Intégrer plusieurs fonctions simultanément avec des mesures de Gibbs

Martin ROUAULT[1] Rémi BARDENET[1] Mylène MAÏDA[2]

[1]Université de Lille, CNRS, Centrale Lille, UMR 9189 – CRIStAL, F-59000 Lille, France

[2]Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France

Résumé – Les méthodes de Monte Carlo comme MCMC sont désormais très répandues en inférence bayésienne, mais fournissent cependant des taux de convergence lents. Inspirés par [14] et motivés par la classification bayésienne, nous illustrons expérimentalement dans ce papier comment des points générés selon des mesures de Gibbs permettent d'obtenir de plus fins intervalles de confiance, comme conséquence de la répulsion entre les points.

Abstract – Monte Carlo methods such as MCMC are now routine in Bayesian inference, yet they come with slow theoretical rates of convergence. Following [14], and motivated by Bayesian classification, we experimentally illustrate in this paper how Gibbs measures can provide better confidence intervals as a trace of their repulsive nature.

1 Introduction

En traitement statistique du signal, les problèmes d'inférence bayésienne se ramènent souvent au calcul de plusieurs intégrales. En effet, le coeur des méthodes bayésiennes est de modéliser un problème de décision dans l'incertain par la maximisation de l'intégrale d'une fonction utilité sous une distribution jointe sur les états du problème [12]. Pour un problème de classification à C étiquettes avec perte 0-1, par exemple, cette maximisation se traduit par la recherche de la plus grande parmi C intégrales. Cette situation de devoir calculer un nombre potentiellement grand d'intégrales n'est pas propre au cas $C > 2$: si le jeu de données à classifier contient M points de données, il s'agit d'estimer simultanément de l'ordre de MC intégrales.

En parallèle, les garanties théoriques des méthodes *Monte Carlo* d'intégration numérique, comme un théorème centrale limite pour un noyau MCMC, sont souvent exprimées pour une seule intégrande [7, Chapter 7]. Lorsque le problème bayésien qui motive l'application de méthodes Monte Carlo requiert l'estimation simultanée d'un grand nombre d'intégrales, il est difficile d'obtenir des intervalles de confiance simultanés de taille raisonnable pour toutes ces intégrales. Il est donc souhaitable de trouver des garanties théoriques pour des méthodes Monte Carlo qui contrôlent l'erreur d'intégration *dans le pire des cas* parmi un ensemble d'intégrandes aussi général que possible.

Nous commençons en Section 2 par détailler l'exemple motivant de la classification bayésienne. Nous rappelons ensuite en Section 3 quelques résultats classiques sur l'erreur d'intégration dans le pire des cas de fonctions Lipschitz pour un tirage i.i.d. ou une chaîne de Markov, avant d'introduire un résultat fondamental de physique statistique [6], qui se traduit par une quadrature Monte Carlo qui donne des intervalles de confiance simultanés plus fins pour les intégrandes lipschitziennes que les résultats connus pour une chaine MCMC de taille équivalente. L'ingrédient principal est que les noeuds d'intégration –les points auxquels on évalue l'intégrande pour construire l'estimateur de l'intégrale– se repoussent comme le feraient des particules élementaires de même charge par l'interaction coulombienne. En Section 4, nous discutons notre résultat récent [14], qui remplace l'intégration coulombienne par un noyau borné, et donne des intervalles de confiance encore plus fins, mais pour un ensemble d'intégrandes restreint et lié au noyau d'interaction. Enfin, en Section 5 et comme contribution de cet article, nous illustrons numériquement l'impact de notre méthode sur l'approximation du classifieur de Bayes dans notre exemple initial de classification bayésienne.

2 La classification bayésienne

Considérons un jeu de données d'entraînement $(x_i, y_i)_{i=1}^N$, où $x_i \in \mathcal{X} = \mathbb{R}^d$ représente les caractéristiques de l'individu i, et $y_i \in \mathcal{Y} = \{1, \dots, C\}$ représente sa *classe*, ou *étiquette*. Soit un nouveau point (x, y), d'étiquette inconnue y. Une *action* est le choix d'un classifieur $g : \mathcal{X}^N \times \mathcal{Y}^N \times \mathcal{X} \to \mathcal{Y}$. Pour une mesure de probabilité jointe p sur $x_1, y_1, \dots, x_N, y_N, x, y$ et la fonction d'utilité dite $0-1$, l'utilité moyenne de l'action g est

$$-\int 1_{y \neq g(x_{1:N}, y_{1:N}, x)} \mathrm{d}p(x_{1:N}, y_{1:N}, x, y), \tag{1}$$

où $u_{1:N} = (u_1, \dots, u_N)$. L'action de Bayes, ici le *classifieur* de Bayes, est la fonction $g^\star$ qui maximise l'utilité moyenne (1). On peut facilement voir qu'il s'agit de

$$g^\star : x_{1:N}, y_{1:N}, x \mapsto \arg\max_{k=1}^{C} p(y = k | x_{1:N}, y_{1:N}, x). \tag{2}$$

Autrement dit, le classifieur de Bayes prédit la classe de plus grande probabilité conditionnelle sous p.

Supposons que la mesure jointe p prenne la forme factorisée classique

$$p(x_{1:N}, y_{1:N}, x, y) = \int p(\theta) \cdot \prod_{i=1}^{N} p(y_i | x_i, \theta) p(x_i) \cdot p(y | x, \theta) p(x) \mathrm{d}\theta, \tag{3}$$

où l'on a introduit un paramètre $\theta \in \Theta$ et une mesure de référence $\mathrm{d}\theta$. On peut alors réécrire la loi conditionnelle dans (2) sous la forme

$$p(y = k | x_{1:N}, y_{1:N}, x) = \int p(y = k | x, \theta) \pi(\theta) \mathrm{d}\theta, \tag{4}$$

où

$$\pi(\theta) \propto \prod_{i=1}^{N} p(y_i|x_i,\theta)p(\theta)$$

est la densité de la mesure de probabilité dite *a posteriori* sur le paramètre θ.

Mettons que l'on utilise une méthode Monte Carlo, i.e. qu'on approche la mesure de densité π par une mesure empirique aléatoire $\pi_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{\theta_j}$, et qu'on approche l'action de Bayes (2) par l'action de Bayes empirique

$$g_n^{\mathrm{MC}} : x_{1:N}, y_{1:N}, x \mapsto \arg\max_{k=1}^{C} \int p(y=k|x,\theta)\mathrm{d}\pi_n(\theta). \tag{5}$$

Une façon naturelle de garantir que l'utilité moyenne (1) de g_n^{MC} est proche de celle de $g^\star$ est de garantir que sur un même événement de grande probabilité sous le tirage de $\theta_{1:n}$, les C intégrales dans l'$\arg\max$ de (5) approximent respectivement les C intégrales dans l'$\arg\max$ de (2). Le tirage de $\theta_{1:n}$, par exemple la loi d'un algorithme MCMC particulier, doit donc idéalement garantir la bonne approximation simultanée de C intégrales. Pire, si le jeu de données sur lequel on doit prédire les étiquettes contient M points et pas un seul, et que l'on souhaite garantir une bonne approximation du classifieur de Bayes dans *toutes* nos prédictions avec le même tirage MCMC $\theta_{1:n}$, il y a de l'ordre de MC intégrales à approcher simultanément.

3 Quelques résultats existants

Noeuds d'intégration i.i.d. Lorsque les noeuds de la quadrature sont tirés i.i.d., on connait l'erreur d'intégration dans le pire des cas pour des intégrandes lipschitziennes.

Théorème 1 ([5]). *Soit $\theta_1, \dots, \theta_n \sim \pi$ i.i.d., où π est supportée dans $\mathbb{R}^d$. Soit $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$. Soit $\epsilon > 0$, $\delta \in (0,1)$. Sous une certaine condition de moment sur π, pour tout $d' > d$, il existe une constante $\alpha > 0$ telle que pour tout n satisfaisant $n^{-1/(d'+2)} \leq \epsilon$ et $\exp\left(-\alpha n\epsilon^2\right) \leq \delta$, avec probabilité $1-\delta$,*

$$\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left| \int f \,\mathrm{d}\mu_n - \int f \,\mathrm{d}\pi \right| \leq \epsilon. \tag{6}$$

Plusieurs commentaires s'imposent. Premièrement, des bornes similaires existent pour des chaines de Markov [8]. Deuxièmement, on peut espérer obtenir le même niveau de précision ϵ et de confiance δ avec un plus petit nombre de points en réduisant l'ensemble des intégrandes possibles, ou en jouant sur le tirage des points. En effet, en considérant des fonctions plus régulières que les fonctions lipschitziennes, on peut espérer intégrer plus rapidement : à l'extrême, si on intègre des polynômes de degré D contre une mesure π dont on connait les D premiers moments, il suffit de D points pour intégrer parfaitement toutes les intégrandes. De façon moins caricaturale, si les intégrandes sont cantonnées dans un espace de Hilbert à noyau reproduisant (RKHS ;[a] [4]), Bach [2] montre par exemple une inégalité de type (6) avec un meilleur taux, bien que non explicite. Si l'on veut au contraire garder des hypothèses faibles sur la régularité des intégrandes possibles, une stratégie peut être d'introduire de la dépendance entre les noeuds d'intégration pour améliorer le taux de convergence de (6). On peut par exemple choisir des noeuds d'intégration dont la distribution jointe est une mesure de Gibbs. Nous détaillons maintenant un résultat récent de physique statistique dont nous pourrons déduire des garanties théoriques meilleures dans le cas gibbsien que dans le cas i.i.d.

[a] Un RKHS est un espace de fonctions dont la régularité est codée par un noyau ; ils sont communs dans l'analyse des méthodes d'apprentissage à noyau, comme les SVM ou les processus gaussiens.

Noeuds suivant une mesure de Gibbs. Pour $\theta_{1:n} = (\theta_1, \dots, \theta_n) \in \left(\mathbb{R}^d\right)^n$, considérons

$$H_n(\theta_{1:n}) = \frac{1}{2n^2}\sum_{i\neq j} K(\theta_i,\theta_j) + \frac{1}{n}\sum_{i=1}^{n} V(\theta_i), \tag{7}$$

pour une fonction $V : \mathbb{R}^d \to \mathbb{R}$ dite *potentiel confinant* et une fonction $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ dite *noyau d'interaction*. Sous des hypothèses générales sur V et K, la mesure de probabilité sur $\left(\mathbb{R}^d\right)^n$ donnée par

$$\frac{\mathrm{d}\mathbb{P}^V_{n,\beta_n}}{\mathrm{d}\theta_{1:n}}(\theta_{1:n}) = \frac{\mathrm{e}^{-\beta_n H_n(\theta_{1:n})}}{Z^V_{n,\beta_n}}, \tag{8}$$

à *température inverse* $\beta_n > 0$ et de constante de normalisation Z^V_{n,β_n}, est bien définie, cf par exemple [14]. On appelle (8) la mesure de Gibbs de paramètres (V, K, β_n). Intuitivement, les tirages $\theta_{1:n}$ de $\mathbb{P}^V_{n,\beta_n}$ correspondent à de faibles valeurs du potentiel d'interaction $K(\theta_i,\theta_j)$ (comme si les points θ_i correspondaient à la position de particules physiques qui se repoussent par une force encodée par le noyau d'interaction), mais qui restent dans une région où le potentiel confinant V n'est pas trop grand.

Dans la suite de l'article, nous supposons π de support borné, inclus dans $B(0,R)$ pour un certain $R > 0$, et nous fixons le potentiel confinant dans (8) à

$$V : \theta \mapsto -\int K(\theta,\cdot)\mathrm{d}\pi + (\|\theta\|^2 - R^2)1_{\|\theta\|>R}, \tag{9}$$

le seul paramètre libre dans la mesure de Gibbs est donc le noyau d'interaction K.

Pour le noyau de Coulomb $K(\theta_i,\theta_j) = 1/\|\theta_i - \theta_j\|^{d-2}$ en dimension $d \geq 3$, un résultat de Chafai et al. [6] implique que la répulsion inhérente aux tirages de $\mathbb{P}^V_{n,\beta_n}$ a bien pour conséquence des intervalles de confiance simultanés plus fins que sous un tirage i.i.d., pour des intégrandes lipschitziennes.

Théorème 2 ([6]). *Soient $\epsilon > 0$ et $\delta \in (0,1)$ fixés. Soit $\beta_n \gg n$. Soit $\theta_1, \dots, \theta_n \sim \mathbb{P}^V_{n,\beta_n}$ définie par (8), avec K le noyau de Coulomb, V le potentiel associé à la mesure cible π par (9) et $\mu_n = n^{-1}\sum_{i=1}^{n}\delta_{\theta_i}$. Il existe une constante $c > 0$ telle que pour $n^{-1/d} \leq \epsilon$ et $\exp\left(-c\beta_n\epsilon^2\right) \leq \delta$, avec probabilité $1-\delta$,*

$$\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left| \int f \,\mathrm{d}\mu_n - \int f \,\mathrm{d}\pi \right| \leq \epsilon. \tag{10}$$

Pour un niveau de confiance δ et une erreur ϵ donnés, dès que $n^{-1/d} \leq \epsilon$, grâce au choix de la température inverse $\beta_n \gg n$, les contraintes du Théorème 2 sont atteintes avec un plus petit n que dans le Theorème 1. On obtient donc la même garantie statistique avec moins de noeuds.

Là aussi, on peut ajouter au bénéfice de la répulsion celui d'une hypothèse sur la régularité des intégrandes : Belhadji

et al. [3] montrent un taux rapide pour des intégrandes dans un RKHS avec une distribution de noeuds différente de (7) appelée *échantillonnage volumique*. Dans [14] et le présent article, nous nous concentrons également sur l'erreur dans le pire des cas sur un RKHS, mais avec un tirage des noeuds selon la mesure de Gibbs (8), où le noyau K de la mesure de Gibbs est précisément celui qui définit le RKHS.

4 Notre résultat récent

Considérons un noyau $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ symétrique, positif, continu et *borné*. On suppose de plus que pour toute mesure signée finie non-nulle μ, $\iint K(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) < \infty$. En particulier, K est alors défini positif (au sens habituel des méthodes d'apprentissage à noyau), et définit un RKHS $\mathcal{H}_K$. Nous avons prouvé dans [14] le résultat suivant.

Théorème 3. *Avec les notations du Théorème 2, supposons que $\beta_n \gg n$, et soit $\theta_1, \dots, \theta_n \sim \mathbb{P}^V_{n,\beta_n}$. Soient $\epsilon > 0$ et $\delta \in (0,1)$ fixés. Il existe une constante $c > 0$ telle que pour $n^{-1/2} \leq \epsilon$ et $\exp\left(-c\beta_n\epsilon^2\right) \leq \delta$, avec probabilité $1-\delta$,*

$$\sup_{\|f\|_{\mathcal{H}_K}\leq 1} \left| \int f \,\mathrm{d}\mu_n - \int f \,\mathrm{d}\pi \right| \leq \epsilon. \tag{11}$$

Comme dans le cas du Théorème 2 pour le noyau coulombien, nous obtenons que, pour un niveau de confiance δ et une erreur ϵ donnés, dès que $n^{-1/2} \leq \epsilon$, grâce au choix de la température inverse $\beta_n \gg n$, les contraintes sont atteintes avec un plus petit nombre de noeuds que dans le cas i.i.d.

Par rapport aux garanties fournies par le Théorème 2, notre résultat présente l'avantage de pouvoir atteindre des erreurs ϵ de l'ordre de $n^{-1/2}$ (contre $n^{-1/d}$ dans le cas Coulomb) mais pour des intégrandes a priori un peu moins générales puisque leur régularité depend du noyau choisi.

5 Application à la classification

Nous illustrons ici le Théorème 3 sur l'exemple de la classification bayésienne introduit en Section 2. Nous considérons la régression logistique binaire. Plus précisément, nous fixons $C = 2$, $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathbb{R}$, $d = 3$, et prenons une loi jointe de factorisation (3), avec

$$p(y = 1|x, \theta) = 1 - p(y = 0|x, \theta) \propto \exp(\theta^\top [x, 1]),$$

où $[x, 1] \in \mathbb{R}^3$ est le vecteur x auquel on ajoute un 1 pour inclure une ordonnée à l'origine (*intercept*). Pour le prior $p(\theta)$, nous prenons une gaussienne isotrope centrée de variance $\sigma = 1$, retreinte à $B(0, R)$, où $R \gg 5\sigma$. On garantit ainsi que le support de la posterior π est inclus dans $B(0, R)$, et que le choix de potentiel confinant (9) est valide.

Nous considérons des données-jouets synthétiques inspirées de la documentation de la librairie d'échantillonnage MCMC *blackjax*.[b] Les données sont issues d'un modèle de mélange de deux gaussiennes isotropes de même poids, et l'étiquette correspond à la variable cachée identifiant la composante du mélange utilisée lors du tirage. On prend $N = 50$ individus pour l'ensemble d'entrainement, et $M = 10$ données test à classifier. La détermination de la valeur du classifieur de Bayes en ces 10 points test requiert alors l'évaluation de $M(C-1) = 10$ intégrales.

Dans un premier temps, pour pouvoir nous comparer à une valeur cible, nous approximons la conditionnelle $p(y = 1|x_{1:N}, y_{1:N}, x)$ pour chaque donnée (x, y) du jeu de M données test en utilisant une longue chaine de Metropolis ciblant π dans (3), avec proposition gaussienne ajustée pour obtenir 25% d'acceptations et $T = 10^5$ itérations. Dans un second temps, nous calculons l'approximation

$$p(y = 1|x_{1:N}, y_{1:N}, x) \approx \frac{1}{n}\sum_{j=1}^{n} p(y = 1|x, \theta_j) \tag{12}$$

lorsque l'échantillon $(\theta_1, \dots, \theta_n)$ est simulé (i) par une chaine de Metropolis ciblant π ou (ii) est approximativement distribué selon la mesure de Gibbs (8). Pour chacune des deux méthodes d'échantillonnage, nous répétons cette expérience sur 50 réalisations de l'échantillon $(\theta_1, \dots, \theta_n)$. Nous en déduisons, pour plusieurs valeurs de $\delta > 0$, les proportions $\mathfrak{p}_{\delta,\text{MCMC}}$ et $\mathfrak{p}_{\delta,\text{Gibbs}}$ de réalisations des noeuds d'intégration pour lesquelles les M estimateurs (12) tombent simultanément dans les M régions de confiance

$$\left[\frac{1}{n}\sum_{j=1}^{n} p(y = 1|x, \theta_j) \pm \delta\right],$$

correspondant aux M valeurs x des points test. Si la mesure de Gibbs donne bien des intervalles de confiance plus petits pour un même nombre de noeuds d'intégration, nous devrions avoir, pour chaque $\delta > 0$, une plus grande proportion d'intégrales estimées qui tombent simultanément à distance δ de leur cible.

Echantillonnage MCMC. Toutes les chaines MCMC sont de Metropolis, de loi stationnaire π, avec 500 itérations de burn-in, et une proposition gaussienne dont la variance est fixée pendant le burn-in pour obtenir un ratio d'acceptation d'environ 0.5. Nous comparons des chaines de longueur allant de $n = 100$ à $n = 10\,000$ points, avec pour chacune 50 réalisations indépendantes.

Echantillonnage par mesure de Gibbs. Le calcul du potentiel V dans (9) nécessite de calculer l'intégrale $\int K(\theta, \tau)\,\mathrm{d}\pi(\tau)$, ce qui est une limitation usuelle des méthodes à noyau pour l'intégration Monte Carlo [11]. Certains outils ont été développés pour outrepasser cette limitation [1], mais nous nous limitons ici à la méthode la plus simple qui consiste à approcher cette intégrale par une longue chaine de Metropolis de 10 000 iterations, dont nous sous-échantillonnons $N_{\text{pot}} = 1\,000$ points de façon uniforme et i.i.d.

Pour simuler $(\theta_1, \dots, \theta_n)$ approximativement selon la mesure de Gibbs (8), nous utilisons $T = 10000$ itérations de MALA (Metropolis-adjusted Langevin algorithm ; [13]), i.e. une chaine de Metropolis–Hastings de proposition

$$\theta|\theta_t \sim \mathcal{N}\left(\theta_t - \alpha\beta_n \nabla H_n(\theta_t), 2\alpha I_{dn}\right), \tag{13}$$

où α est fixé manuellement pour atteindre un taux d'acceptation d'environ 0.5, et la chaine est initialisée avec un échantillon $(\theta^0_{1:n})$ obtenu en sous-échantillonnant de façon i.i.d. n

[b] https://blackjax-devs.github.io/sampling-book/models/logistic_regression.html

points dans la réalisation d'une longue chaine de Metropolis–Hastings. Du fait du calcul des interactions par paires dans (7), chacune des T itérations de MALA a une complexité de $\mathcal{O}(n^2 + N_{\text{pot}} n)$, voir [14] pour plus de discussion sur l'échantillonnage de la mesure de Gibbs.

Comme pour MCMC, nous tirons 50 chaines MALA indépendantes, donc 50 échantillons approximatifs de (8) avec $n = 100$ et $\beta_n = n^2$, avec pour K une interaction de Riesz régularisée, i.e.

$$K(\theta_i, \theta_j) = \frac{1}{(|\theta_i - \theta_j|^2 + 0.01^2)^{1/2}}.$$

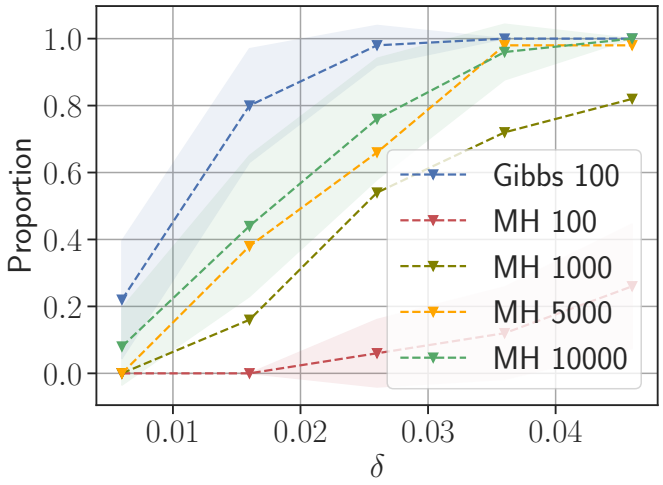

FIGURE 1 : Proportion de fois où les M chaines tombent simultanément à δ de leurs intégrales cibles. Les zones de couleur indiquent les intervalles de confiance à 95%, sous approximation gaussienne, avec correction de Bonferroni pour les 5 mesures effectuées avec les mêmes 50 réalisations.

Sur les résultats de la Figure 1, on confirme que les proportions de succès sont significativement plus hautes pour la mesure de Gibbs que pour l'échantillon MCMC de même taille 100. Cela confirme que pour un nombre n de noeuds d'intégration donné, on a une meilleure approximation simultanée des M intégrales avec la mesure de Gibbs, et donc du classifieur de Bayes. Intuitivement, on s'attend également à ce que la mesure de Gibbs à n noeuds fasse au moins aussi bien qu'une chaine MCMC de taille n^2. En effet, on observe un comportement similaire, avec un léger recouvrement des intervalles de confiance.

6 Discussion

Nous avons vu que simuler des échantillons approximativement selon une mesure de Gibbs permet d'améliorer les intervalles de confiance simultanés par rapport aux méthodes MCMC usuelles pour l'intégration Monte Carlo, notamment pour des exemples-jouets de classification bayésienne. Ces observations viennent illustrer les résultats de convergence rapide des mesures de Gibbs présentés en Sections 3 et 4 comme conséquence de la répulsion entre les points.

Cependant, dans l'état, les résultats théoriques présentés en sections 3 et 4 ne tiennent pas compte de l'approximation effectuée pour le calcul du terme intégral du potentiel confinant (9), et il serait donc important de garantir la convergence rapide des mesures de Gibbs vers l'équilibre malgré cette approximation. Idem, d'un point de vue pratique, il s'agirait de comparer Gibbs à Metropolis à cout computationnel équivalent. Si le cout d'évaluation des intégrandes est élevé, c'est lui qui dominera le cout total, ce qui donne l'avantage à Gibbs. Si les intégrandes sont faciles à évaluer, il est sans doute avantageux de faire tourner une plus longue chaine MCMC pour un budget en temps donné. Enfin, il serait intéressant de comparer ces intervalles de confiance avec d'autres méthodes répulsives comme des méthodes de quasi-Monte Carlo randomisées [9, 10].

Références

[1] A. ANASTASIOU, A. BARP, F-X. BRIOL, B. EBNER, R. GAUNT, F. GHADERINEZHAD, J. GORHAM, A. GRETTON, C. LEY, Q. LIU, L. MACKEY, C. OATES, G. REINERT et Y. SWAN : Stein's Method Meets Computational Statistics : A Review of Some Recent Developments. *Statistical Science*, 38(1):120 – 139, 2023.

[2] F. BACH : On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[3] A. BELHADJI, R. BARDENET et P. CHAINAIS : Kernel interpolation with continuous volume sampling. *In International Conference on Machine Learning (ICML)*, 2020.

[4] A. BERLINET et Ch. THOMAS-AGNAN : *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[5] F. BOLLEY, A. GUILLIN et C. VILLANI : Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.

[6] D. CHAFAÏ, A. HARDY et M. MAÏDA : Concentration for Coulomb gases and Coulomb transport inequalities. *J. Funct. Anal.*, 275(6):1447–1483, 2018.

[7] R. DOUC, É. MOULINES et D. STOFFER : *Nonlinear time series*. Chapman-Hall, 2014.

[8] N. FOURNIER et A. GUILLIN : On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

[9] A. B. OWEN : Scrambled net variance for integrals of smooth functions. *The Annals of Statistics*, 25(4):1541–1562, 1997.

[10] A. B. OWEN : Local antithetic sampling with scrambled nets. *The Annals of Statistics*, 36(5):2319–2343, 2008.

[11] L. PRONZATO et A. ZHIGLJAVSKY : Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):959–1011, 2020.

[12] C. P. ROBERT : *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[13] C. P. ROBERT et G. CASELLA : *Monte Carlo statistical methods*. Springer, 2004.

[14] M. ROUAULT, R. BARDENET et M. MAÏDA : Monte Carlo with kernel-based Gibbs measures : Guarantees for probabilistic herding. *arXiv preprint arXiv :2402.11736*, 2024.

†Avec le soutien du financement ERC Blackjack (ERC-2019-STG-851866) et de la chaire IA Baccarat (ANR-20-CHIA-0002). Les auteurs remercient le CDP C2EMPI pour son soutien, ainsi que l'État Français dans le cadre du programme France-2030, l'Université de Lille, l'Initiative d'Excellence de l'Université de Lille, la Métropole Européenne de Lille pour leur financement et leur appui au projet R-CDP-24-004-C2EMPI.