ISTITUTO ITALIANO DI TECNOLOGIA

Learning Dynamical Systems

A transfer operator approach

Pietro Novelli

Jun 6, 2024

Roadmap

1. Introduction

- Why *learning* dynamical systems?
- An operatorial perspective: Transfer operators.

2. Statistical learning

- Problem formalization and low-rank estimators. (NeurIPS '22 +'23)
- Representation learning. (ICLR '24)

Roadmap

1. Introduction

- Why *learning* dynamical systems?
- An operatorial perspective: Transfer operators.

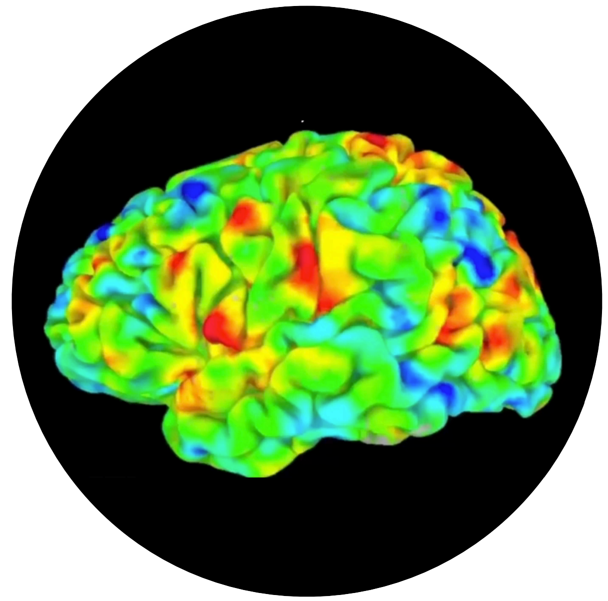
2. Statistical learning

- Problem formalization and low-rank estimators. (NeurIPS ‘22 +‘23)
- Representation learning. (ICLR ’24)

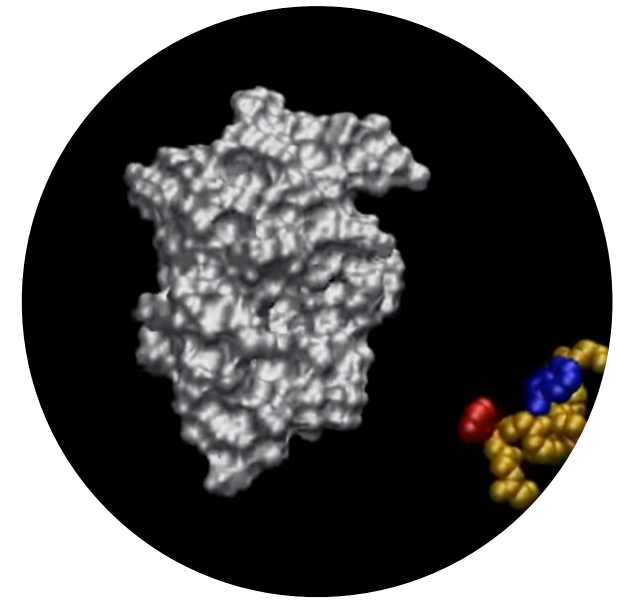
Dynamical Systems

& Machine Learning

Meteorology



Neuroscience



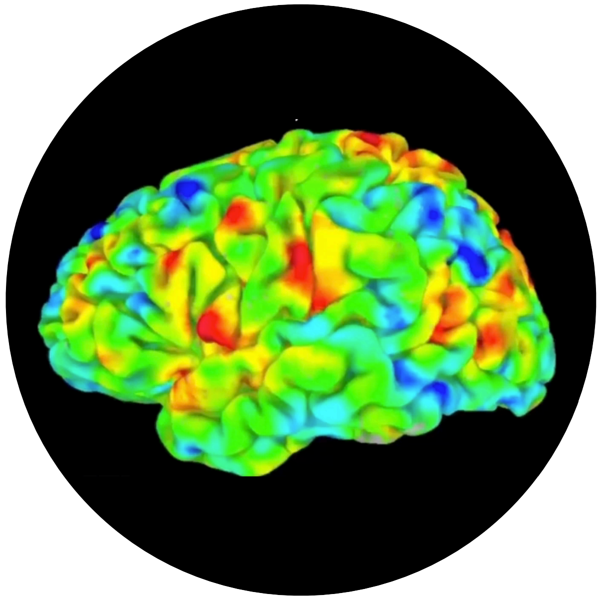
Atomistic Dynamics

- Dynamical Systems are mathematical models of temporally evolving phenomena.
- Data-driven dynamical systems are becoming key in science & engineering.
- Advances in ML lead to better algorithms.

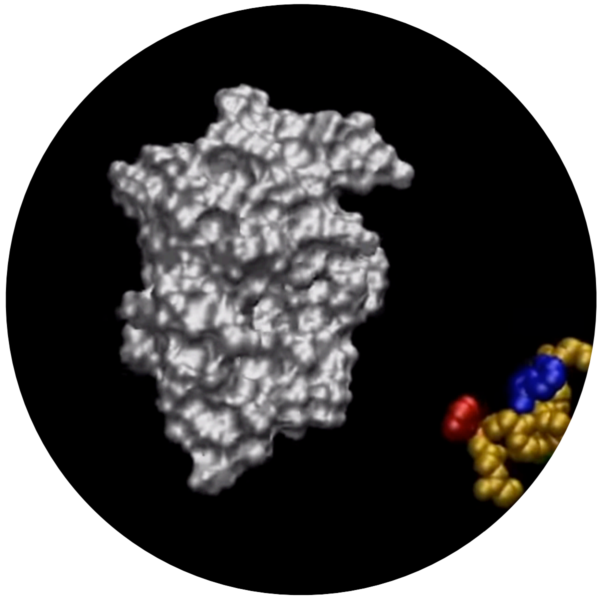
Dynamical Systems

& Machine Learning

Meteorology



Neuroscience



Atomistic Dynamics

- Dynamical Systems are mathematical models of temporally evolving phenomena.
- Data-driven dynamical systems are becoming key in science & engineering.
- Advances in ML lead to better algorithms.

Learning dynamical systems

Transfer operators as alternative to differential equations

- Classical approach: model dynamics with an ODE, PDE, or SDE and learn the unknown equation parameters from data.
- If the system is too complex, or too big, can we build efficient models of dynamics purely from the observed data?
- This is not only possible, but also remarkably elegant via **transfer operator theory**.

Andrej Andreevič Markov

Bernard O. Koopman

Andrej Nikolaevič Kolmogorov

Dynamical Systems

Stochastic setting

- Evolution of a **state** variable over time: $(x_t)_{t\geq 0} \subseteq \mathcal{X}$.
- We focus on discrete, time homogenous, Markov processes:

$$\boxed{\mathbb{P}\left[X_{t+1} \,|\, X_1, \ldots, X_t\right] = \mathbb{P}\left[X_{t+1} \,|\, X_t\right], \text{ independent of } t}$$

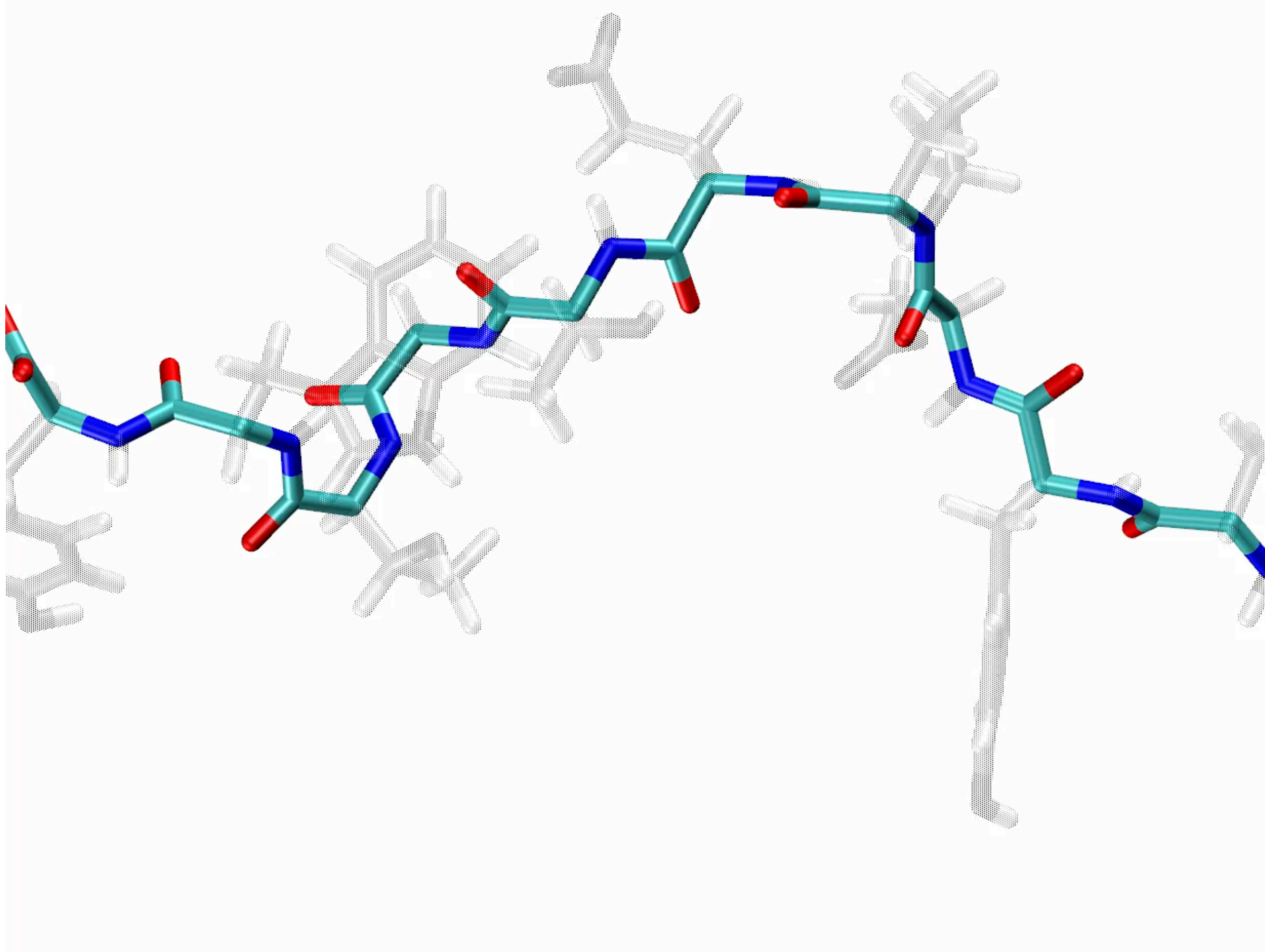
- A prototypical example: $X_{t+1} = F(X_t) + \mathrm{noise}_t$.

Langevin Equation

A model for atoms' dynamics

Overdamped Langevin equation driven by a potential $U: \mathbb{R}^d \to \mathbb{R}$

$$\boxed{dX_t = -\nabla U(X_t)dt + \beta^{-1/2} dW_t}$$



Folding of CLN025 (Chignolin)

Euler—Maruyama discretization

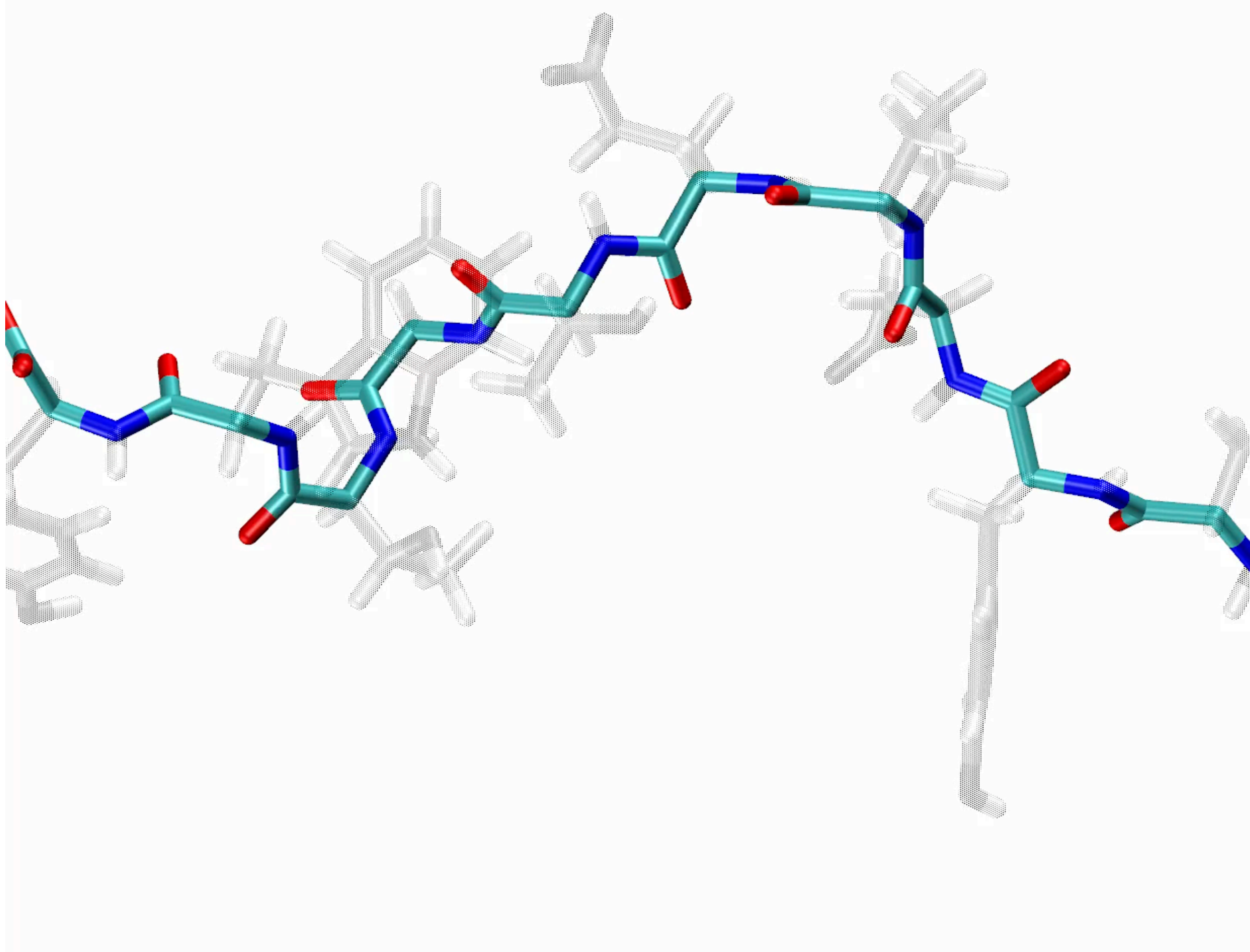
$$X_{t+1} = \underbrace{X_t - \nabla U(X_t)}_{F(X_t)} + \underbrace{\beta^{-1/2}(W_{t+1} - W_t)}_{\text{noise}_t}$$

Langevin Equation

A model for atoms' dynamics

Overdamped Langevin equation driven by a potential $U: \mathbb{R}^d \to \mathbb{R}$

$$dX_t = -\nabla U(X_t)dt + \beta^{-1/2} dW_t$$



Folding of CLN025 (Chignolin)

Euler—Maruyama discretization

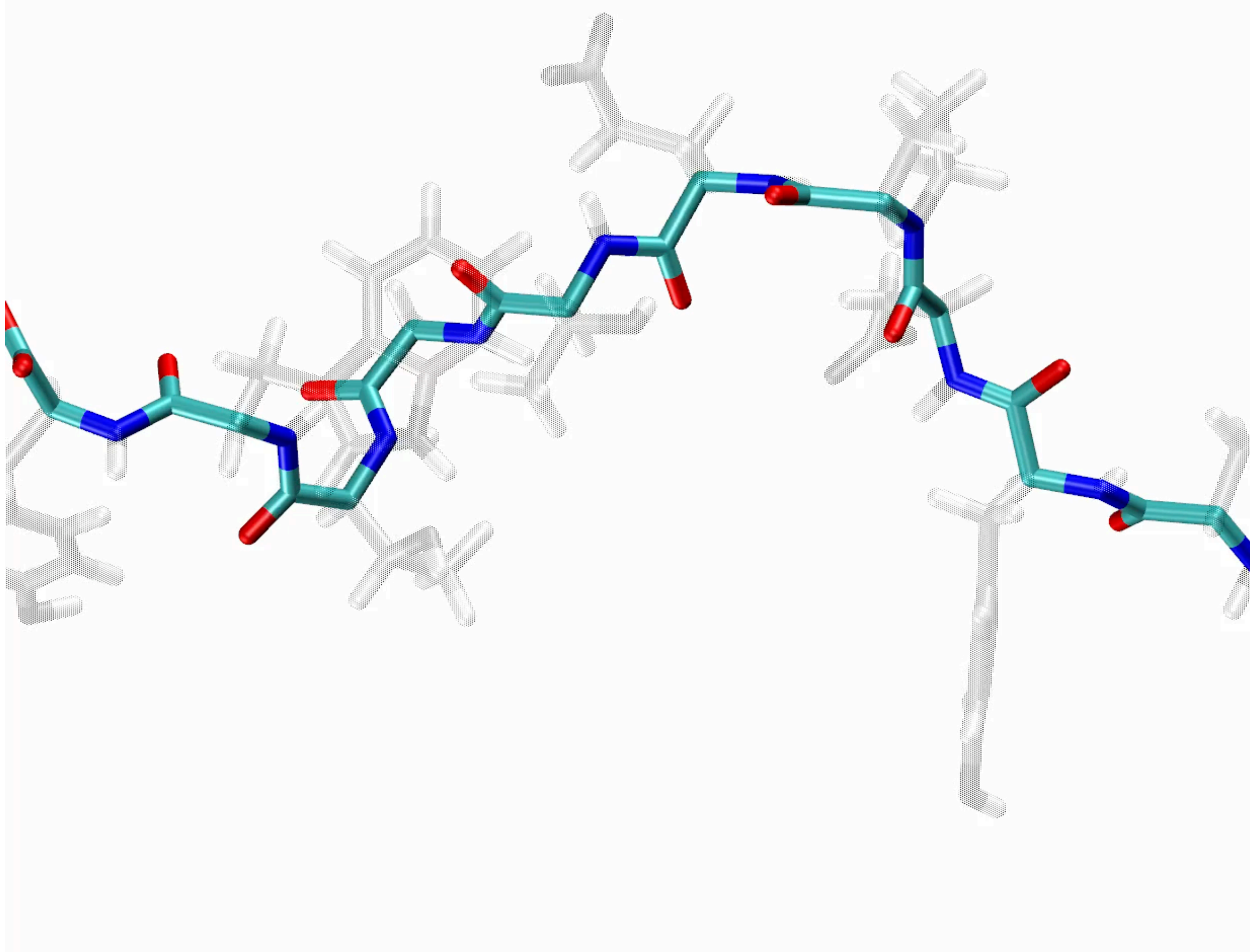
$$X_{t+1} = \underbrace{X_t - \nabla U(X_t)}_{F(X_t)} + \underbrace{\beta^{-1/2}(W_{t+1} - W_t)}_{\text{noise}_t}$$

Langevin Equation

A model for atoms' dynamics

Overdamped Langevin equation driven by a potential $U: \mathbb{R}^d \to \mathbb{R}$

$$\boxed{dX_t = -\nabla U(X_t)dt + \beta^{-1/2}dW_t}$$



Folding of CLN025 (Chignolin)

Euler—Maruyama discretization

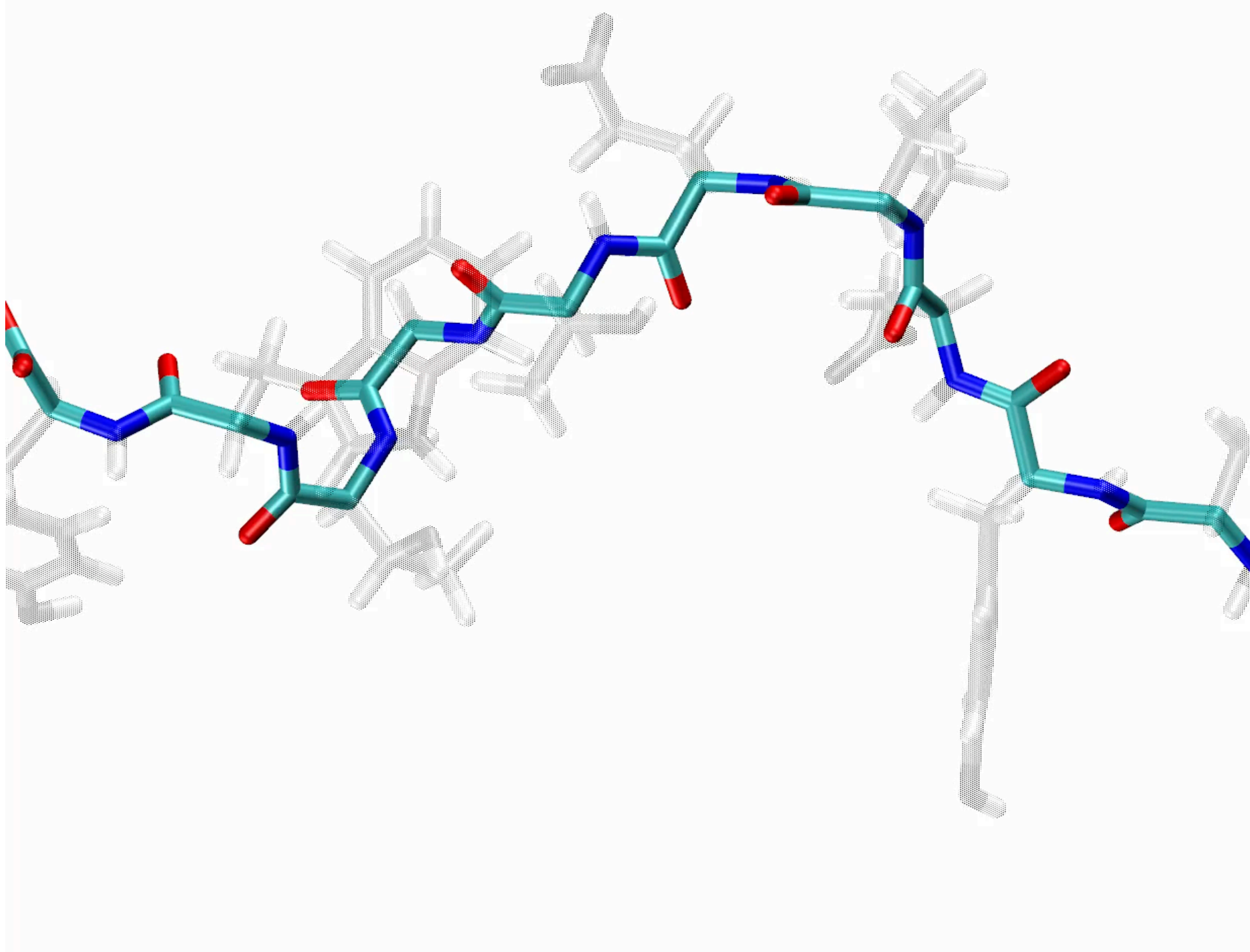
$$X_{t+1} = \underbrace{X_t - \nabla U(X_t)}_{F(X_t)} + \underbrace{\beta^{-1/2}(W_{t+1} - W_t)}_{\text{noise}_t}$$

Langevin Equation

A model for atoms’ dynamics

Overdamped Langevin equation driven by a potential $U: \mathbb{R}^d \to \mathbb{R}$

$$\boxed{dX_t = -\nabla U(X_t)dt + \beta^{-1/2} dW_t}$$



Folding of CLN025 (Chignolin)

Euler—Maruyama discretization

$$X_{t+1} = \underbrace{X_t - \nabla U(X_t)}_{F(X_t)} + \underbrace{\beta^{-1/2}(W_{t+1} - W_t)}_{\text{noise}_t}$$

The Transfer Operator

What does "learning a dynamical system" means, anyway?

- The **transfer operator** T describes the evolution of any scalar function of the state in a suitable set $\mathscr{F}$.

$$\boxed{(\mathsf{T}f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,], \quad f \in \mathscr{F}}$$

- If $\mathscr{F}$ it is large enough, the transfer operator offers a comprehensive characterization of a stochastic process *as a whole*.

- Provides a **global** linearization of the dynamics.

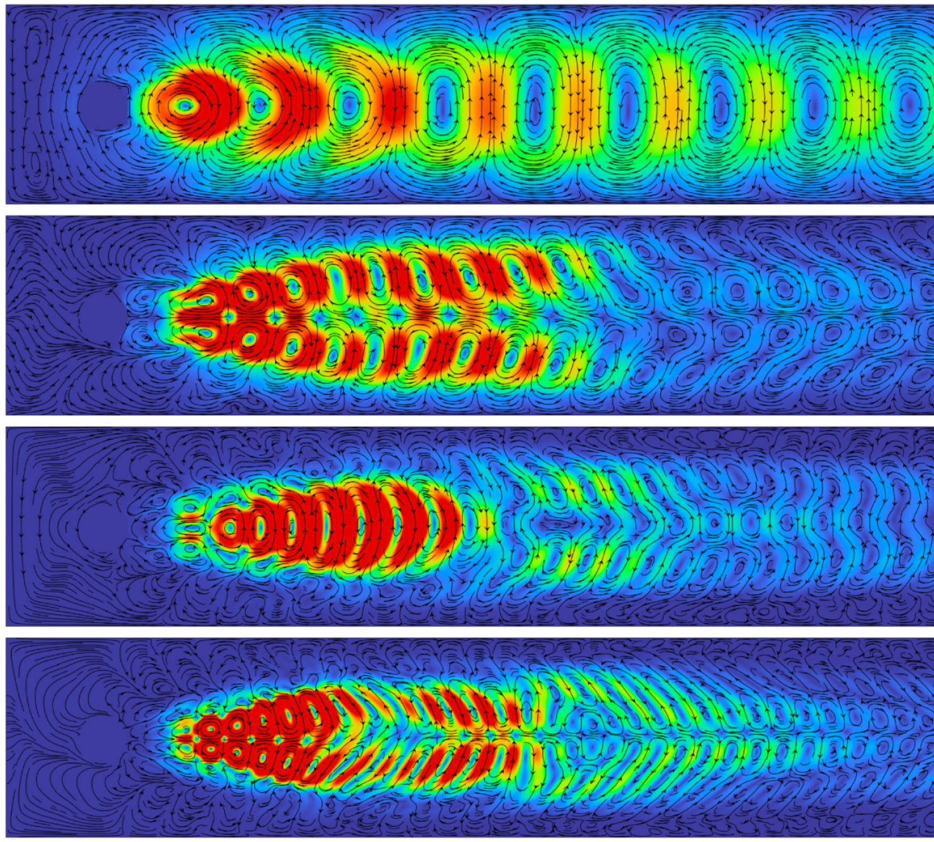
- Its spectral decomposition yield dynamic modes, for interpretability and control.

Dynamical Mode Decomposition

To interpret dynamical systems

- Spectral decomposition: $\mathsf{T} = \sum_{i=1}^{\infty} \lambda_i \psi_i \otimes \psi_i$
  (self-adjoint and compact)

- Scalars λ_i and functions ψ_i are eigenvalues and eigenfunctions
  $\mathsf{T}\psi_i = \lambda_i \psi_i$



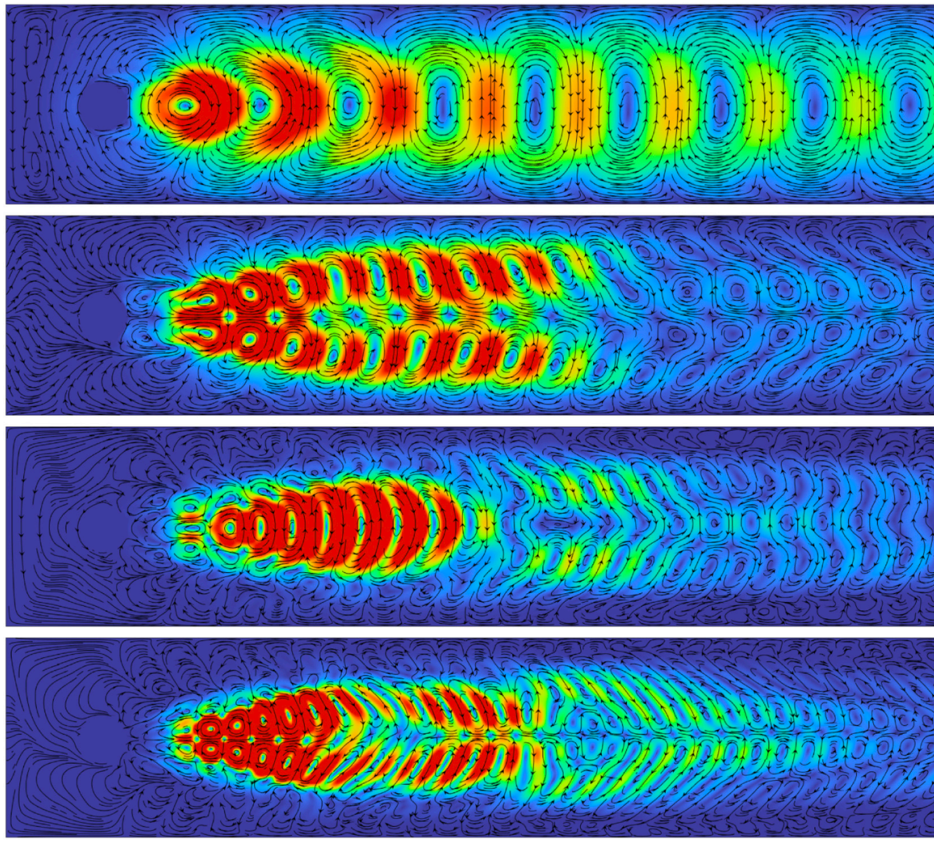
Dynamical modes: 2D Von Karman Vortex Street.
(T. Krake et al. 2021)

- Mode Decomposition disentangles the expected value of an observable into temporal and spatial components.

$$\mathbb{E}[\, f(X_t) \,|\, X_0 = x \,] = (\mathsf{T}^t f)(x) = \sum_i \lambda_i^t \langle \psi_i, f \rangle \psi_i(x)$$

Dynamical Mode Decomposition

To interpret dynamical systems



Dynamical modes: 2D Von Karman Vortex Street.
(T. Krake et al. 2021)

- Spectral decomposition: $\mathsf{T} = \sum_{i=1}^{\infty} \lambda_i \psi_i \otimes \psi_i$
  (self-adjoint and compact)

- Scalars λ_i and functions ψ_i are eigenvalues and eigenfunctions
  $\mathsf{T}\psi_i = \lambda_i \psi_i$

- Mode Decomposition disentangles the expected value of an observable into temporal and spatial components.

$$\mathbb{E}[\, f(X_t) \,|\, X_0 = x\,] = (\mathsf{T}^t f)(x) = \sum_i \boxed{\lambda_i^t} \boxed{\langle \psi_i, f \rangle} \boxed{\psi_i(x)}$$

Learning the Transfer Operator

Statistical analysis of transfer operator regression

Learning the transfer operator

Kostic et al. — NeurIPS '22

$$(\mathsf{T}f)(x) = \mathbb{E}[f(X_{t+1}) \,|\, X_t = x] \quad f \in \mathscr{F}$$

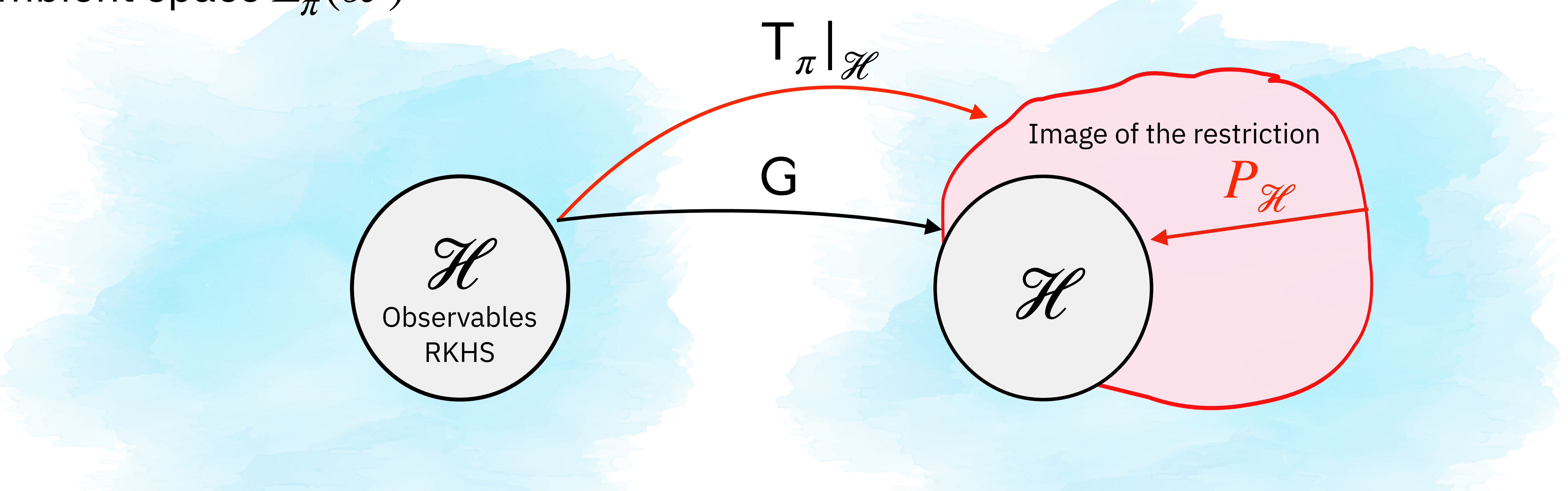
Assumptions:

- **Ergodicity**: there is a unique distribution π s.t. $X_t \sim \pi \Rightarrow X_{t+1} \sim \pi$.
- T is well-defined on $\mathscr{F} = L^2_\pi(\mathscr{X})$, that is $\mathsf{T}[L^2_\pi(\mathscr{X})] \subseteq L^2_\pi(\mathscr{X})$.
- **Challenge**: the operator and its domain are unknown!

Subspace approach

- Idea: approximate T_π at least on a subset $\mathscr{H} \subset L^2_\pi$.
- We choose $\mathscr{H}$ to be a **Reproducing Kernel Hilbert Space**.
- Linearly parametrized functions $\langle w, \phi(x) \rangle$ for some $w \in \mathscr{H}$.
- $\phi : \mathscr{X} \to \mathscr{H}$ is called **feature map**. $\mathscr{H}$ can be finite or infinite dim.

Ambient space $L^2_\pi(\mathscr{X})$



Risk functional

- By the linearity of T_π (conditional expectation is linear).
- And the linearity of **observables' parametrization** $\langle w, \phi(x) \rangle$.

$$\boxed{\mathbb{E}\left[\phi(X_{t+1}) \,|\, X_t = x\right] \approx \mathsf{G}^* \phi(x)}$$

Risk functional

- By the linearity of T_π (conditional expectation is linear).
- And the linearity of **observables' parametrization** $\langle w, \phi(x) \rangle$.

$$\mathbb{E}\left[\phi(X_{t+1}) \,|\, X_t = x\right] \approx \mathsf{G}^* \phi(x)$$

The left side is the **regression function** of this **risk functional**

$$R(\mathsf{G}) = \mathbb{E}_{(X_t, X_{t+1}) \sim \rho} \| \phi(X_{t+1}) - \mathsf{G}^* \phi(X_t) \|^2$$

The risk functional can be interpreted as a **linearization error**.

Empirical risk minimization

And low-rank models

- Given a sample $(x_i, y_i)_{i=1}^{n} \sim \rho$ learn $\mathsf{G} : \mathscr{H} \to \mathscr{H}$ minimizing the **regularised empirical risk:**

$$\boxed{\hat{R}_\gamma(\mathsf{G}) = \sum_{i=1}^{n} \|\phi(y_i) - \mathsf{G}^*\phi(x_i)\|^2 + \gamma\|\mathsf{G}\|_{\mathrm{HS}}^2}$$

Ridge Regression

Full-rank solution.

Principal Component Regression

Low-rank: Minimizes the risk on a feature subspace spanned by the principal components.

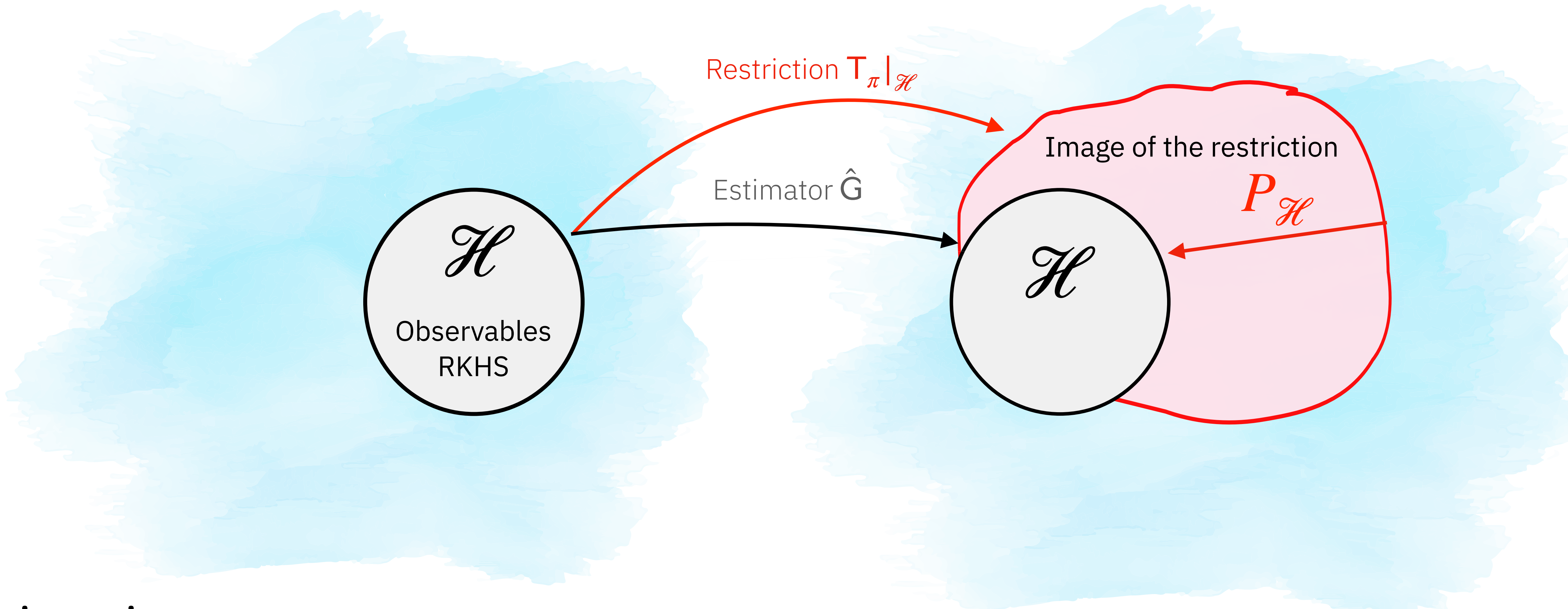
Reduced Rank Regression

Low-rank: Adds an *hard* rank constraint, leading to a generalized eigenvalue problem.

Statistical learning analysis

Justifying every following result

Ambient space $L^2_\pi(\mathscr{X})$



Estimation error

$$\|\mathsf{T}_{\pi|_{\mathscr{H}}} - \hat{\mathsf{G}}\|_{\mathscr{H}\to L^2_\pi} \leq \underbrace{\|(I-P_{\mathscr{H}})\mathsf{T}_{\pi|_{\mathscr{H}}}\|}_{\text{Representation error}} + \underbrace{\|P_{\mathscr{H}}\mathsf{T}_{\pi|_{\mathscr{H}}} - \mathsf{G}\|}_{\text{Estimator bias}} + \underbrace{\|\mathsf{G}-\hat{\mathsf{G}}\|}_{\text{Estimator variance}}$$

Representation Learning

Kostic, Novelli, Grazzi, Lounici, and Pontil — ICLR '24

$$\|\mathsf{T}_{\pi|_{\mathscr{H}}} - \hat{\mathsf{G}}\|_{\mathscr{H}\to L^2_\pi} \leq \underbrace{\boxed{\|(I-P_{\mathscr{H}})\mathsf{T}_{\pi|_{\mathscr{H}}}\|}}_{\text{Representation error}} + \underbrace{\boxed{\|P_{\mathscr{H}}\mathsf{T}_{\pi|_{\mathscr{H}}} - \mathsf{G}\|}}_{\text{Estimator bias}} + \underbrace{\boxed{\|\mathsf{G}-\hat{\mathsf{G}}\|}}_{\text{Estimator variance}}$$

Our approach looks for an empirical estimator of the **representation error** via the following upper and lower bounds (consequence of the norm change from $\mathscr{H}$ to L^2_π)

$$\|(I-P_{\mathscr{H}})\mathsf{T}_\pi P_{\mathscr{H}}\|^2 \lambda^+_{\min}(C_{\mathscr{H}}) \leq \|(I-P_{\mathscr{H}})\mathsf{T}_{\pi|_{\mathscr{H}}}\|^2 \leq \|(I-P_{\mathscr{H}})\mathsf{T}_\pi P_{\mathscr{H}}\|^2 \lambda_{\max}(C_{\mathscr{H}})$$

Representation Learning

Kostic, Novelli, Grazzi, Lounici, and Pontil — ICLR '24

$$\|\mathsf{T}_{\pi|_{\mathscr{H}}} - \hat{\mathsf{G}}\|_{\mathscr{H}\to L^2_\pi} \leq \underbrace{\boxed{\|(I-P_{\mathscr{H}})\mathsf{T}_{\pi|_{\mathscr{H}}}\|}}_{\text{Representation error}} + \underbrace{\boxed{\|P_{\mathscr{H}}\mathsf{T}_{\pi|_{\mathscr{H}}} - \mathsf{G}\|}}_{\text{Estimator bias}} + \underbrace{\boxed{\|\mathsf{G}-\hat{\mathsf{G}}\|}}_{\text{Estimator variance}}$$

Our approach looks for an empirical estimator of the **representation error** via the following upper and lower bounds (consequence of the norm change from $\mathscr{H}$ to L^2_π)

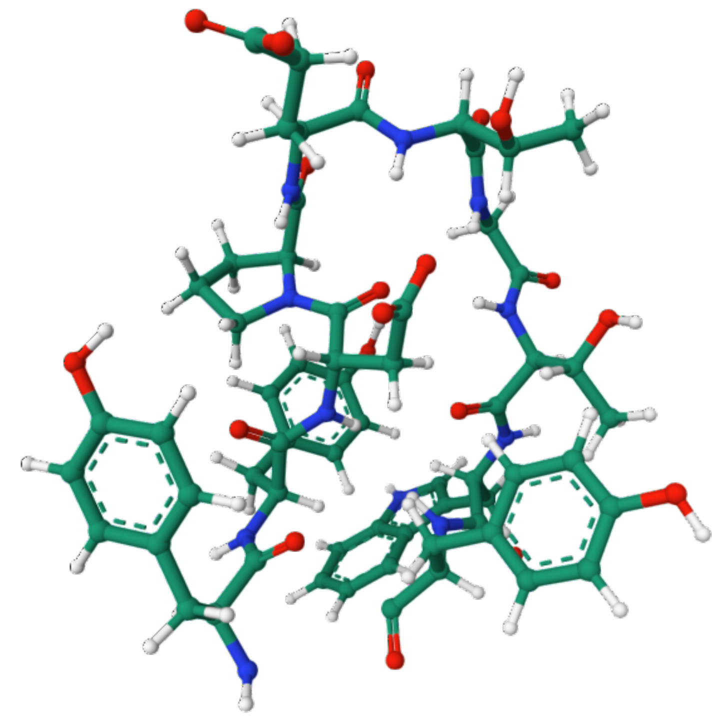
$$\|(I-P_{\mathscr{H}})\mathsf{T}_\pi P_{\mathscr{H}}\|^2 \lambda^+_{\min}(C_{\mathscr{H}}) \leq \|(I-P_{\mathscr{H}})\mathsf{T}_{\pi|_{\mathscr{H}}}\|^2 \leq \|(I-P_{\mathscr{H}})\mathsf{T}_\pi P_{\mathscr{H}}\|^2 \lambda_{\max}(C_{\mathscr{H}})$$

If $C_{\mathscr{H}} = I$ the upper and lower bound match, and the Eckart-Young-Mirsky theorem on $P_{\mathscr{H}}\mathsf{T}_\pi P_{\mathscr{H}}$ assures that the representation error is minimized.

$$\boxed{\frac{\|C^\theta_{XY}\|^2_{\mathrm{HS}}}{\|C^\theta_X\|\,\|C^\theta_Y\|} - \gamma\|I-C^\theta_X\|^2_{\mathrm{HS}} - \gamma\|I-C^\theta_Y\|^2_{\mathrm{HS}}}$$

Application: metastable states of Chignolin

Kostic, Novelli, Grazzi, Lounici, and Pontil — ICLR '24

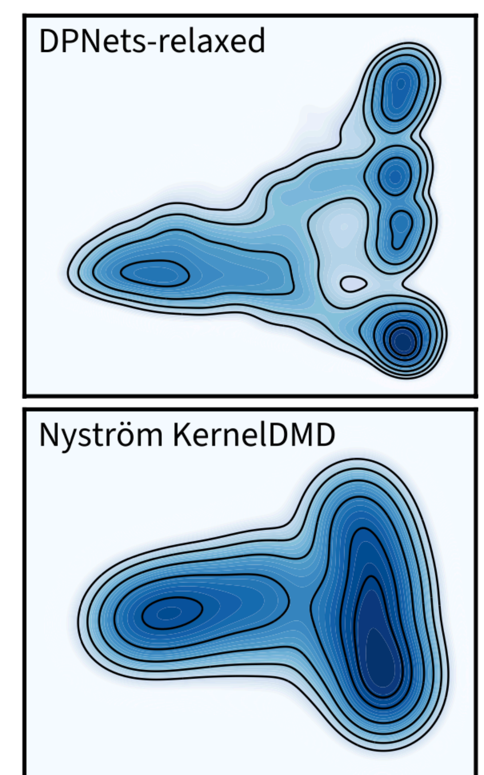


The leading eigenfunctions of T capture the long-term behavior of atomistic dynamics.

A better representation of the data allows a more accurate physical understanding.

Trained DPNets on a Graph Neural Network appropriate for the problmem vs. Fixing $\mathscr{H}$ to be the Gaussian RKHS.

Model	$\mathcal{P}$	Transition	Enthalpy ΔH
DPNets	**12.84**	**17.59 ns**	**-1.97 kcal/mol**
Nys-PCR	7.02	5.27 ns	-1.76 kcal/mol
Nys-RRR	2.22	0.89 ns	-1.44 kcal/mol
Reference	-	40 ns	-6.1 kcal/mol



Free Energy Surface

Conclusions

Additional works

- Sharp spectral rates for Koopman operator learning. (Spotlight @ NeurIPS ’23)
- Estimating Koopman operators with sketching to provably learn large-scale dynamical systems. (NeurIPS’23)
- A randomized algorithm to solve reduced rank operator regression. (Submitted)

Ongoing work

- Operatorial formulation of Reinforcement Learning.
- Neural Conditional Probability models.

kooplearn

Vladimir Kostic

Karim Lounici

Massi Pontil

And also:

- Riccardo Grazzi
- Giacomo Turri
- Daniel Ordoñez-Apraez
- Prune Inzerilli
- Carlo Ciliberto
- Andreas Maurer
- Luigi Bonati
- Michele Parrinello
- Lorenzo Rosasco
- Giacomo Meanti
- Antoine Chatalic

Thank you!