

# **Methodes de Monte Carlo robustes pour la simulation dans des espaces de grande dimension**

## **Mots clés :**

- **Directeur de thèse** : Gersende FORT
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire Traitement et Communication de l'Information
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

## **Résumé du projet de recherche (Langue 1)**

Les méthodes de Monte Carlo sont des méthodes numériques de type probabilistes pour la résolution de problèmes complexes de type calcul d'intégrales (relue comme le calcul d'une espérance sous une loi de probabilité) ou exploration d'une loi de probabilité. Dans ces deux cas, implémenter un algorithme de Monte Carlo consiste à simuler un nuage de points tel que la mesure empirique associée (éventuellement mesure empirique pondérée) est une approximation de la loi cible. Bien que très répandues, ces méthodes doivent être améliorées notamment dans l'optique de les rendre robustes à la dimension de l'espace de simulation (simulation de variables aléatoires à valeur dans  $\mathbb{R}^d$  pour  $d$  grand, ou dans des espaces de dimension infinie (espaces fonctionnels par exemple)). Les méthodes de Monte Carlo combinent (a) tirage de points sous une loi dite de proposition et (b) correction de ces points pour qu'ils définissent une approximation de la loi cible. Dans tous les cas, l'efficacité des algorithmes et notamment l'efficacité vis-à-vis de la dimension dépend de la loi proposition. Dès lors, définir de nouvelles méthodes de Monte Carlo robustes vis-à-vis de la dimension revient à définir des stratégies de construction de la loi de proposition. L'objet de cette thèse est de développer et d'étudier de nouvelles stratégies pour la définition de ces lois de proposition. Nous envisagerons les approches suivantes : 1) Méthodes MCMC en interaction : généralisant les approches de tempering, ces algorithmes autorisent le nuage d'intérêt à interagir avec d'autres nuages de façon à hériter de leurs bonnes propriétés d'exploration du support de leur loi cible ([KZW06, Atc09, AJDDM11, LLC11]). Les questions ouvertes portent notamment sur la construction des processus auxiliaires et les interactions autorisées, celles-ci devant être faites de façon à garantir la convergence du processus d'intérêt vers la loi cible [ARR11, FMP10, FMPV11]. 2) Méthodes parallélisables : il s'agit d'exploiter la force computationnelle disponible pour répartir les simulations sur plusieurs acteurs. Différentes stratégies sont envisageables, comme (a) l'exploration du support de la loi par différents calculateurs indépendants et recombinaison de ces explorations locales pour former l'apprentissage global; (b) partition du support en sous-espaces de petite dimension, exploration de ces sous-espaces par différents calculateurs, et recombinaison de ces explorations. [WL01, Lia05, Wil05, CRY09, AL10, JRS10]. 3) Méthodes de mutation-selection. On distingue deux grandes familles de méthodes de Monte Carlo: les méthodes d'échantillonnage d'importance (IS) et les méthodes d'acceptation-rejet. Ces dernières sont bien plus robustes à la dimension que les techniques d'IS. Notamment, elles favorisent le déplacement vers les modes de la loi cible. Par suite, les lois de proposition qui produisent des nuages de points proches des modes de la loi cible sont à favoriser. Une piste pour construire de telles lois de proposition est de s'inspirer des algorithmes d'optimisation de type ES (evolutionary strategy) qui construisent de façon adaptative, un nuage de point convergeant vers l'optimum de la fonction objectif [Han06, BAH+10]. Les recherches 1 et 2 (respectivement 2 et 3) s'inscriront plus précisément dans le cadre de programmes de recherche ANR auxquels le LTCI participe : • BigMC "Méthodes de Monte Carlo en grande dimension" (ANR BLAN, 2009-2012 - coordinateur Gersende Fort) • et SIMINOLE "Méthodes de simulation pour des applications de grande échelle en physique expérimentale : inférence statistique, optimisation et apprentissage discriminant" (ANR COSINUS, 2010-2014 - resp. LTCI : Olivier Cappé). Les méthodes de Monte Carlo sont en particulier très utilisées dans les approches statistiques du traitement de signal et en particulier, pour la résolution de problèmes inverses par une approche bayésienne. Les développements méthodologiques seront motivés par de telles applications, et notamment, par le traitement de données d'astro-physique issues de l'expérience AUGER (collaboration avec les astro-physiciens du LAL, dans le cadre du programme ANR SIMINOLE). Références : [AJDDM11] C. Andrieu, A. Jasra, A. Doucet, and P. Del Moral. On non-linear Markov chain Monte Carlo. *Bernoulli*, 17(3):987–1014, 2011. [AL10] Y. Atchad'e and J. Liu. The Wang-Landau algorithm in general state spaces: applications and convergence analysis. *Statistica Sinica*, 20:209–233, 2010. [ARR11] Y. Atchad'e, G.O. Roberts, and J.S. Rosenthal. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Stat. and Computing*, 21:555–568, 2011. [Atc09] Y. F. Atchad'e. Resampling from the past to improve on Markov Chain Monte Carlo algorithms. *Far East Journal of Theoretical Probability*, 27(1), 2009. [BAH+10] D. Brockhoff, A. Auger, N. Hansen, D.V. Arnold, and T. Hohm. Mirrored sampling and sequential selection for evolution strategies. In *PPSN*, volume 6238, pages 11–21, 2010. [CRY09] R.V. Craiu, J.S. Rosenthal, and C. Yang. Learn from thy neighbor: parallel-chain adaptive Markov chain Monte Carlo. *J. American Statist. Assoc.*, 488:1454–1466, 2009. [FMP10] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive MCMC algorithms: ergodicity and law of large numbers. Technical report, arXiv, 2010. [FMPV11] G. Fort, E. Moulines, P. Priouret, and P. Vandekerkhove. A Central Limit Theorem for adaptive and interacting Markov Chains. Technical report, arXiv, 2011. [Han06] N. Hansen. Towards a new evolutionary computation. *Advances on estimation of distribution algorithms*, chapter The CMA evolution strategy: a comparing review, pages 75–102. Springer, 2006. [JRS10] P. Jacob, C.P. Robert, and M.H. Smith. Using parallel computation to improve independent Metropolis-Hastings based estimation. Technical report, arXiv, 2010. [KZW06] S. C. Kou, Q. Zhou, and W. H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Stat.*, 34(4):1581–1619, 2006. [Lia05] F. Liang. Generalized Wang-Landau algorithm for Monte Carlo computation. *J. American Statist. Assoc.*, 100(472):1311–1327, 2005. [LLC11] F. Liang, C. Liu, and R.J. Carroll. *Advanced Markov chain Monte Carlo methods: learning from past samples*. Wiley, 2011. [Wil05] D.J. Wilkinson. *Handbook of parallel computing and statistics*, chapter Parallel Bayesian computation, pages 481–512. 2005. [WL01] F. Wang and D. Landau. Determining the density of states for classical statistical models: a random walk algorithm to produce a fast histogram. *Phys. Rev. E*, 64, 2001.

## Résumé du projet de recherche (Langue 2)

Voir ci-dessus