Pierre Monmarché
LJLL & LCT, Sorbonne Université
4 place Jussieu 75005 Paris, France
pierre.monmarche@sorbonne-universite.fr

Report on the doctoral thesis of Martin Šípka, “Machine learning through geometric mechanics and thermodynamics”

Summary. This thesis is concerned with the use of machine learning (ML) techniques for the simulation of dynamical systems in physics. In particular, a question which is transversal to the various specific works of Martin Šípka, and which is more generally a very important topic in current research on applications of ML, is whether some a priori expert knowledge (like physical laws, symmetries, invariances) should be enforced in the ML models by design (instead of learned along an unsupervised training) and how to do that.

The dissertation itself is constituted of a summary presentation of, on the first hand, in particular for readers from a physics or chemistry background, the basic notions of ML that are relevant for the research works carried out during the PhD and, on the other hand, of the results obtained in the 4 articles produced along the PhD (with additional general presentation of the required notions, such as the GENERIC formalism, variational encoders and differentiable simulations), first on Hamiltonian systems and then on the molecular simulation of chemical reactions. The task is, in each case, to learn on data (i.e. a sample of trajectories, e.g. short ab initio molecular simulations) some non-linear feature of interest of a dynamical system, using a neural network representation. Four different features are considered in the four works : reactive potential ; collective variables ; dynamical structure (Hamiltonian and Poisson bivector) ; reactive trajectories. More specifically :

— In *Direct Poisson neural networks : learning non- symplectic mechanical systems*, the goal is to learn the Hamiltonian structure (in terms of the Hamiltonian and Poisson bivector) of a dynamics, based on data constituted of a sample of simulated trajectories. A particular focus is made on the Jacobi identity, which is satisfied by the Poisson bivector in Hamiltonian systems (but not for dissipative dynamics). Three methods are considered : either this identity is built in the model, or it is a part of the loss function (i.e. the model has to make a compromise between fitting the data and approximately satisfying the Jacobi identity), or it is not considered at all. For non-dissipative systems, it is shown that enforcing the identity is better while, conversely, of course, for dissipative systems, models that try to enforce an identity which is in fact not satisfied by the ground truth fail. This is suggested as a way to distinguish dissipative and non-dissipative systems from trajectorial data.

— In *Differentiable Simulations for Enhanced Sampling of Rare Events*, the goal is to learn a biasing potential which minimizes the transition time of a Langevin dynamics (which is a stochastic perturbation of the Hamiltonian dynamics) between two domains of the state space. This question raises a number of numerical and stability issues, which are solved. A crucial point here is that the Langevin dynamics with a sufficiently large friction coefficient is close to a Markovian diffusion dynamics, so that a long trajectories in fact behaves like a series of somehow independent short trajectories (circumventing the issue of vanishing gradients in long and chaotic simulations). The method is applied on standard benchmarks.

- The main idea of *Constructing Collective Variables Using Invariant Learned Representations* is the following : once a ML potential energy based on a graph convolutional network is available for a molecular system (which is the case for many systems as this have been an active area of research over the last years), then it can be used to derive collective variables. Indeed, as a first step, the graph convolutional network learns a suitable representation of the system, invariant by its symmetries and capturing its main features (in terms of computations of its energy). A variational encoder can then be trained for dimension reduction, using directly as input the same representation as the potential energy. This approach is successfully applied on several benchmark problems.
- In *A reactive neural network framework for water-loaded acidic zeolites*, a reactive potential, which is an intermediary in terms of complexity and accuracy between, on the one hand, expensive quantum physics-based DFT simulations and, on the other hand, analytical potentials, which are based on a fixed connectivity between atoms and thus don't allow for chemical reactions, is designed over a class of chemical elements (the acidic zeolites) within water. The main specificity of the work with respect to other ML applications is the construction of the training set, constituted here of ab initio molecular simulations at various temperatures, topologies, concentrations etc. to get a rich range of structures. Once trained, the model shows good properties of generalization and transferability. Moreover, using the methodology of the previous work of Martin Šípka, this potential is used to give data-based collective variables.

Evaluation. The presentation of the results is well-structured and clear on the whole, although, due to the variety of topics, some points remain superficial. The thesis contains sufficient novel research work to justify the award of the PhD. Together with his ability for creative scientific work, it shows that Martin Šípka has reached the required level of expertise in, first, applications and developments of ML methods, second, molecular simulations and, third, the mathematical notions and formalism underlying the two previous aspects. This is illustrated by the journals in which his works have been published : one in computational chemistry (JCTC), one in ML (Proceedings of the 40^{th} ICML), one in mathematical physics (J. Phys. A : Math), the fourth work being still a preprint. It should be emphasized that working at the interface of these different fields and obtaining new methodological contributions of interest for applications, as in the work of Martin Šípka, is not an easy task. This is promising for future research, in particular for applications of the methods developed and tested on academic benchmarks to more challenging systems in material science and chemistry (as, in the thesis, the method for collective variables learning has first been developed on relatively simple reactions and then applied to the whole family of water-loaded acidic zeolites).

Pierre Monmarché Paris, 5th April 2024