Rémi Bardenet, CR CNRS,
PI of ERC grant Blackjack
PI of Artificial Intelligence chair Baccarat
Centre de recherche en informatique, signal et automatique de Lille (CRIStAL)
Université de Lille
rbardenet.github.io
remi.bardenet@gmail.com

Subhroshekhar Ghosh, Assistant professor,
Department of mathematics,
National University of Singapore
https://subhro-ghosh.github.io/

Proposal for PhD thesis

Keywords. Negative dependence in large-scale machine learning, time-frequency signal processing, Monte Carlo integration.

Funding. This thesis will be funded by AI chair "Baccarat" managed by Rémi Bardenet.

Starting date. To be discussed with the advisors. A PhD in France takes 3 years, typically extended by 6 months if needed. It is possible, and actually recommended, to arrange a master's level internship before the thesis starts.

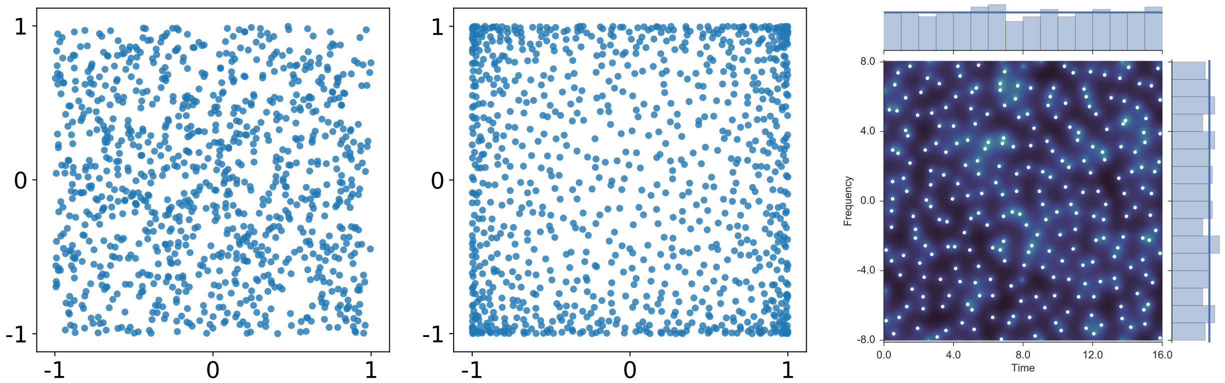
Supervision. To tackle this interdisciplinary project, we offer the co-supervision of a computational statistician developing applications of repulsive point processes (Rémi Bardenet, main supervisor; Univ. Lille, CNRS, École Centrale de Lille), and an expert in probability and statistical physics (Subhro Ghosh: National Univ. Singapore).

Environment. Time will be shared between Lille (France) and Singapore. There is ample funding for travel between the two sites.

In more detail, the first site is CRIStAL, the department of computer science of the University of Lille, where co-supervisor Rémi Bardenet is based. Besides roughly 100 people working on data science and signal processing, CRIStAL hosts a group of 7-10 people dedicated precisely to the applications of repulsive point processes in data science, funded either by the Baccarat AI chair or ERC grant Blackjack. Moreover, we have strong links to the department of mathematics in Laboratoire Painlevé, a few minutes away by foot. A weekly workgroup between the two labs is further devoted to point processes and their applications, with a focus on interacting point processes. This makes Lille a rich scientific environment for the thesis.

The second site is the National University of Singapore (NUS), with co-supervisor Subhro Ghosh at the department of mathematics and the department of statistics. The two departments are very strong across the wide spectrum of mathematics for data science.

Context. A point process is a random discrete set of points in a generic space. A broad interest has recently emerged around point processes that exhibit a regular arrangement of their points in a wide sense. Figure 1(b) shows an example of such a regular arrangement in 2D, while Figure 1(a) shows the same number of i.i.d. points drawn uniformly on the same rectangle, for reference. In condensed matter physics, it has been observed that particle systems like Figure 1(b) are actually so regularly spread that the variance of the number of



(a) (b) (c)

Figure 1: (a) A i.i.d. uniform draw, along with two draws of repulsive point processes: (b) a determinantal point process, and (c) the zeros of the planar Gaussian analytic function.

points in a large window is lower than expected, a phenomenon called *suppressed fluctuations*, or *hyperuniformity*. This has triggered a line of research in mathematical probability around so-called ***repulsive* point processes**, like zeros of Gaussian analytic functions (GAFs; Hough et al. (2009)) and determinantal point processes (DPPs; Hough et al. (2006)). In less than 10 years, machine learners and statisticians have then turned some of these repulsive point processes into **powerful subsampling tools** (Kulesza and Taskar, 2012; Derezinski and Mahoney, 2021; Belhadji, Bardenet, and Chainais, 2020b; Bardenet and Hardy, 2020) and **statistical models** (Kulesza and Taskar, 2012; Lavancier, Møller, and Rubak, 2014).

For instance, Belhadji, Bardenet, and Chainais (2020a) propose to perform feature selection in linear regression using DPPs. A DPP is a point process in which points all interact with each other, and where the interaction is encoded by a kernel, in the same sense as the kernel of support vector machines. It can reasonably be argued that DPPs are the kernel machine of point processes. In the case of (Belhadji, Bardenet, and Chainais, 2020a), the point process draws a set of indices of jointly dissimilar columns of a the (wide) feature matrix. The well-spreadedness of the point process is measured by how well the span of the selected columns approximates the span of the first principal directions. We showed in particular that the resulting set of columns led to performance similar to PCA, while leading to more interpretable features.

As another example, repulsive point processes have helped characterize the behaviour of the zeros of time-frequency transforms of white noises (Bardenet, Flamant, and Chainais, 2018; Bardenet and Hardy, 2019) in **signal processing**. Time-frequency transforms are the mathematical equivalent of musical scores, i.e. they map a signal (say, the sound of an instrument recorded in a noisy instrument) to a function of time and frequency that encodes what frequency is present at what time, like notes on a score. One such transform is the spectrogram, and Figure 1(c) shows the spectrogram of a white Gaussian noise sample. The zeros of the spectrogram, shown as white dots, are a repulsive point process. It turns out that it has the same law as the zeros of the planar GAF, a Gaussian process of particular importance in probability (Hough et al., 2009). The mathematical knowledge we have of this point process in turn leads to improved signal reconstruction algorithms (Bardenet, Flamant, and Chainais, 2018).

Objectives. To get acquainted with the interdisciplinary topic of repulsive point processes, we shall start with a project that fits in ongoing collaboration between the two supervisors. Ideally, this project shall be tackled during a master's level internship prior to starting the PhD. Depending on the student's background and taste, this can be, e.g., (i) topological data analysis applied to the zeros of random spectrograms like Figure 1(c) in statistical signal processing.

Alternately, the internship could revolve around (*ii*) negatively dependent subsampling for large-scale machine learning. For instance, how can we taylor a repulsive point process to sample a *coreset* (Tremblay, Barthelmé, and Amblard, 2019) for some specific ML task? In other words, is there a natural repulsive point process that subsamples a large dataset, while guaranteeing that learning on the subsample leads to the same generalization properties as learning on the initial dataset? This is related to recent work by the two supervisors (Bardenet, Ghosh, and Lin, 2021) on determinantal subsampling for stochastic gradient descent.

After this first project, the three of us will pick an ambitious open problem in line with the objectives of the Baccarat AI chair, according to the student's interest. Candidate problems include identifying and studying repulsive point processes for high-dimensional Monte Carlo integration, fast sampling algorithms for determinantal point processes in machine learning, dictionary learning for signal processing, or studying zeros of wavelet transforms of random signals to use them in filtering tasks.

References

[1] R. Bardenet, J. Flamant, and P. Chainais. "On the zeros of the spectrogram of white noise". In: *Applied and Computational Harmonic Analysis* (2018).

[2] R. Bardenet, S. Ghosh, and M. Lin. "Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[3] R. Bardenet and A. Hardy. "Monte Carlo with Determinantal Point Processes". In: *Annals of Applied Probability* (2020).

[4] R. Bardenet and A. Hardy. "Time-frequency transforms of white noises and Gaussian analytic functions". In: *Applied and Computational Harmonic Analysis* (2019).

[5] A. Belhadji, R. Bardenet, and P. Chainais. "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research (JMLR)* (2020).

[6] A. Belhadji, R. Bardenet, and P. Chainais. "Kernel interpolation with continuous volume sampling". In: *International Conference on Machine Learning (ICML)*. 2020.

[7] M. Derezinski and M. W. Mahoney. "Determinantal point processes in randomized numerical linear algebra". In: *Notices of the American Mathematical Society* 68.1 (2021).

[8] J. B. Hough et al. "Determinantal processes and independence". In: *Probability surveys* (2006).

[9] J. B. Hough et al. *Zeros of Gaussian analytic functions and determinantal point processes.* Vol. 51. American Mathematical Society, 2009.

[10] A. Kulesza and B. Taskar. "Determinantal point processes for machine learning". In: *Foundations and Trends in Machine Learning* (2012).

[11] F. Lavancier, J. Møller, and E. Rubak. "Determinantal point process models and statistical inference". In: *Journal of the Royal Statistical Society, Series B.* B (2014).

[12] N. Tremblay, S. Barthelmé, and P.-O. Amblard. "Determinantal Point Processes for Coresets." In: *Journal of Machine Learning Research* 20.168 (2019), pp. 1–70.