Symbolic Regression for Generic Systems

Vivant Rodolphe

UFR de Mathématiques, University of Strasbourg

August 24, 2025

Plan

Overview

- Symbolic Regression
  - Generates interpretable mathematical models from data with minimal prior assumptions
  - SINDy (Sparse Identification of Nonlinear Dynamical Systems) [2]
    - uses sparse regression to identify parsimonious models
    - rely on a fixed set of basis functions thus limiting their ability to capture complex dynamics
  - ADAM-SINDy (Augmented Method) [2]
    - optimizes nonlinear parameters and selects candidate functions from a larger set
    - overcomes limitations of SINDy by allowing for more complex dynamics

Objective

- Explore the use of the ADAM-SINDy method to identify mathematical models of generic systems.
- Implement the SINDy and ADAM-SINDy methods in the Scimba library.
- Test it with various examples :
  - test dynamical systems
  - GENERIC systems

Scimba

Python package for the implementation of different Scientific Machine Learning methods.

- Some of its features:
  - Networks : Multi Layer Perceptron (MLP), Discontinuous MLP, RBF networks, activation functions, etc...
  - Models of differentials equations : Ordinary differential equations (ODE), Partial (PDE), Spatial PDEs, time-space PDEs,...
  - Specific networks for Physics informed neural networks (PINN) : MLP, Discontinuous MLP, nonlinear RBF networks, Fourier networks, etc.
  - Trainer: Each type of PDE has its own trainer

Theoritical Context

System of Equations type :

$$\dot{x}(t) = f(x(t)) \tag{1}$$

- $x(t) \in \mathcal{R}^n$ is the state vector of the system at time t
- $\dot{x}(t)$ its fist time derivative
- $f(x) : \mathcal{R}^n \to \mathcal{R}^n$ is a nonlinear function that describes the dynamics of the system

Sparse Identification of Nonlinear Dynamical Systems (SINDy)

From a set of observed data :

$$\mathbf{X} = \begin{bmatrix} x(t_1)^T \\ x(t_2)^T \\ \vdots \\ x(t_m)^T \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \cdots & x_n(t_m) \end{bmatrix} \tag{2}$$

and a master functions library :

$$\Omega(\mathbf{X}; \mathbf{\Lambda}) = [1 \quad \mathbf{X}^{\mathbf{A}} \quad sin(B\mathbf{X}) \quad cos(C\mathbf{X}) \quad exp(D\mathbf{X}) \quad \mathbf{X} \otimes sin(E\mathbf{X}) \quad \mathbf{X} \otimes cos(F\mathbf{X}) \quad \mathbf{X} \otimes exp(G\mathbf{X})] \tag{3}$$

- A, B, C, D, E, F, G are the chosen non-linear parameters (Λ)

SINDy

- The SINDy method [2] aims to find a sparse coefficient vector $\mathbf{\Theta} = [\theta_1, \theta_2, ..., \theta_p]$ such that:

$$\dot{\mathbf{X}} = \Omega(\mathbf{X}; \mathbf{\Lambda})\mathbf{\Theta} \tag{4}$$

- Sparsity of the method is guaranteed using a regularization technique as Lasso augmented
- Sparse regression formulation :

$$\mathbf{\Theta} = \min_{\mathbf{\Theta}} \left(\left\| \dot{\mathbf{X}} - \Omega(\mathbf{X}; \mathbf{\Lambda})\mathbf{\Theta} \right\|_2^2 + \lambda \left\| \mathbf{\Gamma\Theta} \right\|_1 \right) \tag{5}$$

  - λ : regularization parameter that control the sparsity
  - Γ : Identity Matrix

Augmented Method : ADAM-SINDy

- Aim to overcome the limitations of the fixed basis functions of the SINDy method
- Simultanuous optimization of the linear and non-linear parameters

$$\boldsymbol{\Theta} = \min_{\boldsymbol{\Theta}} \max_{\boldsymbol{\Gamma}\mathbf{or}\lambda} \left(\left\| \dot{\mathbf{X}} - \Omega(\mathbf{X}; \boldsymbol{\Lambda})\boldsymbol{\Theta} \right\|_2^2 + \lambda \left\| \boldsymbol{\Gamma}\boldsymbol{\Theta} \right\|_1 \right) \tag{6}$$

  - minimize the loss function with Θ
  - maximize with Γ or λ
  - rather use of Γ to controll each candidate functions' contribution individually

Generic formalism

- Description of the evolution of systems [1] [3] in beyond-equilibrium thermodynamics
- systematic way to model the dynamics of systems with both conservative and dissipative systems
- useful for studying complex systems where energy and entropy exchanges play a crucial role

Generic Formalism

$$\begin{cases} \dot{x}(t) = L(x(t))\,\nabla E(x(t)) + M(x(t))\,\nabla S(x(t)) \\ L\nabla S = 0 \\ M\nabla E = 0 \end{cases} \tag{7}$$

- $L\nabla E$ is the conservative part of the system
- $M\nabla S$ is the dissipative part
- E and S : Energy and Entropy of the system
- L is the skex -symmetric Poisson Matrix
- M is the symmetric semi-definite friction matrix

Generic Formalism

- Final Loss :

$$L_{tot} = L_{MSE} + L_{deg} \tag{8}$$

  - with $L_{deg} = \sum_{i,j} \min_{\mathbf{M_{ij}}} M_{ij} \nabla E + \min_{\mathbf{L_{ij}}} L_{ij} \nabla S$
  - Ł$_{deg}$ imposes the GENERIC formalism.

Example equation

- Harmonic Oscillator [2] :
  - Dynamical system equation :

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -x_1(t) + 0.1\, x_2(t)\, cos(0.75\, x_1(t)) \end{cases} \tag{9}$$

- Damped NonLinear Oscillator [1]
  - Dynamical system equation :

$$\begin{cases} \dot{q}(t) = p(t) \\ \dot{p}(t) = -3\, sin(q(t)) - 0.04\, p(t) \\ \dot{S}(t) = -0.04\, p(t)^2 \end{cases} \tag{10}$$

  - Energy equation

$$E(t) = 0.5\, p(t)^2 - 3\, \cos(q(t)) + S \tag{11}$$

SINDy - Harmonic Oscillator

- Canditate functions
  - $\Omega_1 = [x_2]$
  - $\Omega_2 = [x_1, x_2 \otimes cos(0.75\, x_1)]$
- 50000 iterations
- tmax = 50s, dt 0.01
- $\lambda = 0.001$
- initial learning rate of 0.1
- Found equation

$$\begin{cases} \dot{x}_1(t) = 1.00199711322784\, x_2 \\ \dot{x}_2(t) = -0.999280571937561\, x_1 + 0.101232923567295\, x_2 \cos(0.75 x_1) \end{cases} \tag{12}$$

SINDy - Harmonic Oscillator

- Plot :

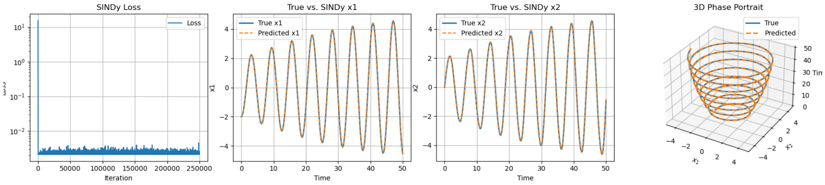

Figure: Identified model for the harmonic oscillator

- Very good approximation, with a small error of 10^{-3}

SINDy - Damped nonlinear Oscillator

- Canditate functions
  - $\Omega_1 = [p]$
  - $\Omega_2 = [p, sin(q)]$
  - $\Omega_3 = [p^2]$
- 50000 iterations
- tmax = 50s, dt 0.01
- $\lambda = 0.001$
- initial learning rate of 0.001
- Found equation

$$\begin{cases} \dot{q}(t) = 0.999506831169128\, p(t) \\ \dot{p}(t) = -0.0397274941205978\, p(t) + -3.00022983551025\, \sin(q(t)) \\ \dot{S}(t) = 0.0395135618746281\, p(t)^2 \end{cases} \tag{13}$$

SINDy - Damped nonlinear Oscillator

- Plot :

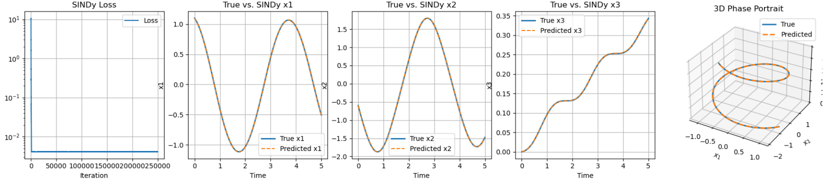

Figure: Identified model for the damped nonlinear oscillator

- Very good approximation, with a small error of 10^{-3}
- Energy found equation :

$$E(t) = 0.500103414058685\, p(t)^2 + 1.0\, S(t) - 2.99728417396545\, \cos(1.0\, q(t)) \tag{14}$$

Adam-SINDy - Harmonic Oscillator

- 50000 iterations
- tmax = 5s, dt 0.01
- $\lambda = 0.001$
- initial learning rate of 0.01
- pruning coefficient $\epsilon = 0.005$
- Found equation

$$\begin{cases} \dot{x}_1(t) = 1.00006556510925\, x_2 + 0.000205039978027344\, x_1 \\ \dot{x}_2(t) = -1.0004680454731\, x_1 + 0.100180670619011\, x_2 \cos(0.75 x_1) \end{cases} \tag{15}$$

Adam-SINDy - Harmonic Oscillator

- Plot :

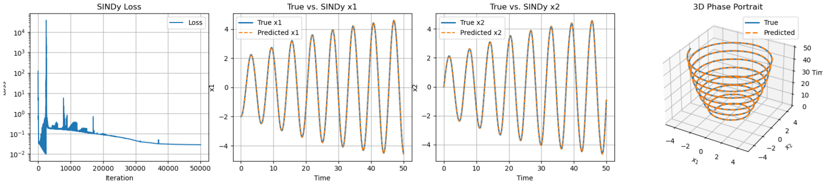

Figure: Identified model for the harmonic oscillator with ADAM-SINDy

- Very good approximation, with a small error from 10^{-3} to 10^{-5}
- Succesfull identification of non-lineat parameters.

Adam-SINDy - Damped nonlinear Oscillator

- 50000 iterations
- tmax = 5s, dt 0.001
- $\lambda = 0.001$
- initial learning rate of 0.01
- pruning coefficient $\epsilon = 0.005$
- Found equation

$$\begin{cases} \dot{q}(t) = 0.999236464500427\, p(t) \\ \dot{p}(t) = -1.98613214492798\, p(t)\, \exp(0.01\, S(t)) \\ \qquad -0.923614144325256\, p(t)\, \cos(0.92\, p(t)) \\ \qquad +0.0422132685780525\, p(t)\, \cos(1.82\, S(t)) \\ \qquad -0.0287085622549057\, S(t)\, \exp(-0.9\, p(t)) \\ \dot{S}(t) = 0.0395686700940132\, p(t)^2 \end{cases} \tag{16}$$

Adam-SINDy - Damped nonlinear Oscillator

- Plot :

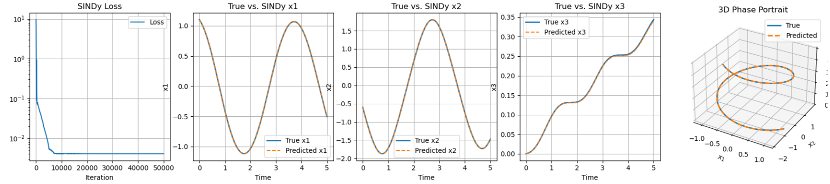

Figure: Identified model for the damped nonlinear oscillator with ADAM-SINDy

- Very good approximation of the main dynamic
- Unexpected terms for the symbolic formula of $p(t)$

Adam-SINDy - Damped nonlinear Oscillator

- Plot of Energy :

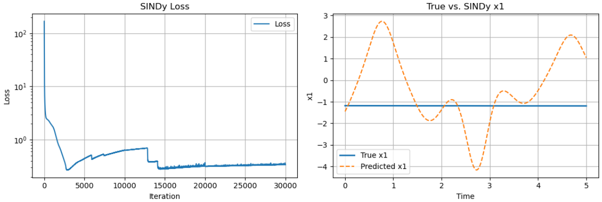

Figure: Identified Energy formula for the damped nonlinear oscillator with ADAM-SINDy

- Difficulty to capture the right behaviour for the energy

Conclusion

During this internship :

- Implementation of SINDy and Adam-SINDy methods into the SCIMBA library
- Extension with the structure-preserving parametrization, the GENERIC formalism
- Successful modelisation of the dynamic with both methods
- Failure to mix Adam-SINDy and GENERIC formalism

Perspectives

In the future :

- Fix this failure
  - looking for other training parameters
  - Coding structural error
- Optimisation of the algorithms since computation time is between 350s and 550s.
- Adding the treatment of noise in the case of real signals

Special thanks to my supervisors Dr. Emmanuel Franck, Dr Andrea Thomann and Dr Michel Victor-Dansac.

References

[1] Kookjin Lee, Nathaniel Trask, and Panos Stinis. "Structure-preserving Sparse Identification of Nonlinear Dynamics for Data-driven Modeling". In: (2021).

[2] Siva Viknesh, Younes Tatari, and Amirhossein Arzani. "ADAM-SINDy: An Efficient Optimization Framework for Parameterized Nonlinear Dynamical System Identification". In: (2025).

[3] Zhen Zhang, Yeonjong Shin, and George Em Karniadakis. "GFINNs: GENERIC Formalism Informed Neural Networks for Deterministic and Stochastic Dynamical Systems". In: (2021).