Lab.: Inria Sophia-Antipolis-Méditerranée
Group: Algorithms-Biology-Structure
Supervisor: Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`

PhD thesis proposal

Analysis of high dimensional energy surfaces, with applications to molecular science and deep learning

Keywords: high-dimensional energy functions, statistical physics, (deep) learning, dimensionality reduction, randomized algorithms, importance sampling, concentration phenomena, molecular simulation.

Context: In molecular science, a molecule is an example high dimensional system ($d = 3n$ Cartesian coordinates with n the number of atoms), whose behavior is encoded by the graph of a function defining the system's potential energy. In supervised learning, a key goal it to minimize a training error by choosing suitable parameters of a model, which often requires exploring a *loss* landscape. In unsupervised learning, choosing a model often requires optimizing the (log) likelihood of the data with respect to a model. From a computer science / applied mathematics standpoint, these classes of problems can be cast in terms of statistical physics, namely describing the topography of a high dimensional energy surface, characterizing the geometry and the *mass* of the catchment basins, understanding the dynamics of a process exploring the landscape. See e.g. [1] for the molecular science perspective, or [2, 3] for the ML perspective.

Goals: The goal of this PhD thesis is to develop novel algorithms to characterize and explore complex high dimensional energy surfaces defined in molecular science and machine learning. Using concepts from computational topology [4], it has been noticed recently that such landscapes often have specific properties in terms of basins and barriers separating them [5]. We ambition to combine ideas from randomized exploration algorithms [6], dimensionality reduction [7], importance sampling [8], and random walks/Monte Carlo methods [9], to design algorithms with controlled complexity and accuracy to explore certain classes of landscapes.

Such insights would pave the way towards very efficient / optimal algorithms for certain tasks in molecular science and machine learning. They would also contribute to shed light on two fundamental questions, namely understanding how proteins accomplish their functions, and why (deep) learning works well for complex tasks.

Software developments will be integrated to the Structural Bioinformatics Library (`https://sbl.inria.fr`), a state-of-the-art environment providing both low level methods (in generic C++) and specific applications in molecular modeling and beyond.

Background. Master in theoretical computer science, or applied mathematics, or (bio-)physics.

Main activities. They will consist of:

- Designing and analyzing algorithms,
- Implementing (C++, python) and testing the algorithms,
- Writing scientific publications.

References

[1] Paraskevi Gkeka, Gabriel Stoltz, Amir Barati Farimani, Zineb Belkacemi, Michele Ceriotti, John D Chodera, Aaron R Dinner, Andrew L Ferguson, Jean-Bernard Maillet, Hervé Minoux, et al. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Journal of Chemical Theory and Computation*, 16(8):4757–4775, 2020.

[2] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 2020.

[3] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, Addis Ababa, 2020.

[4] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. Comp. Chem.*, 36(16):1213–1231, 2015.

[5] Philipp C Verpoort, David J Wales, et al. Archetypal landscapes for deep neural networks. *Proceedings of the National Academy of Sciences*, 117(36):21857–21864, 2020.

[6] A. Roth, T. Dreyfus, C.H. Robert, and F. Cazals. Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes. *J. Comp. Chem.*, 37(8):739–752, 2016.

[7] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[8] A. Chevallier and F. Cazals. Wang-Landau algorithm: an adapted random walk to boost convergence. *J. of Computational Physics*, 410(1):1–19, 2020.

[9] A. Chevallier, S. Pion, and F. Cazals. Improved polytope volume calculations based on Hamiltonian Monte Carlo with boundary reflections and sweet arithmetics. *Under revision*, 2021.