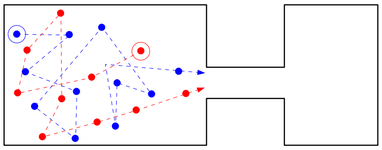
Supervisor: Frédéric Cazals
Inria - Université Côte d'Azur
3IA Côte d'Azur
Email: Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`



PhD thesis proposal: **On the convergence of iterative sampling methods and the detection of meta-stable states in protein science**

Keywords: dynamical systems, meta-stable states, Monte Carlo Markov Chain, two-sample tests, high dimensional geometry, tree-like decompositions, random forests, molecular conformations.

Context. In the theory of dynamical systems, loosely speaking, a *meta-stable state* is a region in phase or conformational space where the system remains sufficiently long before jumping to another such state, via some *transition* which is in general a *rare* event. Equivalently, such a state may be characterized by *local* ergodicity, meaning that for such a region and at the relevant time scale, spatial averages equal time averages [1]. A key difficulty for complex systems, for example a molecule undergoing conformational changes, is to understand the multiple scales at which the system is meta stable.

In statistics, the convergence of iterative methods in general and Monte Carlo Markov chains in particular relies on techniques related to r-hat ($\hat{R}$) and effective sample sizes [2, 3]. In the theory of statistical hypothesis testing [4], a two-sample test is a statistical test aiming at detecting whether two collections of samples (*e.g.* in a high dimensional space, on a manifold, etc) have the same underlying distribution [5].

Goals. The goal of this thesis is to develop a novel approach for the detection of meta-stable states in dynamical systems, using ideas from geometry [6], information theory [7, 8], and statistical hypothesis testing [5]. The line pursued will be to use properties from spatial decompositions yielded by trees (cf random projections trees and random forests), to exploit local properties in the sample space, and the ability to update these data structures dynamically. Tests will be conducted on classical test systems and proteins undergoing conformational changes [9].

The work envisioned encompasses the design and the mathematical analysis of algorithms, their coding (C++ and python), as well their experimental evaluation.

Training. Master 2 or equivalent degree in Computer science (algorithms) or machine learning or statistics or statistical physics.

Conditions. Position at Centre Inria at Université Côte d'Azur, France.

References

[1] J.C. Schön and M. Jansen. Prediction, determination and validation of phase diagrams via the global study of energy landscapes. *Int. J. of Materials Research*, 100(2):135, 2009.

[2] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[3] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

[4] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[5] A. Gretton, K. M. Borgwardt, J.R. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[6] N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 565–574. AUAI Press, 2009.

[7] A. Lhéritier and F. Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018.

[8] F. Cazals and A. Lhéritier. Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In P. Gallinari, J. Kwok, G. Pasi, and O. Zaiane, editors, *IEEE/ACM International Conference on Data Science and Advanced Analytics*, Paris, 2015.

[9] T. O'Donnell and F. Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *J. Comp. Chem.*, 44(11):1094–1104, 2023.

- `http://www-sop.inria.fr/teams/abs/positions/thesis24dynamical-systems-mcmc-convergence.pdf`
- On the convergence of iterative sampling methods and the detection of meta-stable states in protein science
- Etude de la convergence de méthodes de sampling itératives, et detection d'états meta-stables pour les protéines

In the theory of dynamical systems, loosely speaking, a *meta-stable state* is a region in phase or conformational space where the system remains sufficiently long before jumping to another such state, via some *transition* which is in general a *rare* event. Equivalently, a state is characterized by *local* ergodicity, meaning that for such a region and at the relevant time scale, spatial averages equal time averages. In the context of biophysics, meta-stable states are especially important for proteins, as they provide insights on the stable conformations and important intermediates on reaction pathways accounting for the function of proteins. Alas, while the prediction of folded protein structures is now possible thanks to Alphafold, the study of dynamics remains a major open problem.

The work envisioned is to develop novel multi-scale methods to detected meta-stable state in protein simulations. The line pursued will be to use spatial decompositions yielded by (random) trees and forests, to exploits local properties and also make the detection methods dynamic.