PhD thesis proposal

Learning poorly known and observed large scale complex systems

Objective

"*Governing is forecasting*". This proverbial saying is relevant to many situations of engineering interest where decisions must be taken based on predictions or when devising a suitable sequence of actions to achieve some goal requires a good knowledge of the effect of these actions onto the system under consideration. Such predictions usually rely on a simulation of a model of the system at hand and/or observations collected over time. A reliable model may however not be available, or be too computationally costly to be useful. Observations, on the other hand, are often scarce and do not provide a complete picture of the state of the system.

In this thesis, ***we aim at deriving a principled approach to predict the time-evolution of quantities of interest associated with a system observed only via a few noisy sensors active at unpredictable times***. To this end, we leverage the history of the information one can collect. This paradigm of predicting the future from whatever available knowledge over a past horizon is rigorously justified by the Mori-Zwanzig framework developed in the statistical physics community in the late 60s. A particular focus will be on developing scalable approaches, suited for large-scale systems, such as those encountered in haemodynamics, cf. Fig 1.

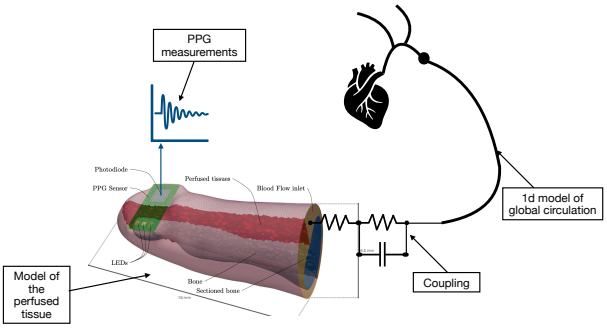

Figure 1: Modelling of the global circulation and the perfuseed tissue.

Scientific context

Describing and predicting the dynamics of complex systems remains a fundamental challenge across many scientific domains. These systems are commonly described by dynamical systems in the form of differential equations. While this formulation is principled, it assumes that the model is known and tractable. In practice, however, the dynamics are often partially unknown, computationally expensive, or only valid within limited regimes. This limitation has led to the development of data-driven approaches that infer system dynamics directly from observations. A key difficulty arises from partial observability. In many applications, only a subset of the system variables is accessible, and observations are often noisy, sparse, or irregular. As a result, the system cannot be accurately described as a Markovian process depending solely on the current observation. Instead, its evolution depends on past states, leading naturally to a non–Markovian formulation. Several modeling strategies explicitly incorporate memory effects, such as autoregressive models such as ARMAX [5], while recurrent neural networks (RNNs), including LSTMs [9, 17, 7], introduce latent memory variables. Reservoir computing and echo state networks [8, 11] offer computationally efficient alternatives capable of capturing long-term dependencies [19]. More recent developments include Latent ODEs [16], which combine Neural ODEs with RNN encoders, augmented Neural ODEs [3], and Transformer architectures [18]. Despite their empirical success, these approaches inherently involve a trade-off between expressivity and interpretability or tend to operate as black boxes. A natural first approach to incorporate

non–Markovian effects is by explicitly including past states, leading to delay differential equations (DDEs). Neural State-Dependent Delayed Differential Equations [8] introduced a flexible framework allowing multiple delays that depend on both time and state.

Research challenges

While these approaches are purely data-driven, they do not explicitly exploit the physical structure of the underlying system. We aim at leveraging a theoretically grounded approach to efficiently predict quantities of interest or (approximation of) the state of a system. We rely on the Mori-Zwanzig framework developed in the statistical physics community in the late 60s, [13, 20]. In a nutshell, it formalizes the time-evolution of a set of variables $\mathbf{x}(t)$ related to the system as a function of their history, without requiring knowledge of the other variables describing the system and, under suitable assumptions, simplifies in:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = M\,\mathbf{x}(t) + \int_{t_0}^{t} \Omega(s)\,\mathbf{x}(t + t_0 - s)\,\mathrm{d}s,$$

where the dynamics is driven by a Markovian operator M and a second contribution (the integral term) depends on the whole history of the considered variables since an initial condition t_0. Accounting for the past essentially allows to isolate the dynamics of these *observables*. This framework is general and applies widely. For instance, when the whole state of the system is not accessible, the dynamics of the observables can be described with a non-Markovian model via this framework. It similarly provides a principled closure for coarse models which can be effectively complemented with a history-based term, [14, 12, 6].

In this thesis, ***we will explore the potential of* Signatures** to efficiently approximate the history of the observations, [2, 4, 15]. The *Signature transform* introduced in [1, 10] has recently been used in several areas, including rough path theory, finance, stochastic control, and machine learning. It has proven to be an effective tool to summarize the information of paths and dependencies across different dimensions, with high computational efficiency. Signatures consist of iterated integrals of the history of its inputs and enjoys interpretability, see Fig. 1 for a sketch. They provide a way to *linearize all possible functions of their input* and exhibit nice theoretical properties. In particular, owing to tensor algebra, they can be efficiently updated when new observations become available, without recomputing the whole object.

Many open questions however remain and will be the focus of this thesis. In particular, how are the different time scales of the physical system preserved across the Signature of its observations? What are the properties of the time series to retain in order to allow for a reliable and efficient prediction based on Signatures? How large should the truncation order be for a

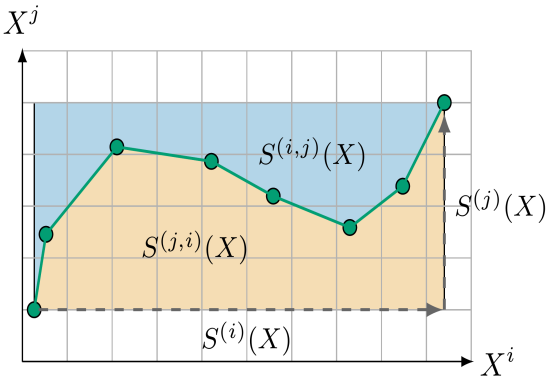

Figure 2: Orders 1 and 2 of the Signature correspond to increments and areas. From [4].

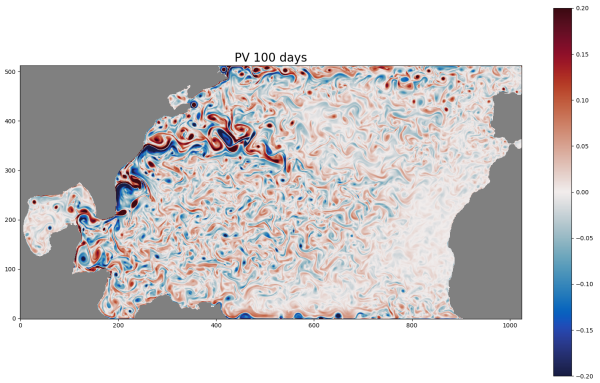

Figure 3: Potential vorticity of the upper layer of the North Atlantic.

given performance? How frugal can the Signature-based term in the Mori-Zwanzig framework be in terms of training data, a critical point in many situations? Does the Mori-Zwanzig solution has a structure that can be exploited, such as low rankness, sparsity or multi-dependence which can be captured with tensor formats, etc.? These methodological developments will first be illustrated on low-dimensional dynamical systems before, if time allows, being demonstrated on large scale real data from geophysics, see Fig. 3.

Team and research environment

The work will take place at the *Laboratoire Interdisciplinaire des Sciences du Numérique* (LISN – `https://www.lisn.upsaclay.fr/`) on the campus of Université Paris-Saclay, benefiting from expertise of the research team in machine learning, applied mathematics, computer science, statistical physics, fluid mechanics and dynamical systems.

The PhD student will be integrated in a vibrant research team focused on scientific machine learning, deep learning, applied mathematics and statistical physics. He/She will be advised by Lionel Mathelin and Onofrio Semeraro, both CNRS researchers involved in the topic for several years. In addition to the rich scientific environment of the Paris-Saclay, the student will benefit from the numerous interactions within the team, in particular with other PhD students and postdocs, and from the weekly seminars which provide exposition to a wide state-of-the-art research.

The candidate should ideally have a solid background in machine learning, applied maths and/or statistics. Knowledge in machine learning numerical framework (for instance, Pytorch, Jax or Julia) is a plus.

Contacts:

- Lionel Mathelin, `lionel.mathelin@cnrs.fr`
- Onofrio Semeraro, `onofrio.semeraro@cnrs.fr`

References

[1] Chen K.-T., Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula, *Annals of Mathematics. 2nd ser.*, **65**, p. 163–178, 1957.

[2] Chevyrev Ilya & Kormilitzin Andrey, 2025 A Primer on the Signature Method in Machine Learning.

[3] Dupont E., Doucet A. & Teh Y.W., Augmented neural ODEs, *Adv. Neural Inf. Process. Syst.*, **32**, p. 3140–3150, 2019.

[4] Fermanian A., Learning time-dependent data with the signature transform, Theses, Sorbonne Université, 2021.

[5] Guidorzi R., *Multivariable system identification: from observations to models*, Bononia University Press, 2003.

[6] Gupta P., Schmid P., Sipp D., Sayadi T. & Rigas G., Mori–Zwanzig latent space Koopman closure for nonlinear autoencoder, *Proc. R. Soc. A*, **481** (2313), p. 20240259, 2025.

[7] Hochreiter S. & Schmidhuber J., Long short-term memory, *Neural Comput.*, **9** (8), p. 1735–1780, 1997.

[8] Jaeger H. & Haas H., Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, *Science*, **304** (5667), p. 78–80, 2004.

[9] Jordan M.I., Serial order: a parallel distributed processing approach. Technical report, California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, *Tech. Rep.*, 1986.

[10] Lyons T., Caruana M. & Lévy T., Differential equations driven by rough paths, In *Lecture notes in Mathematics, École d'été de probabilités de Saint-Flour XXXIV-2004*, 2007.

[11] Maass W., Natschläger T. & Markram H., Real-time computing without stable states: A new framework for neural computation based on perturbations, *Neural Comput.*, **14** (11), p. 2531–2560, 2002.

[12] Menier E., Bucci M.A., Yagoubi M., Mathelin L. & Schoenauer M., CD-ROM: Complemented Deep - Reduced Order Model, *Computer Methods in Applied Mechanics and Engineering*, **410**, p. 115985, 2023.

[13] Mori H., A Continued-Fraction Representation of the Time-Correlation Functions, *Prog. Theor. Phys.*, **34** (3), p. 399–416, 1965.

[14] Parish E. J. & Duraisamy K., Non-Markovian closure models for large eddy simulations using the Mori–Zwanzig formalism, *Phys. Rev. Fluids*, **2** (1), p. 014604, 2017.

[15] Pradeleix E., Hosseinkhan-Boucher R., Shilova A., Semeraro O. & Mathelin L., 2025 *Learning non-Markovian dynamical systems with signature-based encoders*. ECAI 2025 - 2nd ECAI Workshop on "Machine Learning Meets Differential Equations: From Theory to Applications".

[16] Rubanova Y., Chen R.T.Q. & Duvenaud D.K., Latent ODEs for irregularly-sampled time series, In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (ed. H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. B. Fox & R. Garnett), p. 5320–5330, 2019.

[17] Rumelhart D. E., Hinton G. E. & Williams R. J., 1986 *Learning internal representations by error propagation*, p. 318–362. Cambridge, MA, USA: MIT Press.

[18] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L. & Polosukhin I., Attention is All you Need, In *Advances in Neural Information Processing Systems*, , vol. 30, 2017.

[19] Vlachas P.-R., Pathak J., Hunt B.R., Sapsis T.P., Girvan M., Ott E. & Koumoutsakos P., Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics, *Neural Netw.*, **126**, p. 191–217, 2020.

[20] Zwanzig R., Nordholm K.S. J. & Mitchell W.C., Memory Effects in Irreversible Thermodynamics: Corrected Derivation of Transport Equations, *Phys. Rev. A*, **5**, p. 2680–2682, 1972.