(x,y_j)}{\varepsilon}\Big)\Big) - \varepsilon & \text{pour} \quad \varepsilon > 0. \end{cases} \tag{2.3}$$

Deux problématiques seront alors au centre de projet de thèse :

- `Adam` Construire et étudier les propriétés de converge d'algorithmes stochastiques pour résoudre le problème d'optimisation minimax (2.2) et obtenir une bonne approximation numérique de $\hat{\theta}$. Pour cela, il est envisagé d'utiliser des algorithmes de type descente de gradient stochastique sur l'espace de paramètre Θ. L'optimisation des paramètres d'un réseau de neurones est un problème d'optimisation stochastique non-convexe qui est délicat. Actuellement, les méthodes numériques les plus utilisées sont basées sur les algorithmes `Adam` [15] ou `RMS-Prop` qui sont des algorithmes du second ordre non-linéaires, adaptant les vitesses d'évolutions des dynamiques gradient au paysage de la fonction à optimiser. Il conviendra bien sûr d'adapter au cadre du problème d'optimisation (2.2) l'une de ces deux méthodes récentes. Par ailleurs, les propriétés théoriques de l'algorithme `Adam` sont toutefois assez peu étudiées. L'un des enjeux de ce projet de thèse est d'arriver à une meilleure compréhension des propriétés de convergence de l'algorithme d'`Adam` dans le contexte des GAN basés sur le transport optimal. Dans ce but, une piste intéressante serait d'introduire une version en temps continue d'ADAM telle que proposée récemment dans [3, 9] sous la forme d'une equation différentielle ordinaire

non-autonome. Cette partie de la thèse se placerait dans le cadre de l'analyse de systèmes déterministes à temps continu [6], comme moyen de comprendre la dynamique des algorithmes d'optimisation numérique, et leurs versions stochastiques [12].

- `Approximation` étudier les propriétés d'approximation de la mesure inconnue ν (loi de génération des images observées $Y_1, \dots, Y_J$) par la mesure $\mu_{\hat{\theta}}$ dans un cadre non-asymptotique. Il conviendra d'étudier l'influence des choix du paramètre de régularisation ε et de la classe de générateur $\mathcal{G}$ paramétrée par des réseaux de neurones profonds, voire de faire dépendre ϵ du nombre d'observations n avec $\epsilon_n \to 0$. De nombreux travaux récents existent sur ce sujet (voir par exemple [18]). Dans cette autre partie de la thèse, il sera proposé de se placer du point de vue de la statistique mathématique (approche nonparamérique, inégalités oracles, convergence optimale au sens minimax) en s'inspirant de travaux récents [16] dans cette direction pour l'analyse des modèles GAN.

References

[1] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), pp. 214–223.

[2] Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research 15* (2014), 595–627.

[3] Barakat, A., and Bianchi, P. Convergence and dynamical behavior of the adam algorithm for non convex stochastic optimization. *Preprint, arXiv:1810.02263* (2019).

[4] Bercu, B., and Bigot, J. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *Preprint, arXiv:1812.09150* (2019).

[5] Bercu, B., Godichon-Baggioni, A., and Portier, B. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *Preprint, arXiv:1904.07908* (2019).

[6] Cabot, A., Engler, H., and Gadat, S. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society 361*, 11 (2009), 5983–6017.

[7] Cardot, H., Cénac, P., and Godichon-Baggioni, A. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *Ann. Statist. 45*, 2 (04 2017), 591–614.

[8] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2292–2300.

[9] Da Silva, A., and Gazeau, M. A general system of differential equations to model first order adaptive algorithms.

[10] Duflo, M. *Random iterative models*, vol. 34 of *Applications of Mathematics, New York*. Springer-Verlag, Berlin, 1997.

[11] Gadat, S., and Panloup, F. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *Preprint* (2019).

[12] Gadat, S., Panloup, F., and Saadane, S. Stochastic heavy ball. *Electron. J. Statist. 12*, 1 (2018), 461–529.

[13] Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3440–3448.

[14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada* (2014), pp. 2672–2680.

[15] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[16] Liang, T. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *Preprint, arXiv:1811.03179* (2019).

[17] Liang, T., and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *Proceedings of Machine Learning Research* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 907–915.

[18] Liu, S., Bousquet, O., and Chaudhuri, K. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA* (2017), pp. 5551–5559.

[19] Polyak, B. T., and Juditsky, A. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim. 30*, 4 (1992), 838–855.

[20] Villani, C. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.