When deep learning meets ergodic theory

Exploiting the physics to go beyond the ergodic limit

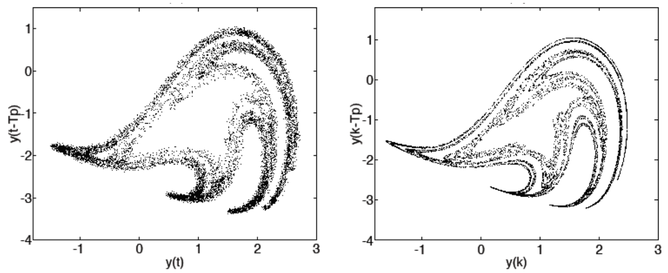
Michele A. Bucci[1], **Onofrio Semeraro**[2], **Sergio Chibbaro**[3], **Alex Allauzen**[4], **Lionel Mathelin**[2]

1) TAU, Inria, Université Paris-Saclay, CNRS, LRI, Orsay, (FR)
2) LIMSI (Decipher), CNRS, Université Paris-Saclay, Orsay (FR)
3) Institut Jean Le Rond d'Alembert, CNRS, Sorbonne Université, Paris (FR)
4) MILES team, LAMSADE, ESPCI, Paris (FR)

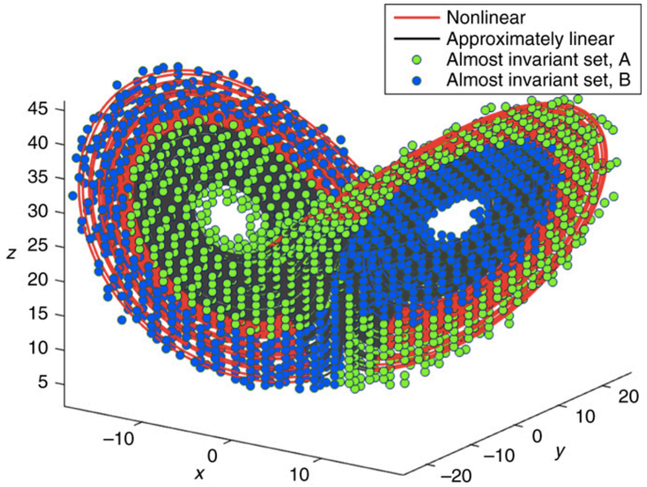
Mail: michele-alessandro.bucci@inria.fr

Data Driven model for chaotic systems

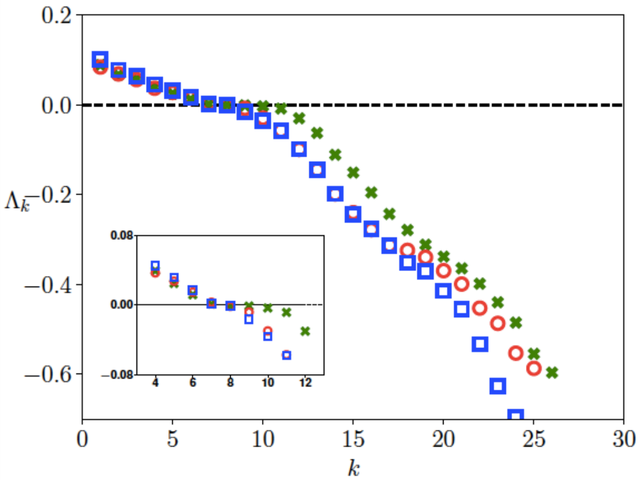
Retrieving the invariants of a chaotic systems *a-posteriori* from a model obtained by data-driven strategies is important to: i) **assess the underlying dynamical systems** and ii) **evaluate performances of new algorithms** in machine learning



Recover Poincaré section of modified van der Pol oscillator with NARMAX



Linear and approximately linear regions compared to quasi-invariant sets. (HAVOK)



Lyapunov exponents in Kuramoto-Sivashinsky with RNN

Aguirre, L. A., and S. A. Billings. "*Retrieving dynamical invariants from chaotic data using NARMAX models.*" International Journal of Bifurcation and Chaos (1995)

Brunton, S.L., et al. "*Chaos as an intermittently forced linear system.*"Nature communications (2017)

Vlachas P.R., et al. "*Backpropagation Algorithms and Reservoir Computing in Recurrent Neural Networks for the Forecasting of Complex Spatiotemporal Dynamics.*" (2019)

Ergodic theory
convergence criteria for dataset assessment

I.I.D. assumption

Machine learning models are valid under the **i.i.d.** (independent identically distributed) **assumption** of the dataset.

- sufficient but not necessary requirement to ensure consistency and error-bounds of the model
- a large **i.i.d**. dataset is **representative** of the whole **data distribution**

Low of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \mathbb{E}[X]$$

Shalizi, Cosma. "*Advanced data analysis from an elementary point of view*." 2013

Ergodic assumption

The Markov assumption states that: the future is independent of the past given the present. This assumption provides insights on the learnability of time series but does not provide insights on the amount of data required to learn.

In time-series, the i.i.d. assumption is substituted with the ergodic assumption:

- a single **long time series** becomes representative of the whole data-generating process

Low of large numbers

$$\frac{1}{n}\sum_{t=1}^{n} f(X_t^{t+k-1}) \rightarrow \mathbb{E}[f(X^k)]$$

for any reasonable function f. Markov assumption is weaker than ergodic one.

Shalizi, Cosma. "*Advanced data analysis from an elementary point of view*." 2013

Ergodic time-series

The ergodic theorem asserts that for every continuous function φ, for (almost) all initial condition $x(0)$ with respect to the *invariant measure* ρ.

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \varphi[f^t x(0)]dt = \int \rho(dx)\varphi(x)$$

Ergodic theory says that a time average equals a space average.

We get ergodic convergence if

Sensibility

- Time-series are long enough to converge on the largest Lyapunov exponent independent to the initial condition $x(0)$. **Require the Jacobian computation**

Geometry

- Time-series are long enough to converge the computation of the dimension of the probability measure ρ. **Fully data-driven**

Information

- Time-series are long enough to converge he computation of the entropy $h(\rho)$ of the probability measure. **Require partition of the support** ρ

Eckmann, J-P., and David Ruelle. "Ergodic theory of chaos and strange attractors." The theory of chaotic attractors. Springer, 1985.

Correlation dimension

Measure of the dimensionality is defined by different, yet equivalent, criteria: fractal dimension, Hausdorff dimension, Lyapunov dimension (chaos theory) or intrinsic dimension (machine learning).

Correlation dimension D_2 estimates the dimension of the system as the effective space occupied by the time-series.

$$C(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \Phi(\varepsilon - \|x_i - x_j\|)$$

Correlation integral

mean probability that the states at two different times are ε-close

$$C(\varepsilon) = \varepsilon^{D_2}$$

Grassberger-Procaccia algorithm

Grassberger, Peter, and Itamar Procaccia. "*Measuring the strangeness of strange attractors*." *The Theory of Chaotic Attractors*. Physica 9D, 1983

Minimum length of the time-series

Given a precision ε, a time-series of known correlation dimension D_2 requires N points for being fully explored

$$N > \left(\frac{D}{\varepsilon}\right)^{D_2/2} \quad [1]$$

$$N > 2(D_2 + 1)^{D_2} \quad [2]$$

$$N > \frac{R(2-Q)}{2(1-Q)}^{2D_2+1} \quad [3]$$

Regardless of the criteria, N **always depends exponentially on** D_2

[1] Eckmann, J-P., and David Ruelle. "*Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems.*" Physica D: Nonlinear Phenomena (1992)

[2] Essex, Christopher. "Correlation dimension and data sample size." Non-linear Variability in Geophysics. Springer, Dordrecht, 1991. 93-98.

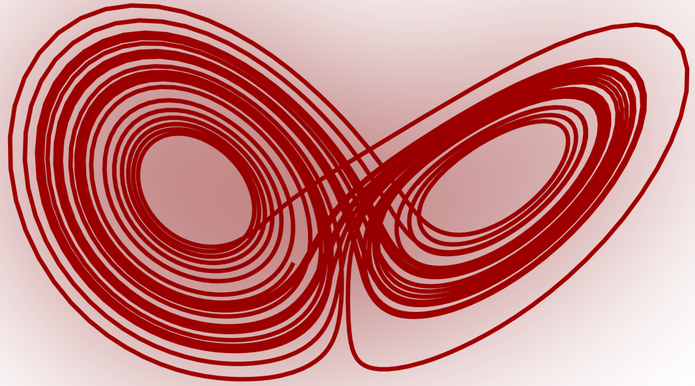
[3] Baker, Gregory L., Gregory L. Baker, and Jerry P. Gollub. Chaotic dynamics: an introduction. Cambridge university press, 1996.

Machine learning modelling

Introduce invariants to go beyond the ergodic limit

Lorenz system test case

Developed by Edward Lorenz in 1963 to describe the atmospheric convection. One of the **simplest chaotic dynamical system**.



$$(\sigma, \beta, \rho) = (10, 8/3, 28)$$

$$\begin{cases} \dot{x} = \sigma(y - x), \\ \dot{y} = x(\rho - z) - y, \\ \dot{z} = xy - \beta z\,. \end{cases}$$

Correlation dimension
$D_2 = 2.06$

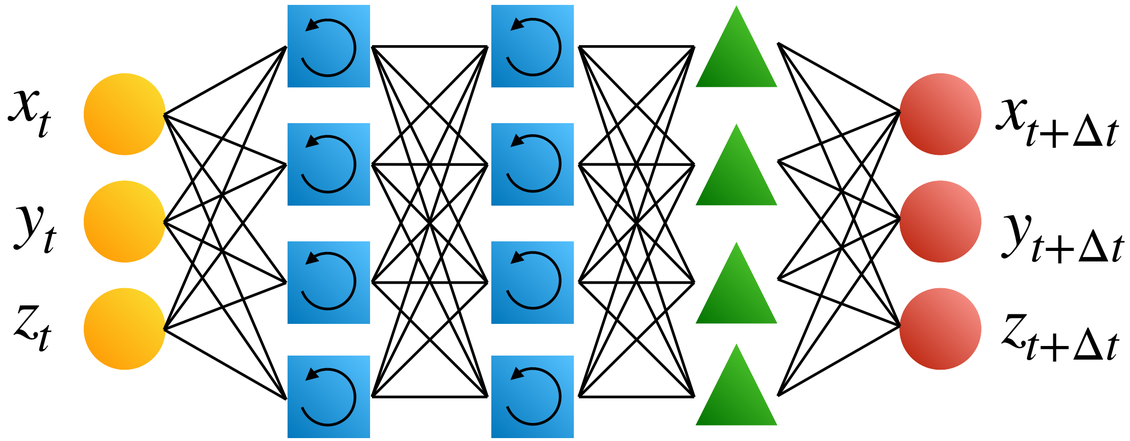
Eckmann & Ruelle

$D/\varepsilon = 5 \cdot 10^{-4}$

Minimum length
$N \approx 27000$

Data driven model

Long Short Time Memory (**LSTM**[1]) architecture state-of-art in physics to learn sequential informations*. Two LSTM layers with 50 neurons in each layer and one fully connected layer (~ 31000 parameters) employed to learn Lorenz chaotic attractor



LSTM uses **past information to predict** the future state. The past information is condensed in a vector (memory) and updated with new, incoming observations at each iteration.

Training $\min_{\theta} ||NN_{\theta}(x_t) - x_{t+\Delta t}||_2$ with **Adam** optimizer and adaptive learning rate

[1]Hochreiter, and Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997)

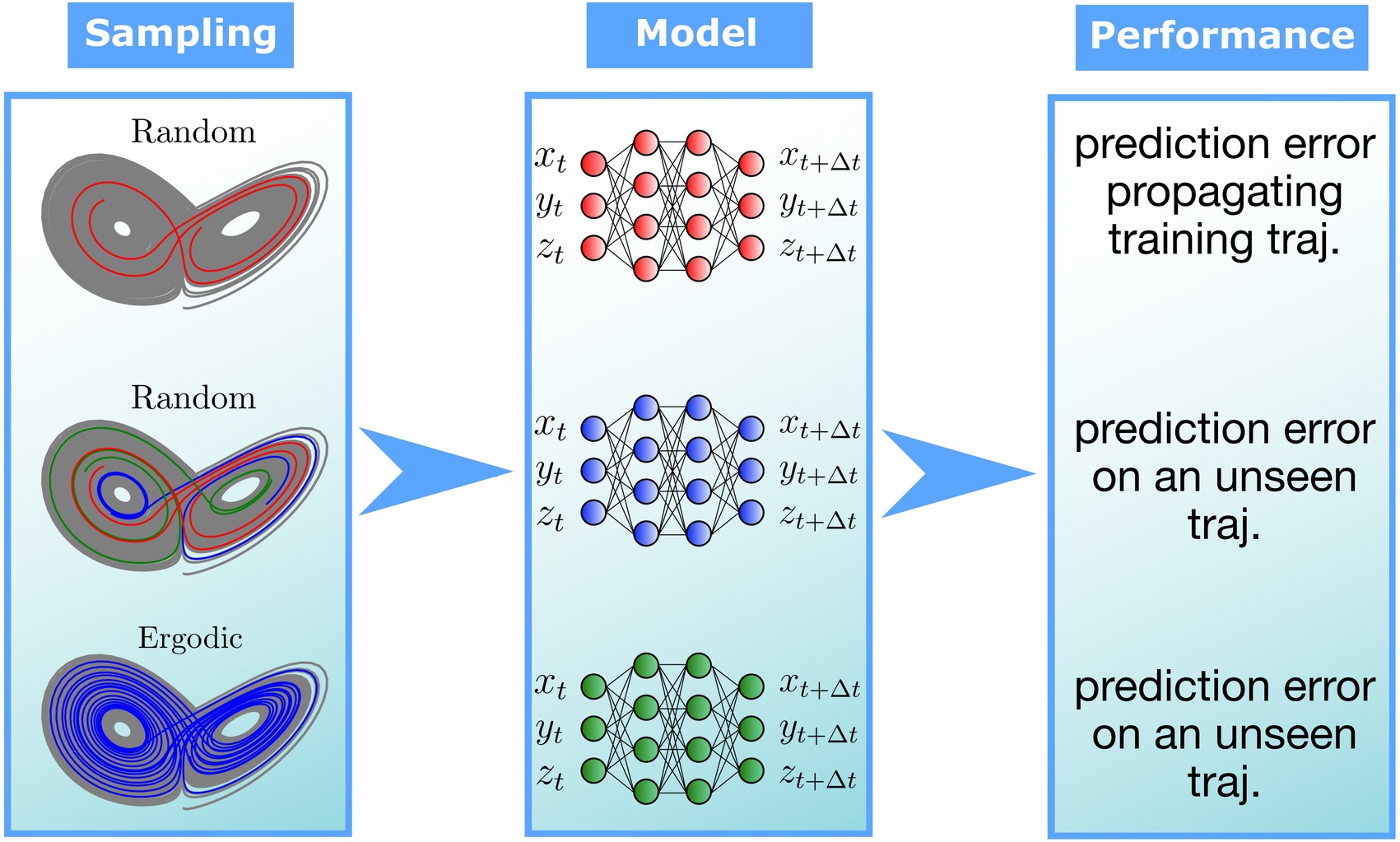
*A new algorithm is now used to handle sequential informations in Natural Language Processing (Attentional Neural Network) but it has been never tested for physical systems

Numerical experiment

1 dataset with 1 randomly sampled traj. $N = 3000$ and 100 models with different initialisation of the NN parameters.

100 dataset with 1 randomly sampled traj. $N = 3000$ and 1 model for each dataset.

100 dataset with 1 ergodic traj. $N = 27000$ and 1 model for each dataset.

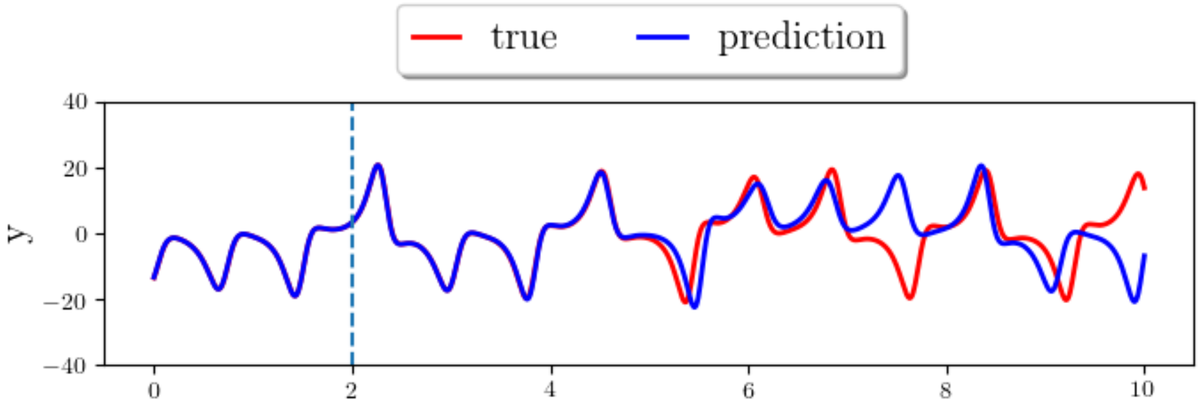


Random vs Ergodic sampling

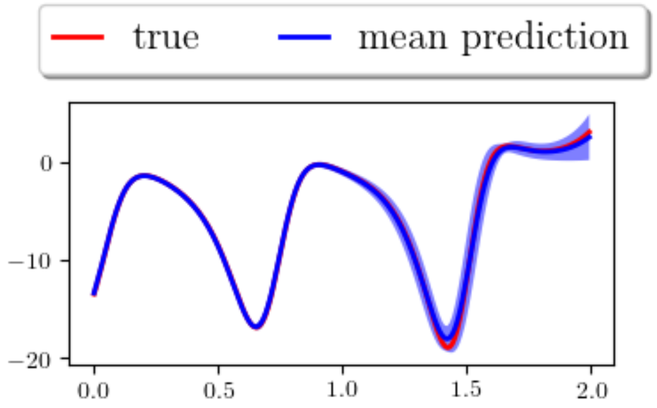
LSTM-based model is able to predict a chaotic dynamics (prediction time horizon $> 1/\lambda_1 \approx 1.1$) also when trained on non ergodic time-series ($N < 27000$).

Random vs Ergodic sampling

Short w/ propagation



Comparison between true dynamics and predicted one. In this numerical experiment the NN is used to **propagate the time-series observed in the training** stage. Training data: random time-series with $N = 3000$

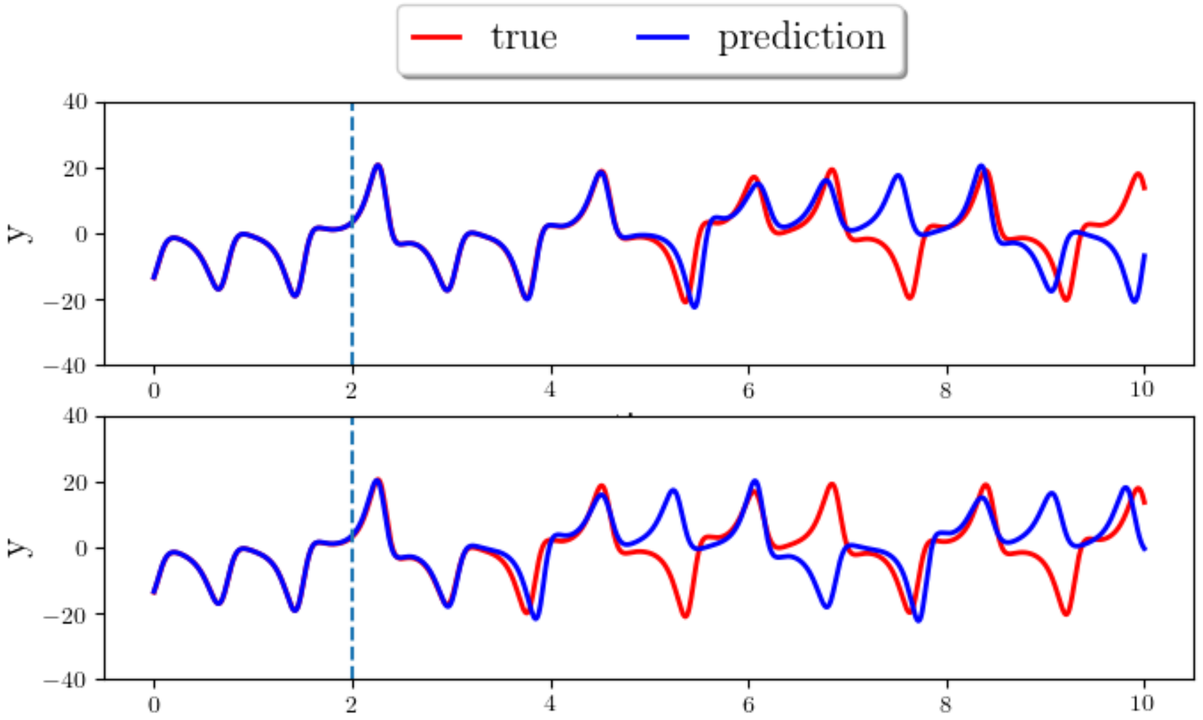


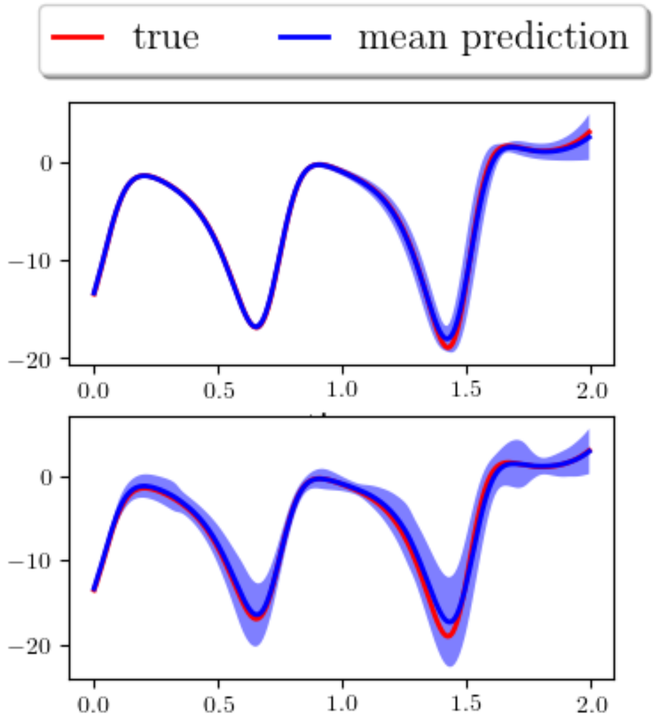
Same experiment repeated 100 times with random initialization of the NN parameters

Random vs Ergodic sampling

Short w/ propagation

Short w/o propagation





Repeating the numerical experiment but without propagation of the training time series, the **resulting models are no longer satisfactory**

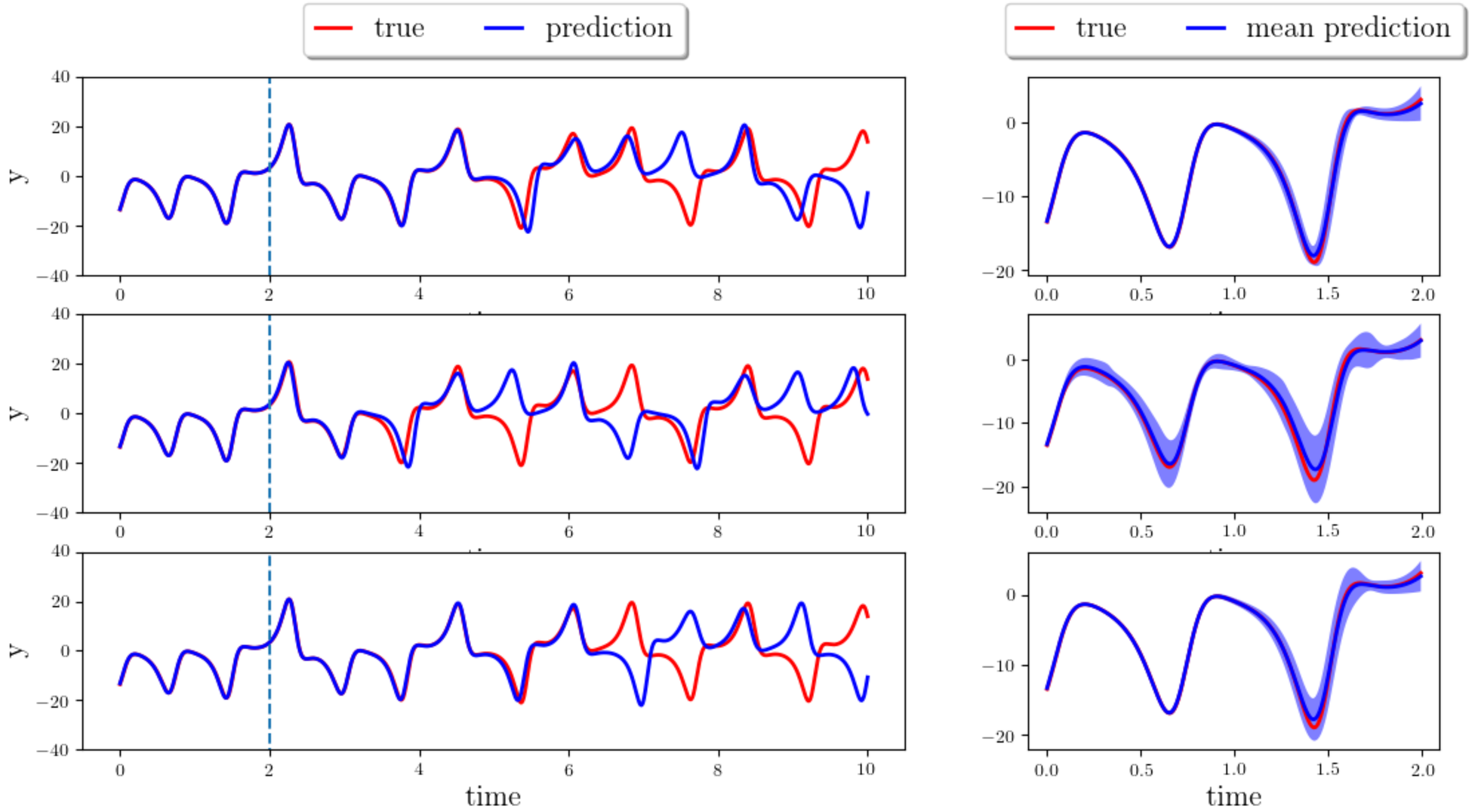
Even with a loss $< 10^{-4}$, many models are neither chaotic nor unsteady

Random vs Ergodic sampling

Short w/ propagation

Short w/o propagation

Ergodic



With an **ergodic time-series** and w/o propagation, the model **generalise** ∀ **i.c.**

Random vs Ergodic sampling

Short w/ propagation

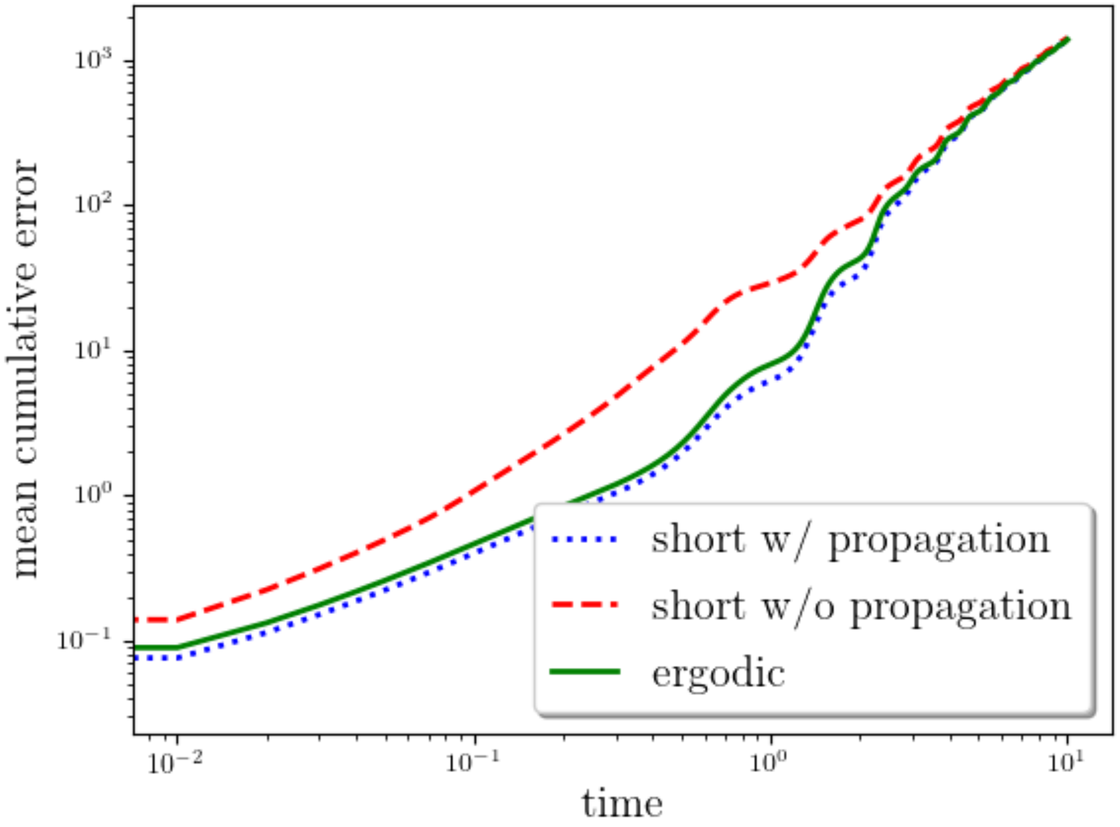
100 models trained with different i.c. of the parameters and the same trajectory ($N = 3000$)

Short w/o propagation

100 models trained with different i.c. of the parameters and different trajectories ($N = 3000$)

Ergodic

100 models trained with different i.c. of the parameters and different trajectories ($N = 27000$)



Without an ergodic dataset and w/o propagation, the model **errors** are **an order of magnitude higher**

Introducing invariants in the dataset

An **ergodic** window of **observation** of the system is **rarely guaranteed** in real-life applications. The number of time-steps required to reconstruct an high dimensional dynamics ($D_2 > 7$) rapidly increases to several order of magnitude. Furthermore the back propagation of the gradients through an infinitely long time-series is affected by the **vanishing gradient** issue.

IID assumption is a sufficient but not necessary condition to ensure error bounds.

Is there a way to sample a chaotic system which allows us to use less data?

"*if x is an unstable **fixed point** of the evolution, then the δ function at x is an **invariant measure**, …*" Eckmann & Ruelle (1985)

The dynamics around fix points is important to "interpret" chaotic behaviours[1,2].

[1]Kawahara G., Uhlmann M., & Van Veen, L. (2012). "*The significance of simple invariant solutions in turbulent flows"*. Annual Review of Fluid Mechanics

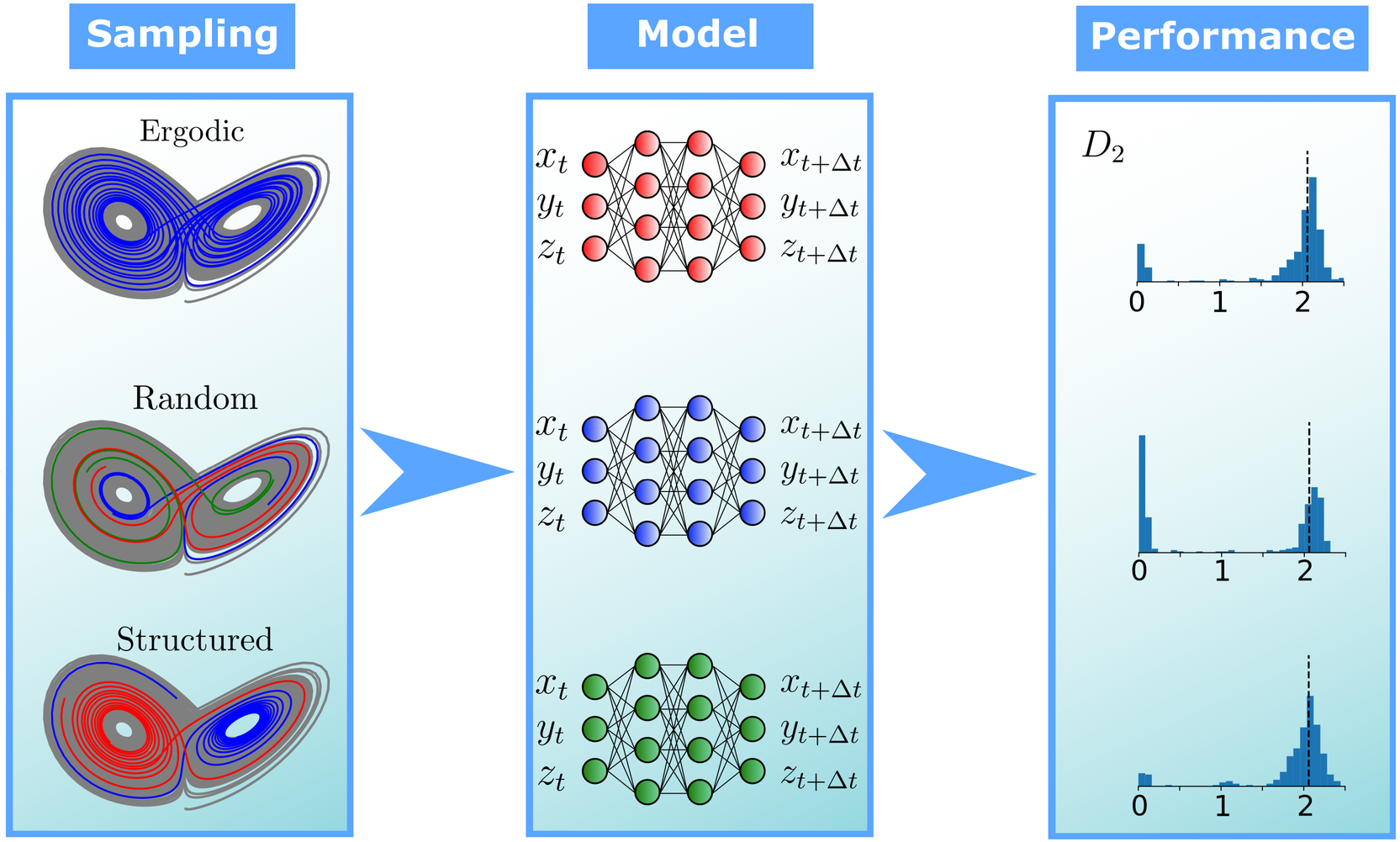
[2]Cvitanović, Predrag, Ruslan L. Davidchack, and Evangelos Siminos. "On the state space geometry of the Kuramoto–Sivashinsky flow in a periodic domain." *SIAM Journal on Applied Dynamical Systems* (2010).

Sampling strategy

300 dataset with 1 ergodic traj. $N = 27000$ and 1 model for each dataset.

300 dataset with 9 randomly sampled traj. $N = 3000$ and 1 model for each dataset.

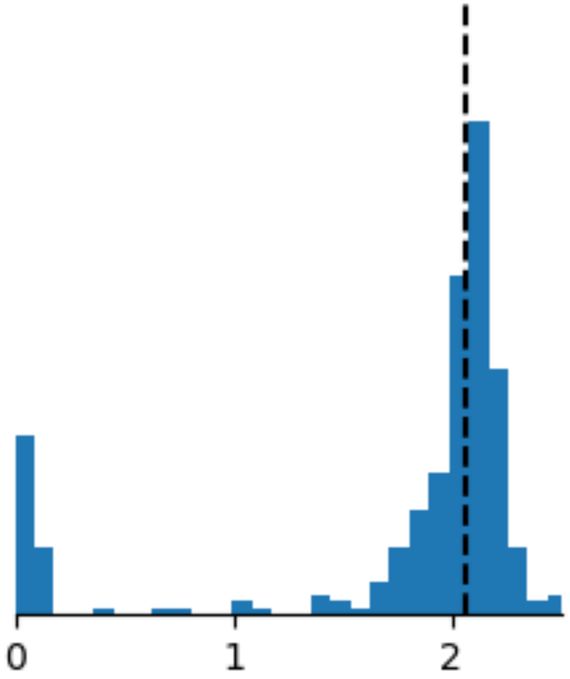
300 dataset with 9 traj. $N = 3000$ initialised from fix points (3 for each fixed point) and 1 model for each dataset.



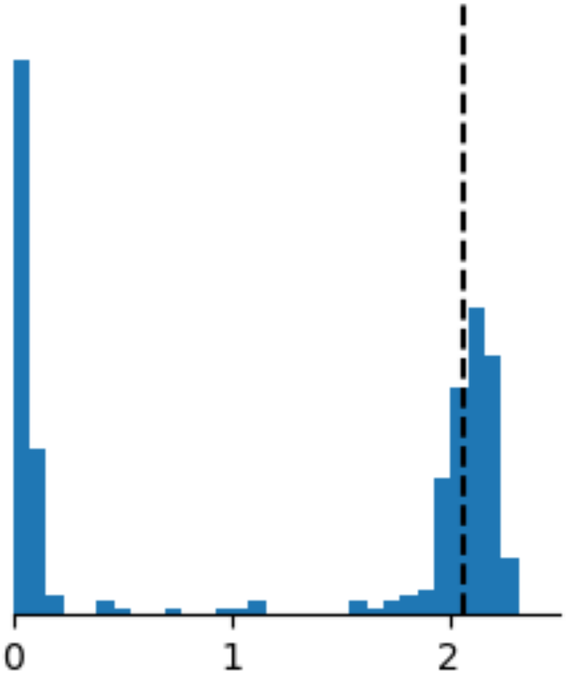
$$\{E_1, E_2, E_3\} = \{(0,0,0), (\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, (\rho-1)), (-\sqrt{\beta(\rho-1)}, -\sqrt{\beta(\rho-1)}, (\rho-1))\}$$

Model quality

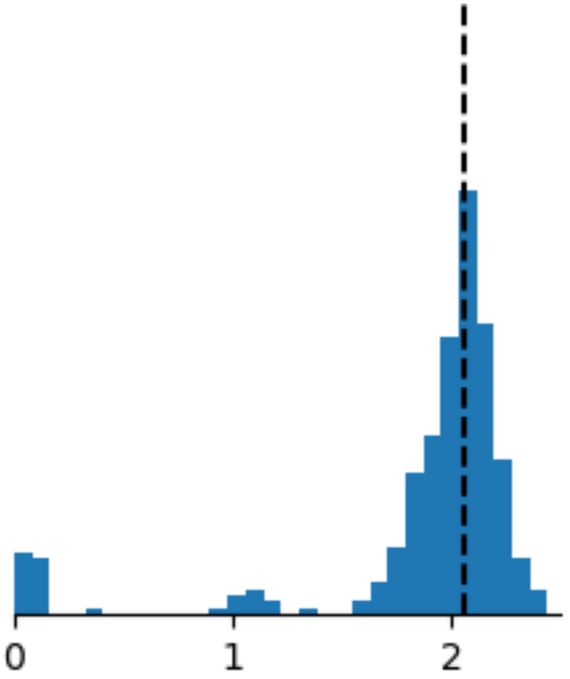
Ergodic



Random



Structured

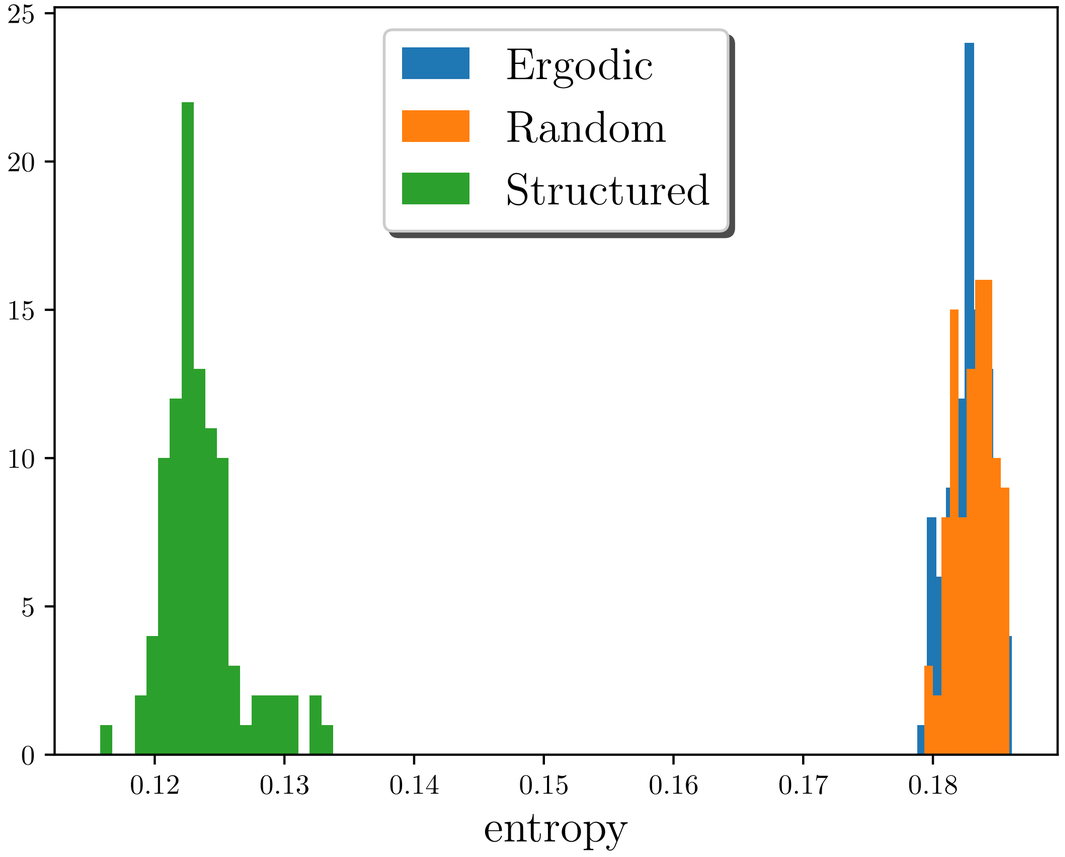


Outliers: models out

$D_2 = 2.06 \pm 0.3$

Ergodic	Random	Structured
16%	42%	**6%**

Dataset assessment



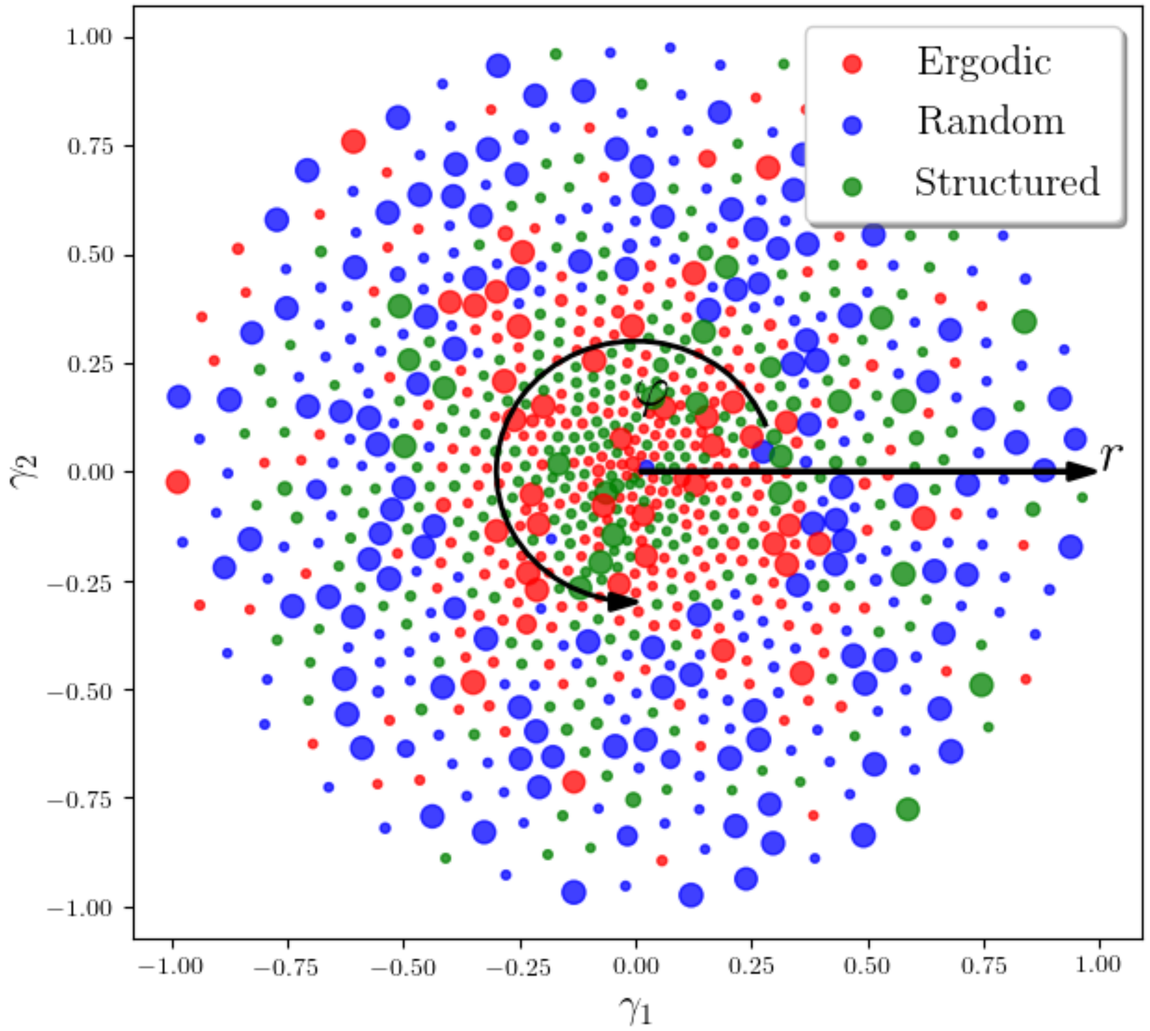
The SVD entropy of the trajectories in the three dataset strategies reveals how the **Structured dataset** has the **lowest entropy**.

The entropy for a time series is defined on the **probability transition** between embedded (through SVD) states. Low entropy:

- high predictability
- less information

[1]Varshavsky, Roy, et al. "Novel unsupervised feature filtering of biological data." Bioinformatics 22.14 (2006)

Model parameter assessment



The NN parameters are **initialised with a gaussian distribution**. After the training stage the parameters move to fit the data distribution.

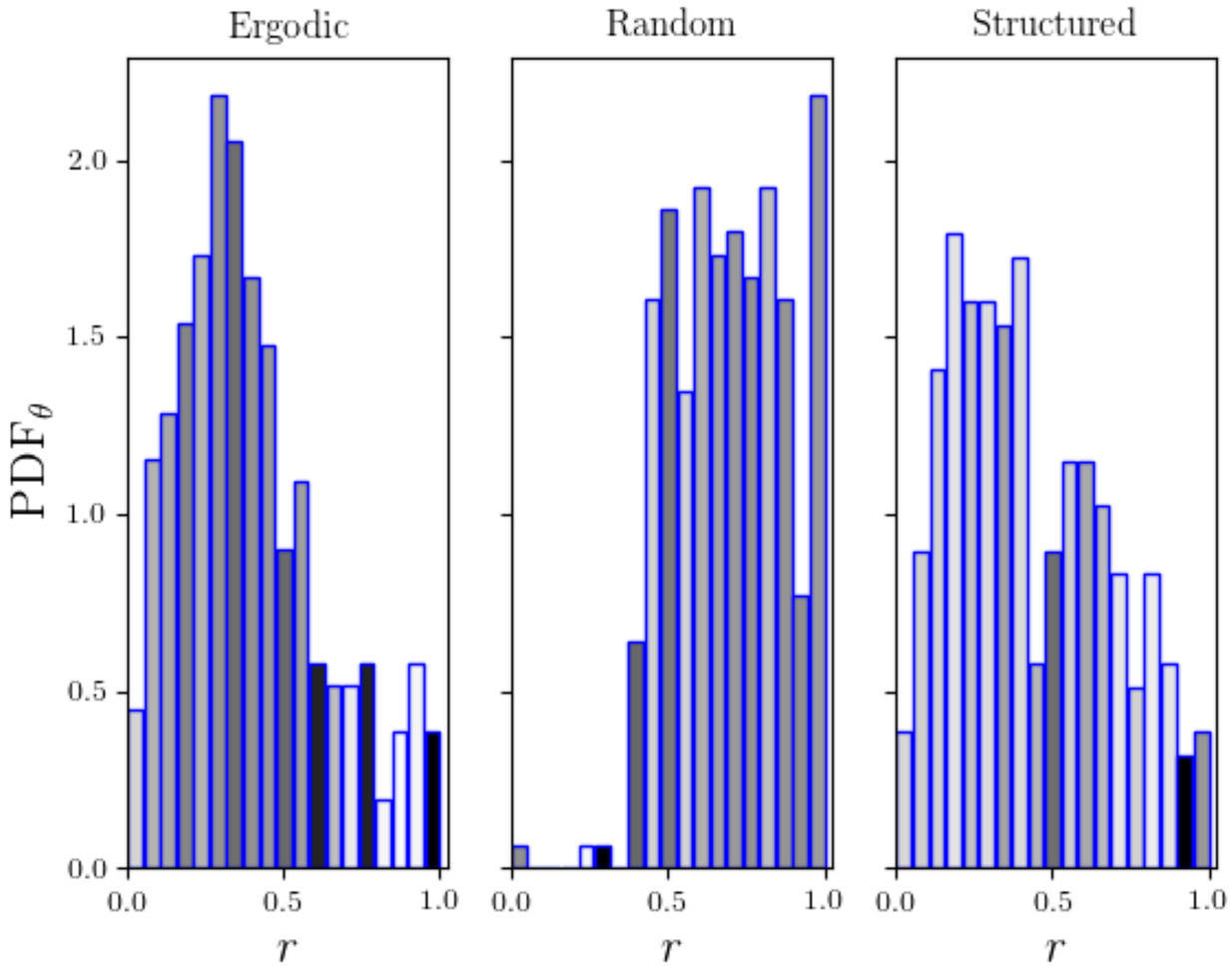
The parameter vector $\mathbb{R}^n$ with $n = 31000$ is projected in $\mathbb{R}^2 = (\gamma_1, \gamma_2)$ with t-SNE[1].

Each point corresponds to a model and the point size is proportional to the D_2 error.

Homogeneous distribution of the embedded parameters in the azimuthal direction φ…

[1]Van Der Maaten, Laurens. "Learning a parametric embedding by preserving local structure." *Artificial Intelligence and Statistics*. 2009.

Model parameter assessment



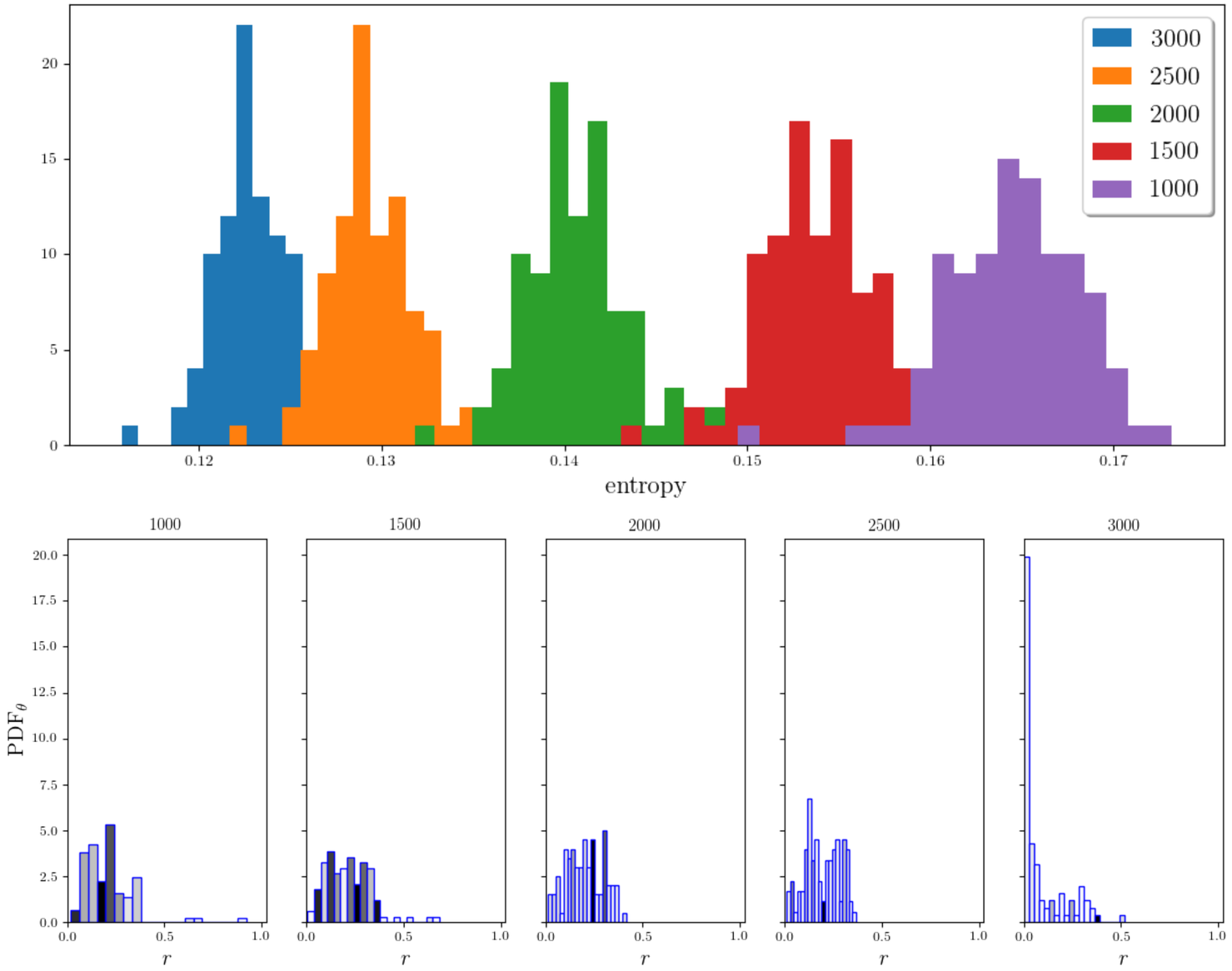
PDF of the radial distribution of the models. Ergodic and Structured dataset have almost the same distribution with a peak at $r \approx 0.3$.

The Random dataset has a complete different distribution. **Non-gaussianity** of the parameter distribution is symptomatic of **biased models**[1].

With Structured dataset, high D_2 errors occur with models in the tail distribution.

[1] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019

Smaller dataset



The **linear region has redundant informations** (exponential amplification of the same eigenvector).

In order to provide more **effective dataset** the linear part might be reduced increasing the entropy of the dataset.

Until $N = 18000$ $(-33\,\%)$ the model is not overfitting prone with low probability to recover biased models.

Discussion

	Ergodic assumption	Low dimensionality assumption
Hypotesis	The dataset is representative of the data distribution and we might be sure on the **model generalisation**	We can apply **manifold reduction** techniques to project the problem in the “effective dimension”, then apply regression for the temporal dynamics
Risk	**high probability to recover biased models**	**manifold reduction filters out useful informations**
Possible solution	**Our work**	**Physical constraints**[1]

[1] Loiseau, Jean-Christophe, and Steven L. Brunton. "Constrained sparse Galerkin regression.” JFM (2016)

Conclusions

Introducing information in the dataset on the invariant measures of the system it is possible to design **smart dataset** which engender more **robust learning process**:

"*if x is an unstable **fixed point** of the evolution, then the δ function at x is an **invariant measure**, …*" Eckmann & Ruelle (1985)

Conclusions

Introducing information in the dataset on the invariant measures of the system it is possible to design **smart dataset** which engender more **robust learning process**:

"*if x is an unstable **fixed point** of the evolution, then the δ function at x is an **invariant measure**, **but is not observed**.*" Eckmann & Ruelle (1985)

Conclusions

Introducing information in the dataset on the invariant measures of the system it is possible to design **smart dataset** which engender more **robust learning process**:

"*if x is an unstable **fixed point** of the evolution, then the δ function at x is an **invariant measure**, **but is not observed**.*" Eckmann & Ruelle (1985)

Unstable fixed points are never observed and recover them is not straightforward (e.g. non-linear systems requires Newton's iterations).

Conclusions

Introducing information in the dataset on the invariant measures of the system it is possible to design **smart dataset** which engender more **robust learning process**:

"*if x is an unstable **fixed point** of the evolution, then the δ function at x is an **invariant measure**, **but is not observed**.*" Eckmann & Ruelle (1985)

Unstable fixed points are never observed and recover them is not straightforward (e.g. non-linear systems requires Newton's iterations).

In many problems the equations are known but their solution is computationally demanding. In this scenario machine learning models can be a solution to recover real-time predictions or optimal control policy (i.e. reinforcement learning framework).

Tackle invariants to explore the manifold of the equation is a way to get less data-hungry machine leanrning models