

**Nov. 19-25  
2025**

# Hybrid-Cooperative Learning for PDEs



Context

# From First Principles to Data: Opposing Paradigms

## Partial Differential Equations

$$\Delta\phi = 0$$

**Laplace Equation**

$$\frac{\partial T}{\partial t} = \Delta T$$

**Heat Equation**

$$\frac{\partial^2 u}{\partial t^2} = \Delta u$$

**Wave Equation**

# The Ubiquity of AI

## The Nobel Prize in Chemistry 2024

### They cracked the code for proteins' amazing structures

The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools. David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins. Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins' complex structures. These discoveries hold enormous potential.

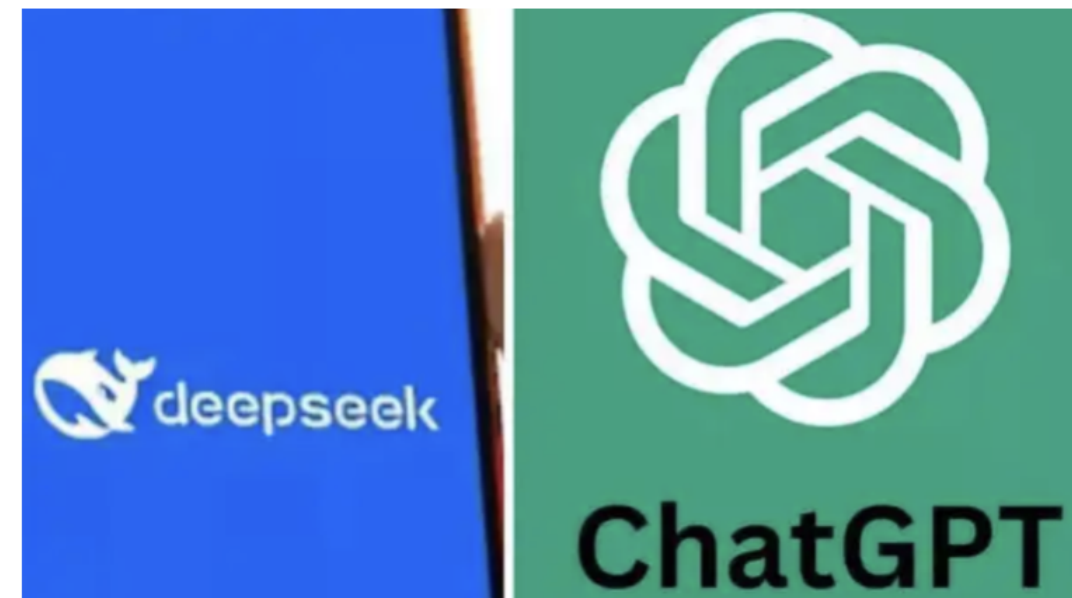


Geoffrey Hinton  
Nobel Prize in Physics 2024

Born: 6 December 1947, London, United Kingdom

Affiliation at the time of the award: University of Toronto, Toronto, Canada

Prize motivation: “for foundational discoveries and inventions that enable machine learning with artificial neural networks”





# Main objectives

- **Express / Represent**

- Capture, describe, or encode patterns and relationships present in data.

- **Generalize / Forecast**

- Use learned patterns to make predictions or draw conclusions about new, unseen situations.
- Example: predicting tomorrow's weather, or how a system will evolve.

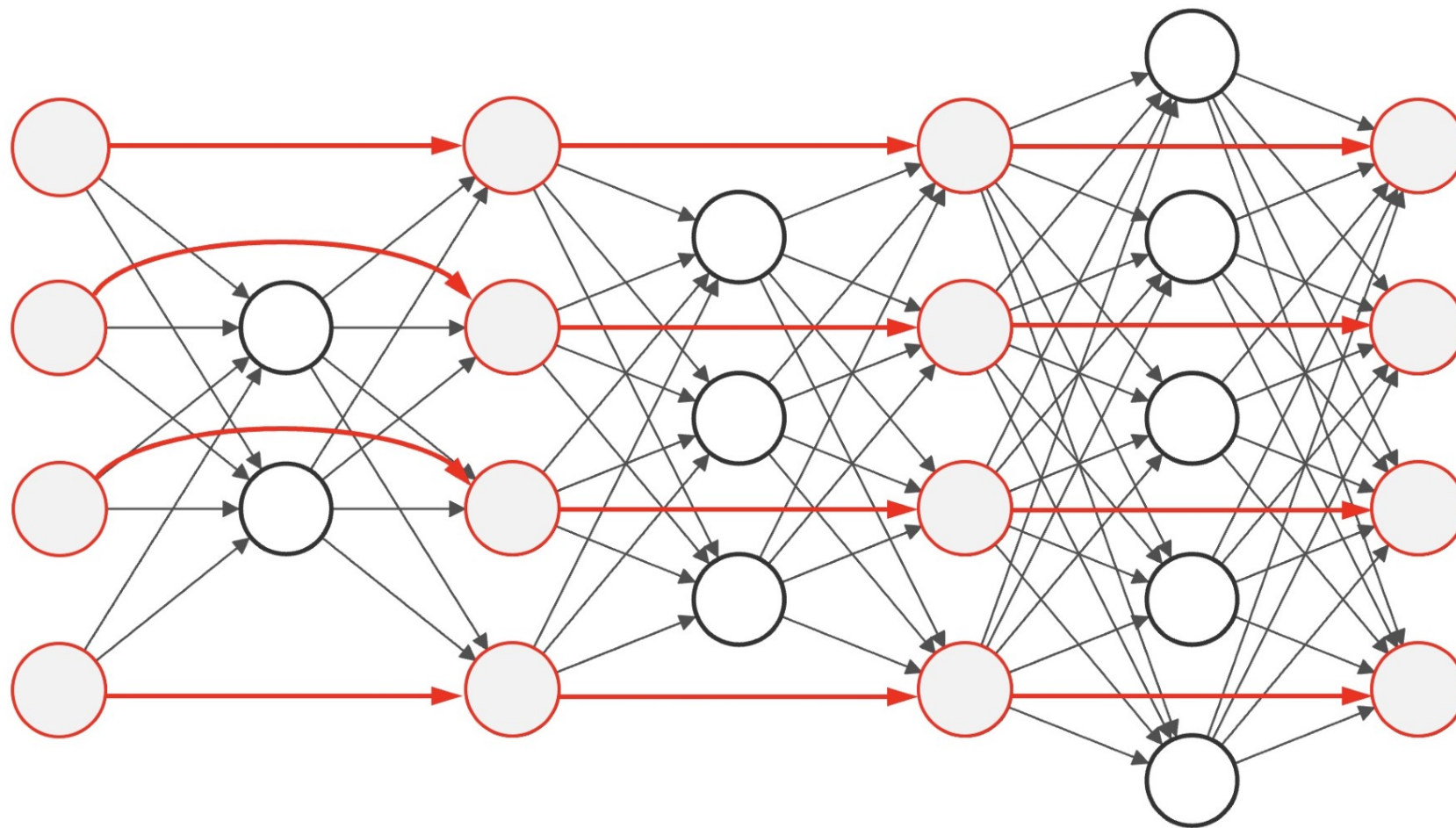
- **Generate / Create**

- Produce new, coherent outputs consistent with what has been learned, going beyond imitation.
- Example: generating realistic images, music, or solutions to design problems.

Express

# Supervised Learning: Computational practice

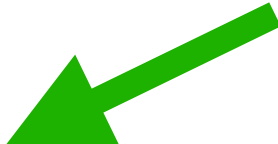
$$\underbrace{\frac{1}{N} \sum_{i=1}^N \text{loss}(x^i, \ell^i)}_{\text{empirical risk} := E(x(\cdot))} + \alpha \sum_{j=1}^K \|(\mathbf{a}_j, \mathbf{w}_j, b_j)\|^2$$



Supervised Learning

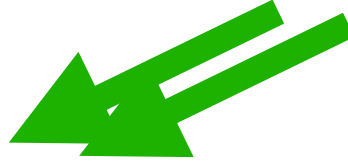
# Why does it work? Universal Approximation

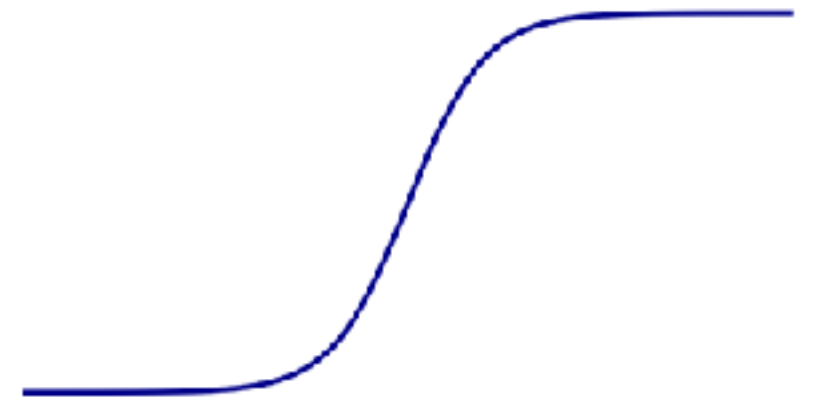
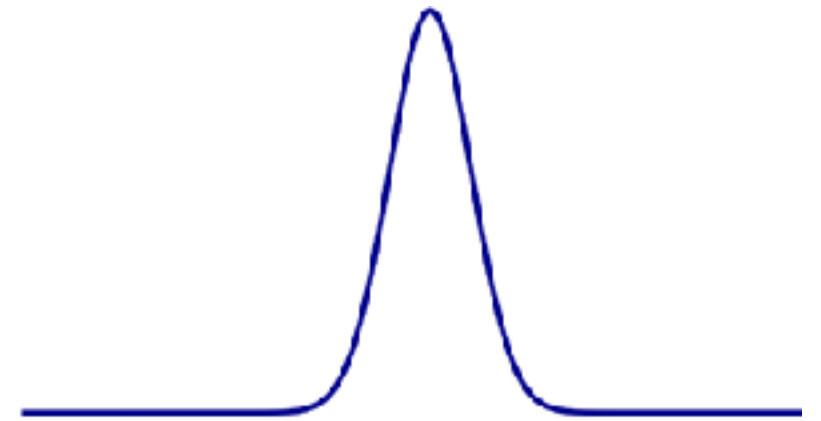
**N. Wiener**, Tauberian Theorems,  
Annals of Mathematics, 33 (1) (**1932**), 1-100.

$$f(x) \sim \sum_{j=1}^K w_j G(x + b_j)$$


**G. Cybenko**, Approximation by superpositions  
of a sigmoidal function,  
Mathematics of Control, Signals and Systems,  
(**1989**), 2: 303-314.

$$\lim_{x \rightarrow -\infty} \sigma(x) = 0, \quad \lim_{x \rightarrow +\infty} \sigma(x) = 1$$

$$f(x) \sim \sum_{j=1}^K w_j \sigma(a_j \cdot x + b_j)$$




# Cybernetics, Norbert Wiener, 1948

The science of control and communication in animals and machines



The linear finite  $d$ -dimensional system

$$x'(t) = Ax(t) + Bu(t), \quad t \in (0, T); \quad x(0) = x^0 \quad (1)$$

with  $m \ll d$  controls.

$A \in M_{d \times d}$ ,  $B \in M_{d \times m}$  and  $x^0 \in \mathbb{R}^n$ ;  $x : [0, T] \rightarrow \mathbb{R}^d$  represents the *state* and  $u : [0, T] \rightarrow \mathbb{R}^m$  the *control*.

Can we control  $d$  states with only  $m$  controls, even if  $n \gg m$ ?

## Theorem

(1958, Rudolf E. Kálmán) System (1) is controllable iff

$$\text{rank}[B, AB, \dots, A^{d-1}B] = d.$$



**DeepMind breaks 50-year math record using AI; new record falls a week later**

AlphaTensor discovers better algorithms for matrix math, inspiring another improvement from afar.



$$\dot{\mathbf{x}}(t) = \mathbf{w}(t) \sigma(\mathbf{a}(t) \cdot \mathbf{x}(t) + b(t))$$



Generalize

# A hallway discussion on generalization

**Mathematician A:** Generalization in Machine Learning is overrated. In the end it's just Gronwall's inequality.



**Mathematician B:** So with “just Gronwall” you can tell me how all the parameters of a large language model (LLM) will change if we simply drop a few web pages from the training set?

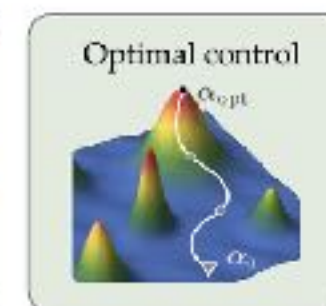
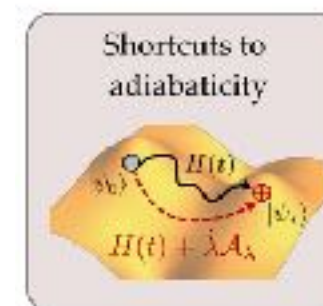
**Mathematician A:** In principle, yes. If the loss is smooth, at the global minimizer, we get an exponential bound with a constant  $C$  that depends on the model.

**Mathematician B:**  $C = C(\text{nonconvex loss}, 10^{11} \text{ parameters})$ ?  
That's not a constant, that's a cry for help.

**Mathematician A:** Well, assuming we are close to a global minimizer...

**Mathematician B:** A global minimizer? In that landscape?  
That's not an assumption, that's *science fiction*.

**Mathematician A:** So what do you suggest?



**Mathematician B:** First, we admit we are doing “Gronwall in the fog”.  
Then we call it a new research program on *robust generalization bounds* and apply for an ERC.

# PDE Approximation

# NN version of variational PDEs

Warning: Lack of convexity!

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

$$u \in H_0^1(\Omega) : \int_{\Omega} \nabla u \cdot \nabla \varphi dx = \int_{\Omega} f \varphi dx \quad \forall \varphi \in H_0^1(\Omega)$$

$$u \in H_0^1(\Omega) : \min_{v \in H_0^1(\Omega)} \left[ \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right]$$

FEM approximation (Galerkin): Replace the search and test infinite-dimensional space  $H_0^1(\Omega)$  by a FEM finite-dimensional one  $V_h$

$$u_h \in V_h : \min_{v \in V_h} \left[ \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right]$$

$$\|u - u_h\|_{H_0^1(\Omega)} \leq Ch \|f\|_{L^2(\Omega)}$$



# The NN version

What can NN do?

Replace  $V_h$  by a NN finite-dimensional manifold  $\mathcal{M}_K$ :

$$\mathcal{M}_K = \left\{ v(x) = \sum_{j=1}^K w_j \sigma(\mathbf{a}_j \cdot x + b_j) \right\}$$

$$\dim(\mathcal{M}_K) = K(d + 2), \quad d = \dim(\Omega)$$

Then

$$u_K \in \mathcal{M}_K : \min_{v \in \mathcal{M}_K} \left[ \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right]$$

And letting  $K \rightarrow \infty \dots$  one can develop a  $\Gamma$ -convergence like theory. <sup>3</sup>

**But the problem of minimising Dirichlet's energy in  $\mathcal{M}_K$  is non-convex!**

---

<sup>3</sup>(1) W. E & B. Yu, (2017). The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems.

(2) Luo, T. & Yang, H., (2020). Two-layer neural networks for partial differential equations: Optimization and generalization theory.

# Mean-field relaxation Back to $\infty$ -dimensions...

[K. Liu & E. Zuazua, Representation and regression problems in NN: Relaxation, Generalisation and Numerics, M3AS, 2025.]

## Shallow NN

The original Shallow NN writes:

$$\sum_{j=1}^K w_j \sigma(\mathbf{a}_j \cdot \mathbf{x} + b_j) \quad \rightarrow$$

where  $(w_j, \mathbf{a}_j, b_j) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$  for all  $j$ .

As the number of neurons  $K$  tends to infinity, and gets denser, the ansatz evolves into its relaxed/convexified version.



## Mean-field shallow NN

The **mean-field** shallow NN writes:

$$v_\mu(\mathbf{x}) = \int_{\mathbb{R}^{d+1}} \sigma(\mathbf{a} \cdot \mathbf{x} + b) d\mu(\mathbf{a}, b),$$

where  $\mu \in \mathcal{M}(\mathbb{R}^{d+1})$ .

The outcome is **linear** with respect to  $\mu$ !

This leads to the minimisation problem

$$\min_{\mu \in \mathcal{M}} \left[ \frac{1}{2} \int_{\Omega} |\nabla v_\mu|^2 dx - \int_{\Omega} f v_\mu dx \right].$$

### Warning!

- Loss of Hilbertian structure
- Loss of coercivity.

# Loss of coercivity even in finite-d

Consider the minimization of the Dirichlet energy over the 2-neuron Wiener ansatz:

$$\mathcal{M}_2 = \{v(x) = w_1 G(x - \mathbf{b}_1) + w_2 G(x - \mathbf{b}_2)\}.$$

This corresponds to a  $2(d + 1)$ -dimensional ansatz.

Coercivity can be rephrased as follows: *Does the  $H^1$  bound of  $v$  imply a bound on the parameters  $((w_1, \mathbf{b}_1), (w_2, \mathbf{b}_2))$ ?*

And the answer is NO!!!!

$$v_\epsilon = \frac{1}{\epsilon} [G(x - \epsilon) - G(x + \epsilon)]$$

generates a “wave packet” in which two neurons collapse or condensate so that  $v_\epsilon$  is bounded in  $H^1$  but the corresponding coefficients blow-up

$$|b_1| = |b_2| = \frac{1}{\epsilon}$$

Then,

$$v_\epsilon \rightarrow v^* \in H^1$$

with

$$v^*(x) = \partial_x G(x) = -\frac{x}{2} G(x) \notin \mathcal{M}_2.$$

Coercivity then requires the addition of some Tikhonov regularization:

- One can penalize the amplitude  $b$  adding a penalty term

$$\lambda[|b_1|^2 + |b_2|^2]$$

tion:

with  $\lambda > 0$ .

- Or the distance between the centers

$$\lambda \left[ \frac{1}{|b_1 - b_2|^2} \right].$$

Then,

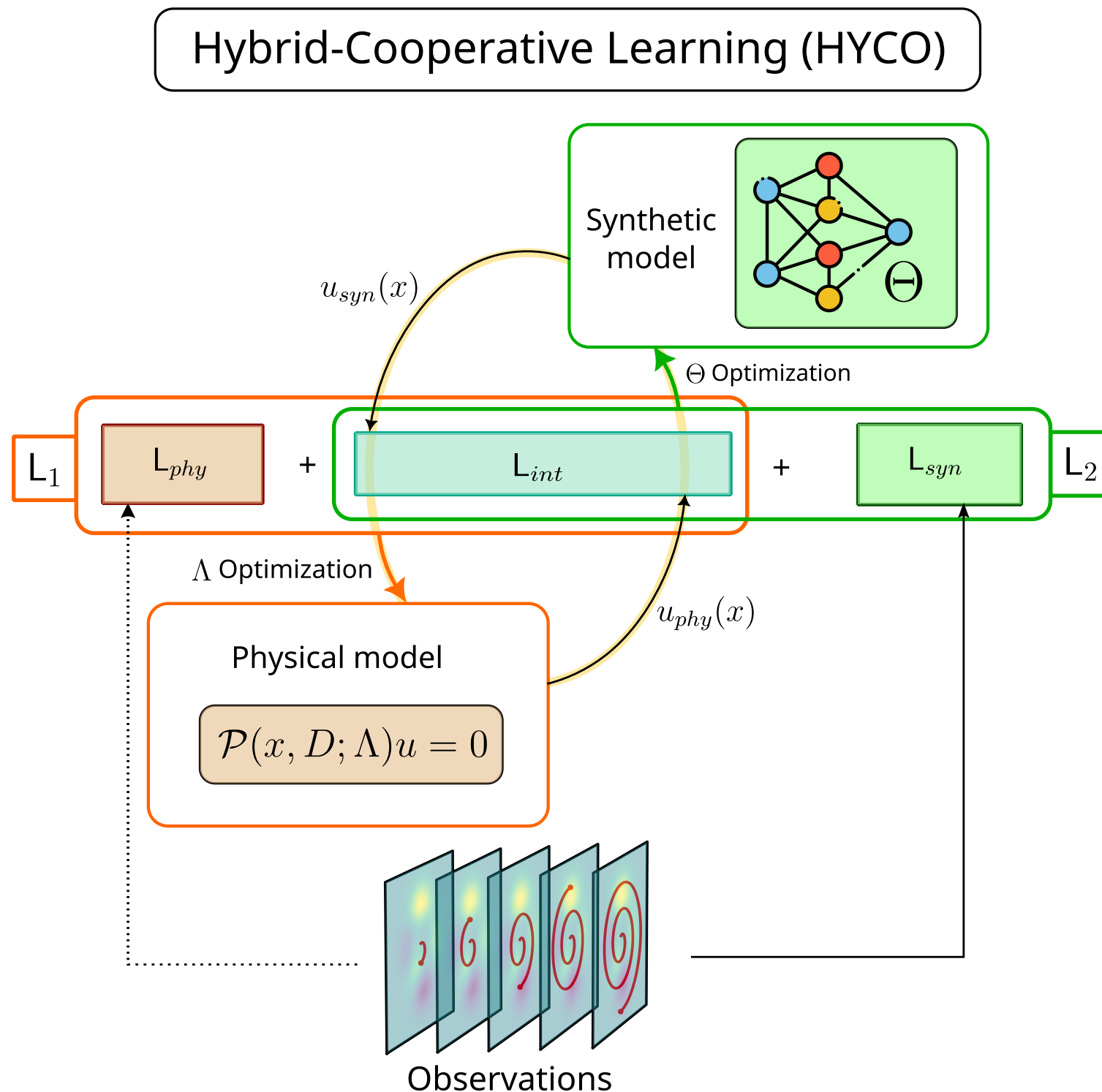
$$v_\epsilon \rightarrow v^* \in H^1$$

with

$$v^*(x) = \partial_x G(x) = -\frac{x}{2} G(x) \notin \mathcal{M}_2.$$

# HYCO: A Hybrid-Cooperative Strategy for Data-Driven PDE Model Learning

joint work with Lorenzo Liverani and Thys Steynberg





Generate

# Li–Yau estimate

Joint work with Kang Liu, Dijon

If  $u \geq 0$  solves the  $N$ -dimensional heat equation,

$$u_t - \Delta u = 0 \quad \text{in } \mathbb{R}^N \times [0, \infty)$$

then

$$\Delta \log(u) \geq -\frac{N}{2t}.$$

- The inequality is sharp and is saturated by the Gaussian heat kernel  $G$ .
- The bound deteriorates as  $N \rightarrow \infty$ , an expression of the intrinsic difficulties of understanding heat diffusion in infinite-dimensional settings.
- It guarantees that  $U = \log(u)$  (which solves the viscous Hamilton-Jacobi equation) is semiconcave, i.e.,

$$\Delta U \geq -\frac{N}{2t}$$

Given a dataset  $\{x_i\}_{i=1}^I$  sampled from an unknown distribution, we define the corresponding empirical measure

$$u_0(x) = \frac{1}{I} \sum_{i=1}^I \delta(x - x_i).$$

This leads to the solution

$$u(x, t) = \frac{1}{I} \sum_{i=1}^I G(x - x_i, t),$$

which diffuses the information of the initial empirical measure through  $\mathbb{R}^N \times [0, \infty)$ .

Generative diffusion models aim to reverse the diffusion process, thereby generating new samples from the same distribution.

However, this backward process is severely ill-posed, which paradoxically underpins their strong generative capacity.

The backward heat equation can be rewritten as

$$u_t + \Delta u - 2 \operatorname{div} \left( u \frac{\nabla u}{u} \right) = u_t + \Delta u - 2 \operatorname{div} (u \nabla \log(u)) = 0$$

With the **score function**

$$s(x, t) = \nabla \log(u),$$

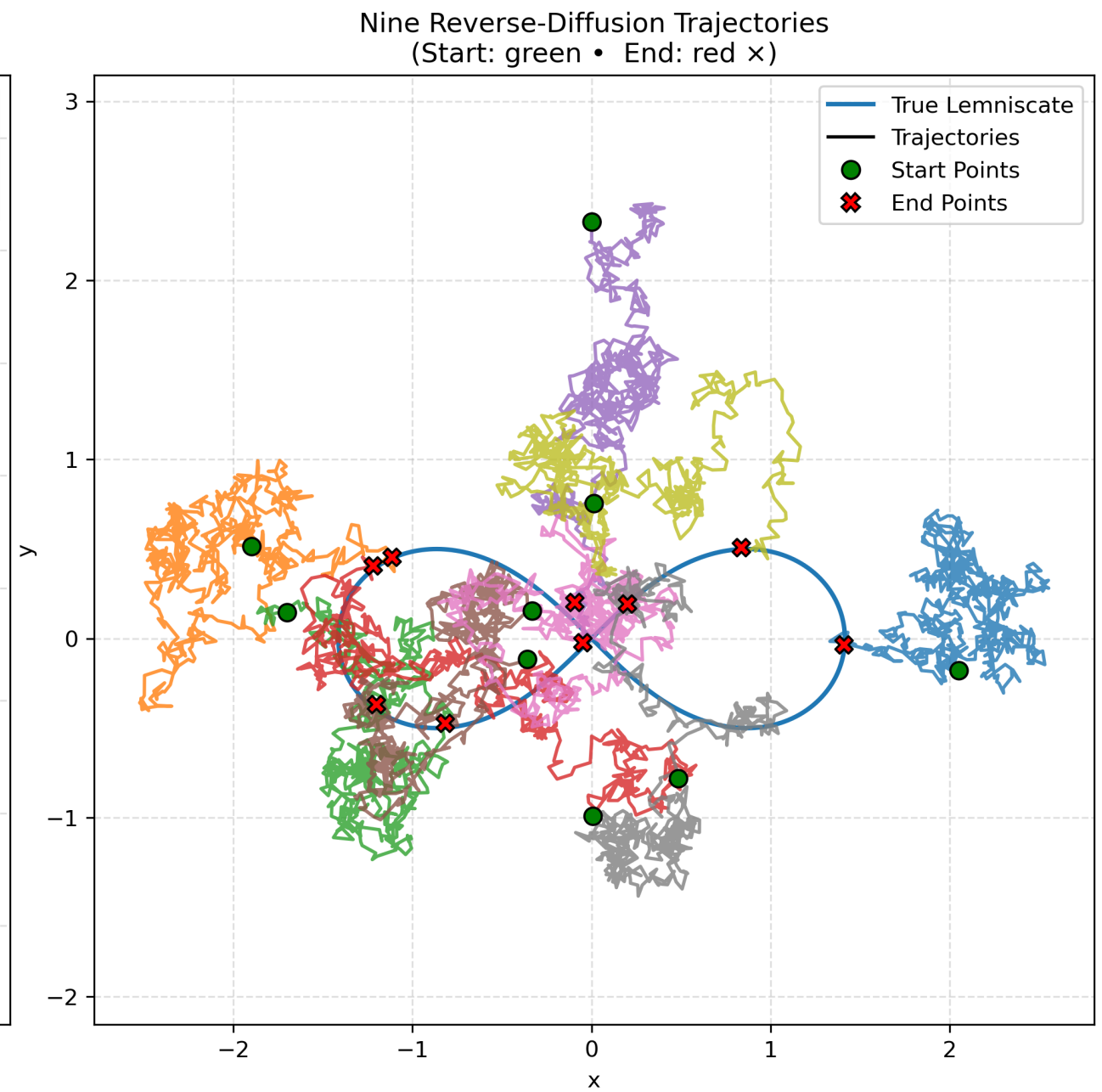
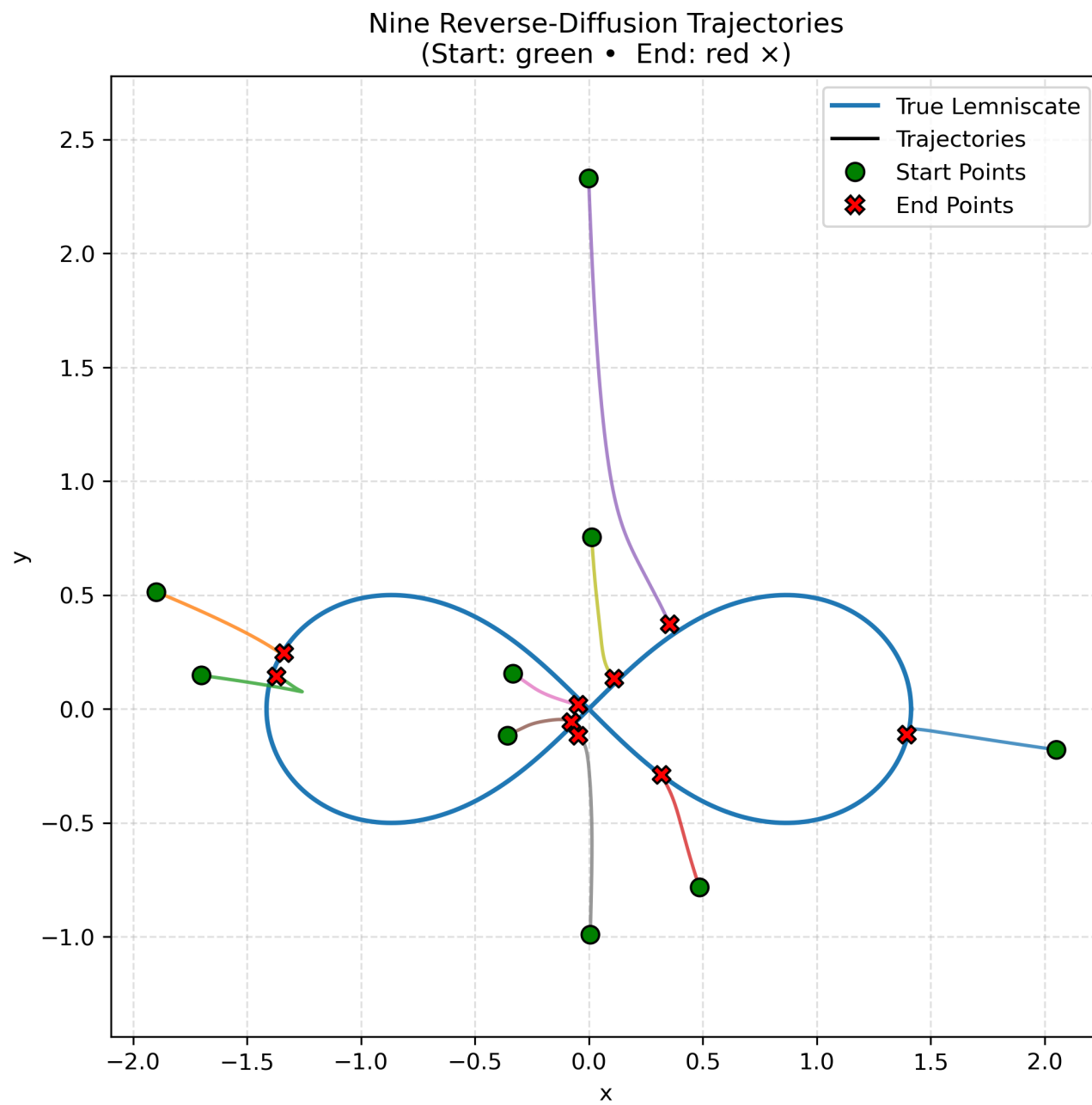
it takes the form of a convection-diffusion or Fokker–Planck model:

$$u_t + \Delta u - 2 \operatorname{div} (s(x, t) u) = 0$$

which, unlike the original backward heat equation, is well-posed backward in time thanks to the Li–Yau inequality:

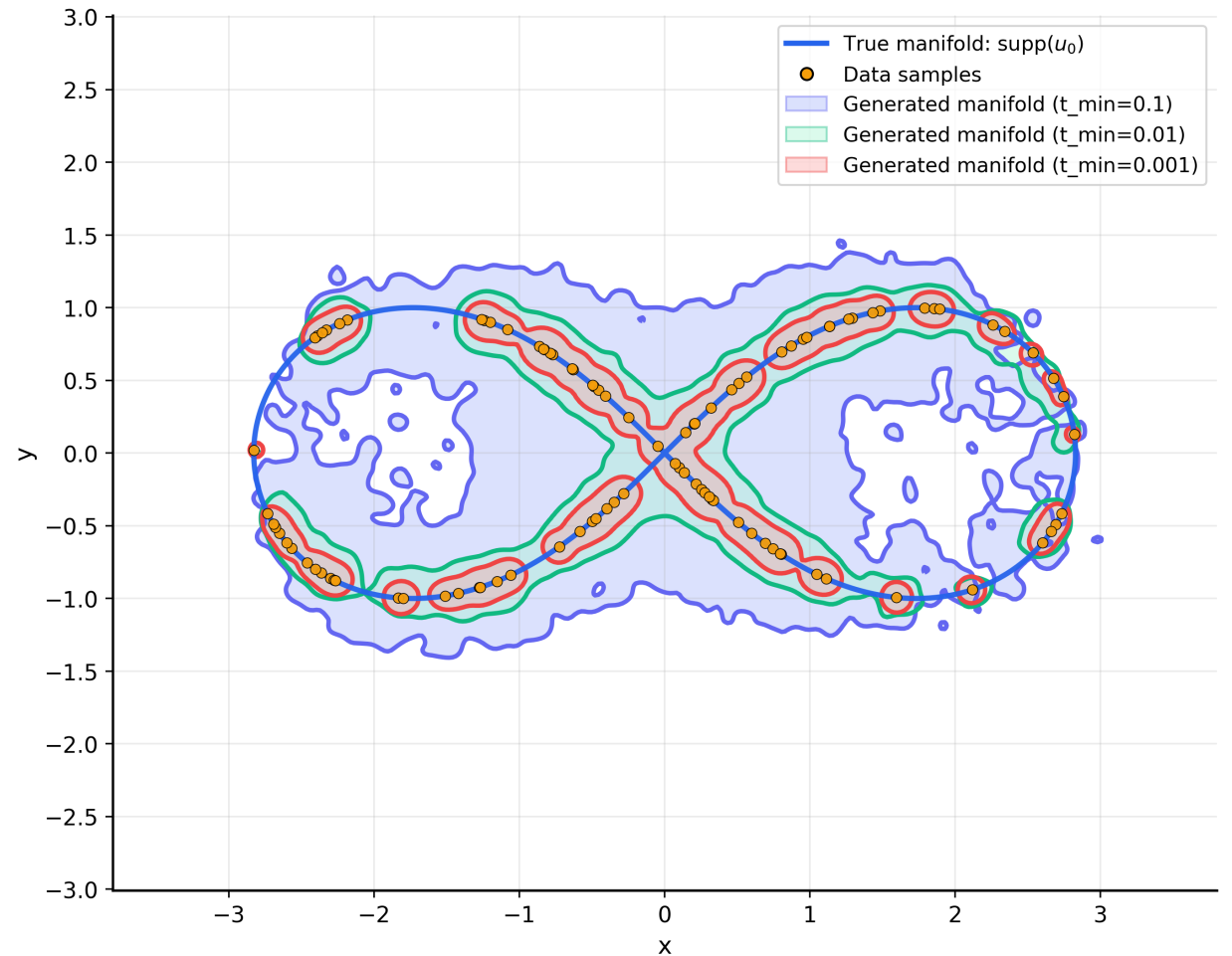
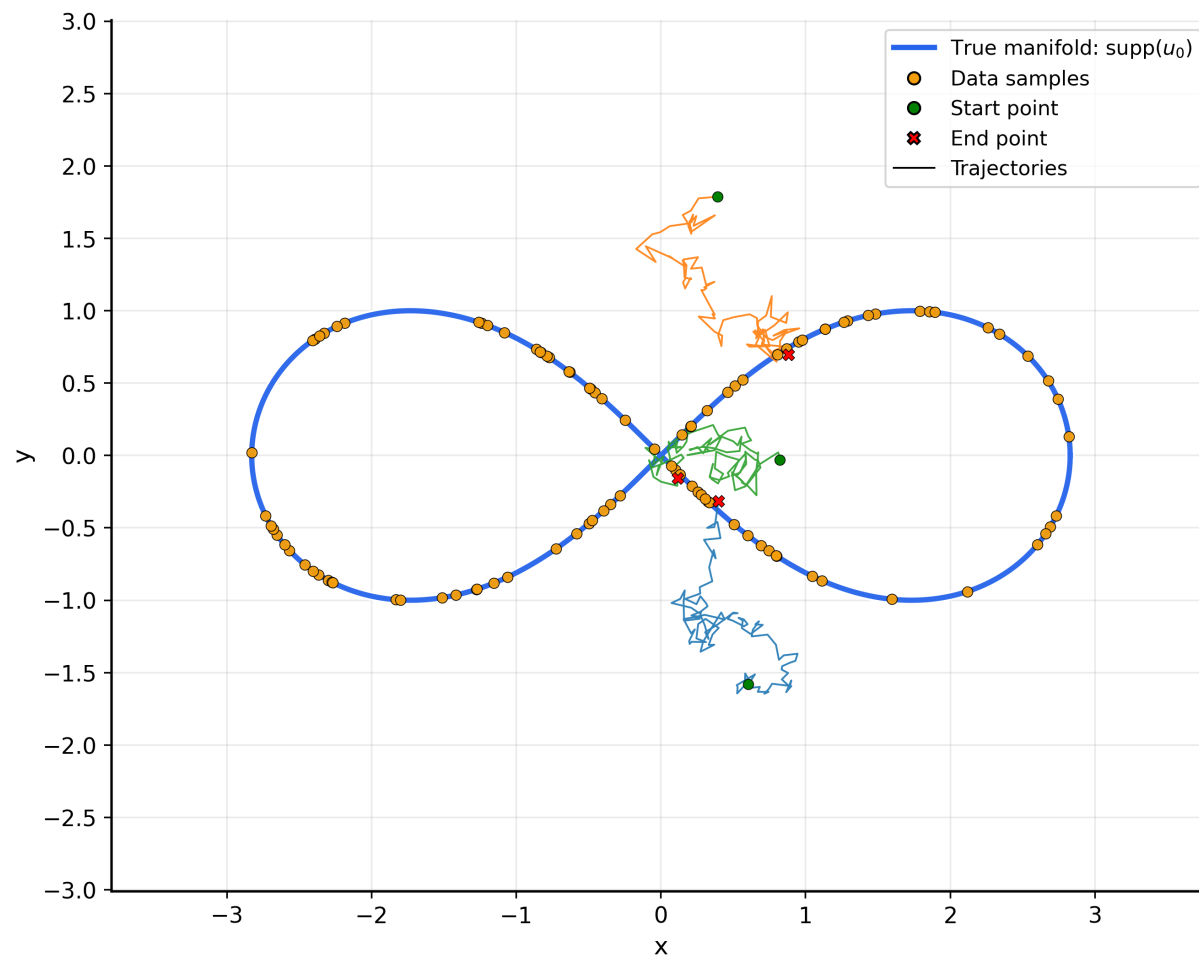
$$\operatorname{div} s(x, t) \geq -\frac{N}{2t}.$$

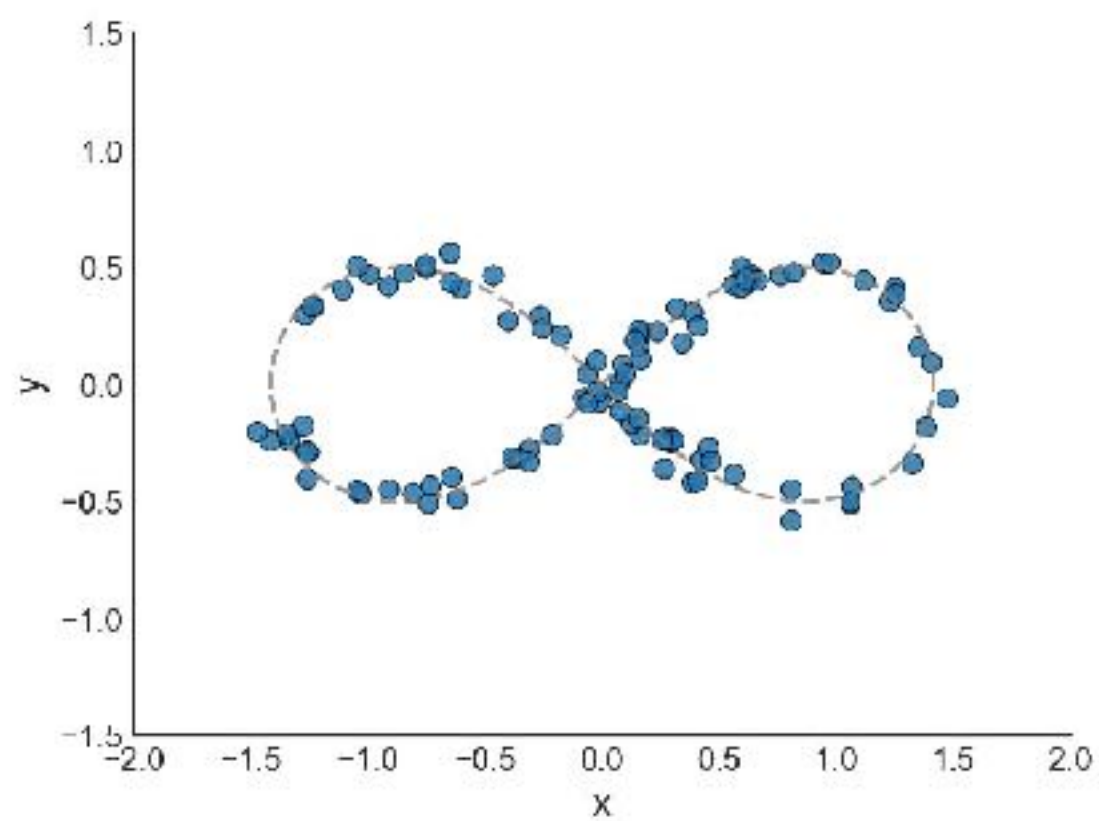
# Deterministic versus stochastic score dynamics



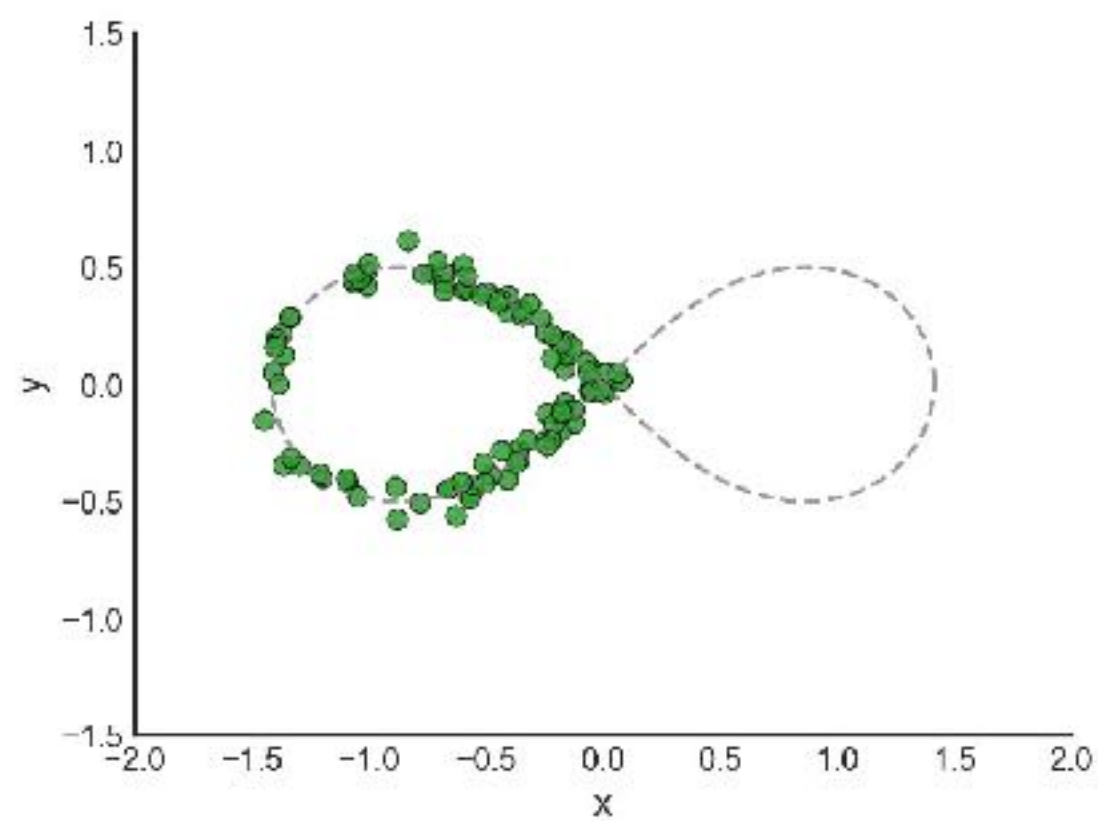


# Stopping time regulation

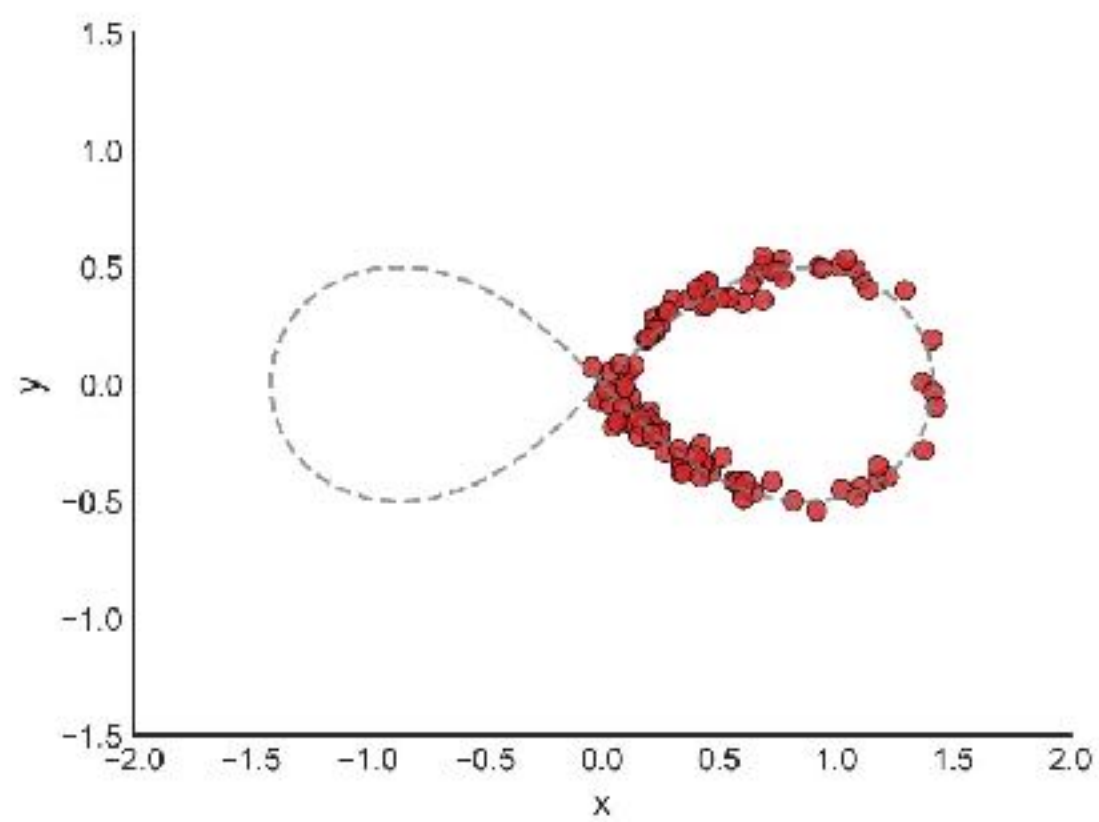




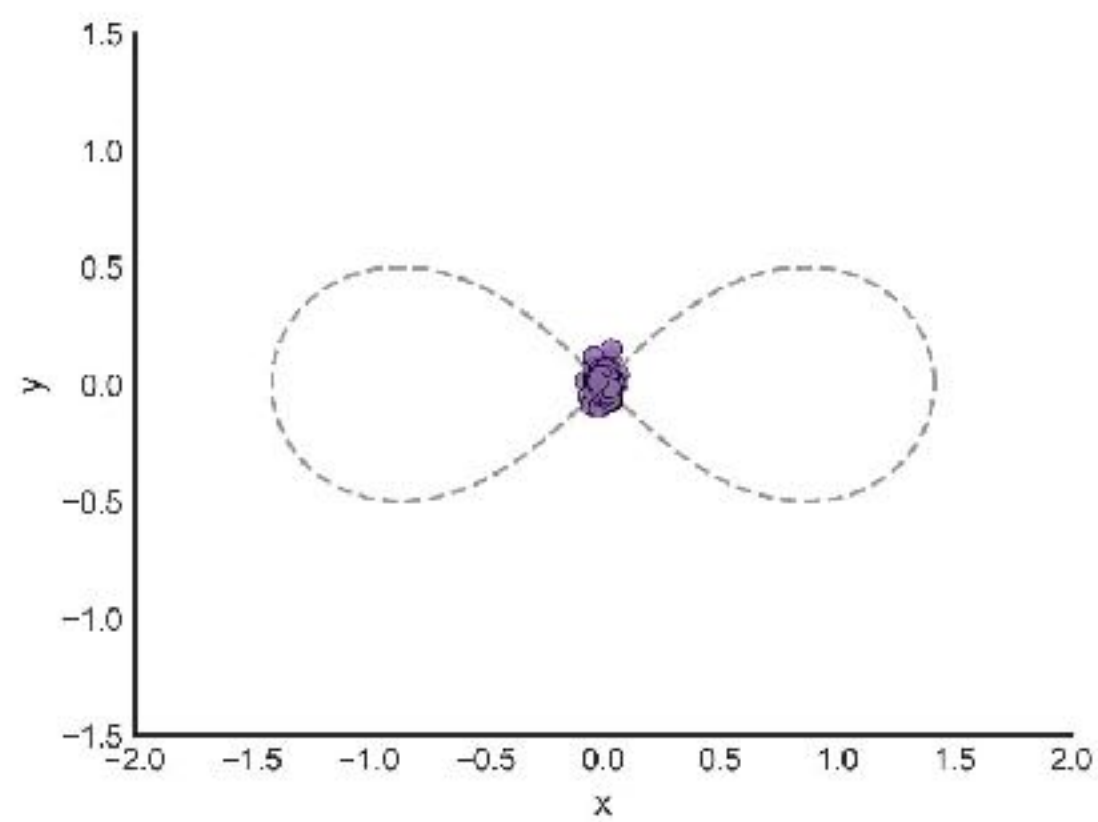
**(a) True score**



**(b) Left-only score**



**(c) Right-only score**



**(d) Left score + Right score**



# Welcome to the Future

