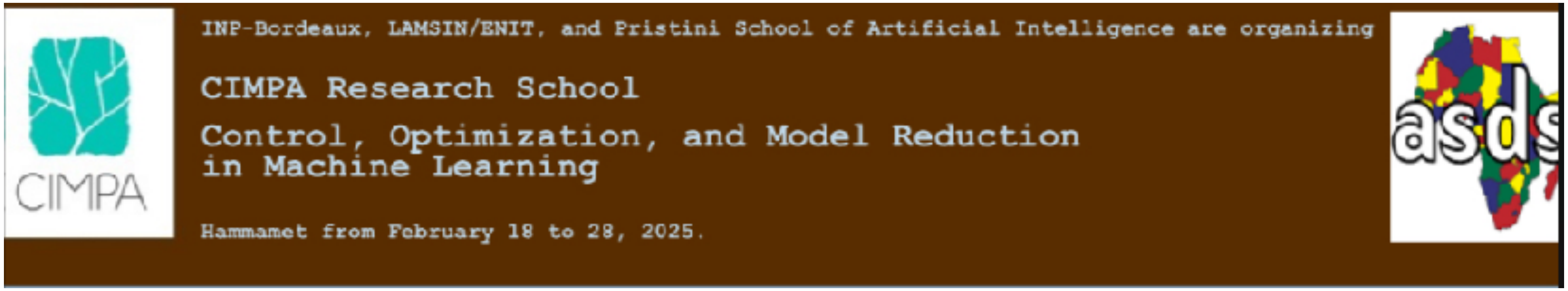
Control and Machine Learning

Enrique Zuazua

FAU & AvH, Erlangen, Germany



INP-Bordeaux, LAMSIN/ENIT, and Pristini School of Artificial Intelligence are organizing

CIMPA Research School

Control, Optimization, and Model Reduction in Machine Learning

Hammamet from February 18 to 28, 2025.

Outline

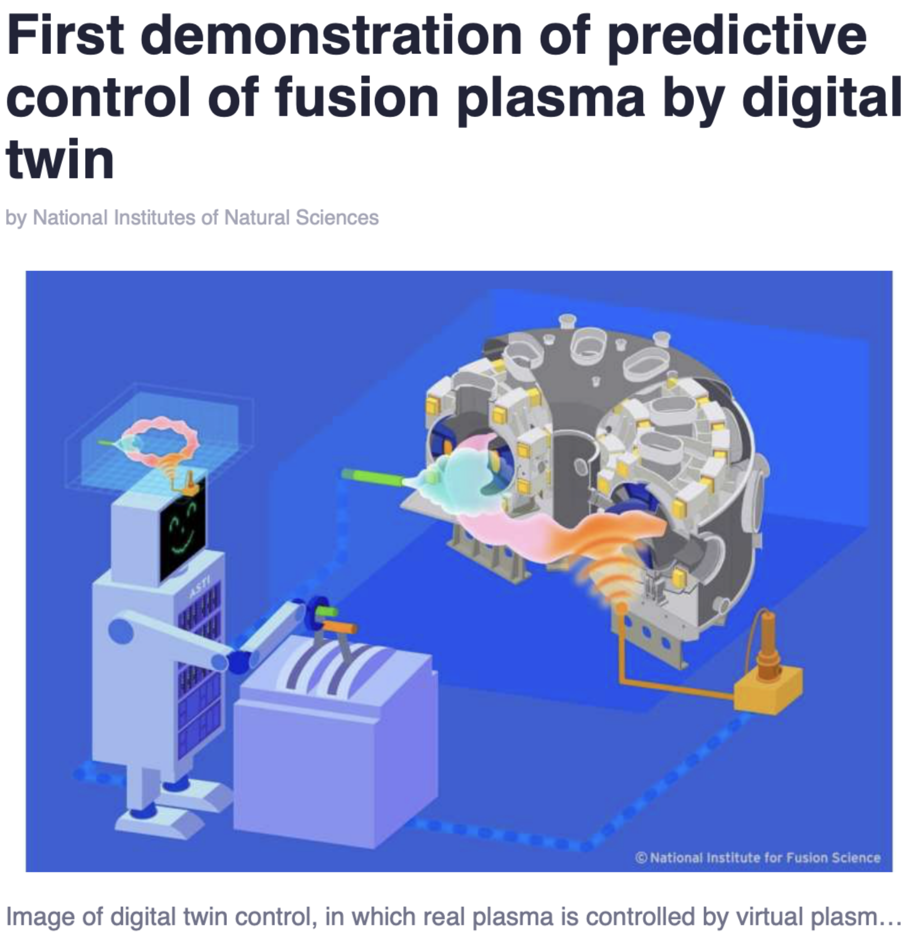
1. Context: Applied Mathematics + Machine Learning

2. Why does it work?

3. NN for PDE approximation

4. PDE+D

Nowadays AI: small and big



First demonstration of predictive control of fusion plasma by digital twin

by National Institutes of Natural Sciences

© National Institute for Fusion Science

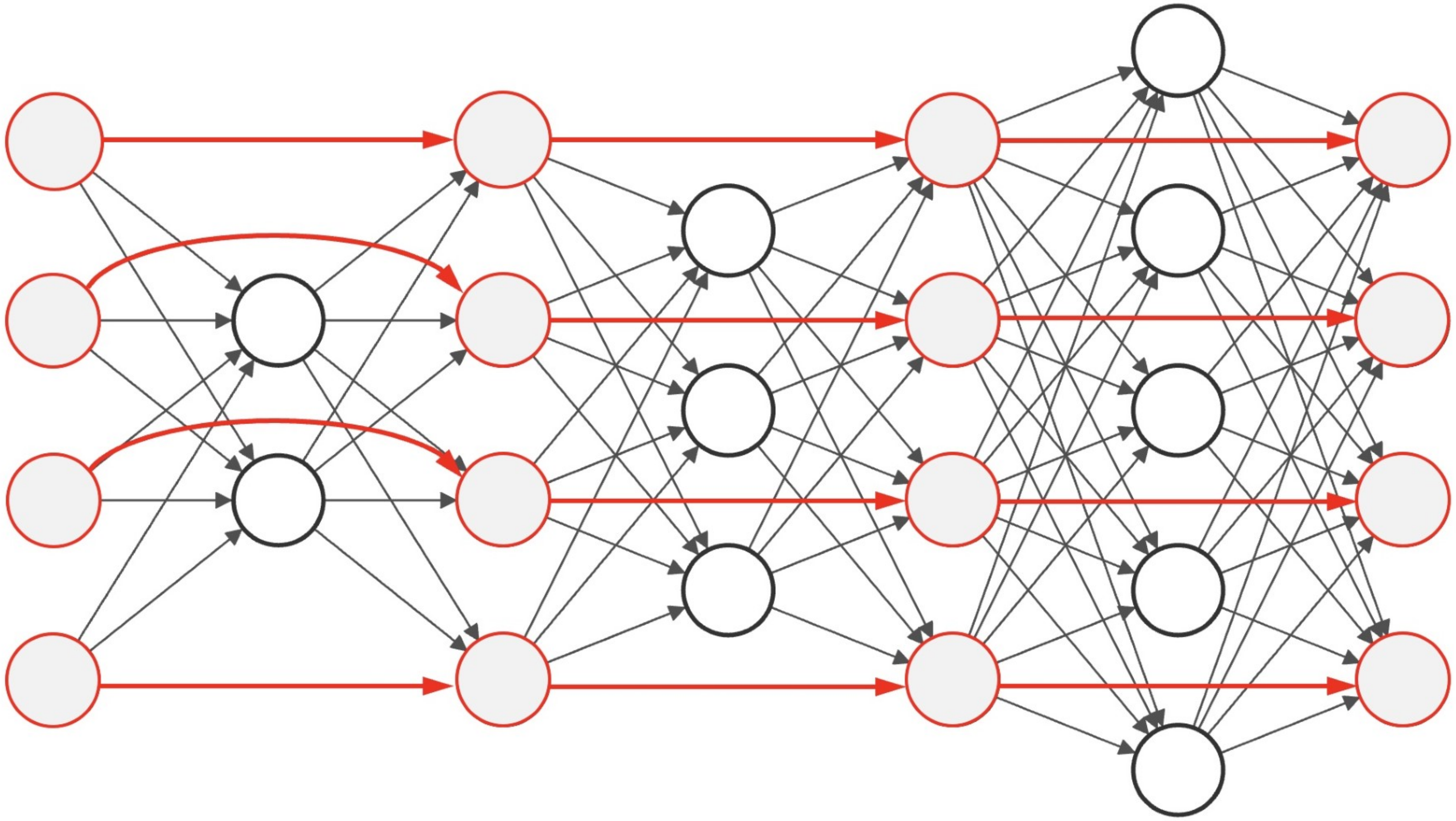
Image of digital twin control, in which real plasma is controlled by virtual plasm…

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.

How does it work? Computational practice

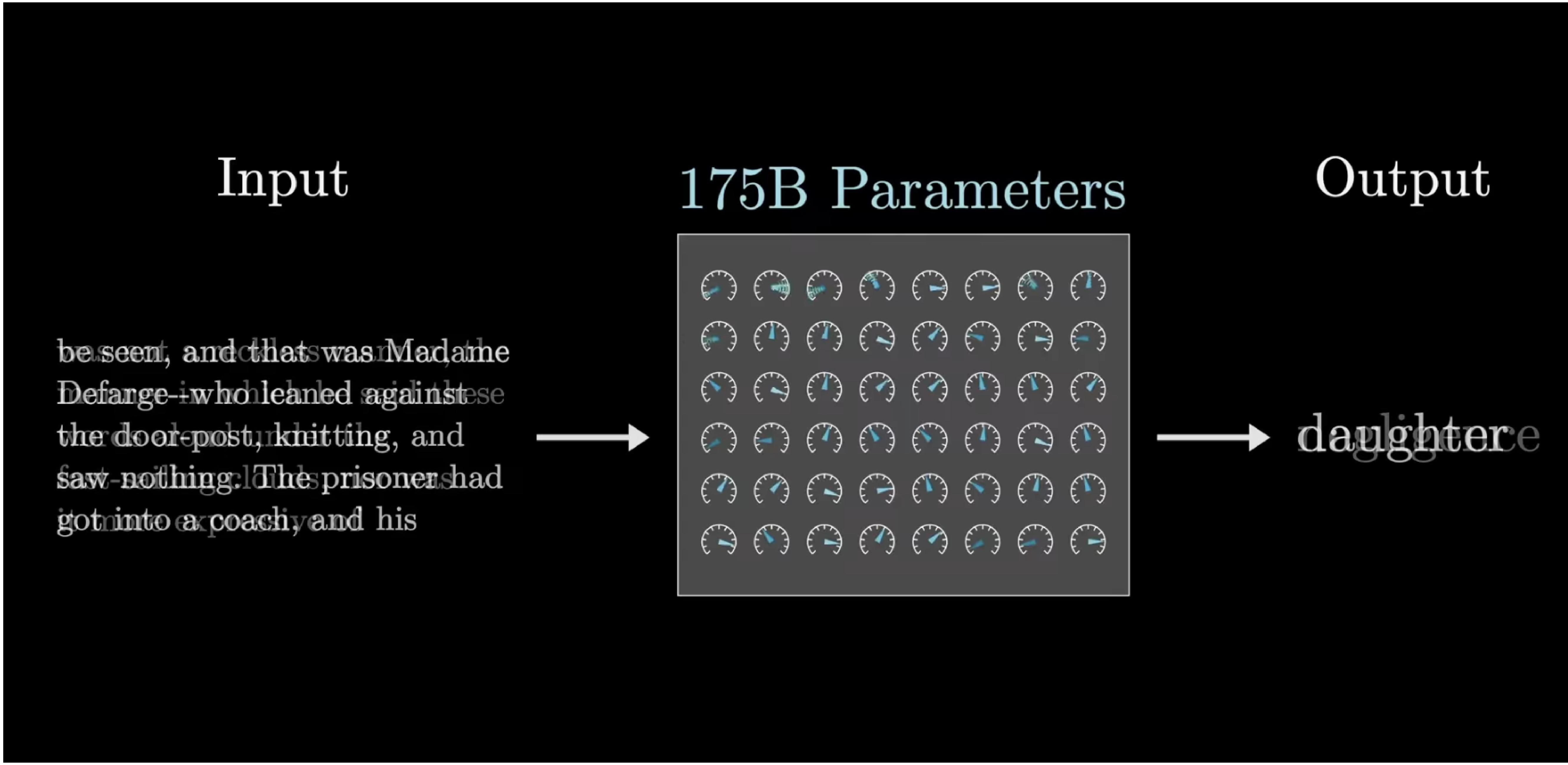
$$\underbrace{\frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(x_K^i, \ell^i\right)}_{\text{empirical risk } := E(x(\cdot))} + \alpha \sum_{j=1}^{K} \|(\mathbf{a_j}, \mathbf{w_j}, b_j)\|^2$$



Supervised Learning

Complexity: Curse of dimensionality
+ Devil of non-convexity

3Blue1Brown


Input
175B Parameters
Output
be seen, and that was Madame Defarge—who leaned against the door-post, knitting, and saw nothing. The prisoner had got into a coach, and his
daughter

Some relevant questions

- **Why does it work?**

  Can traditional applied mathematics contribute to explain the theoretical foundations of this success?

- **Use NN for PDE approximation**

  Replace the classical linear ansätze (finite differences, spectral, FEM) by a NN nonlinear one.

  (Devil of non-convexity!)

- **What can Applied Maths learn from these new tools?**

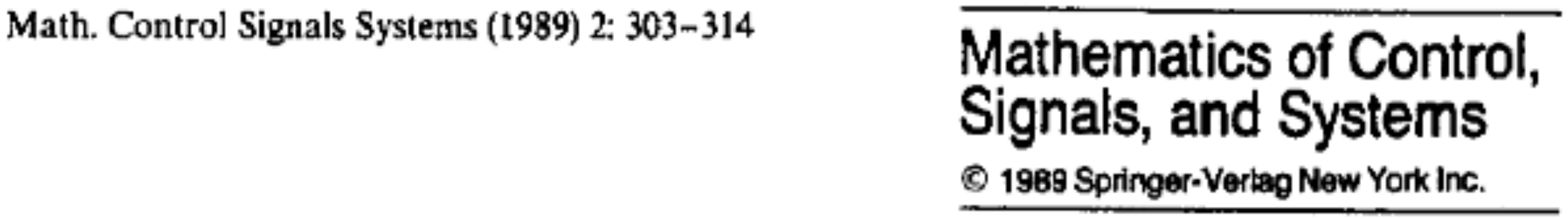
  Merging: PDE+D(ata)

  "Digital Twins: Where Data, Mathematics, Models and Decisions Collide"

Outline

1. Context: Applied Mathematics + Machine Learning
2. Why does it work?
3. NN for PDE approximation
4. PDE+D

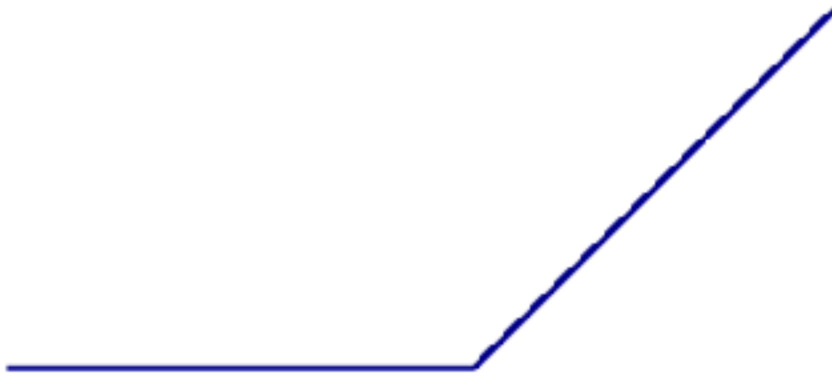
Why does it work? Universal Approximation

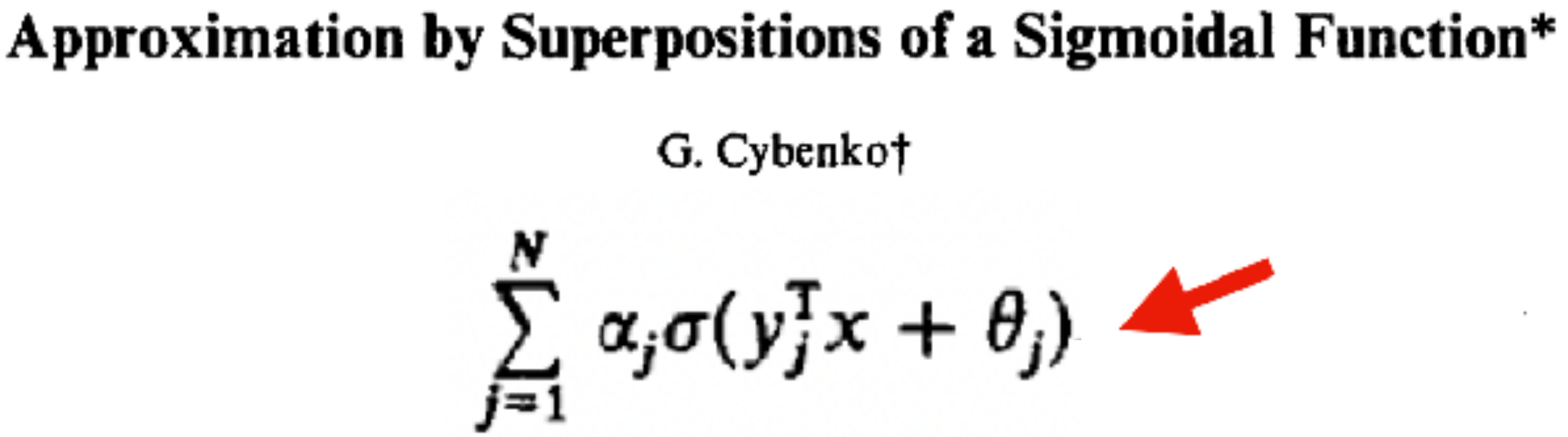


Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

© 1989 Springer-Verlag New York Inc.

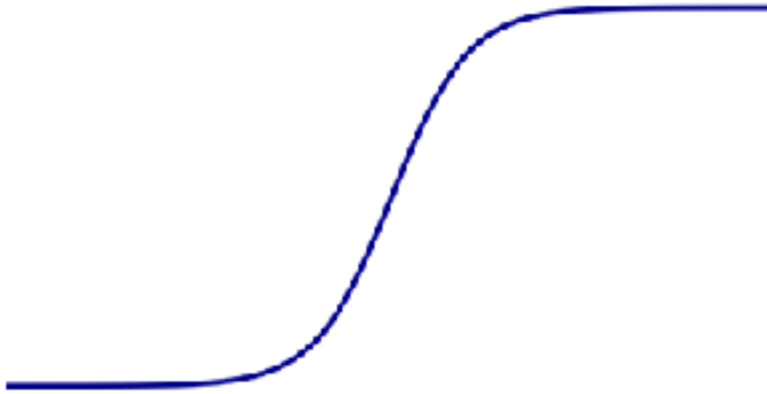


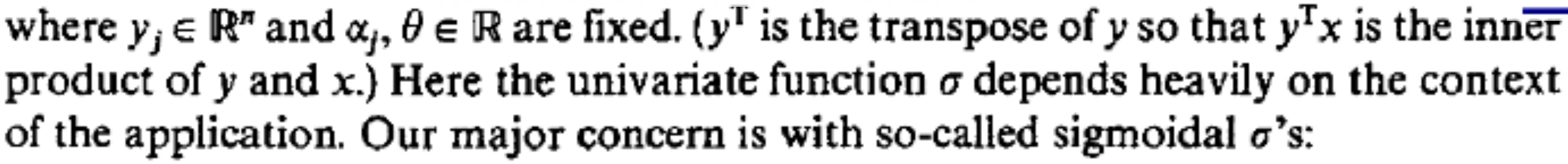


Approximation by Superpositions of a Sigmoidal Function*

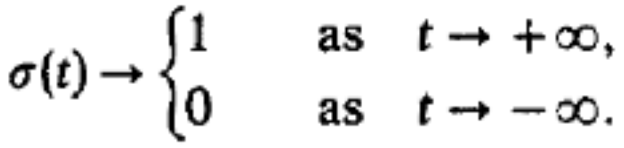
G. Cybenko†

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j)$$

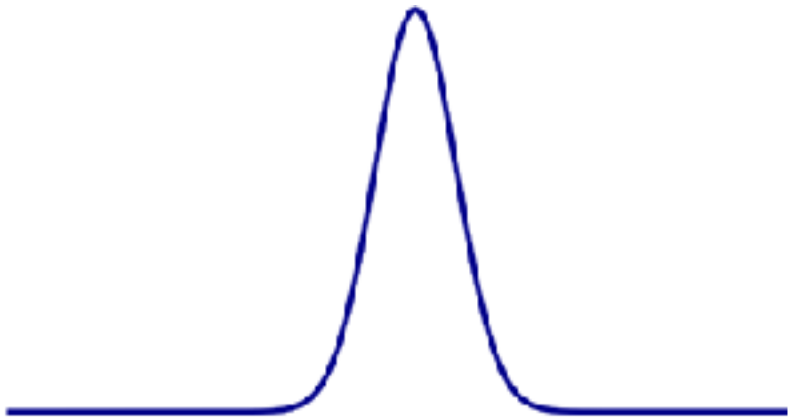


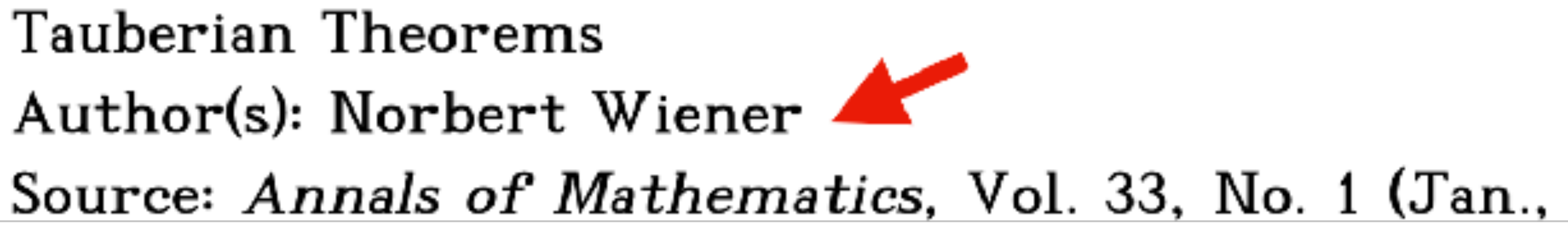


where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^{T} is the transpose of y so that $y^{\mathrm{T}}x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:



$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty. \end{cases}$$



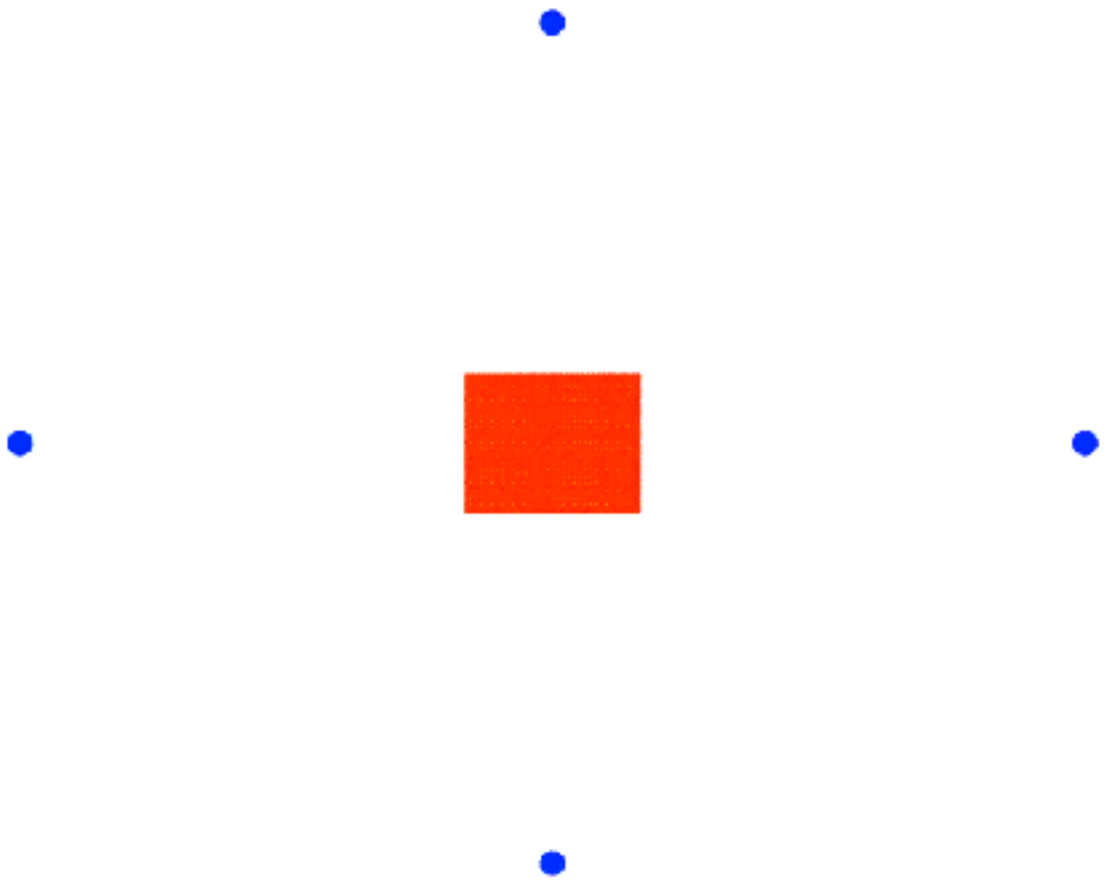


Tauberian Theorems

Author(s): Norbert Wiener

Source: *Annals of Mathematics*, Vol. 33, No. 1 (Jan., 1932), pp. 1-100

Why does it work so well? Control $\leftrightarrow$ ML



Control: Dogs-Sheep

Supervised Learning

Cybernetics, Norbert Wiener, 1948

The science of control and communication in animals and machines

The linear finite d-dimensional system

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

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

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

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.

NN Modelling

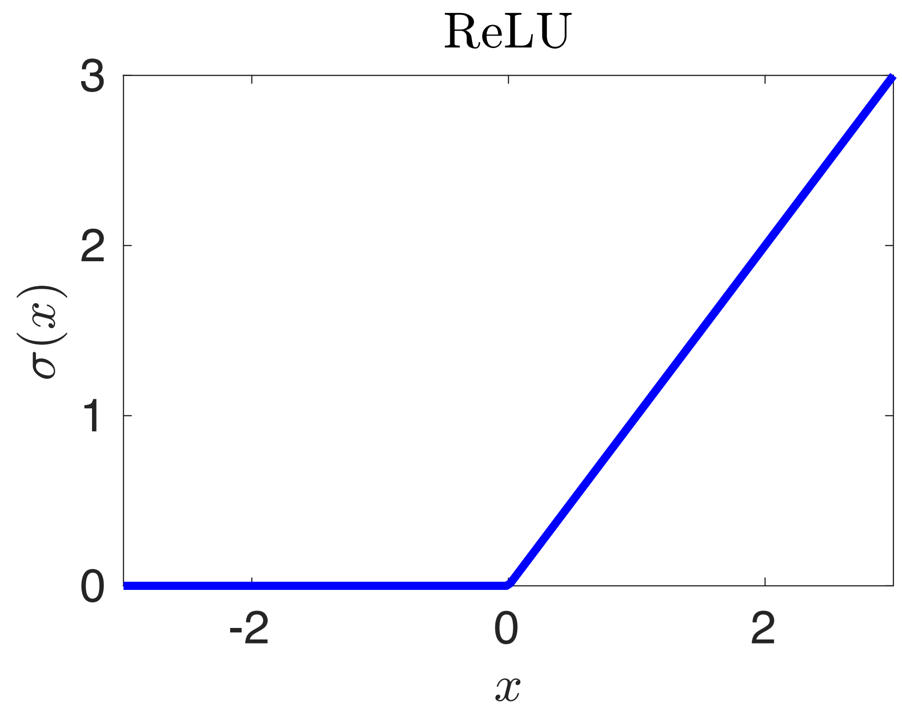
$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t)+b(t))$$

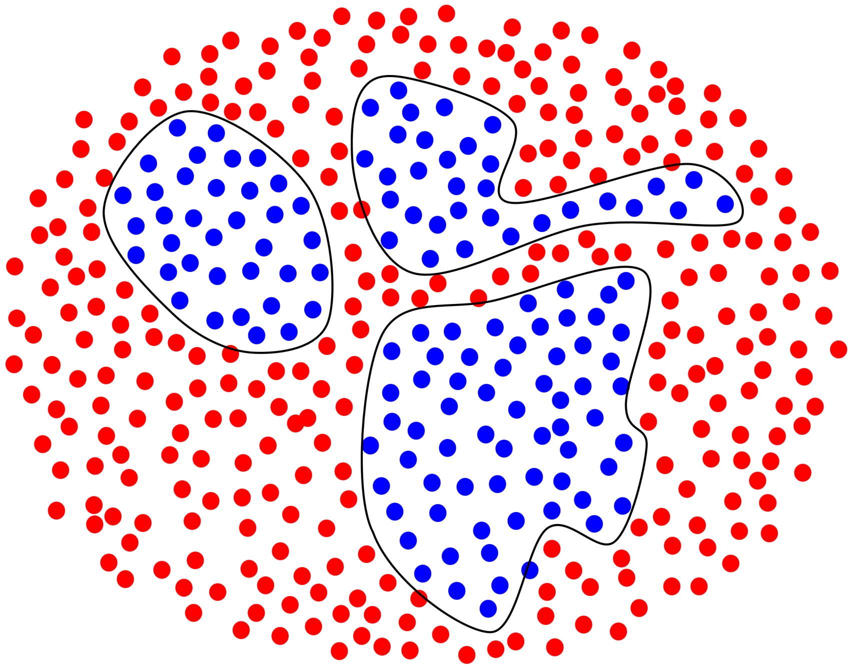
$$\Updownarrow$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h\,\mathbf{w}^k\sigma(\mathbf{a}^k\cdot\mathbf{x}^k+b^k)$$

$$\Updownarrow$$

$$f(x) \sim \sum_{j=1}^{K}\mathbf{w}_j\sigma(\mathbf{a}_j\cdot x+b_j)$$





Supervised learning by control

Goal: Find an approximation of a function $f_\rho : \mathbb{R}^d \to \mathbb{R}^n$ from a dataset

$$\{\vec{x}_i, \vec{y}_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^n$$

drawn from an unknown probability measure ρ on $\mathbb{R}^d \times \mathbb{R}^n$.

Classification: match points (images) to respective labels (cat, dog).

This is typically done by **training a neural network**. We will do it through the **simultaneous or ensemble control of Neural ODEs.**

ResNets / Neural ODEs in action (Borjan Geshkovski)

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

[1] K. He, X Zhang, S. Ren, J Sun, 2016: Deep residual learning for image recognition

[2] E. Weinan, 2017. A proposal on machine learning via dynamical systems.

[3] R. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, 2018.

[4] E. Sontag, H. Sussmann, 1997.

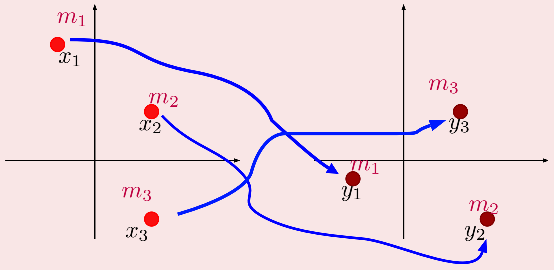
Classification by simultaneous or ensemble control of Neural ODEs

Theorem (Classification, Domènec Ruiz-Balet & EZ, SIREV, 2023)

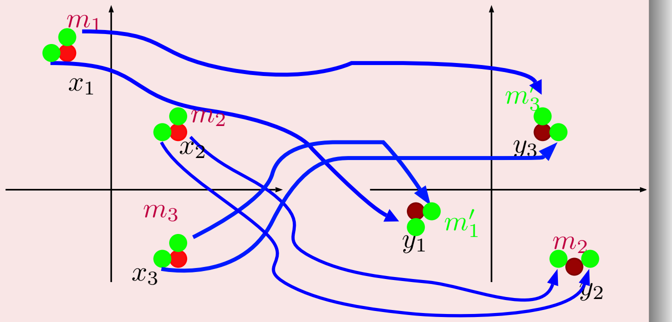
In dimension $d \geq 2$, in any time horizon $[0, T]$, a finite number of arbitrary items can be driven to pre-assigned open subsets of the Euclidean space, corresponding to its labels, by piece-wise constant controls.

Generative Neural Transport

Neural ODEs $\quad \dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t)+b(t)), \quad$ interpreted as the characteristics of the transport equation:



$$\partial_t \rho + \mathrm{div}_x\Big[\underbrace{\big(\mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot x + b(t)\big)}_{V(x,t)}\rho\Big] = 0$$



allow transporting atomic measures and constitute a tool for generative transport.

2

[2] Related results for smooth sigmoids using Lie brackets: A. Agrachev and A. Sarychev, arXiv:2008.12702, (2020); Li, Q., Lin, T., & Shen, Z. (2022), JEMS.

What is the ResNet doing? Basic control actions

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

Control functions $(\mathbf{w}, \mathbf{a}, \mathbf{b}) \longrightarrow$ Piecewise constant.
Each time discontinuity $\sim$ change of layer.

- $\mathbf{a}(t), b(t)$ define a hyperplane $H(\mathbf{x}) = \mathbf{a}(t)\cdot\mathbf{x}(t) + b(t) = 0$ in $\mathbb{R}^d$.
- $\sigma(z) = \max\{z, 0\}$ "activates" the halfspace $H(\mathbf{x}) > 0$ and "freezes" $H(\mathbf{x}) \leq 0$.
- $\mathbf{w}(t)$ determines the direction of the field in the active halfspace.

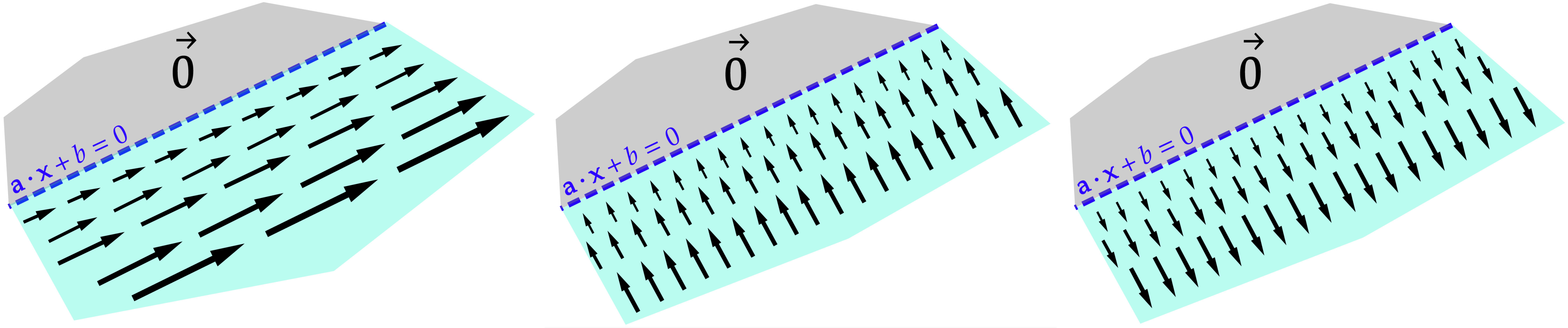
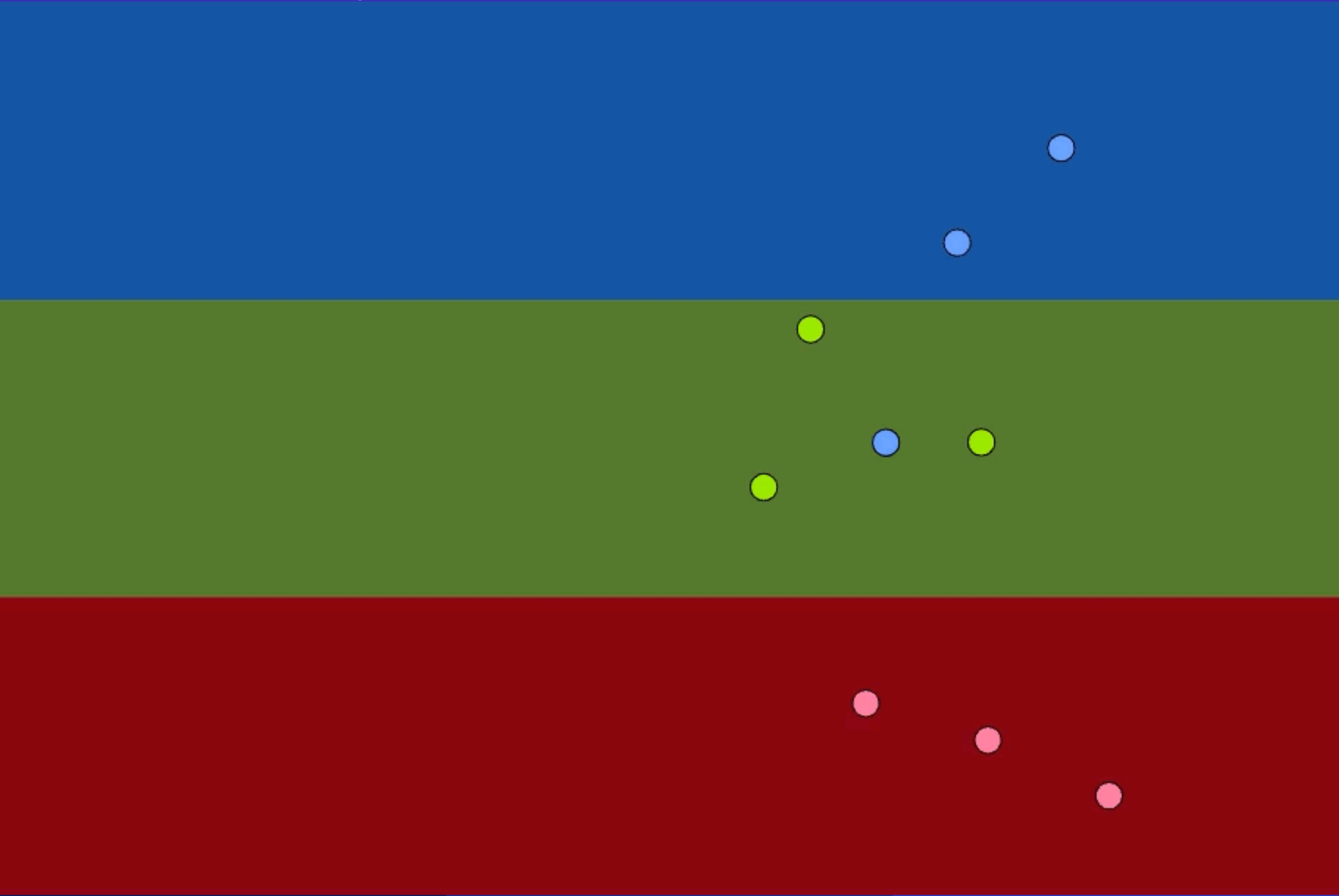

$\vec{0}$
$\mathtt{a}\cdot\mathtt{x}+b=0$
$\vec{0}$
$\mathtt{a}\cdot\mathtt{x}+b=0$
$\vec{0}$
$\mathtt{a}\cdot\mathtt{x}+b=0$

Figure: Parallel (left); Contraction (center); Expansion (right).

Classification by Control of ResNets: One step + Induction



Outline

1. Context: Applied Mathematics + Machine Learning

2. Why does it work?

3. NN for PDE approximation

4. PDE+D

NN version of variational PDEs

Warning: Lack of convexity!

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

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

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

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

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

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

The NN version

What can NN do?
Replace V_h by a NN finite-dimensional manifold $\mathcal{M}_{\mathcal{K}}$:

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

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

Then

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

And letting $K \to \infty$... one can develop a Γ-convergence like theory. [3]

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

[3] (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 ∞-dimensions...

Mean-field relaxation is commonly employed in shallow NNs. [4]

Shallow NN

The original Shallow NN writes:

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

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 densifies the ansatz evolves into its relaxed version.

Mean-field shallow NN

The mean-field shallow NN writes:

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

where $\mu \in \mathcal{M}(\mathbb{R}^{d+1})$.
The outcome is linear with respect to μ! This leads to the minimisation problem

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

Is it well-posed? Does the minimiser exist? Does it coincide with the weak solution of the Dirichlet problem?

[4] [Mei-Montanari-Nguyen, 2018], [Chizat-Bach, 2018], [K. Liu & E. Zuazua, (2024). Representation and regression problems in NN: Relaxation, Generalisation and Numerics.]

Outline

1. Context: Applied Mathematics + Machine Learning

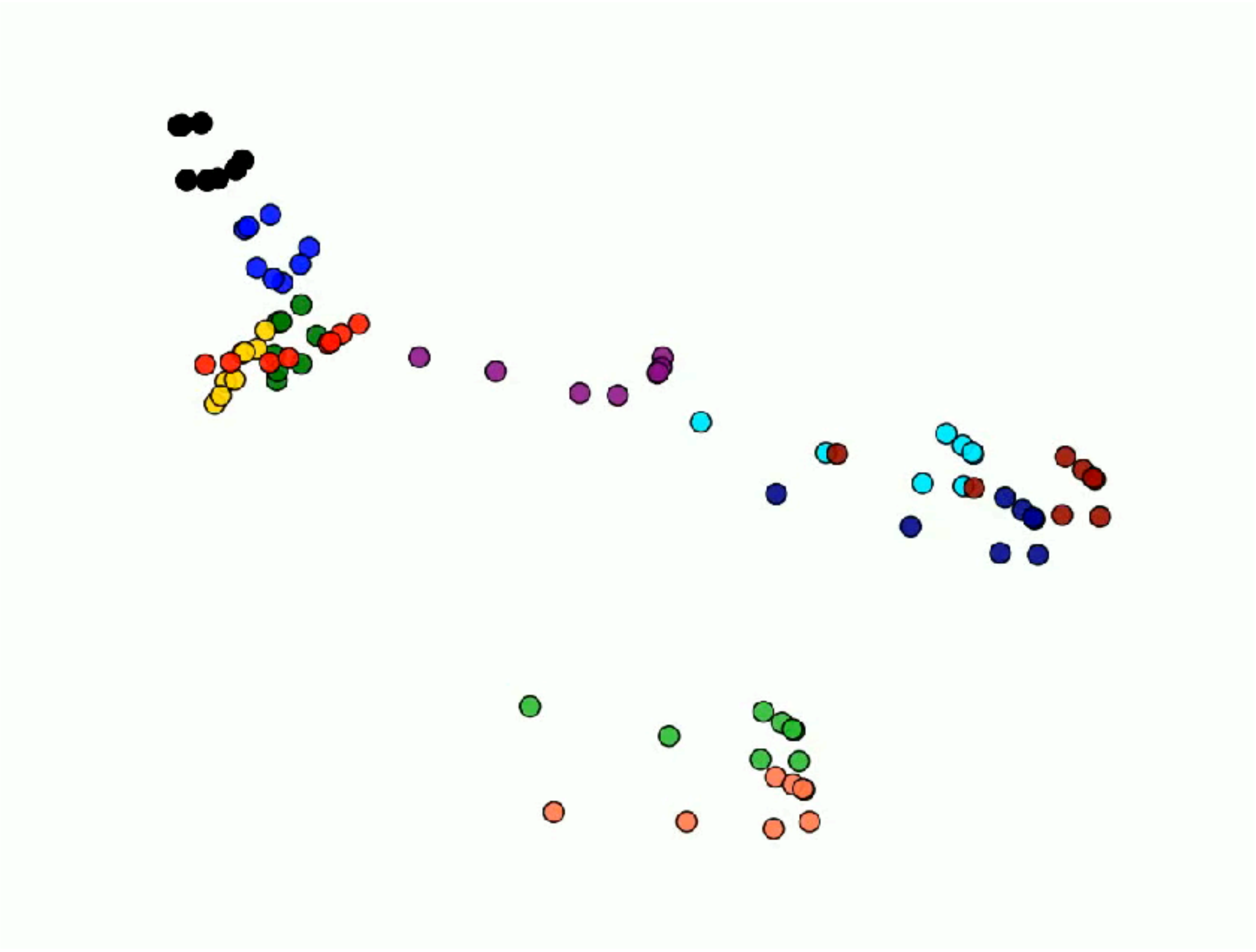
2. Why does it work?

3. NN for PDE approximation

4. PDE+D

Tracking dynamic data

Joint work with K. Liu, L. Liverani and Z. Li



Semi-autonomous NODEs

- The structure is motivated by the Universal Approximation property of ReLU activation functions (Pinkus, 1999)

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \;\rightarrow\; \mathbf{f}(\mathbf{x}, t) \sim \sum_{j=1}^{K} \mathbf{w}_j \, \sigma(\mathbf{a}_j^1 \cdot \mathbf{x} + a_j^2 t + b_j)$$

- Complexity reduction
- Anticipate future evolution of trajectories.

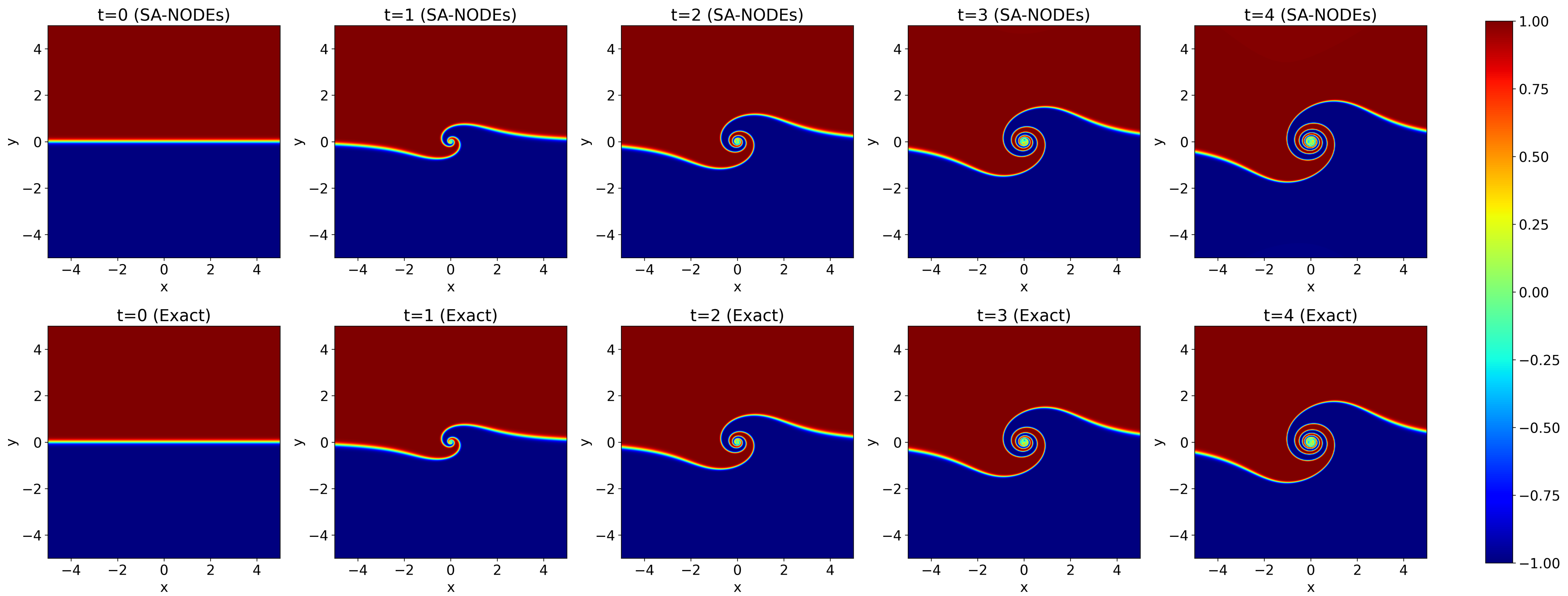
A time-independent choice of the parameters leads to a non-autonomous dynamics, but with a trivial time-dependence,

$$\dot{\mathbf{x}}(t) = \sum_{j=1}^{K} \mathbf{w}_j \, \sigma(\mathbf{a}_j^1 \cdot \mathbf{x}(t) + a_j^2 t + b_j)$$

To be complemented with Modelm Predictive Control (MPC)?

Doswell Frontogenesis

Ongoing work with Weiwei Hu (Atlanta) on optimal fluid mixing



SA-NODEs and exact solution of the transport equation modeling Doswell frontogenesis

$$\partial_t \rho(x, y, t) + \operatorname{div}\left(\rho(x, y, t)\left(-yg(r), xg(r)\right)\right) = 0,$$

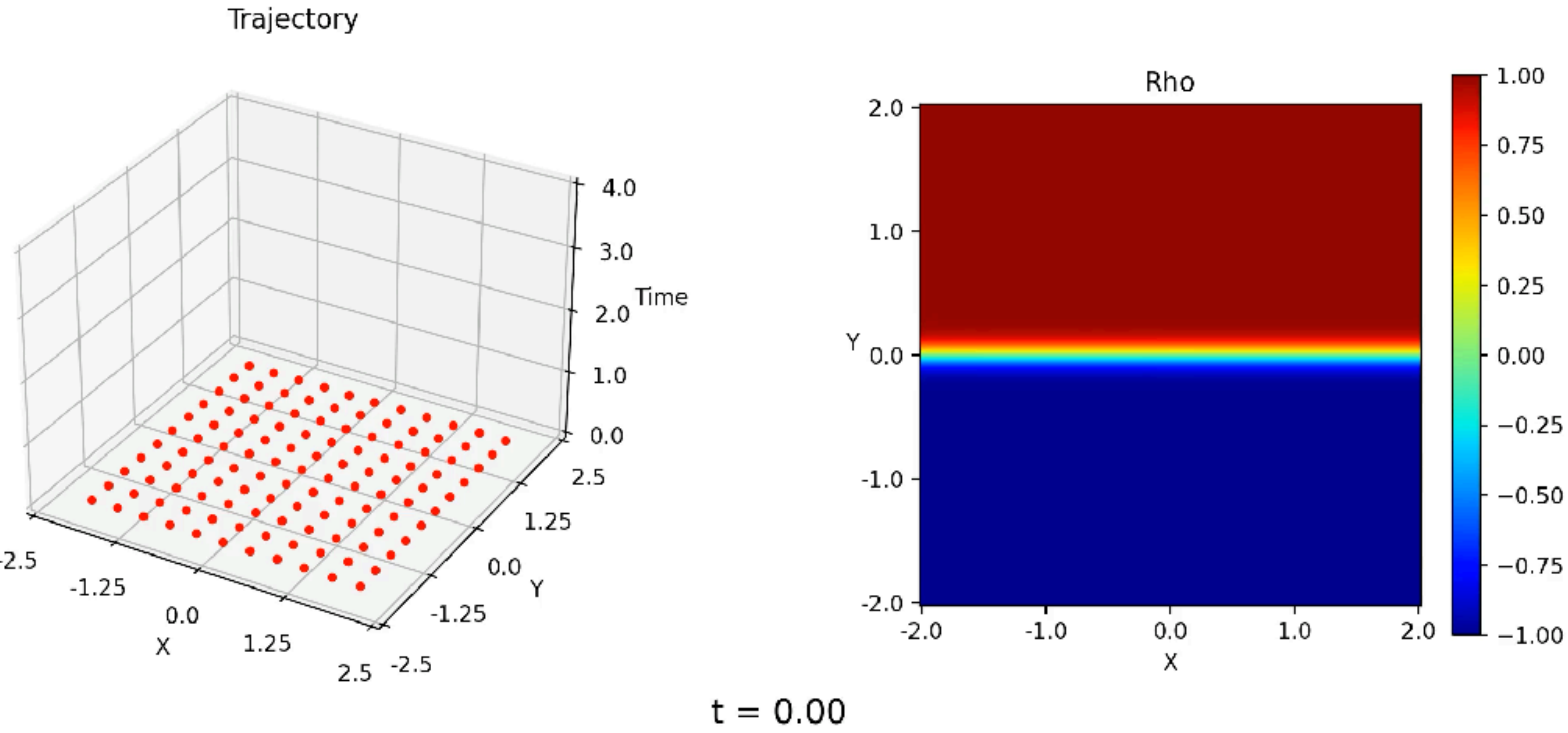
where $(x, y, t) \in \mathbb{R}^2 \times [0, T]$ and,

$$g(r) = c\, r^{-1}\, \text{sech}^2 r \tanh r, \quad \rho_0(x, y) = \tanh\left(y/\delta\right).$$

The exact solution:

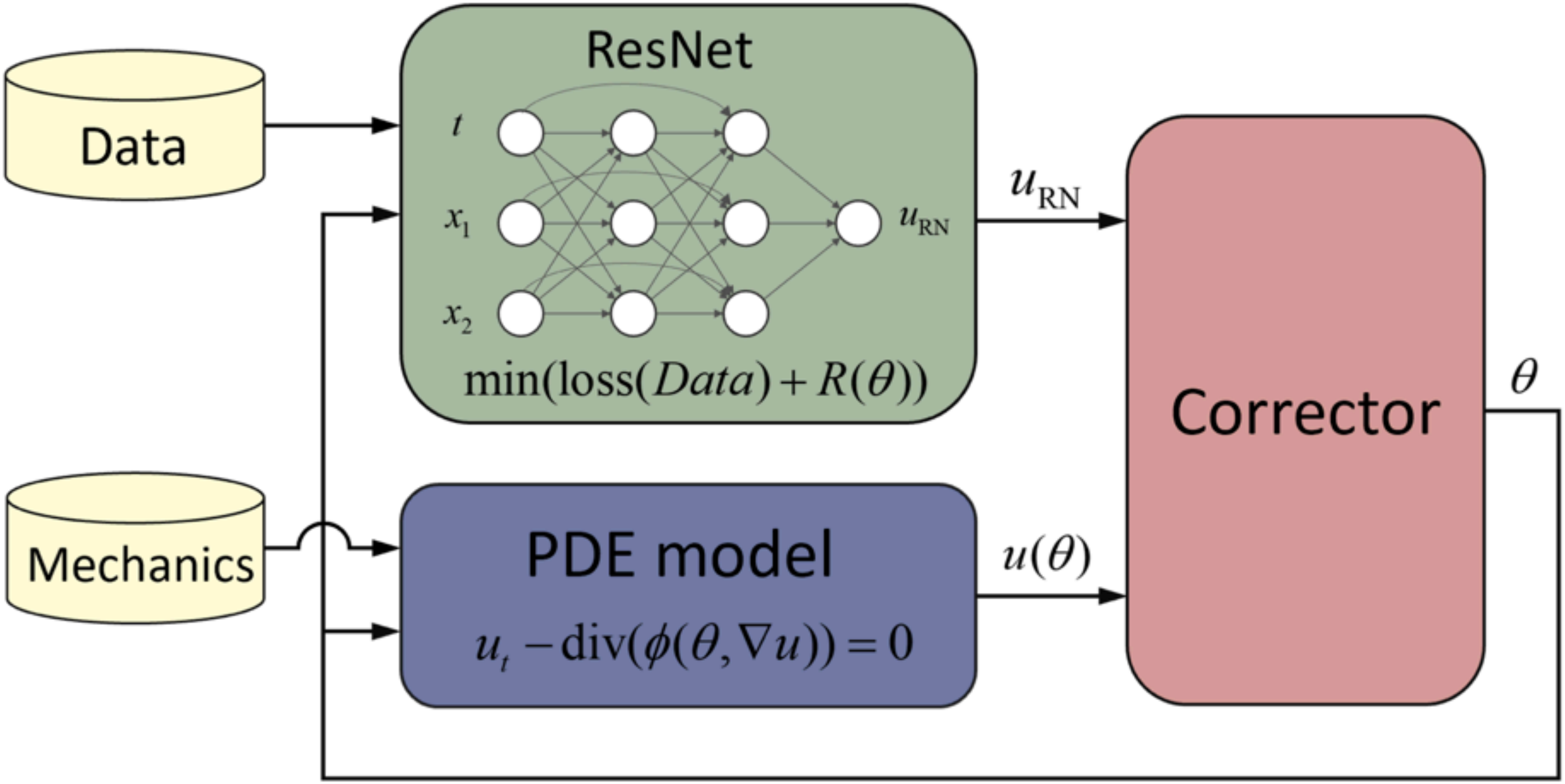
$$\rho(x, y, t) = \tanh\left(\frac{y\cos(gt) - x sin(gt)}{\delta}\right).$$

Tornado Emergence



PDE+D

Hybrid Methodology: Data Driven + PDE modelling + Collapse



Data
ResNet
t
x_1
x_2
u_{RN}
$\min(\mathrm{loss}(Data) + R(\theta))$
u_{RN}
Corrector
θ
Mechanics
PDE model
$u_t - \mathrm{div}(\phi(\theta, \nabla u)) = 0$
$u(\theta)$

Our recent contributions

E. Zuazua, *Control and Machine Learning*, SIAM News, October 2022

D. Ruiz-Balet, E. Zuazua, *Neural ODE control for classification, approximation and transport*, SIAM Review, 65 (3)3 (2023), 735-773.

B. Geshkovski, E. Zuazua, *Turnpike in optimal control of PDEs, ResNets, and beyond*, Acta Numer., 31 (2022), 135–263

D. Ruiz-Balet, E. Zuazua, *Control of neural transport for normalizing flows, Journal de mathématiques pures et appliquées*, 181 (2024), 58-90.

...And more to appear

Z. Wang, Y. Song, E. Zuazua, *Approximate and Weighted Data Reconstruction Attack in Federated Learning*, arXiv:2308.06822 (2023)

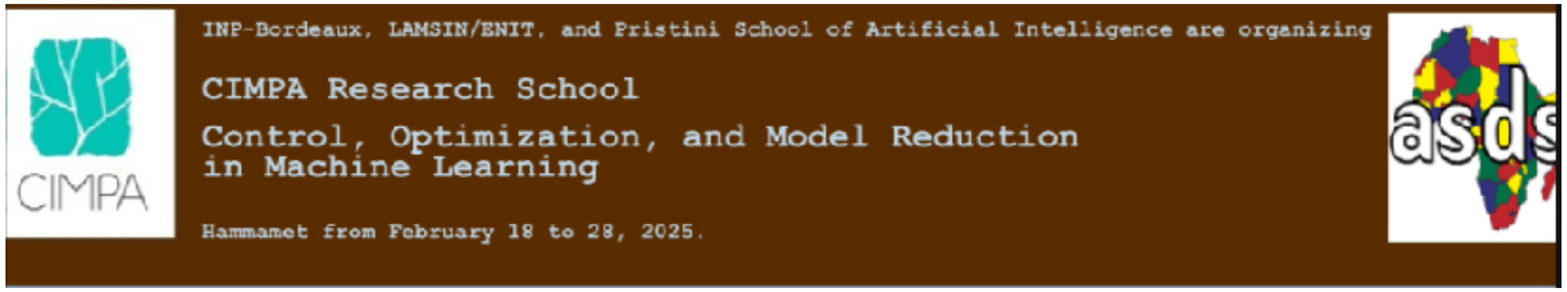
A. Álvarez-López, R. Orive-Illera, E. Zuazua, *Optimized classification with neural ODEs via separability*, arXiv:2312.13807 (2023)

A. Álvarez-López, A. H. Slimane, E. Zuazua, *Interplay between depth and width for interpolation in neural ODEs*, NEUNET, 180 (2024), 106640.

M. Hernández, E. Zuazua, *Deep neural networks: multi-classification and universal approximation*, arXiv preprint arXiv:2409.06555.

Conclusions and Perspectives

Fantastic horizon for mathematical research



INP-Bordeaux, LAMSIN/ENIT, and Pristini School of Artificial Intelligence are organizing

CIMPA Research School

Control, Optimization, and Model Reduction in Machine Learning

Hammamet from February 18 to 28, 2025.

- **Maths for Learning**
  - Gradient descent dynamics
  - Generalization
  - Generation
  - Width/Depth... Architectures
  - Dimensionality and probabilities
  - Attention mechanisms
  - Federated Learning
  - .... Curse of dimensionality + Devil of non-convexity.
- **Digital Twins Methodologies** pose specific challenges
  - Scalability / Adaptivity / Personalised / Goal oriented (Model Predictive Control?)
  - Control of control for DT modelling
  - Reliability / generalisation / synthetic data
  - Merging with Physics and Mechanics
  - Applications: Personalised Medicine, Environment, Climate, Energy,...

Thank you for the invitation and attention