Rémi Bardenet, CR CNRS,
PI of ERC project Blackjack
Centre de recherche en informatique, signal et automatique de Lille
rbardenet.github.io
remi.bardenet@gmail.com

Mylène Maida, Prof. at Univ. Lille
Laboratoire de mathématiques Paul Painlevé
math.univ-lille1.fr/~maida/
mylene.maida@univ-lille.fr

Proposal for a master's internship in computational statistics, with an option to follow on an ERC-funded PhD.

Keywords. Markov chain Monte Carlo, random matrix theory, applications to biology and ML.

Markov chain Monte Carlo algorithms (MCMC; [5]) are numerical integration algorithms that are ubiquitous in **high-dimensional statistical inference** and Bayesian machine learning. The crux is to sample a carefully-chosen Markov chain in the domain of integration, and average the evaluations of the integrand along that chain. However, **MCMC is slow**: the resulting estimators have a mean squared error that decreases as $1/N$, where N is the number of time steps in the Markov chain sample. One natural workaround is to launch P independent copies of the same chain for N steps, and average the results. Unfortunately, the mean square error still decreases as $1/N$: large P will reduce variance, but the bias of the combined estimator is still the same as that of a single chain. Although nontrivial parallel versions of MCMC have been considered [2, 4, 3], none has provably beaten the naive *convergence rate* of averaging the results of P independent copies of the chain [6]. Following the intuition of [1] that repulsiveness brings qualitative variance reduction, we propose to leverage repulsive stochastic processes, to **build parallel MCMC algorithms that quickly and jointly explore the domain of integration**.

Dyson's Brownian motion (DBM; [7, Chapter 3]) is an example of repulsive stochastic process appearing in random matrix theory, which can lead to dramatic variance reduction. Intuitively, think of P Brownian motions forced to stay close to each other but never to intersect. In Figure 1, the mean (dashed blue line, right panel) across all Dyson chains and time steps is much closer to its limit (dashed green line) than the corresponding mean for P independent chains (dashed blue line, left panel). The rationale is that with repulsiveness comes better exploration of the domain and smaller Monte Carlo error. Starting with simple repulsive models, the candidate will progressively find, analyze and implement algorithms that get closer

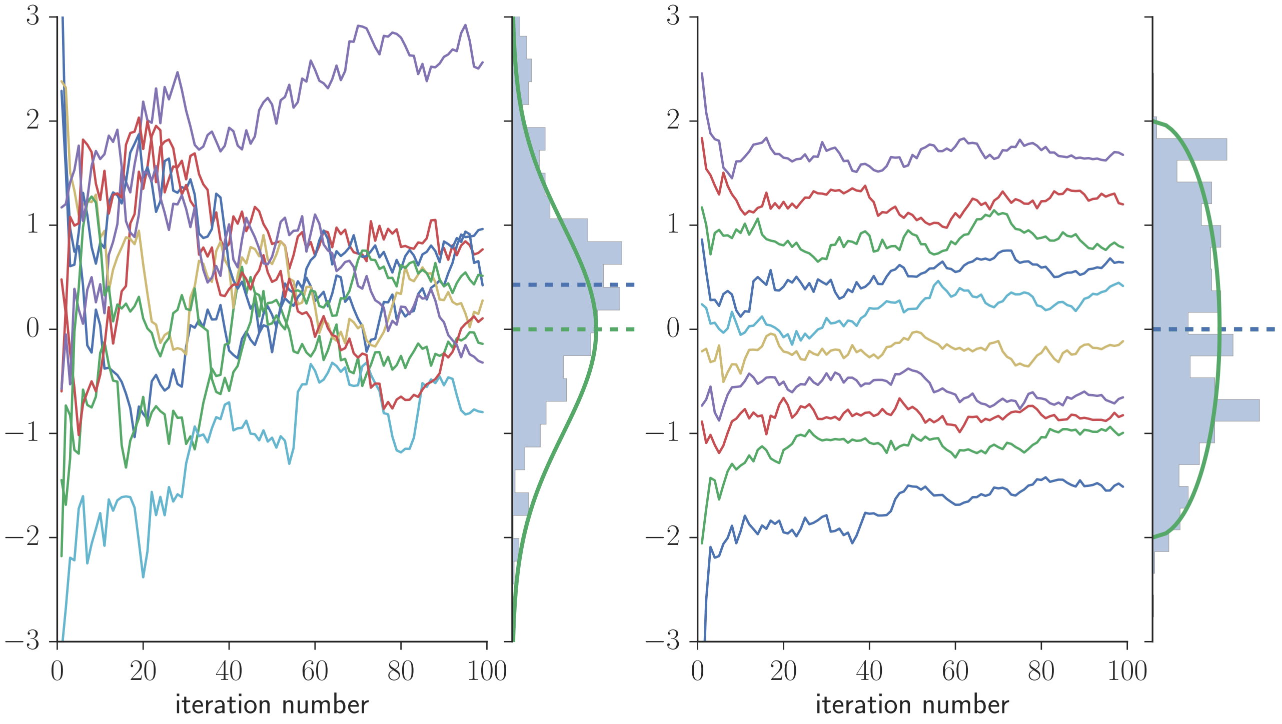

Figure 1: **Left**: The trajectories of $P = 50$ rescaled independent Brownian motions for $N = 50$ time steps. **Right:** The trajectories of a rescaled DBM with $P = 50$ and $N = 50$ time steps. In both panels, the rightmost marginal plot is a histogram of the M positions at the last time step, superimposed with the theoretical limiting pdf in each case.

to a generic parallel MCMC sampler. To be clear, **a computationally efficient parallel MCMC sampler with a fast decreasing error would be a small revolution in scientific computing**. There are many computational and mathematical obstacles that need to be overcome, which means the playground is large and diverse.

The ideal candidate should have a strong background in either probability, statistics, or computer science, and be willing to tackle an interdisciplinary challenge. Depending on the applicant's tastes and skills, the project can move to either a more theoretical or more applied direction. **If everyone is happy at the end of the internship, it is meant to lead to a 3-year PhD with us (Mylène and Rémi) as co-supervisors**. ERC funding is there, so there is no uncertainty on getting a grant.

Besides having a strong research component around data science, and point processes in particular, Lille is a great place to live, with lots of culture and food. Housing is relatively cheap, and the city is extremely well connected by train: Paris is 1h away with trains every half hour, CDG airport 50min, Brussels 30min, London 90min.

References

[1] R. Bardenet and A. Hardy. Monte Carlo with determinantal point processes. *To appear in Annals of Applied Probability*, 2019.

[2] L. Bornn, P. E. Jacob, P. Del Moral, and A. Doucet. An adaptive interacting Wang–Landau algorithm for automatic density exploration. *Journal of Computational and Graphical Statistics*, 22(3):749–773, 2013.

[3] B. Calderhead. A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.

[4] G. Fort, É. Moulines, P. Priouret, and P. Vandekerkhove. A central limit theorem for adaptive and interacting Markov chains. *Bernoulli*, 20(2):457–485, 2014.

[5] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

[6] J. S. Rosenthal. Parallel computing and Monte Carlo algorithms. *Far East journal of theoretical statistics*, 4(2):207–236, 2000.

[7] T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.