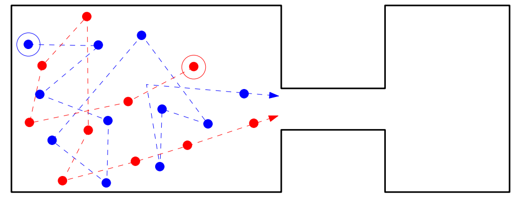
Supervisor: Frédéric Cazals
Inria - Université Côte d'Azur
3IA Côte d'Azur
Email: Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`

Co-supersivor: Alix Lhéritier,
Amadeus SAS and 3IA Côte d'Azur
Email: alix.lheritier@amadeus.com



PhD thesis proposal

Online detection of meta-stable states in dynamical systems: a two-sample test based approach

Keywords: dynamical systems, meta-stable states, two-sample tests, high dimensional geometry, deep learning, Bayesian models, protein structures.

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 a *local* ergodicity property, 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 protein molecular undergoing conformational changes, or a dynamical system modeling the climate, is to understand the multiple scales at which the system is meta stable.

In the theory of statistical hypothesis testing [2], 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. The test is termed *online* if the number of samples is not fixed a priori, but the test accommodates two sources of data providing samples online.

Goals. The goal of this PhD thesis is to develop a novel approach for the detection of meta-stable states in dynamical systems, using ideas from geometry, information theory, and statistical hypothesis testing [3, 4, 5]. Two applications will be considered. The first one is the problem of sampling protein conformations using advanced *proposals* or *move sets* studied [6], an especially challenging case since protein motions span 15 orders of magnitude in time scales [7][8]. The second one is the assessment of the convergence of MCMC processes used in machine learning to learn generative models, including Boltzmann samplers [9] and Bayesian models [10, 11], to make such models more robust and sustainable.

The work envisioned encompasses the design and 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.

References

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

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

[4] 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.

[5] A. Lhéritier and F. Cazals. Low-complexity nonparametric bayesian online prediction with universal guarantees. In *NeurIPS*, Vancouver, 2019.

[6] T. O'Donnell and F. Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *J. Comp. Chem.*, 2023.

[7] S.A. Adcock and A.J. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615, 2006.

[8] Pietro Novelli, Luigi Bonati, Massimiliano Pontil, and Michele Parrinello. Characterizing metastable states with the help of machine learning. *Journal of Chemical Theory and Computation*, 18(9):5195–5202, 2022.

[9] Aurélien Decelle and Cyril Furtlehner. Exact training of restricted boltzmann machines on intrinsically low dimensional data. *Physical Review Letters*, 127(15):158303, 2021.

[10] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. *Advances in neural information processing systems*, 28, 2015.

[11] Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):543–600, 2020.