Control and Machine Learning

Enrique Zuazua

FAU - Erlangen
Alexander von Humboldt Proffesorship

Funded by the ERC-AdvG DyCon

European Research Council
Established by the European Commission

Unterstützt von / Supported by
Alexander von Humboldt
Stiftung/Foundation

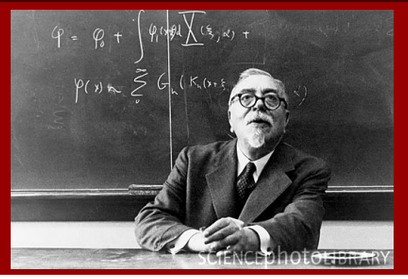
Universidad Autónoma de Madrid

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Deusto

Cybernetics, Norbert Wiener, 1948

The science of control and communication in animals and machines



Let $n, m \in \mathbb{N}^*$ and $T > 0$ and consider the following linear finite-dimensional system

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

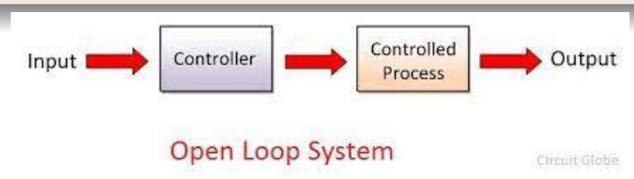
In (1), A is a $n \times n$ real matrix, B is of dimensions $n \times m$ and x^0 is the initial sate of the sytem in $\mathbb{R}^n$. The function $x : [0, T] \longrightarrow \mathbb{R}^n$ represents the *state* and $u : [0, T] \longrightarrow \mathbb{R}^m$ the *control*.

Can we control the state x of n components with only m controls, even if $n >> m$?

Theorem

(1958, Rudolf Emil Kálmán (1930–2016)) System (1) *is controllable iff*

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



An example: Nelson's car.

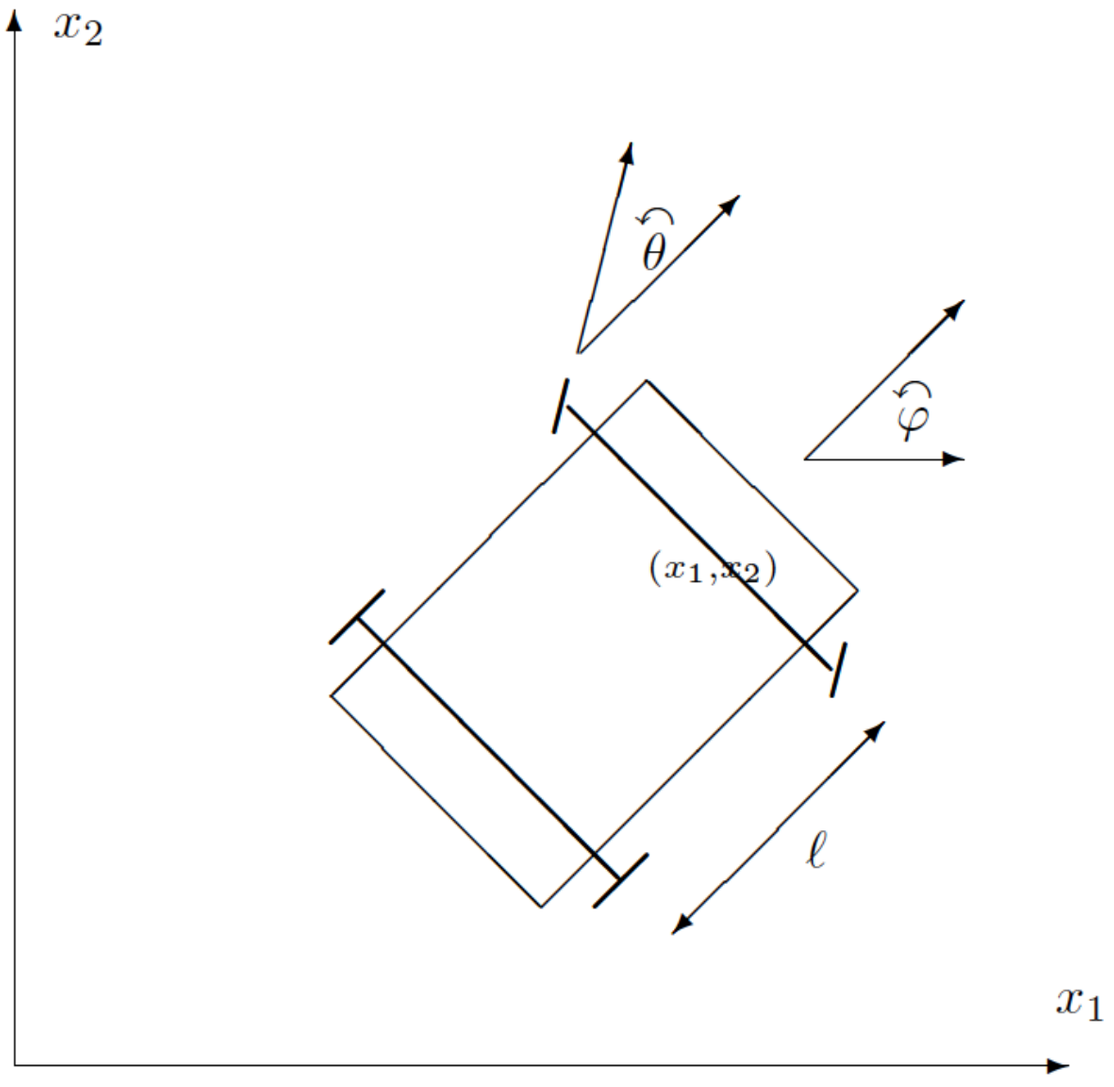
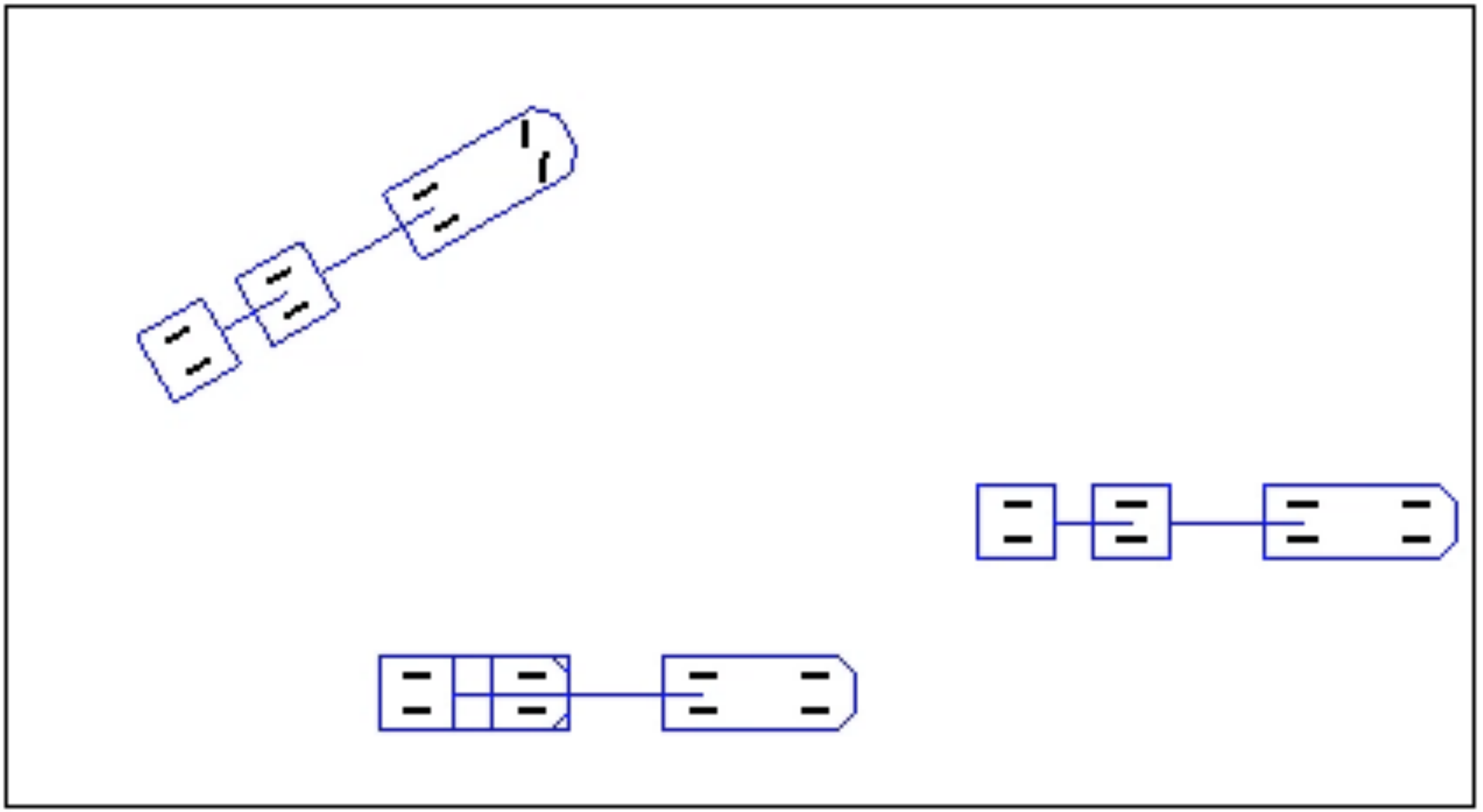


Figure 4.1: *4-dimensional car model.*

Two controls suffice to control a four-dimensional dynamical system.

E. Sontag, *Mathematical control theory*, 2nd ed.,Springer-Verlag, NewYork,1998.

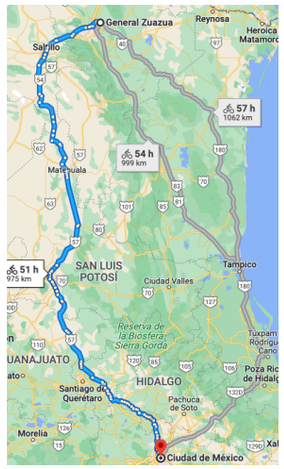
Computational implementation (Y. Privat)



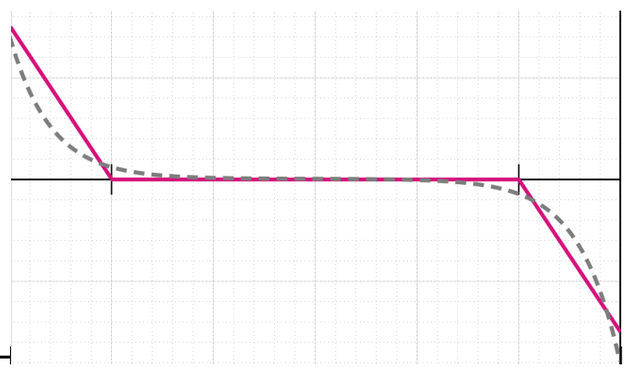
Virtuoso solution

Turnpike Control (time matters!)

Although the idea goes back to John von Neumann in 1945, Lionel W. McKenzie traces the term to Robert Dorfman, Paul Samuelson, and Robert Solow's "Linear Programming and Economics Analysis" in 1958, referring to an American English word for a Highway:[2] [3] [4]



... There is a fastest route between any two points; and if the origin and destination are close together and far from the turnpike, the best route may not touch the turnpike. But if the origin and destination are far enough apart, it will always pay to get on to the turnpike and cover distance at the best rate of travel, even if this means adding a little mileage at either end.

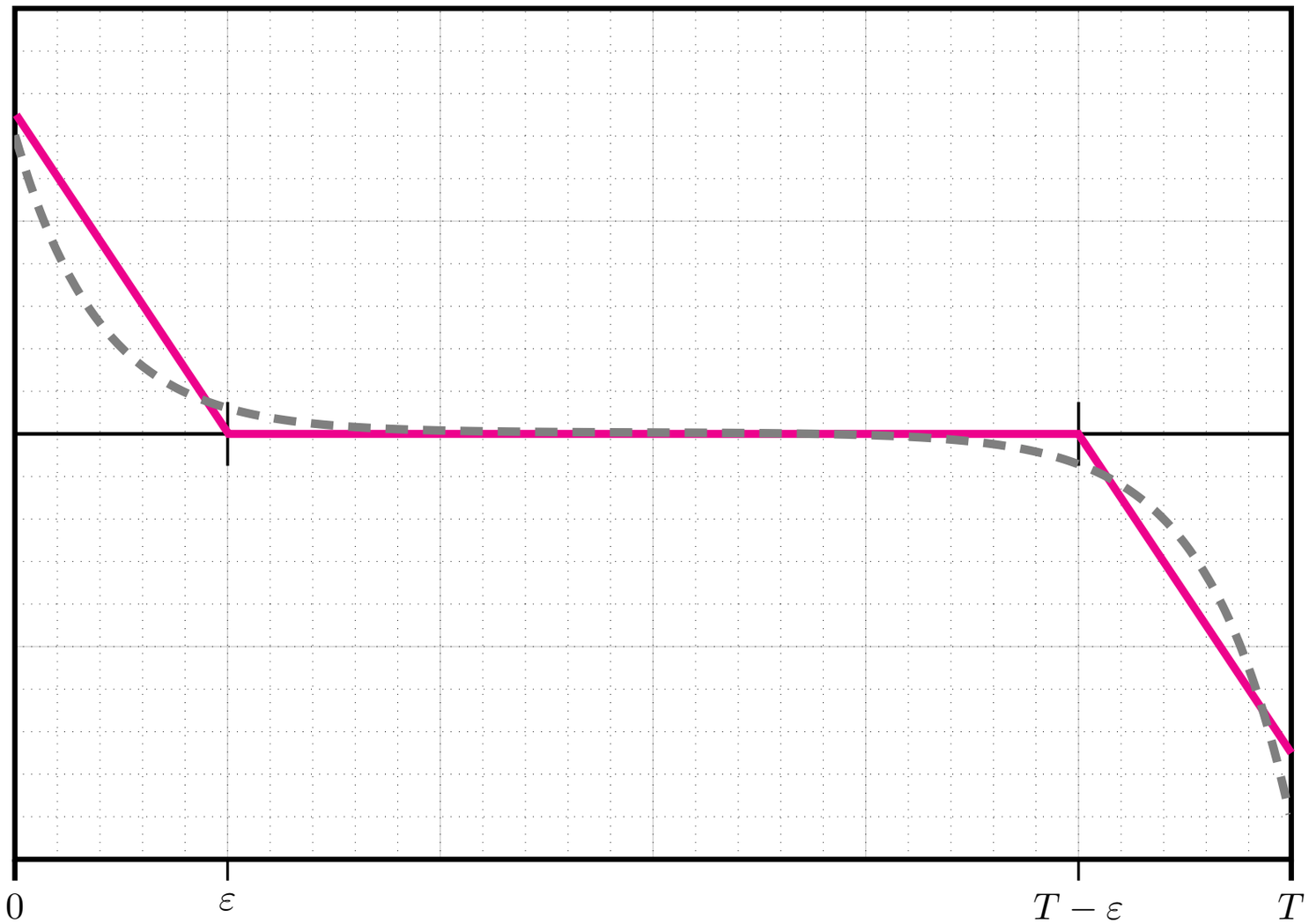


[2] Porretta, A., Z., E. (2013). SIAM J. Control and Optimization, 51(6), 4242-4273.
[3] Trélat, E., Z., E. (2015). J. Differential Equations, 258(1), 81-114.
[4] A. J. Zaslavski, Springer, New York, 2006.

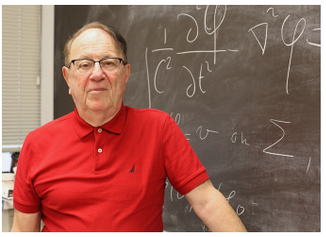
Turnpike pattern

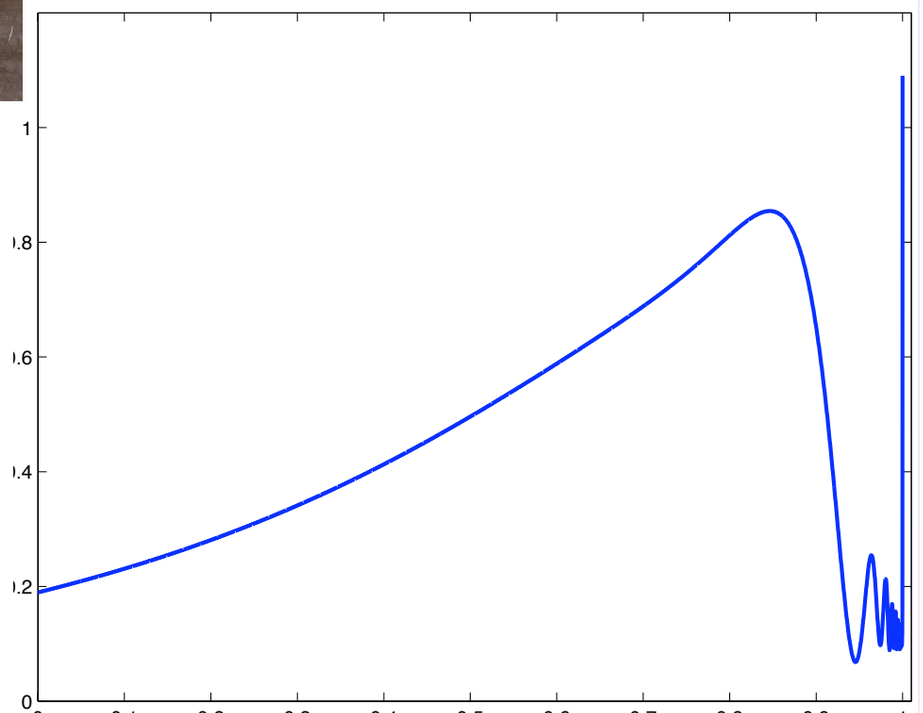
Applications in:

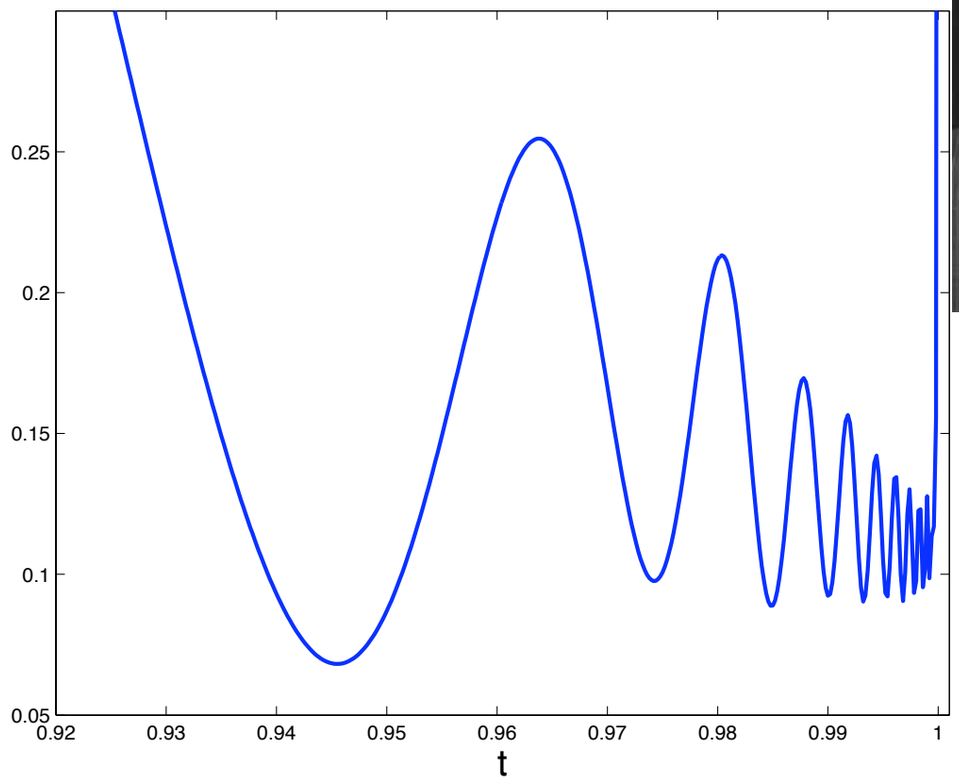
- Sustainable economic growth planning.
- Chronic deseases.



But... oscillatory patterns in heat control







Observed by R. Glowinski and J. L. Lions in the 80's in their works in the numerical analysis of controllability problems for heat and wave equations: Typical controls for the heat equation exhibit **unexpected** oscillatory and concentration effects.

Why? Lazy controls? Turnpike theory does not apply in the PDE context?

Heat-diffusion control : A closer look

Let $n \geq 1$ and $T > 0$, Ω be a simply connected, bounded domain of $\mathbb{R}^n$ with smooth boundary Γ, $Q = (0, T) \times \Omega$ and $\Sigma = (0, T) \times \Gamma$:

$$\begin{cases} y_t - \Delta y = f 1_\omega & \text{in} \quad Q \\ y = 0 & \text{on} \quad \Sigma \\ y(x, 0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2}$$

$1_\omega =$ the characteristic function of ω of Ω where the control is active.

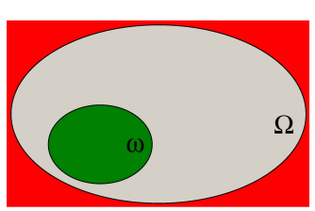
We know that $y^0 \in L^2(\Omega)$ and $f \in L^2(Q)$ so that (2) admits a unique solution

$$y \in C\left([0, T]; L^2(\Omega)\right) \cap L^2\left(0, T; H_0^1(\Omega)\right).$$

$y = y(x, t) = \textit{solution} = \textit{state}$, $f = f(x, t) = \textit{control}$

Goal: Drive the dynamics to equilibrium by means of a suitable choice of the control

$$y(\cdot, T) \equiv y^*(x).$$

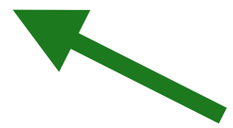


We address this problem from a classical optimal control / least square approach:

$$min\,\frac{1}{2}\left[\int_0^T\int_\omega |f|^2 dxdt + \int_\Omega |y(x,T)-y^*(x)|^2 dx\right].$$

According to Pontryagin's Maximum Principle the Optimality System (OS) reads

$$y_t - \Delta y = p1_\omega \text{ in } Q$$
$$-p_t - \Delta p = 0 \text{ in } Q$$
$$y = 0 \text{ on } \Sigma$$
$$y(x,0) = y^0(x) \text{ in } \Omega$$
$$p(x,T) = y(x,T) - y^*(x) \text{ in } \Omega$$
$$p = 0 \text{ on } \Sigma.$$



And the optimal control is:

$$f(x,t) = p(x,t) \quad \text{in}\,\omega \times (0,T).$$

Remedy: Better balanced controls

Let us now consider the control f minimising a compromise between the norm of the state and the control among the class of admissible controls:

$$min\, \frac{1}{2}\left[\int_0^T\int_\Omega |y|^2 dxdt + \int_0^T\int_\omega |f|^2 dxdt + \int_\Omega |y(x,T) - y^*(x)|^2\right].$$

Then the Optimality System reads

$$y_t - \Delta y = -p1_\omega \text{ in } Q$$

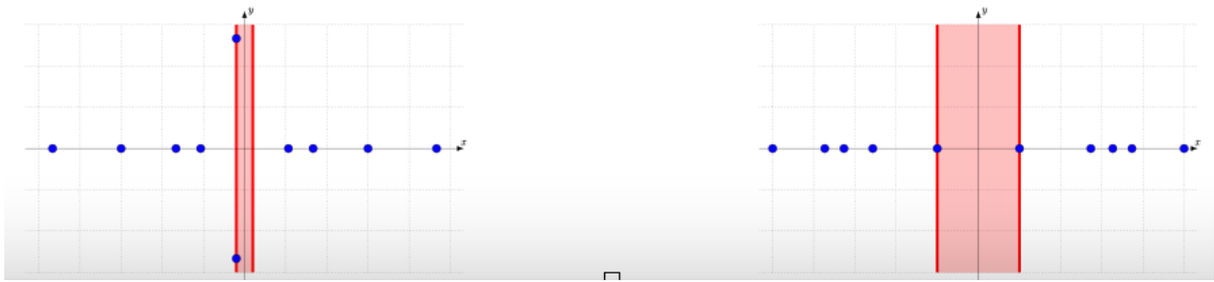
$$-p_t - \Delta p = y \text{ in } Q$$

$$y = p = 0 \text{ on } \Sigma$$

$$y(x,0) = y^0(x) \text{ in } \Omega$$

$$p(x,T) = y(x,T) - y^*(x) \text{ in } \Omega$$

We now observe a coupling between p and y on the adjoint state equation!

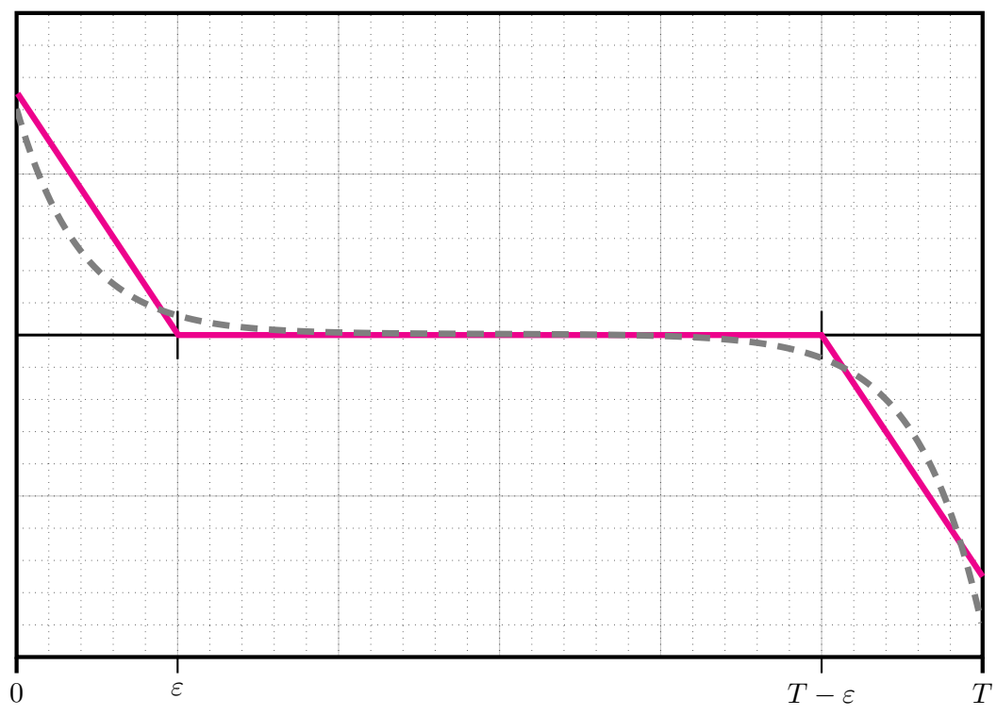


Conclusion

The turnpike principle holds for PDEs as well but under the condition that:

- Controls are characterized as minima of control functionals penalizing sufficiently the control and the state.
- The system is controllable, so that all trajectories might get to the turnpike (steady optimal) gate.

Under these conditions the time-depending controls and controlled trajectories remain most of the time in a steady optimal configuration.



Supervised Learning

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

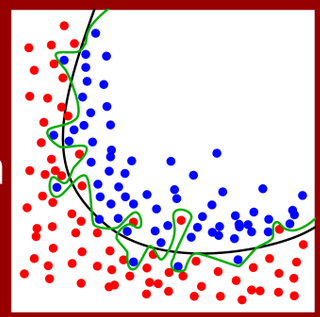
$$\left\{\vec{x}_i, \vec{y}_i\right\}_{i=1}^{N} \subset \mathbb{R}^{d\times N} \times \mathbb{R}^{m\times N}.$$

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

Popular method: **training a neural network**. For instance using Cybenko's ansatz, with σ a sigmoidal activation function or a ReLU:

$$f(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j \cdot x + \alpha_j).$$

Residual neural networks

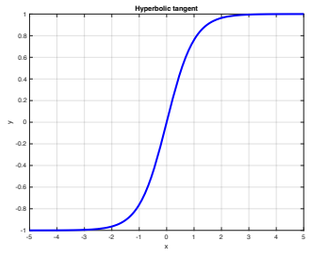


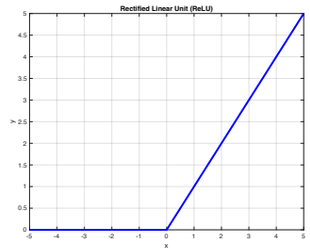
[1] K. He, X Zhang, S. Ren, J Sun, 2016: Deep residual learning for image recognition
[2] E. Weinan, 2017. A proposal on machine learning via dynamical systems.
[3] R. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, 2018.
[4] E. Sontag, H. Sussmann, 1997.

ResNets: for all items (= initial data), $1 \leq i \leq N$

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + hA^k\sigma(w^k\mathbf{x}_i^k + b^k) & \text{for } \; k \in \{0, \ldots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \end{cases}$$

σ globally Lipschitz & $\sigma(0) = 0$. **layer = timestep**; $h = T/N_{layers}$ for given $T > 0$.





Neural ODEs: for all items (= initial data), $1 \leq i \leq N$

$$\begin{cases} \dot{\mathbf{x}}_i(t) = A(t)\sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } \; t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i. \end{cases}$$

A giant simultaneous control problem: each item $\mathbf{x}_i(0)$ needs to the driven to the corresponding destination (label) with the same control.

Our contributions

- The ensemble or simultaneous control property holds [5][6]

  D. Ruiz-Balet & E. Z., Neural ODE control for classification, approximation and transport, arXiv:2104.05278.

- Consequently, the turnpike phenomenon holds as well, under the condition of penalizing trajectories all along time:

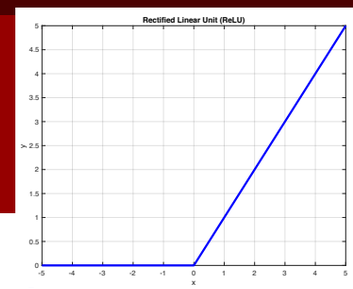
  B. Geshkovski, C. Esteve, D. Pighin, E. Z., Turnpike in Lipschitz nonlinear optimal control, Nonlinearity, 35 (2022), 1652-1701.

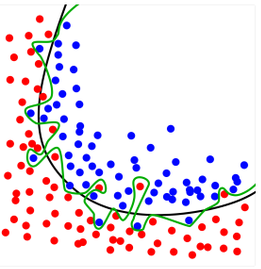
  B. Geshkovski, E. Z. Turnpike in optimal control of PDEs, ResNets, and beyond, Acta Numerica, 2022, pp. 135-263.

[5] Agrachev, A., Sarychev, A. (2021). Control on the Manifolds of Mappings with a View to the Deep Learning. Journal of Dynamical and Control Systems, 1-20.

[6] Li, Q., Lin, T., Shen, Z. (2022). Deep learning via dynamical systems: An approximation perspective. Journal of the European Mathematical Society.

Special features of the control of ResNets

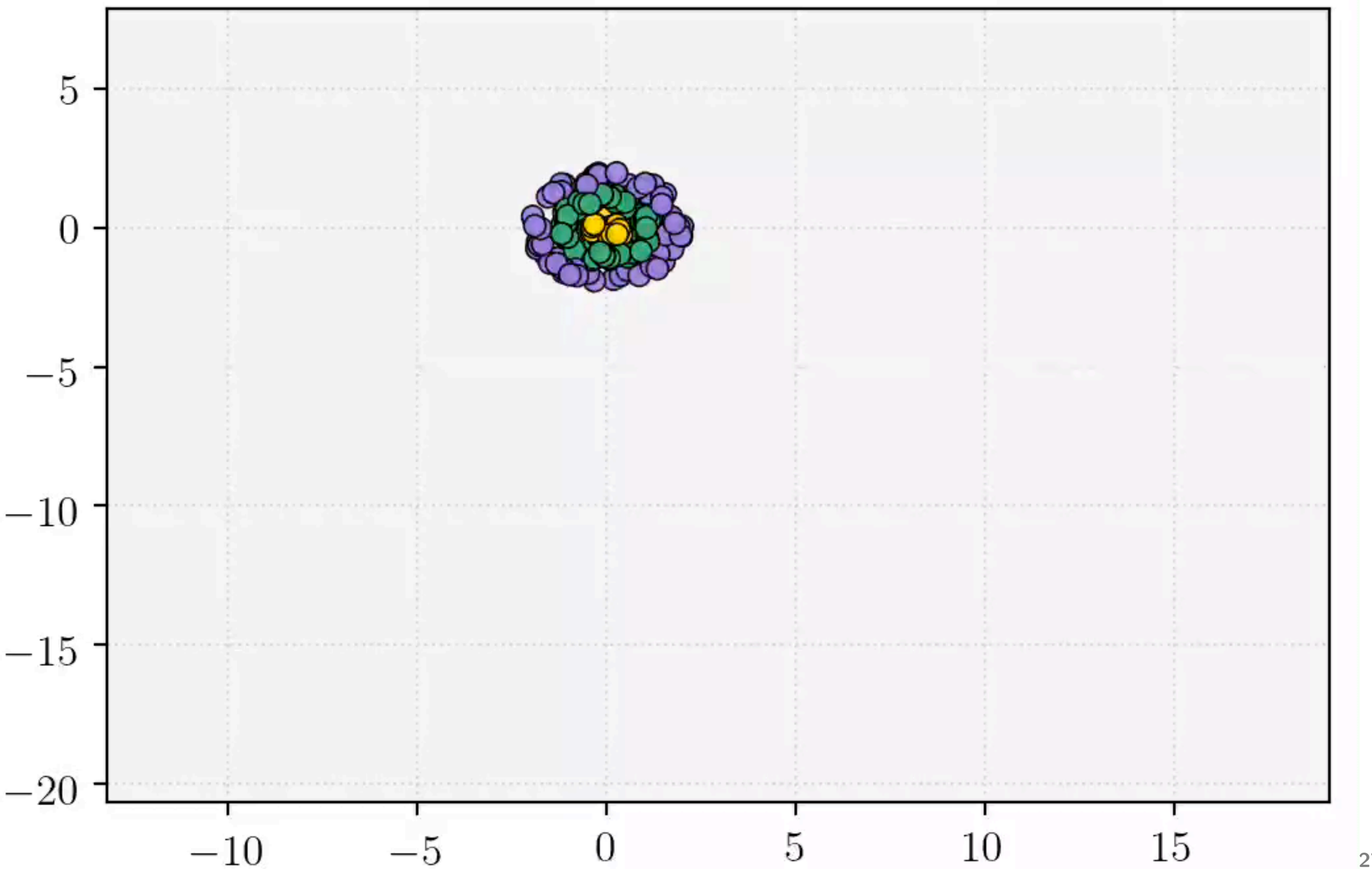




- Nonlinearities are unusual in Mechanics: σ is flat in half of the phase space.
- We need to control many trajectories (one per item to be classified) with the same control!

The very features of the activation function σ allow to achieve this monster simultaneous control goal. The fact that σ leaves half of the phase space invariant while deforming the other one, allows for dynamics that are not encountered in the classical ODE systems in mechanics for which such kind of simultaneous control property is unlikely or even impossible.

ResNets in action (Borjan Geshkovski)



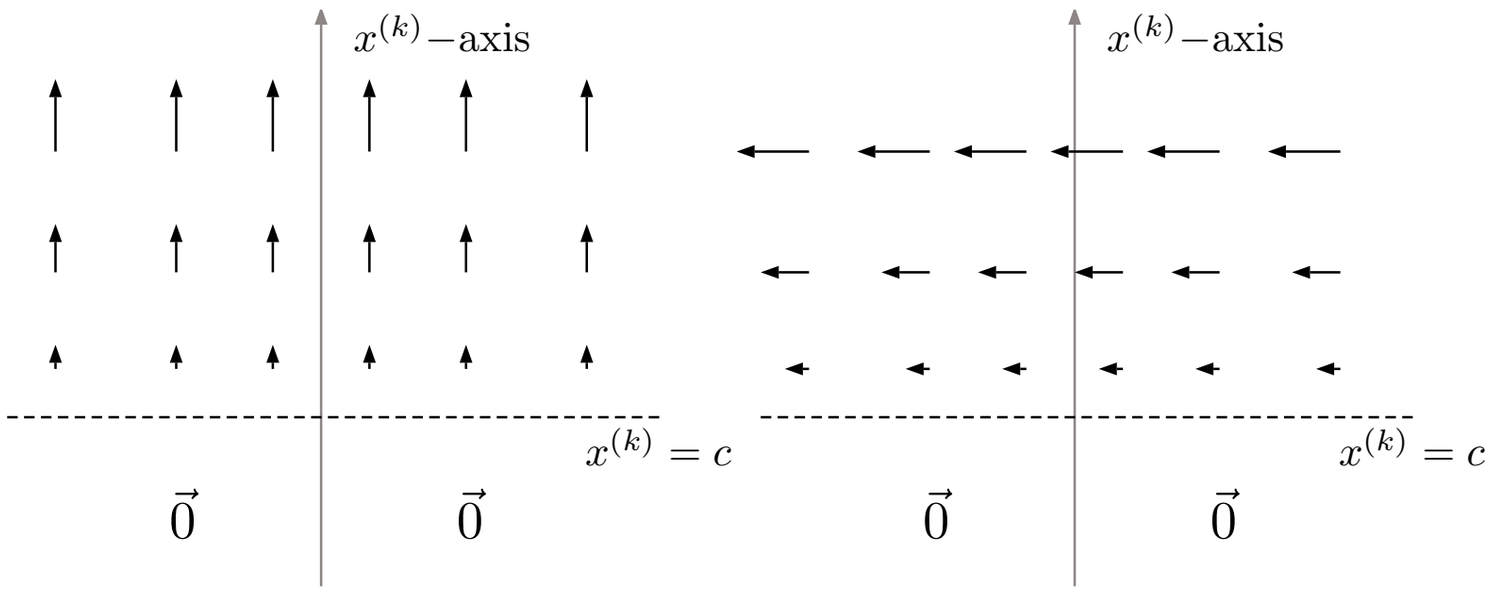
What is the ResNet doing? Basic control actions

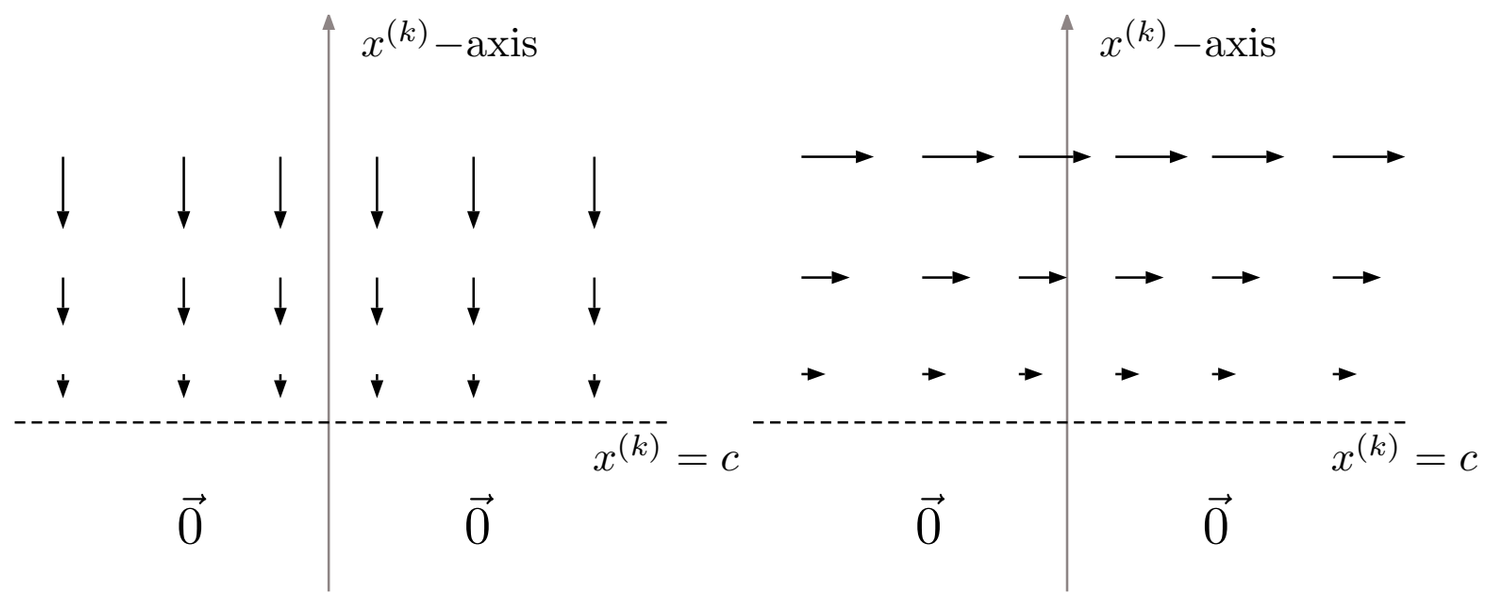
$$\dot{\mathbf{x}}(t) = W(t)\sigma(A(t)\mathbf{x}(t) + b(t))$$

- $b(t)$ induces a time-dependent translation of the Euclidean space. It plays an important role to determine the center of the action of the sigmoid.
- $A(t)$ compresses, expands, and induces rotations in the euclidean space.
- $(A(t), b(t))$ determine a hyperplane in the space, the equator, diving space into the active and the inactive half-spaces.
- $W(t)$ determines the direction and intensity with which the flow will evolve in the active hemisphere.

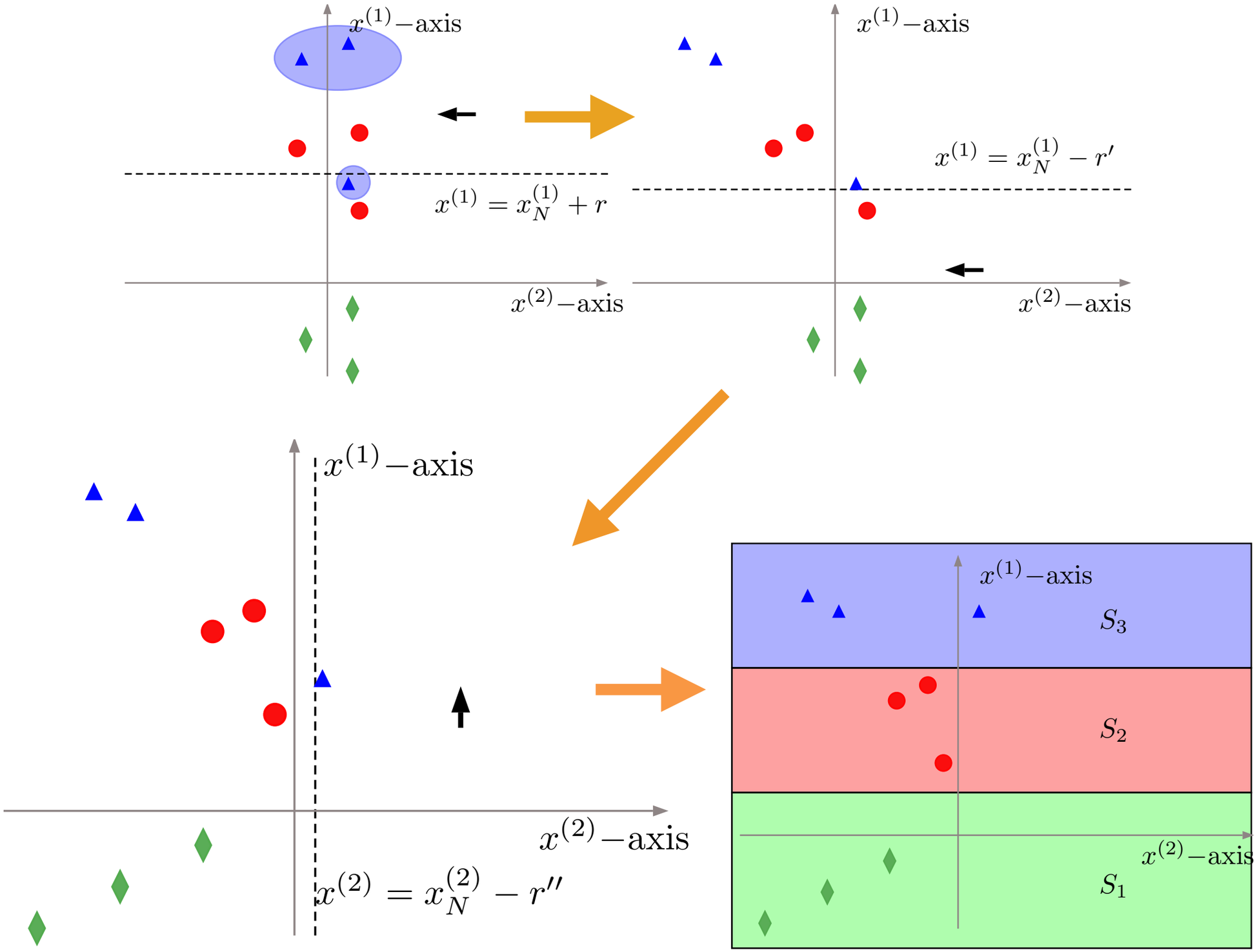
An intelligent piecewise constant choose of controls, by induction, assures the needed simultaneous control property.

Some canonical flows induced by nODE





Classifying / controlling one datum



Classification by control

Theorem (Classification, Domènec Ruiz-Balet EZ, 2021)

[a] *Let σ be the ReLU, $d \geq 2$, and $N, M \geq 2$. Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ be data to be classified into disjoint open non-empty subsets S_m, $m = 1, ..., M$ with labels $m = m(i)$, $i = 1, ..., N$.*
Then, for every $T > 0$, there exist control functions $A, W \in L^\infty \cap BV\left((0,T); \mathbb{R}^d\right)$ and $b \in L^\infty \cap BV\left((0,T), \mathbb{R}\right)$ such that the flow associated to the Neural ODE, when applied to all initial data $\{x_i\}_{i=1}^N$, classifies them simulatenously, i.e.

$$\phi_T(x_i; A, W, b) \in S_{m_i}, \quad \forall i = 1, ..., N.$$

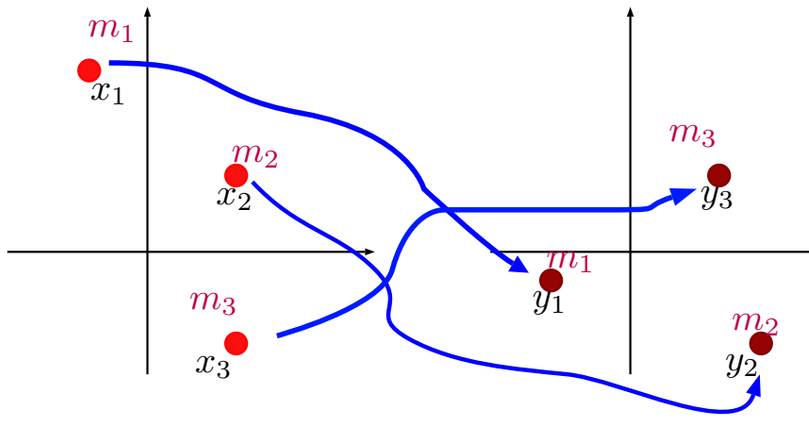
Furthermore,

- *Controls are piecewise constant with a maximal finite number of switches of the order of $\mathcal{O}(N)$. They also lie in BV.*
- *The control time $T > 0$ can be made arbitrarily small (scaling).*
- *The complexity of controls diminishes when initial data are structured in clusters or the control requirement is relaxed.*

Neural transport equations

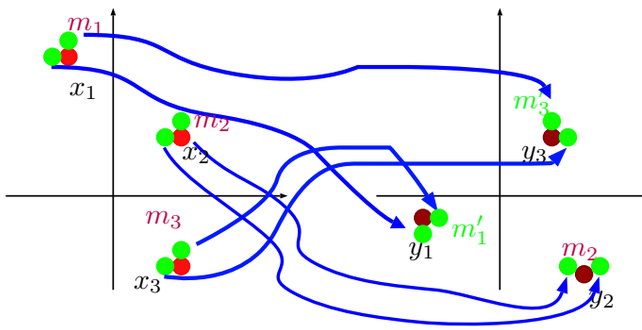
Note that the differential equation

$$\begin{cases} \dot{x} = W(t)\sigma(A(t)x + b(t)) \\ x(0) = x_0 \end{cases}$$



corresponds to the characteristics of the transport equation:

$$\begin{cases} \partial_t \rho + \mathrm{div}_x \left[(W(t)\sigma(A(t)x + b(t)))\rho \right] = 0 \\ \rho(0) = \rho^0 \end{cases}$$



Atomic initial data can be driven to atomic final targets.

This establishes a link to the Theory of Optimal Transport: Neural Transport?

Neural transport

Conclusions and Perspectives

- Control Theory and Machine Learning share in part origins and goals.
- Mutual cross-fertilization offers great opportunities.
- Some of the problems are rather challenging.

We can understand analytically how and why algorithms work in the ResNet context. But we can hardly explain and anticipate the optimal configurations and strategies that emerge computationally.

Plenty to be done to better understand the fully nonlinear discrete dynamics of deep neural networks.

Chair
DYNAMICS, CONTROL
AND NUMERICS
FAU