Generative and unsupervised methods for enhanced non-reversible stochastic exploration in statistical physics (Funded Master internship and PhD project)

Motivation: In statistical physics, a complex system is described at the mesoscopic level by the probability distribution to find it into one of its numerous microscopic configurations. A configuration's probability is directly based on its energy, leading to the picture of an energetic landscape on which the system evolves. This probabilistic modeling is common to the Bayesian approach in statistics, where a priori information is taken into account and the full probability distribution of parameters is inferred instead of only macroscopic quantities, as e.g. point estimates. Such a description requires the computation of high-dimensional integrals to recover the statistics of the macroscopic observables. It is most of the time an intractable task and massive efforts have thus been devoted to the development of efficient computational methods in order to explore those energy landscapes. These numerical schemes, called Monte Carlo (MC) methods, revolve around the approximation of the continuous target integral into a discrete summation over a set of samples generated by a stochastic process [1, 2]. This approximation is exact in the limit of an infinite size of the sample set and the convergence speed is directly ruled by the decorrelations between the samples. This decorrelation can be slowed down by a diffusive dynamics in the stochastic process, which arises most of the time from the reversible nature of the enforced dynamics itself and is worsened with an increasing dimensionality and ruggedness of the energetic landscape to explore. Historically, numerous numerical advances have stemmed from statistical physics. Physical systems indeed present symmetries and already known a priori information, which allows to develop innovative approaches, which conversely shades a new light on the studied physical systems. A careful theoretical characterization can then be carried on to produce robust guarantees on the generality and the convergence of the stochastic process.

Goal: This PhD thesis aims to develop stochastic sampling methods which present better scaling properties in terms of dynamics and computational complexity in the context of high-dimensional and massive datasets in Bayesian inference on one hand and of the need to alleviate finite-size effects in simulations of physical systems by going to large system sizes on another. We propose three research axes:

- developing adaptive solutions through generative and non-supervised algorithms: In this framework, the main goal is to extract the maximum information from data by learning the underlying distribution [3]. We will particularly explore restricted Boltzmann machines [4] and their generalisation to deep undirected graphical models [5].

- upgrading reversible stochastic dynamics to non-reversible versions: First developed for bidimensional sphere systems, non-reversible MCMC algorithms are now under a growing interest in statistics [6]. Their general implementation as well as their behavior characterization [7] are still under study, while the methods are under a continuous improvements [8]. We will study how to set the dynamical and complexity trade-off in those algorithms by the above adaptive methods and how to in turn use these schemes to enhance the sampling of generative models, as they require efficient MCMC algorithms for their training and inference phases [9].

- numerical simulations of spin glass systems and inference algorithms as graphical models: Predicting the capacity of an algorithm to recover a signal given a noisy realization identifies with characterizing the topological properties of a rough random function. This question identifies with the study of energy landscapes of spin glasses [10, 11]. We will study under this light graphical models, which constitute a flexible and appealing framework to model complex and rich structures and in particular can capture dependencies in data [12]. However, sampling from the corresponding posterior/Gibbs

distribution is still a computational challenge, hence the need to develop faster stochastic methods [13, 14].

Environment: The PhD candidate will work at the Université Clermont-Auvergne and will be part of the interdisciplinary ANR project *SuSa*. This thesis lies in the interface between physics, mathematics and computer science, will be supervised by Arnaud Guillin and Manon Michel from Laboratoire de Mathématiques Blaise Pascal (UMR 6620) and will join a local dynamics revolving around stochastic and learning algorithms for physics (high energy physics, cosmology, statistical and chemical physics) and stochastic process analysis (mean-field approach, metastability, optimal transport, functional inequalities).

Profile: We are looking for an enthusiastic individual with a clear interest in interdisciplinary research and a strong background either in theoretical, computational and/or statistical physics or in probability and/or data science to apply. The PhD position is already funded by the *SuSa* project, will start in 2021 and remains open until filled. Interested candidates can contact Manon Michel (manon.michel@uca.fr) for more details. An internship taking place in the beginning of 2021 can precede the PhD and is encouraged.

Webpages:
Manon Michel's webpage: `http://manon-michel.perso.math.cnrs.fr`
Arnaud Guillin's webpage: `http://math.univ-bpclermont.fr/~guillin`
SuSa project's webpage: `https://anr-susa.math.cnrs.fr`

References:

[1] D. Frenkel and B. Smit. *Understanding Molecular Simulation - From Algorithms to Applications*. Academic Press, San Diego, 1996.

[2] W. Krauth. *Statistical Mechanics: Algorithms and Computations*, Oxford University Press, 2006.

[3] I. Goodfellow and Y. Bengio. *Deep learning*, MIT Press, 2016.

[4] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. *Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics* *Journal of Statistical Physics*, 172(6):1576–1608, 2018.

[5] Yoshua Bengio, Gregoire Mesnil, Yann Dauphin, and Salah Rifai. *Better Mixing via Deep Representations*. In *PMLR*, volume 28, 552–560, 2013.

[6] M. Michel, S. Kapfer, and W. Krauth. *Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps*, *J. Chem. Phys*, 140, 054116, 2014.

[7] Alain Durmus, Arnaud Guillin, and Pierre Monmarché. *Piecewise Deterministic Markov Processes and their invariant measure*. preprint, 2018.

[8] M. Michel, A. Durmus, and S. Sénécal. *Forward Event-Chain Monte Carlo: Fast Sampling by Randomness Control in Irreversible Markov Chains*, *JCGS*, 2020.

[9] G. Desjardins, A. Courville, Y. Bengio,P. Vincent and O. Delalleau. *Tempered Markov Chain Monte Carlo for training of Restricted Boltzmann Machines*, In *PMLR*, volume 9, 145–152, 2010.

[10] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., New York, NY, USA, 2009.

[11] L. Zdeborová and F. Krzakala. *Statistical physics of inference: thresholds and algorithms* Advances in Physics, 65(5):453–552, 2016.

[12] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. *A Tutorial on Energy-Based Learning*. Predicting structured data, 1(0), 2006.

[13] R. Arkaprava and D. B. Dunson. *Nonparametric graphical model for counts* arXiv preprint arXiv:1901.00886, 2019.

[14] P. Jalali, K. Khare, and G. Michailidis. *A Bayesian Approach to Joint Estimation of Multiple Graphical Models* arXiv:1902.03651, 2019.