Control and Machine Learning

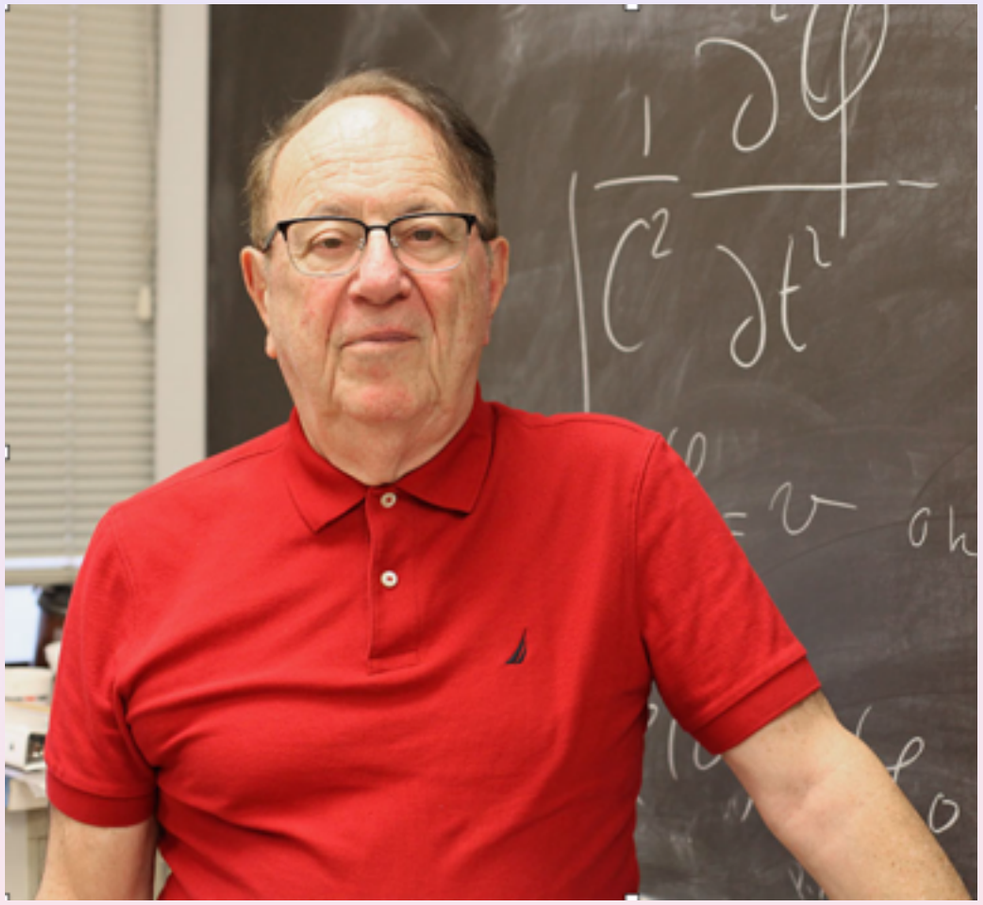
Enrique Zuazua

FAU-Erlangen
UAM-Madrid
Fundación-Deusto-Bilbao
https://caa-avh.nat.fau.eu

https://dcn.nat.fau.eu

European Research Council
Established by the European Commission

Unterstützt von / Supported by
Alexander von Humboldt Stiftung/Foundation



Roland Glowinski
1937-2022

Ireneo Peral
1946-2021

The origins

"... if every instrument could accomplish its own work, obeying or anticipating the will of others . . . if the shuttle weaved and the pick touched the lyre without a hand to guide them, chief workmen would not need servants, nor masters slaves."

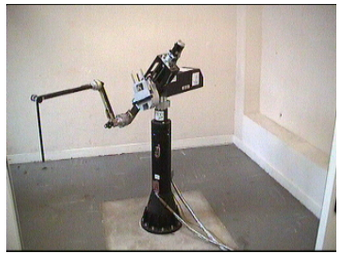
Book I, Chapter II, of the monograph "Politics" by Aristotle (384-322 B. C.).

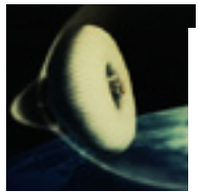
Main motivation: The need of automatizing processes to let the human being gain in liberty, freedom, and quality of life.

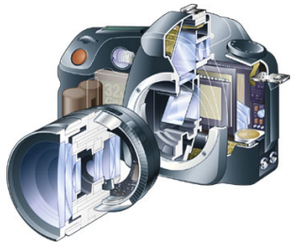
Control theory and applications

Mechanics

Vehicles (guidance, dampers, ABS, ESP, ...),
Aeronautics, aerospace (shuttle, satellites), robotics

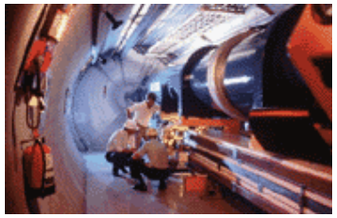


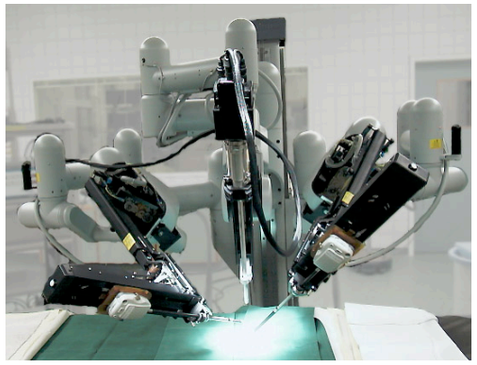


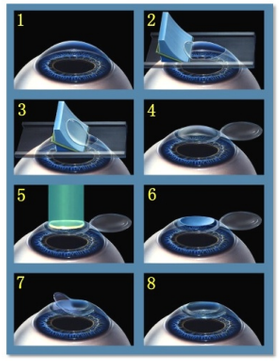


Biology, medicine

Predator-prey systems, bioreactors, epidemiology,
medicine (peacemakers, laser surgery)

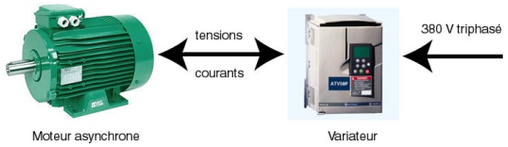


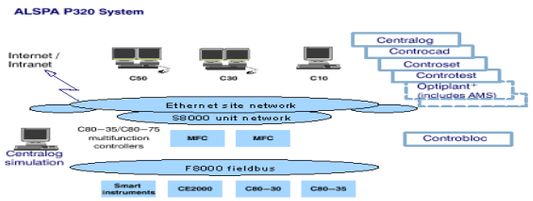




Electricity, electronics

RLC circuits, thermostats, regulation, refrigeration, computers, internet
and telecommunications in general, photography and digital video



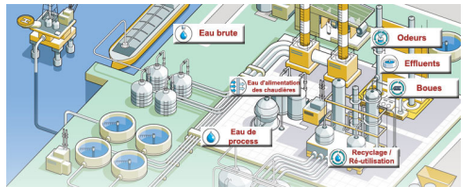


Economics

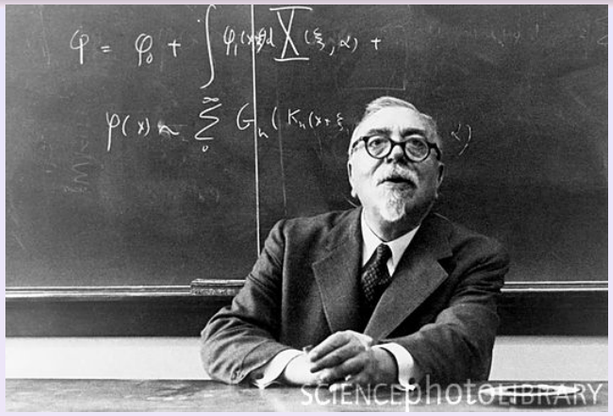
Gain optimization, control of financial flux,
Market prevision

Chemistry

Chemical kinetics, engineering process, petroleum, distillation, petrochemical industry



Historical perspective: Cybernetics



"Cybernétique" was proposed by the French physicist **A.-M. Ampère** in the XIX Century to design the nonexistent science of process controlling. This was quickly forgotten until 1948, when **Norbert Wiener** (1894–1964) chose **"Cybernetics"** as the title of his famous book.

Wiener defined Cybernetics as " the science of control and communication in animals and machines".[1]

In this way, he established the connection between Control Theory and Physiology and anticipated that, in a desirable future, engines would obey and imitate human beings.


Available online at www.sciencedirect.com

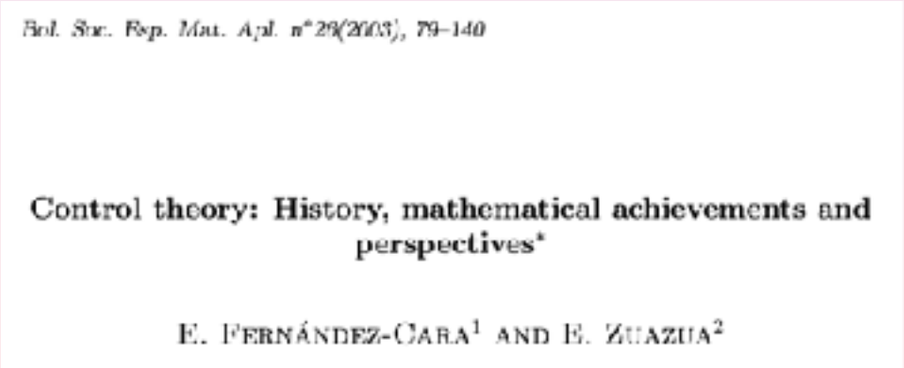
ScienceDirect

IFAC PapersOnLine 53-2 (2020) 1385–1390

Early History of Machine Learning

Alexander L. Fradkov*

* Institute for Problems in Mechanical Engineering, Russian Academy of Sciences, Saint-Petersburg (e-mail: Alexander.Fradkov@gmail.com)


Bol. Soc. Esp. Mat. Apl. n° 29(2003), 79–140

Control theory: History, mathematical achievements and perspectives*

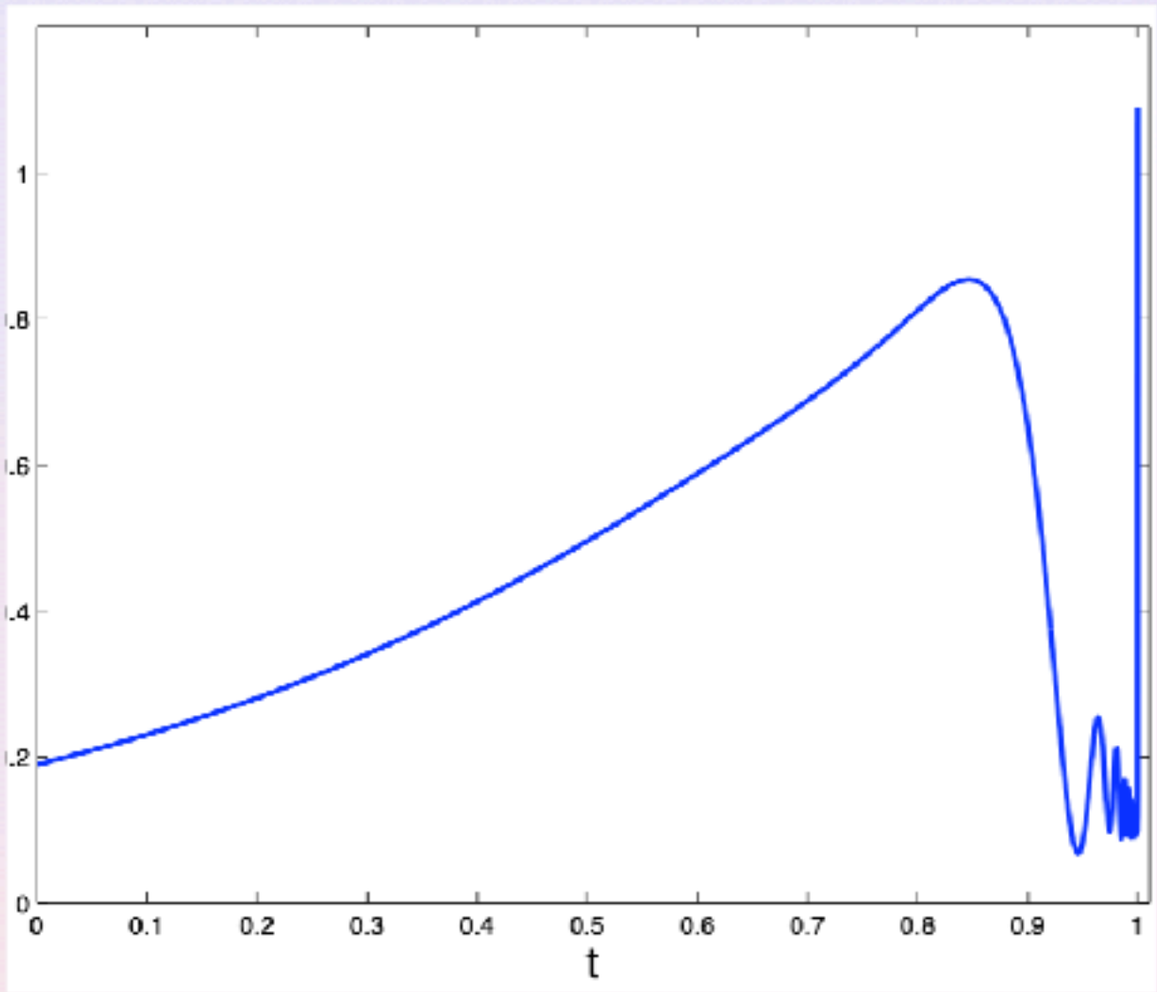
E. Fernández-Cara[1] and E. Zuazua[2]

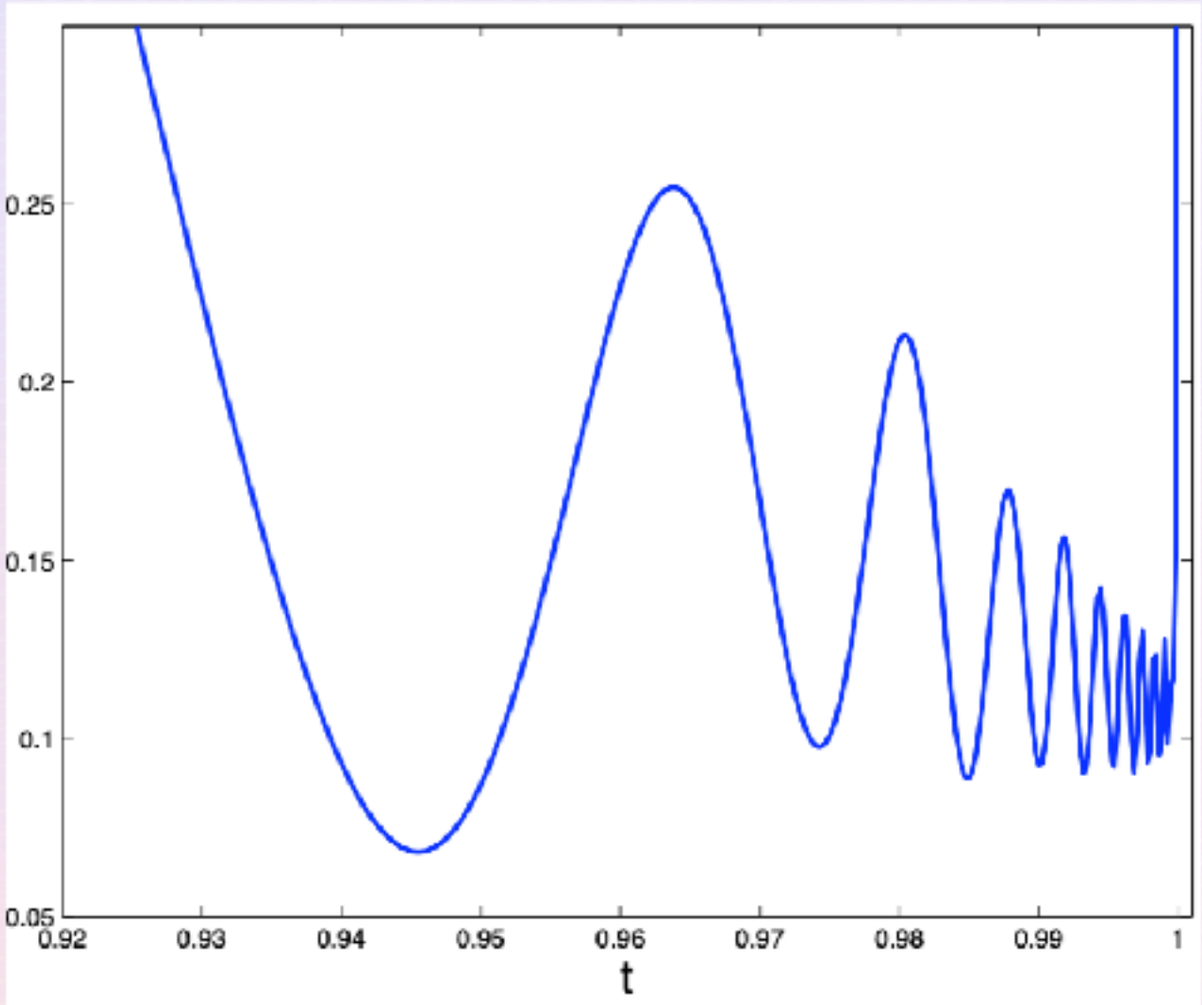
[1] "What we want is a machine that can learn from experience." Alan Turing, 1947.

Heat and diffusion processes

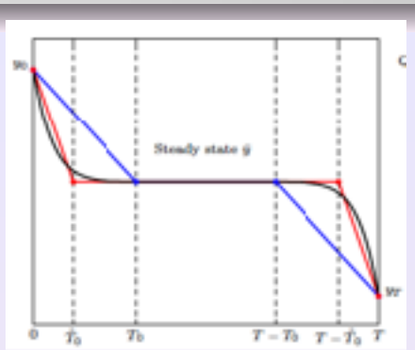
Typical controls for the heat equation exhibit **unexpected** oscillatory and concentration effects. This was observed by R. Glowinski and J. L. Lions in the 80's in their works in the numerical analysis of controllability problems for heat and wave equations.

Why? Lazy controls?



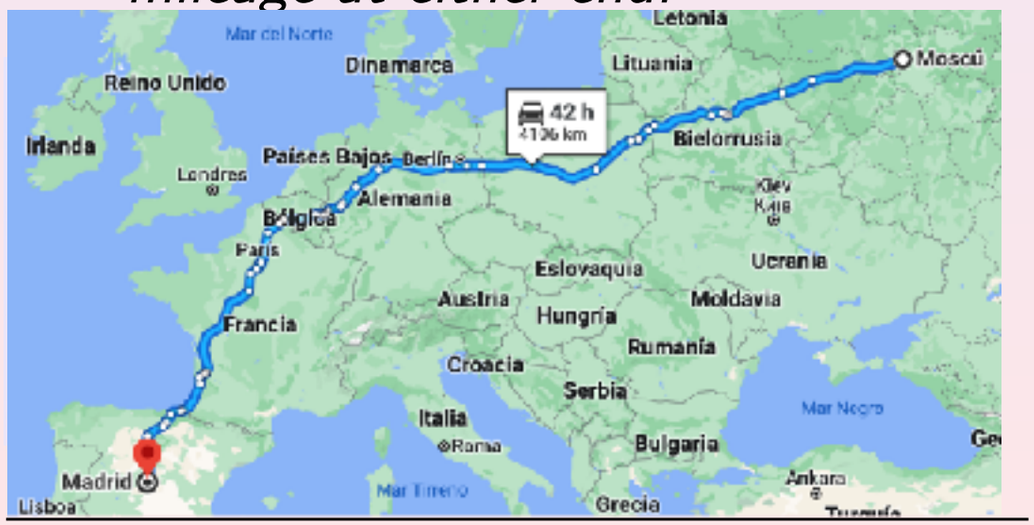


Turnpike



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:[5] [6]

> *... 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.*



[5] A. J. Zaslavski, Springer, New York, 2006.

[6] L. Grüne, Automatica, 49, 725-734, 2013

The control problem for diffusion : 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 $\Sigma = (0, T) \times \Gamma$:

$$\begin{cases} y_t - \Delta y = f 1_\omega & \text{in} \quad Q \\ y = 0 & \text{on} \quad \Sigma \\ y(x,0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{3}$$

1_ω = the characteristic function of ω of Ω where the control is active.

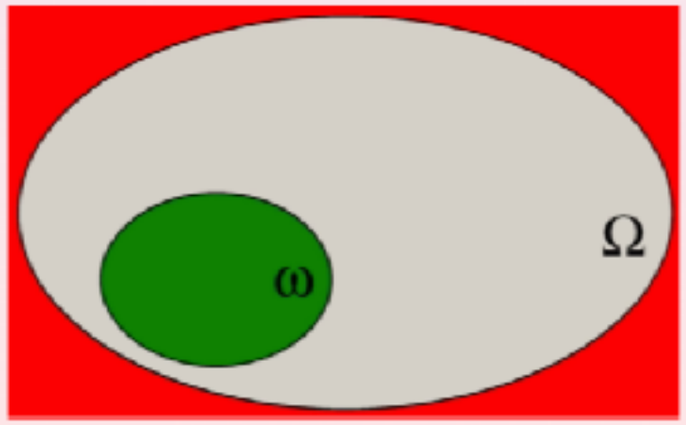
We know that $y^0 \in L^2(\Omega)$ and $f \in L^2(Q)$ so that (3) admits a unique solution

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

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

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 fro 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

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

$$-p_t - \Delta p = 0 \text{ in } Q$$

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

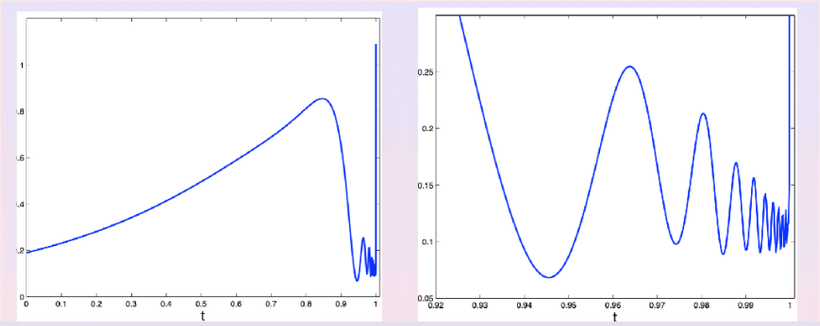
$$y(x,0) = y^0(x) \text{ in } \Omega$$

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

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

And the optimal control is:

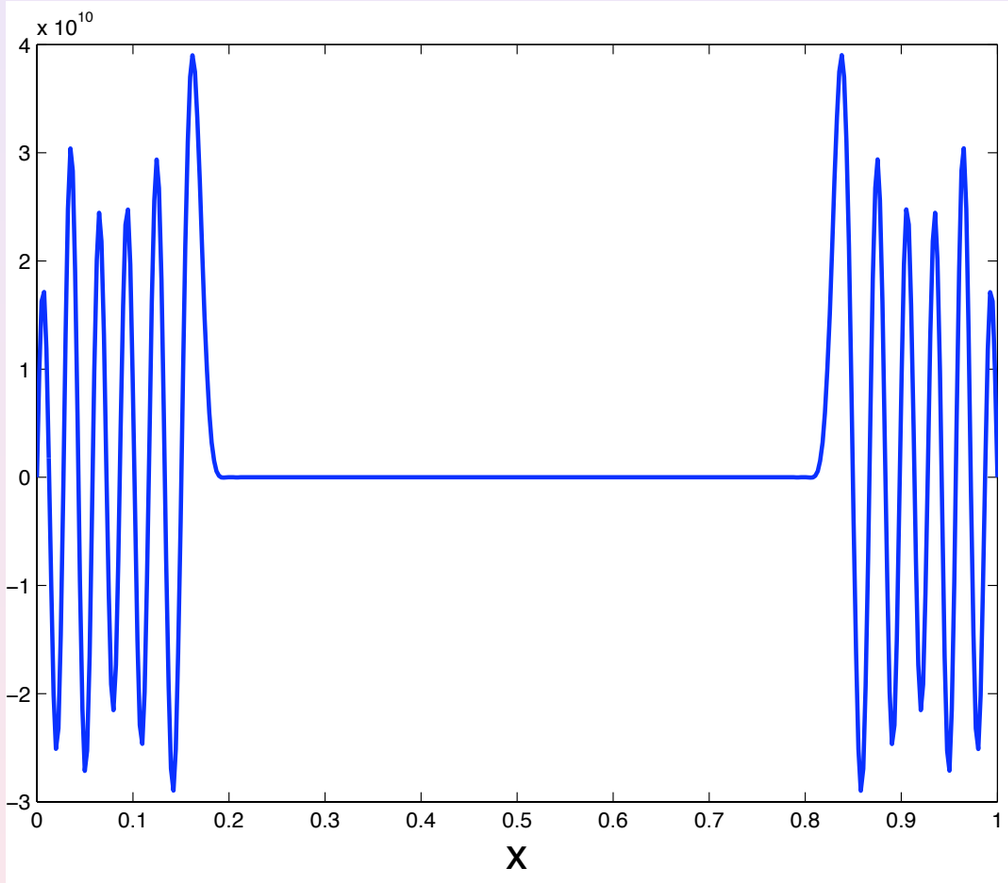
$$f(x,t) = p(x,t) \quad \text{in}\, \omega \times (0,T).$$



By duality (Fenchel-Rockafellar) the adjoint p at time $t = T$, p^T saturates the regularity properties required to assure the well-posedness of the functional:

$$\mathcal{H} = \{p^T : p(x, 0) \in L^2(\Omega)\}$$

This is a huge space, allowing an exponential increase of Fourier coefficients at high frequencies. And, because of this, we observe the tendency of the control to concentrate all the action in the final time instant $t = T$, incompatible with turnpike effects[4]



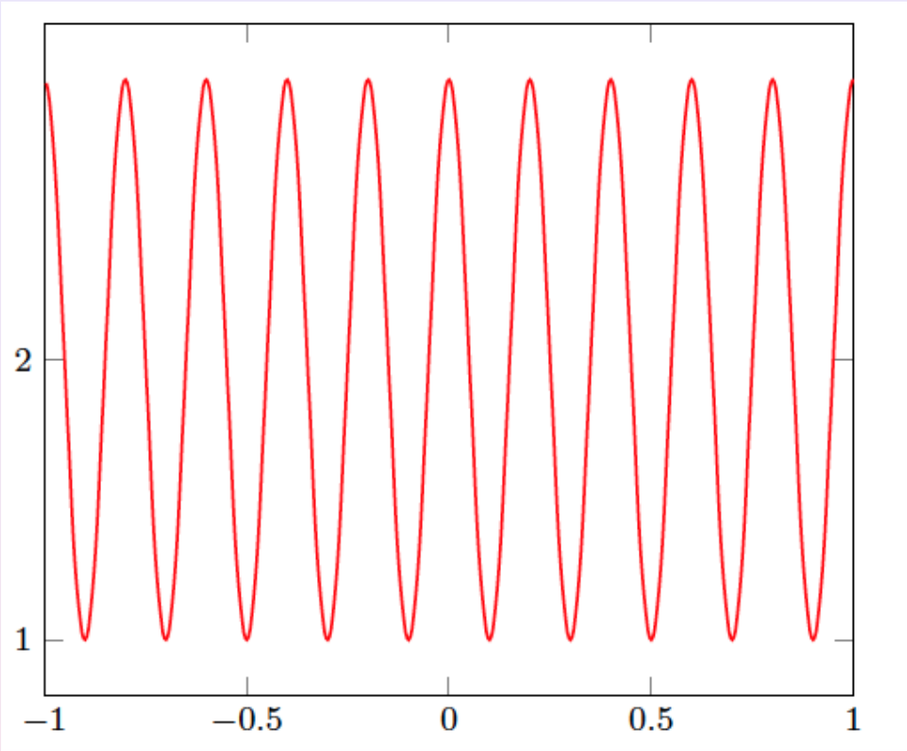
Tychonoff's monster (1935)

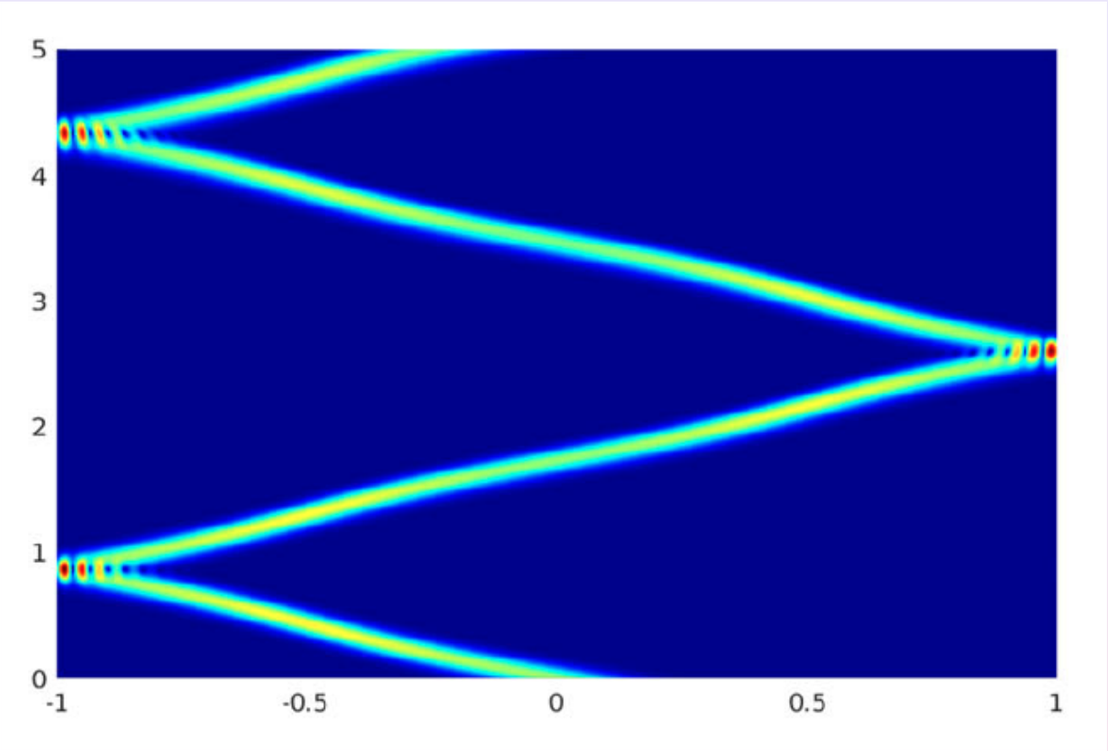
Théorèmes d'unicité pour l'équation de la chaleur

A. Tychonoff (Moscou)

[4] A. Münch & E. Z., Inverse Problems, 2010

Wave propagation: Why do not we see the turnpike?





- Typical controls for the wave equation exhibit an oscillatory behaviour, and this independently of the length of the control time-horizon.
- This fact is intrinsically linked to the oscillatory nature of wave propagation.
- Waves are controlled through anti-waves, reproducing an oscillatory pattern.

Optimal controls are boundary traces of solutions of the **adjoint problem** through the optimality system or the **Pontryagin Maximum Principle**, and solutions of the adjoint heat equation

$$-p_t - \Delta p = 0$$

look precisely this way.

Large and oscillatory near $t = T$ they decay and get smoother when t gets down to $t = 0$. And this is independent of the time control horizon $[0, T]$.

For wave-like equations ontrols are given by the solutions of the adjoint system

$$p_{tt} - \Delta p = 0$$

that exhibit endless oscillations.

First conclusion:
Typical control problems for wave and heat equations do not seem to exhibit the turnpike property.
These are the controls of L^2-minimal norm. There are many other possibilities for successful control strategies (sparse controls by L^1-minimisation, bang-bang controls...)

May be the Turnpike Principle does not hold for for Partial Differential Equations (PDE), i. e. Infinite-Dimensional Dynamical Systems?

Remedy: Better balanced controls

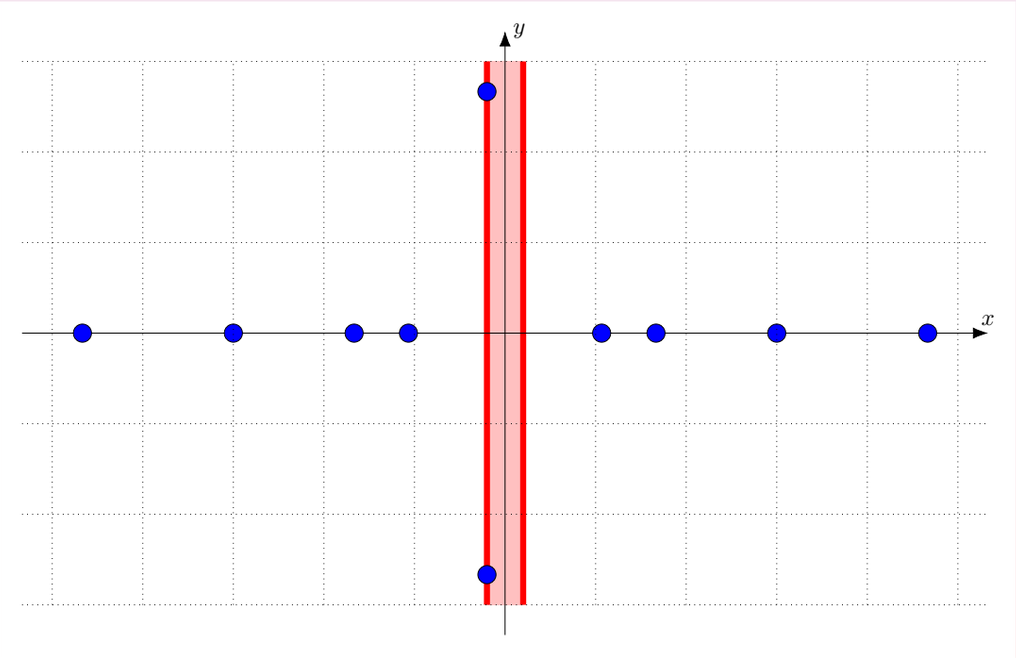
Let us now consider the control f minimising a compromise between the norm of the state and the control among the class of admissible controls:

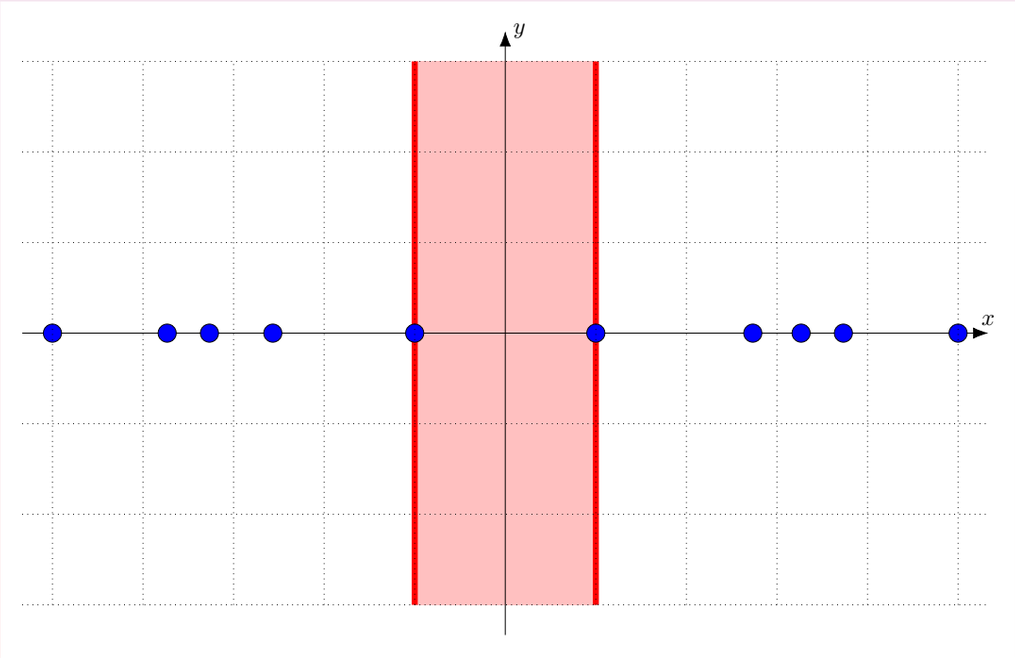
$$min\, \frac{1}{2}\left[\int_0^T \int_\Omega |y|^2 dxdt + \int_0^T \int_\omega |f|^2 dxdt + \int_\Omega |y(x,T) - y^*(x)|^2\right].$$

Then the Optimality System reads

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

We now observe a coupling between p and y on the adjoint state equation![7]

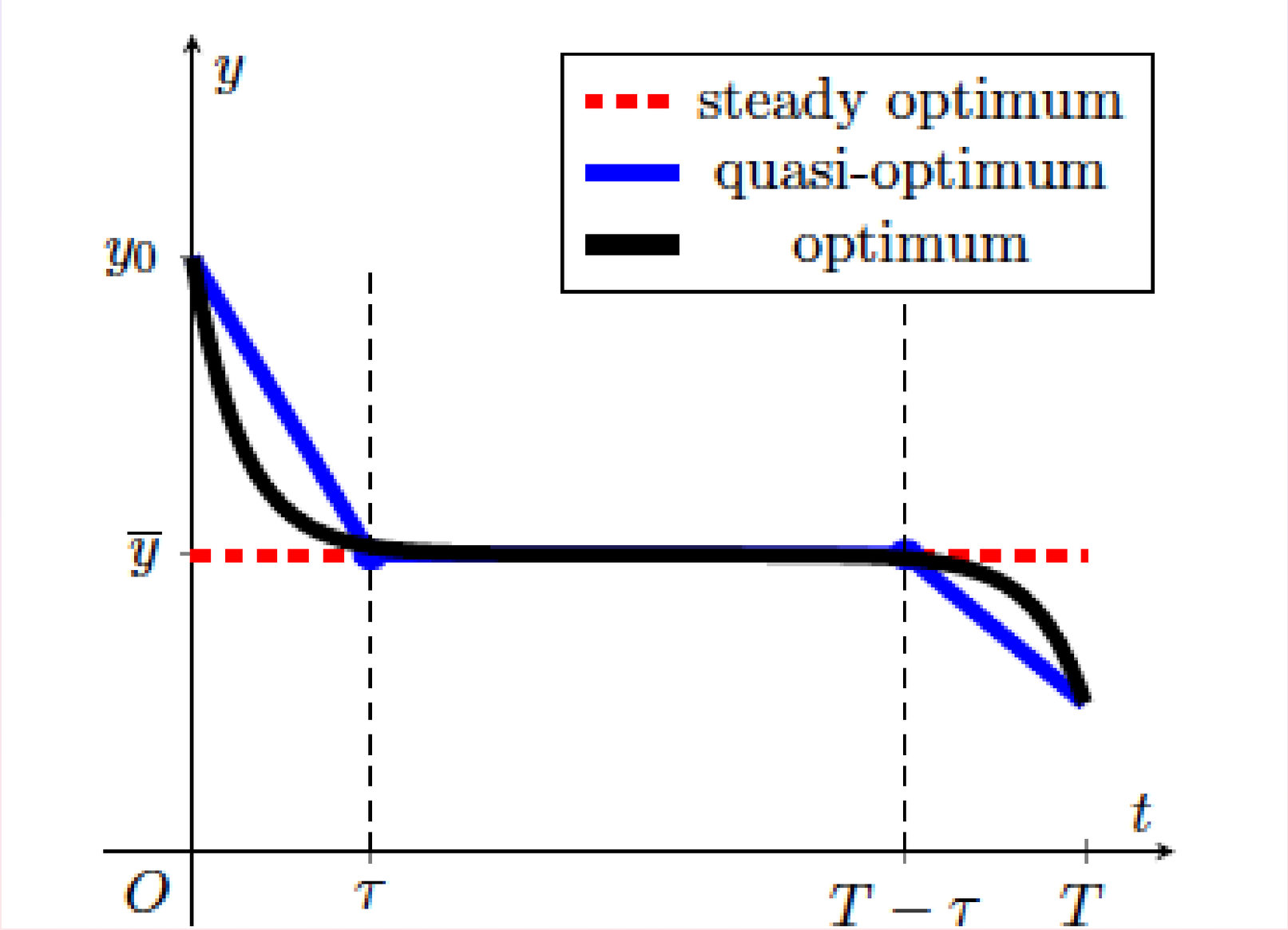




[7] A. Porretta & E. Z., SIAM J. Cont. Optim., 2013

Substantiation

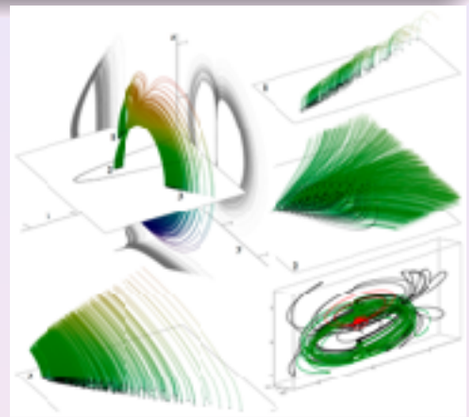
- The turnpike property emerges in long the horizons
- Relevant in economic planning, chronical deseases, sustainable growth, global change, biodiversity, multiculturalism, etc.



- We implement this strategy unconsciently in our daily life
- Turnpike requires the system to be controllable

Linear theory. Joint work with A. Porretta, SIAM J. Cont. Optim., 2013.

The same methods apply in the inifinite-dimensional context, covering in particular linear heat and wave equations



Consider the finite dimensional dvnamics

$$\begin{cases} x_t + Ax = Bu \\ x(0) = x_0 \in \mathbb{R}^N \end{cases}$$

where $A \in M(N,N)$, $B \in M(N,M)$, with control $u \in L^2(0,T;\mathbb{R}^M)$.
Given a matrix $C \in M(N,N)$, and some $x^* \in \mathbb{R}^N$, consider the optimal control problem

$$\min_u \; J^T(u) = \frac{1}{2}\int_0^T (|u(t)|^2 + |C(x(t)-x^*)|^2)dt\,.$$

There exists a unique optimal control $u(t)$ in $L^2(0,T;\mathbb{R}^M)$, characterized by the optimality condition

$$u = -B^*p\,, \qquad \begin{cases} x_t + Ax = -BB^*p \\ x(0) = x_0 \end{cases}, \qquad \begin{cases} -p_t + A^*p = C^*C(x-x^*) \\ p(T) = 0 \end{cases} \tag{3}$$

The steady state control problem

The same problem can be formulated for the steady-state model

$$Ax = Bu.$$

Then there exists a unique minimum $\bar{u}$, and a unique optimal state $\bar{x}$, of the stationary control problem

$$\min_u \; J_s(u) = \frac{1}{2}(|u|^2 + |C(x - x^*)|^2) \tag{4}$$

which is nothing but a constrained minimization in $\mathbb{R}^N$.
The optimal control $\bar{u}$ and state $\bar{x}$ satisfy

$$\bar{u} = -B^* \bar{p}\,, \qquad A\bar{x} = B\bar{u}\,, \quad \text{and} \quad A^* \bar{p} = C^* C(\bar{x} - x^*)\,.$$

We assume that

$$(A, B) \text{ is controllable,} \tag{5}$$

or, equivalently, that the matrices A, B satisfy the Kalman rank condition

$$Rank \left[B \ AB \ A^2 B \dots A^{N-1} B \right] = N \ . \tag{6}$$

Concerning the cost functional, we assume that the matrix C is such that (void assumption when $C = Id$)

$$(A, C) \text{ is observable} \tag{7}$$

which means that the following algebraic condition holds:

$$Rank \left[C \ CA \ CA^2 \dots CA^{N-1} \right] = N \ . \tag{8}$$

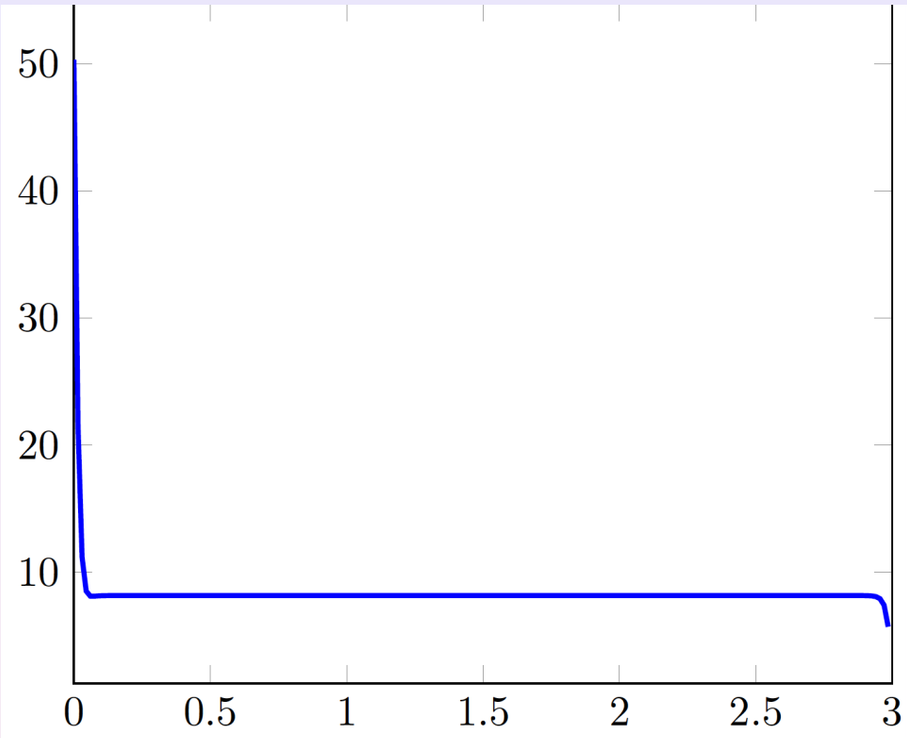
$$x_t + Ax = Bu$$

$$J^T(u) = \frac{1}{2} \int_0^T (|u(t)|^2 + |C(x(t) - x^*)|^2) dt$$

$$\begin{cases} x_t + Ax = Bu \\ -p_t + A^* p = C^* Cx \end{cases}$$

The turnpike property for the heat equation

New dynamics = combination of exponentially stable and unstable branches ≡ compatible with turnpike



The turnpike behaviour is ensured when $T \to \infty$ when the cost functional penalizes sufficiently state and control.

[Controllability] + [Coercive in state + control cost] → Turnpike

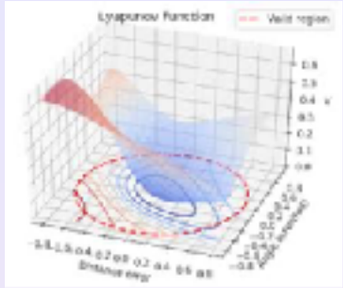
The same occurs for wave propagation[5]

[5] M. Gugat, E. Trélat, E. Zuazua, Systems and Control Letters, 90 (2016), 61-70.

Proofs

Proof # 1: Dissipativity

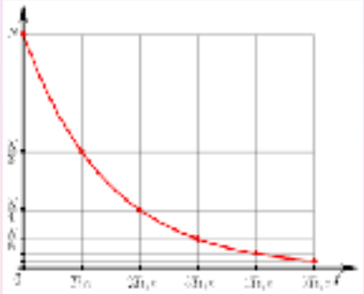
$$\frac{d}{dt}[(x-\bar{x})(p-\bar{p})] = -\left[B^*(p-\bar{p})|^2 + |C(x-\bar{x})|^2\right]$$



That is the starting point of a turnpike proof. Note however that it is much trickier than the classical Lyapunov stability: Two boundary layers at $t = 0$ and $t = T$, moving time-horizon $[0, T]$...

Proof #2 : Riccati

- Consider the Infinite Horizon Linear Quadratic Regulator (LQR) problem in $[0, \infty)$ with null target $x^* \equiv 0$.
- Employ Riccati feedback exponential stabilizator.
- Cut-it-off onto $[0, T]$.
- Correct the boundary layer at $t = T$ to match the terminal conditions .



Proof # 3: Singular perturbations Implement the change of variables $t \to sT$:

$$t \in [0, T] \Longleftrightarrow s \in [0, 1].$$

System

$$x_t + Ax = Bu, \quad t \in [0, T]$$

becomes

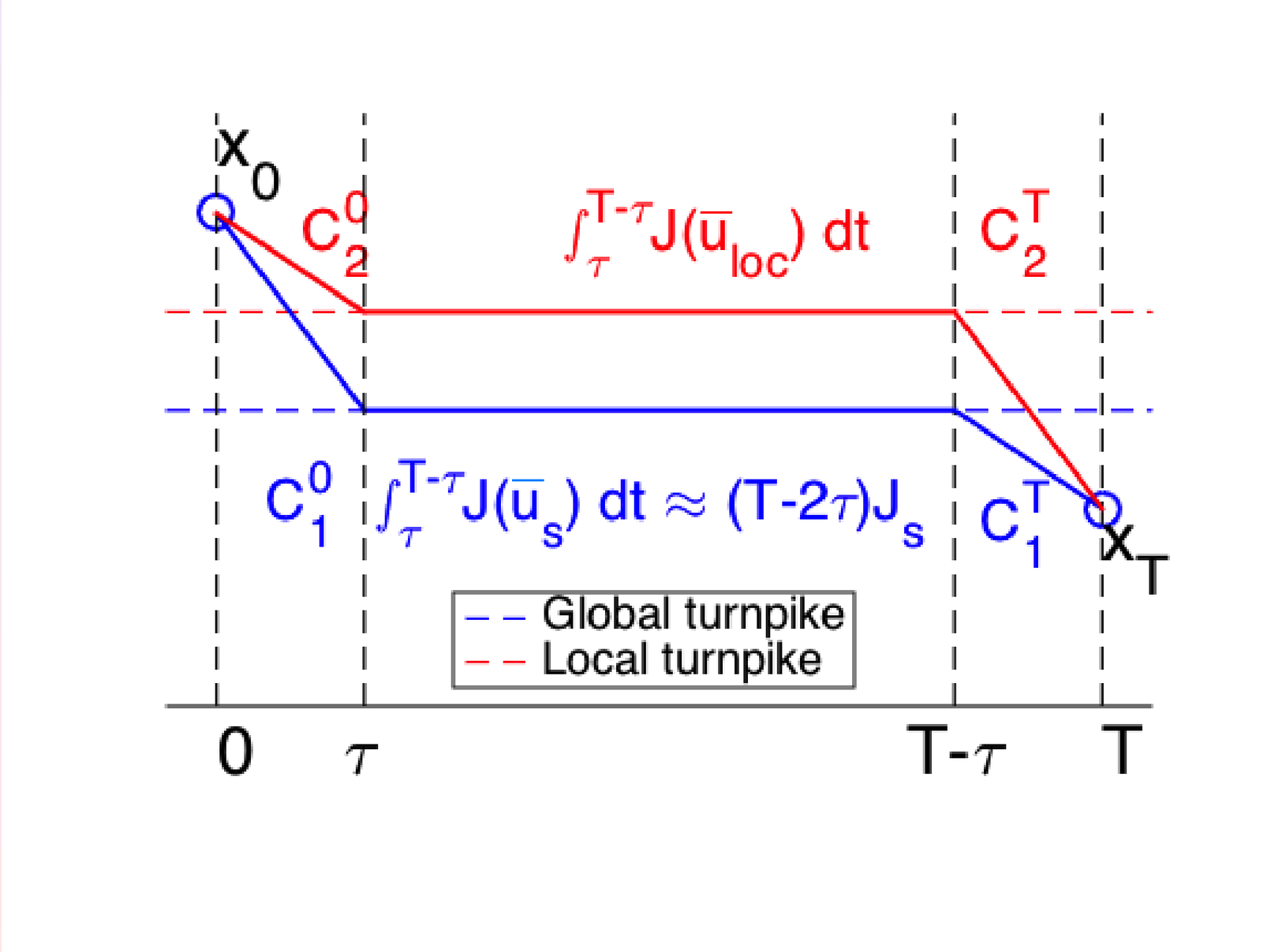
$$\frac{1}{T}x_s + Ax = Bu, \quad s \in [0, 1]$$

As $T \to \infty$, $\varepsilon = 1/T \to 0$.

Heuristic explanation and practical use

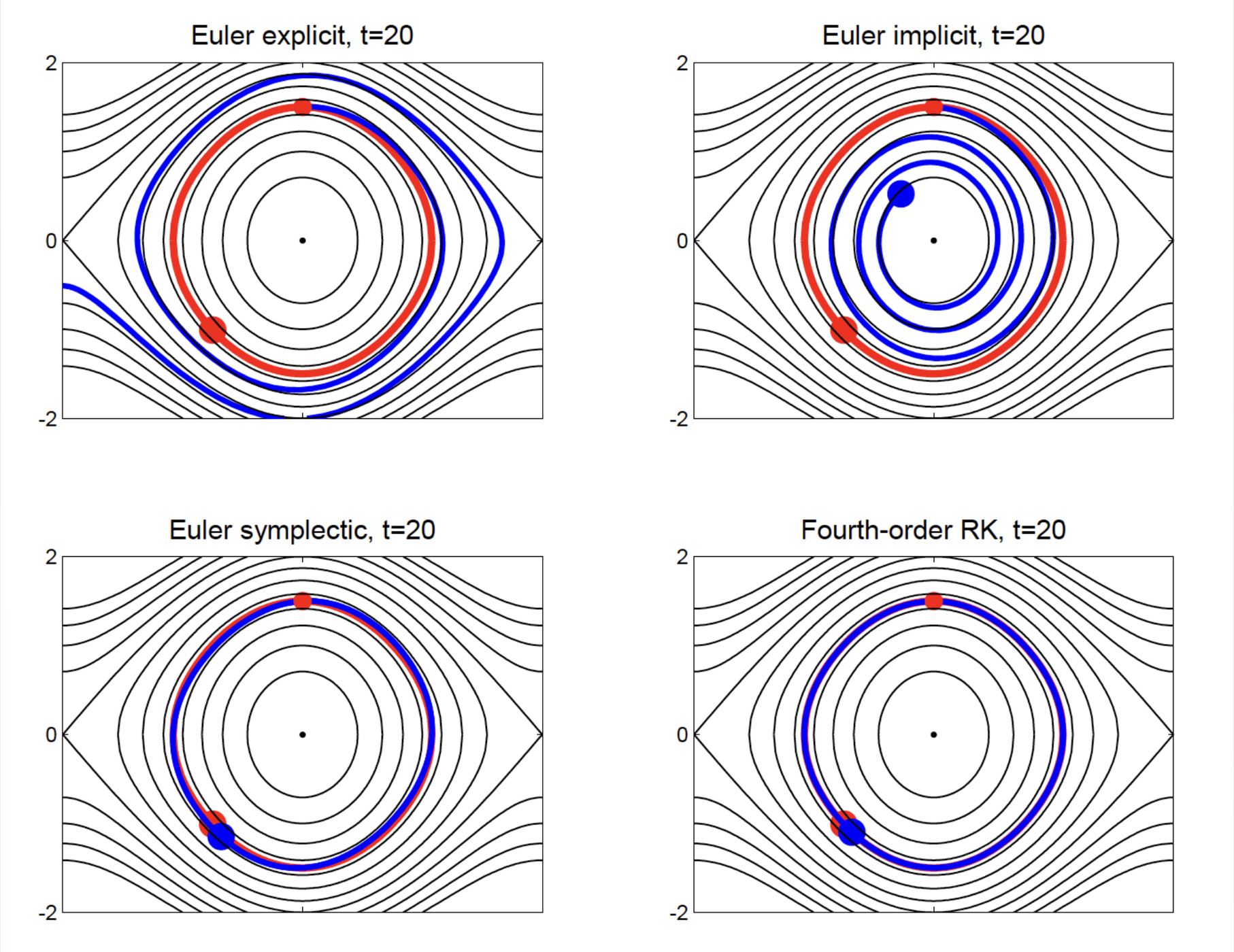
In applications and daily life we use a quasi-turnpike principle, which is very robust and ubiquitous, even in the context of multiple steady optima (local or global):

- Step 1: Compute the optimal steady optimal control and state,
- Step 2: Drive the system from the initial configuration to this steady state one;
- Step 3: Remain in this steady configuration as long as possible;
- Step 4: Exit this configuration if a terminal condition is to be met.

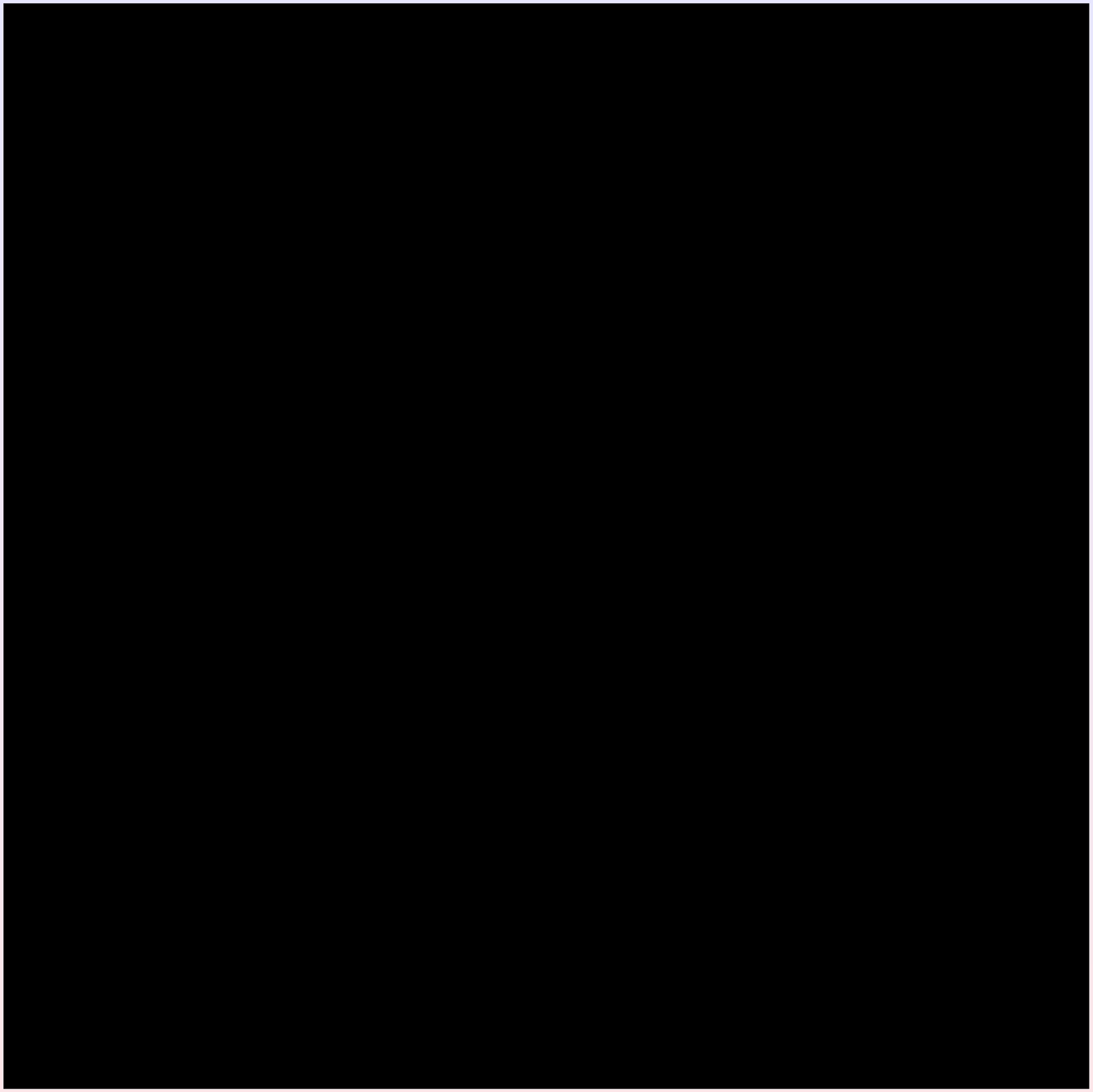


Warning! Long time numerics plays a key role: Geometric/Symplectic integration; Well balanced numerical schemes...

Numerical integration of the pendulum (A. Marica)



In shorter time-horizons, rather than turnpike we observe bang-bang and spike controls



R. Escobedo, A. Ibañez, E. Zuazua, 2016

Universal approximation theorem I

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

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

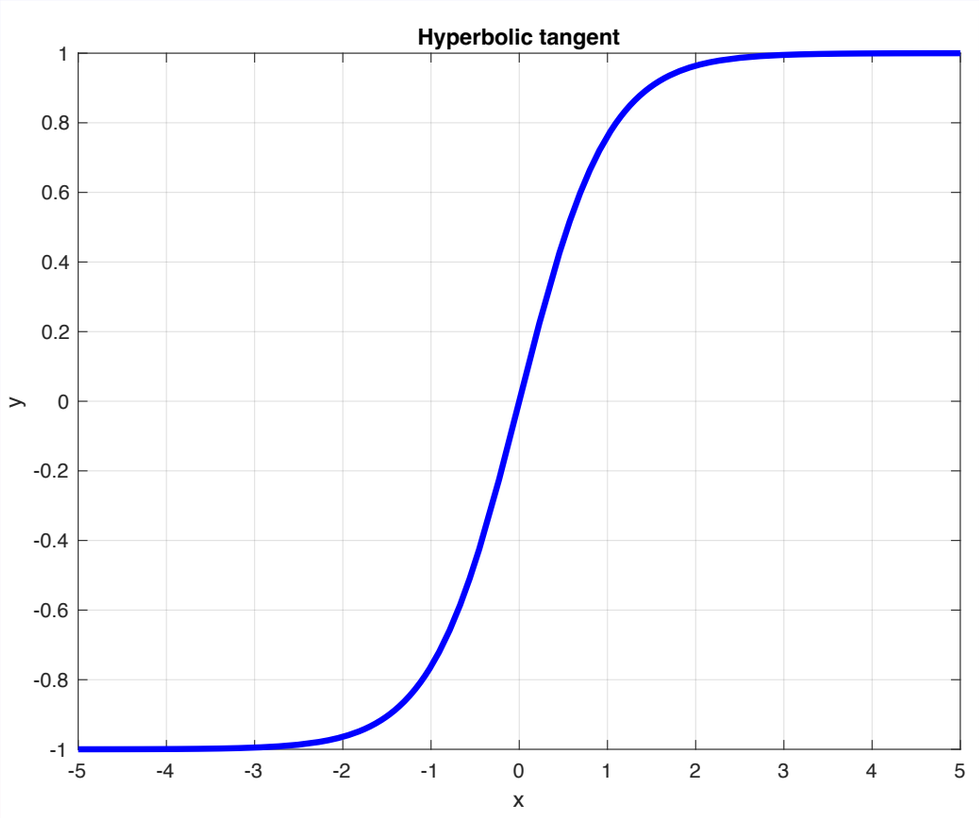
G. Cybenko†

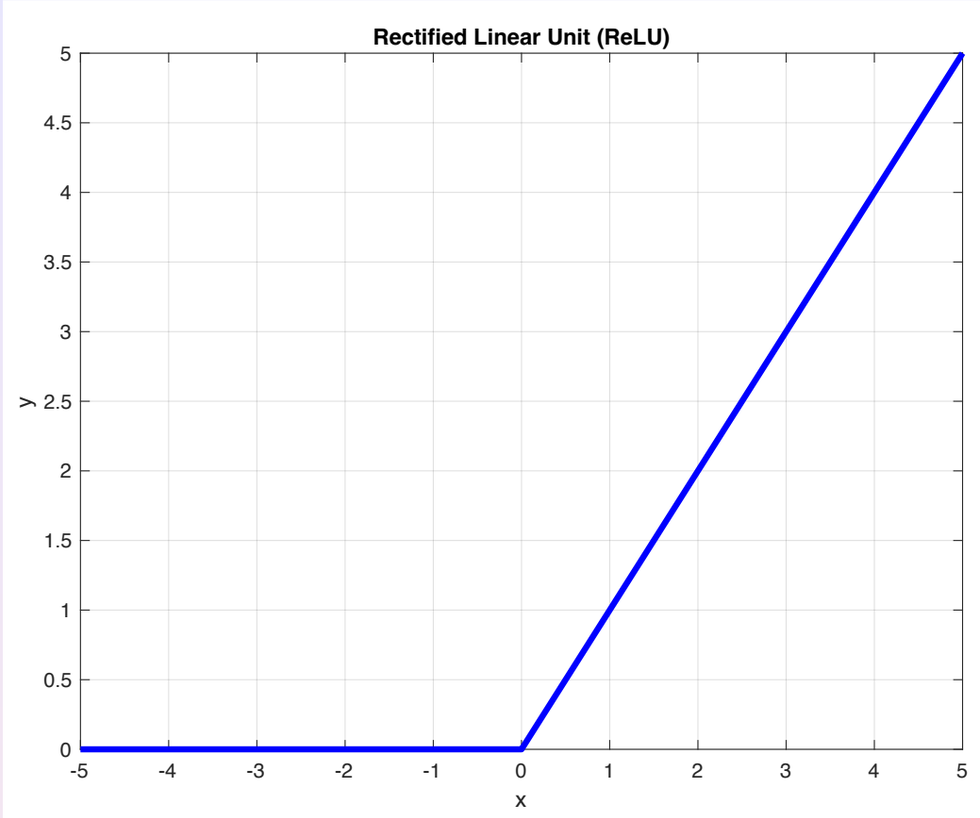
$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j), \tag{1}$$

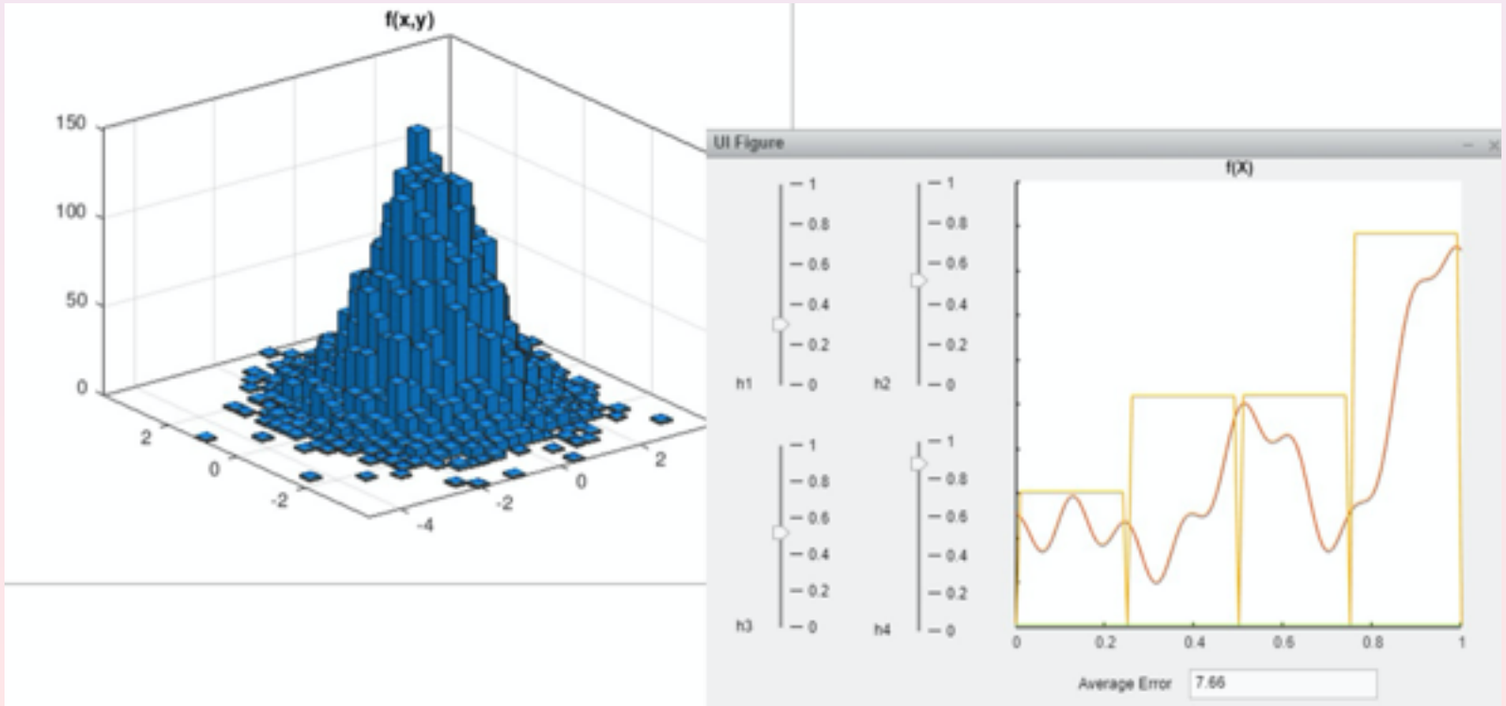
where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^{T} is the transpose of y so that $y^{\mathrm{T}}x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

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

Universal approximation theorem II







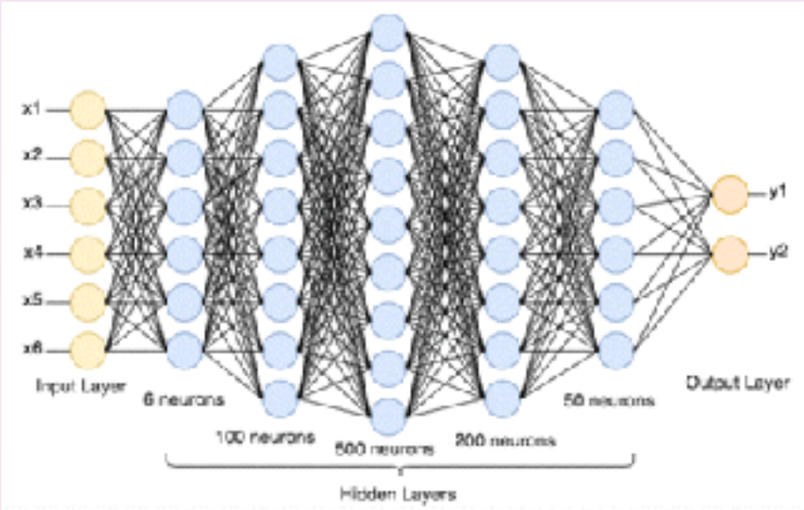
Supervised learning

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

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

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

Classification: match points (images) to respective labels (cat, dog).
$\longrightarrow$ Popular method: **training a neural network**.



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. Neural ordinary differential equations.
[4] E. Sontag, H. Sussmann, 1997, Complete controllability of continuous-time recurrent neural networks.

ResNets

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + hW^k\sigma(A^k\mathbf{x}_i^k + b^k), & k \in \{0, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \tilde{\mathbf{x}}_i, \quad i = 1, ..., N \end{cases}$$

where $h = 1$, σ globally Lipschitz $\sigma(0) = 0$.

nODE

Layer = timestep; $h = \frac{T}{N_{layers}}$ for given $T > 0$

$$\begin{cases} \dot{\mathbf{x}}_i(t) = W(t)\sigma(A(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{\mathbf{x}}_i, \quad i = 1, ..., N \end{cases}$$

The problem becomes then a giant simultaneous control problem in which each initial datum $\mathbf{x}_i(0)$ needs to the driven to the corresponding destination for all $i = 1, ..., N$ with the same controls:

- What happens when $T \to \infty$, i.e. in the deep, high number of layers regime?[8] [9]

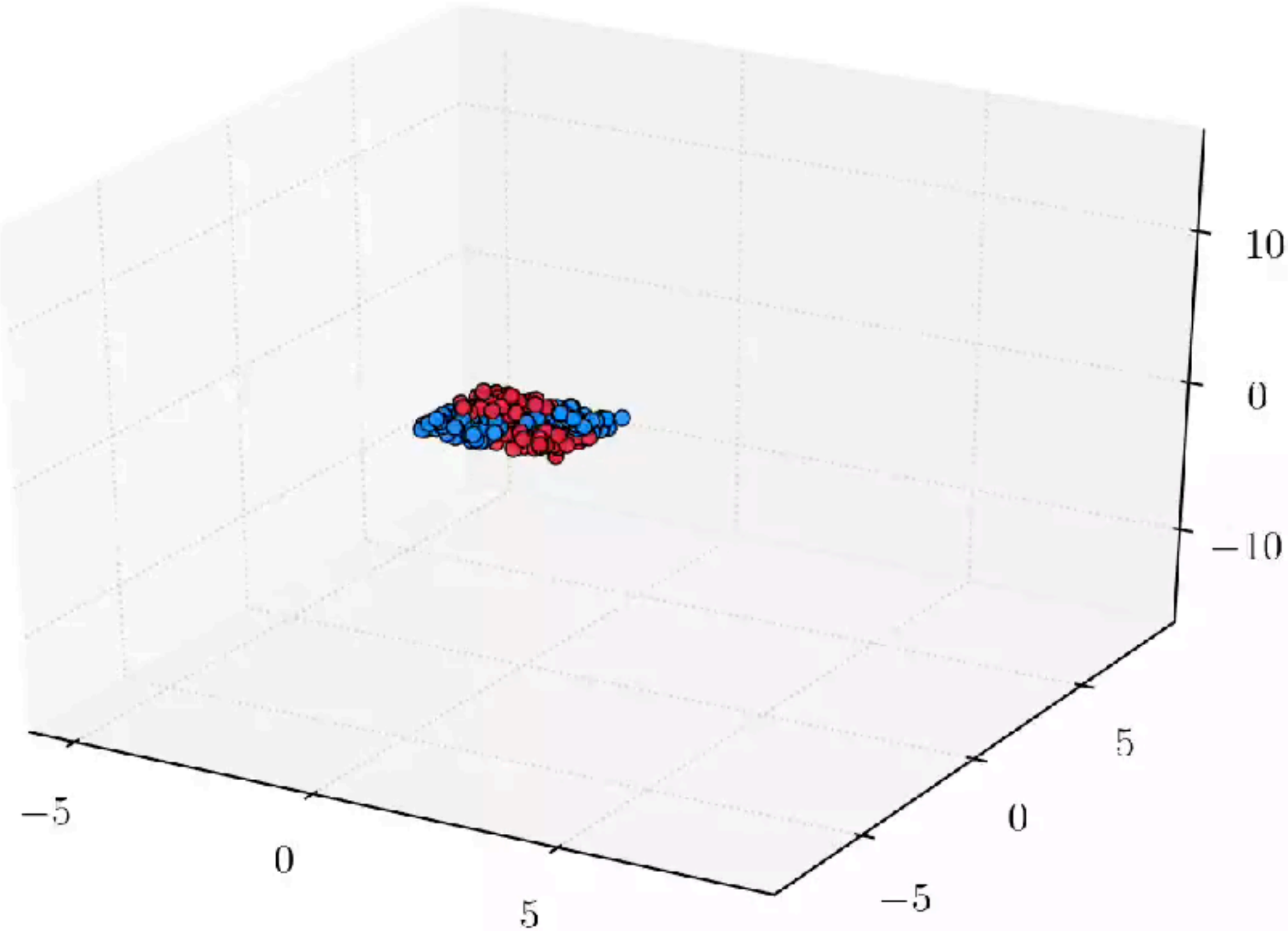
[8] C. Esteve, B. Geshkovski, D. Pighin, E. Zuazua, Large-time asymptotics in deep learning, arXiv:2008.02491
[9] D. Ruiz-Balet & Zuazua, Neural ODE control for classification, approximation and transport, arXiv:2104.05278

Special features of the control of ResNets

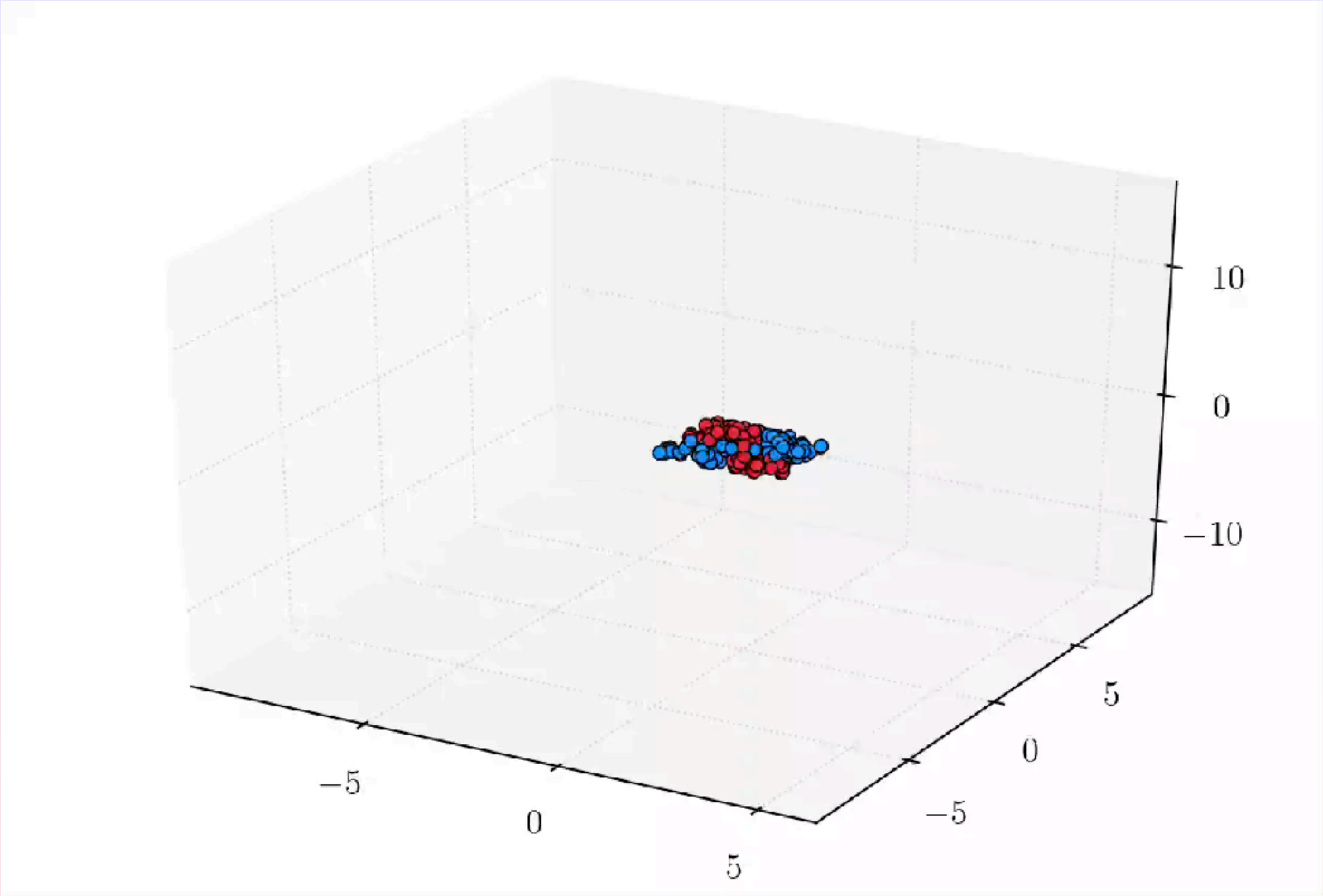
- Nonlinearities are unusual in Mechanics: σ is flat in half of the phase space.
- We need to control many trajectories (one per item to be classified) with the same control! [10]

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

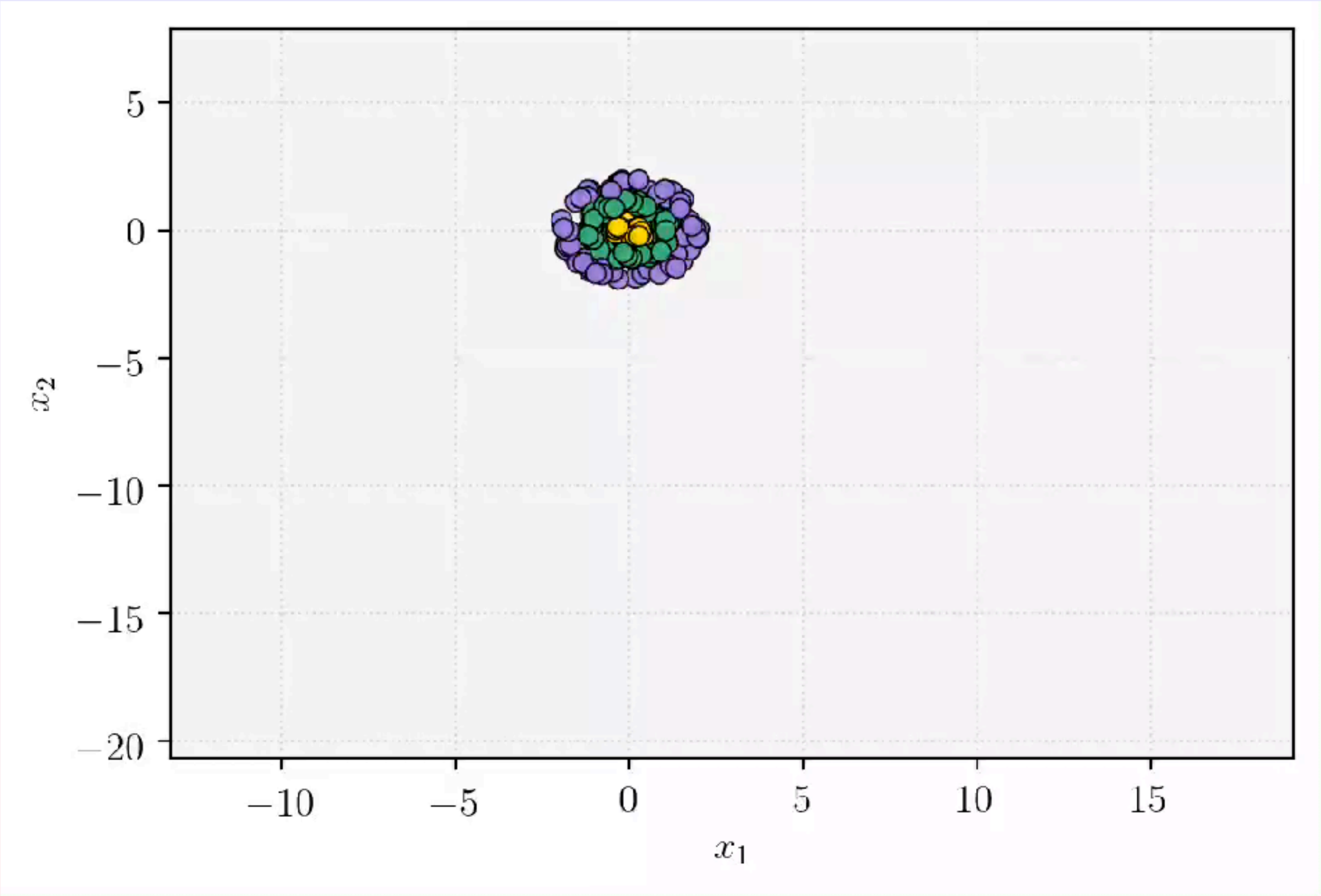
[10] This would be impossible for instance, for the standard linear system $x' = Ax + Bu$.



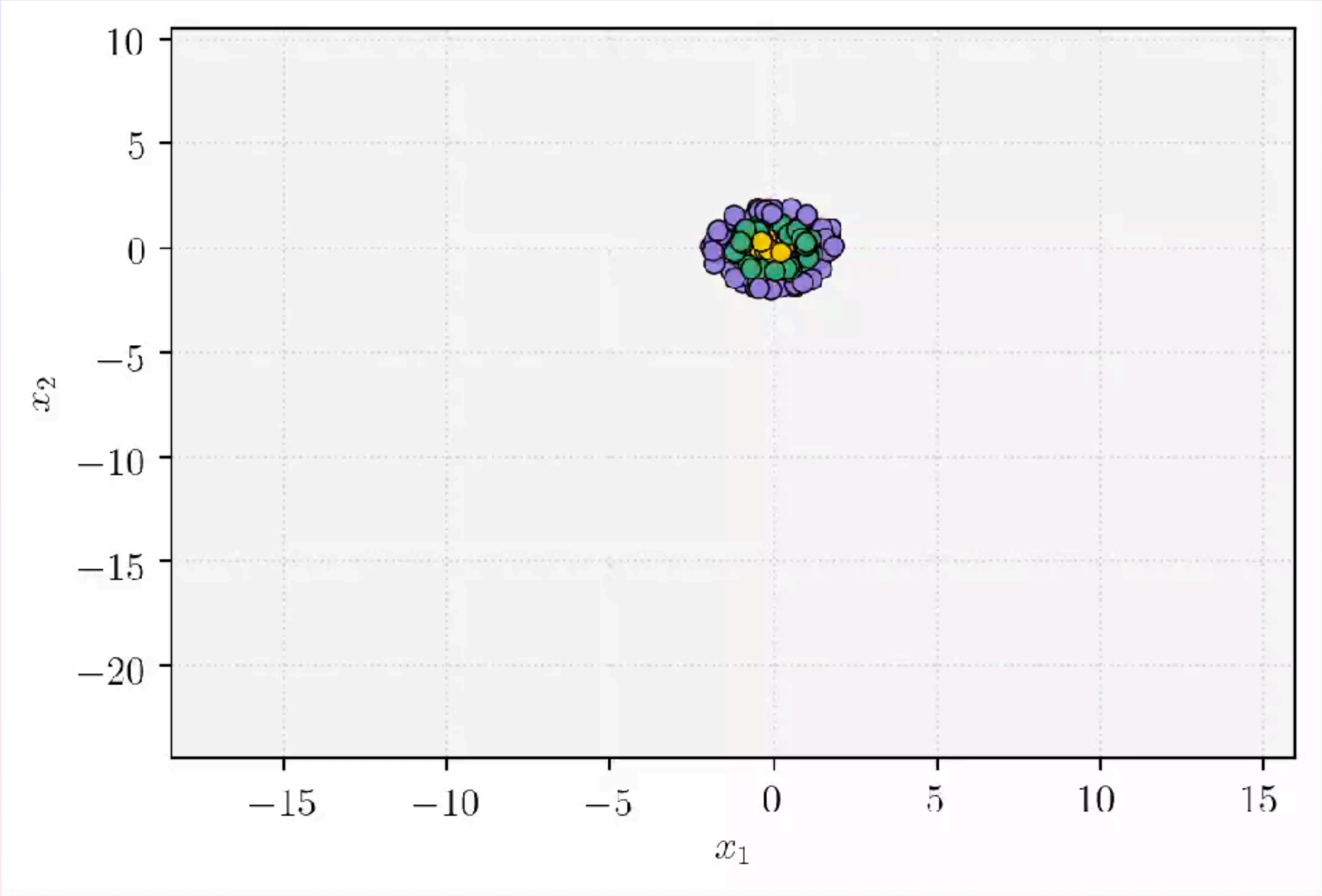
Borjan Geshkovski, MIT



Borjan Geshkovski, MIT



Borjan Geshkovski, MIT



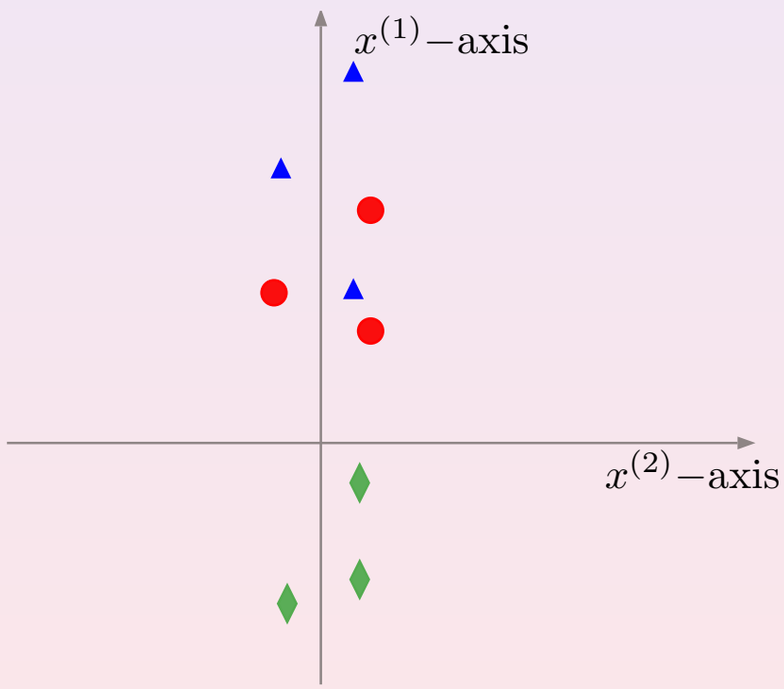
Borjan Geshkovski, MIT

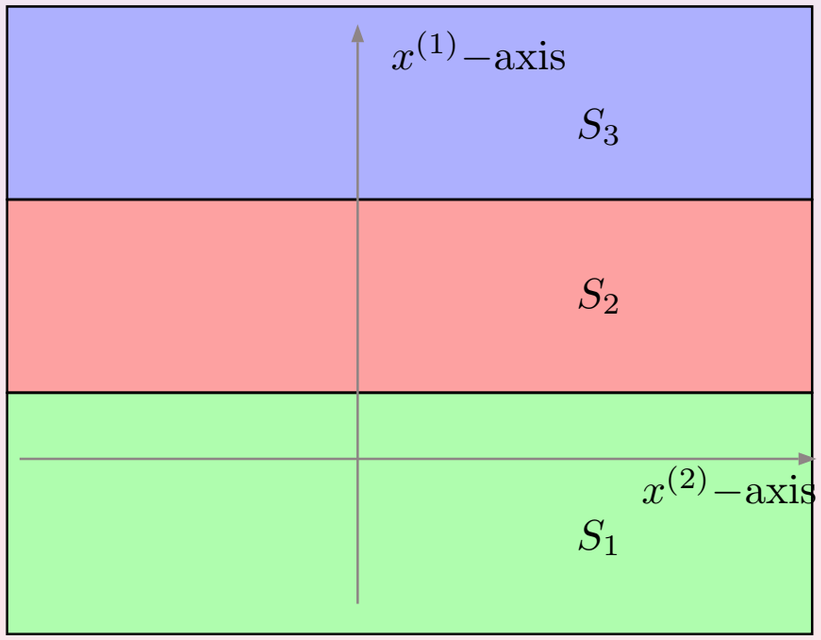
What is actually the ResNet doing?

The classification problem is a relaxed version of the simultaneous control problem. We are given N points in $\mathbb{R}^d$ and M classes $y_i \in \{1, ..., M\}$.

We then proceed as follows:

1. We identify a region in the euclidean space corresponding to each class of data.
2. Look for a control strategy (A, W, b) bringing simultaneously all points to the location corresponding to its class.



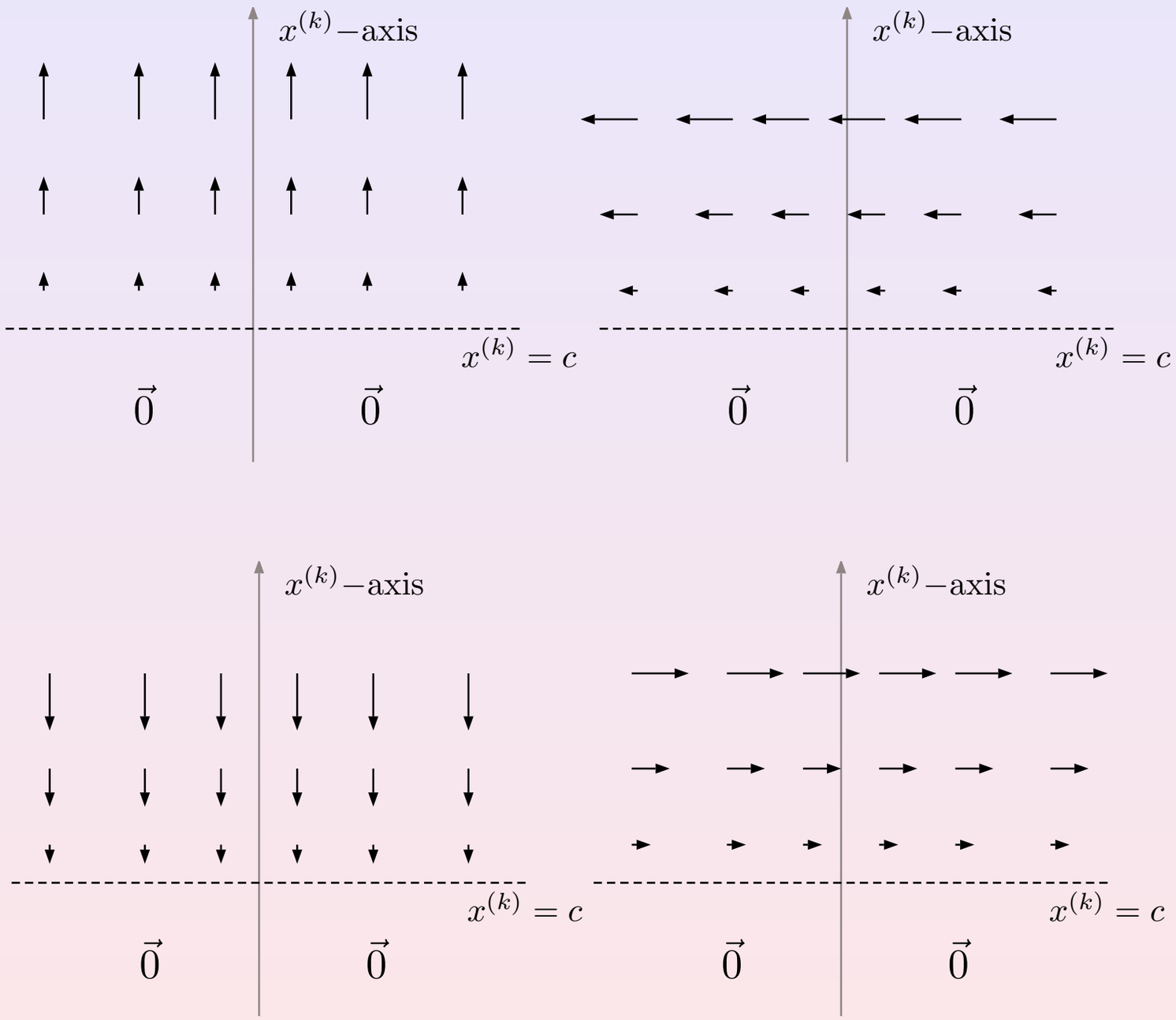


Basic control actions

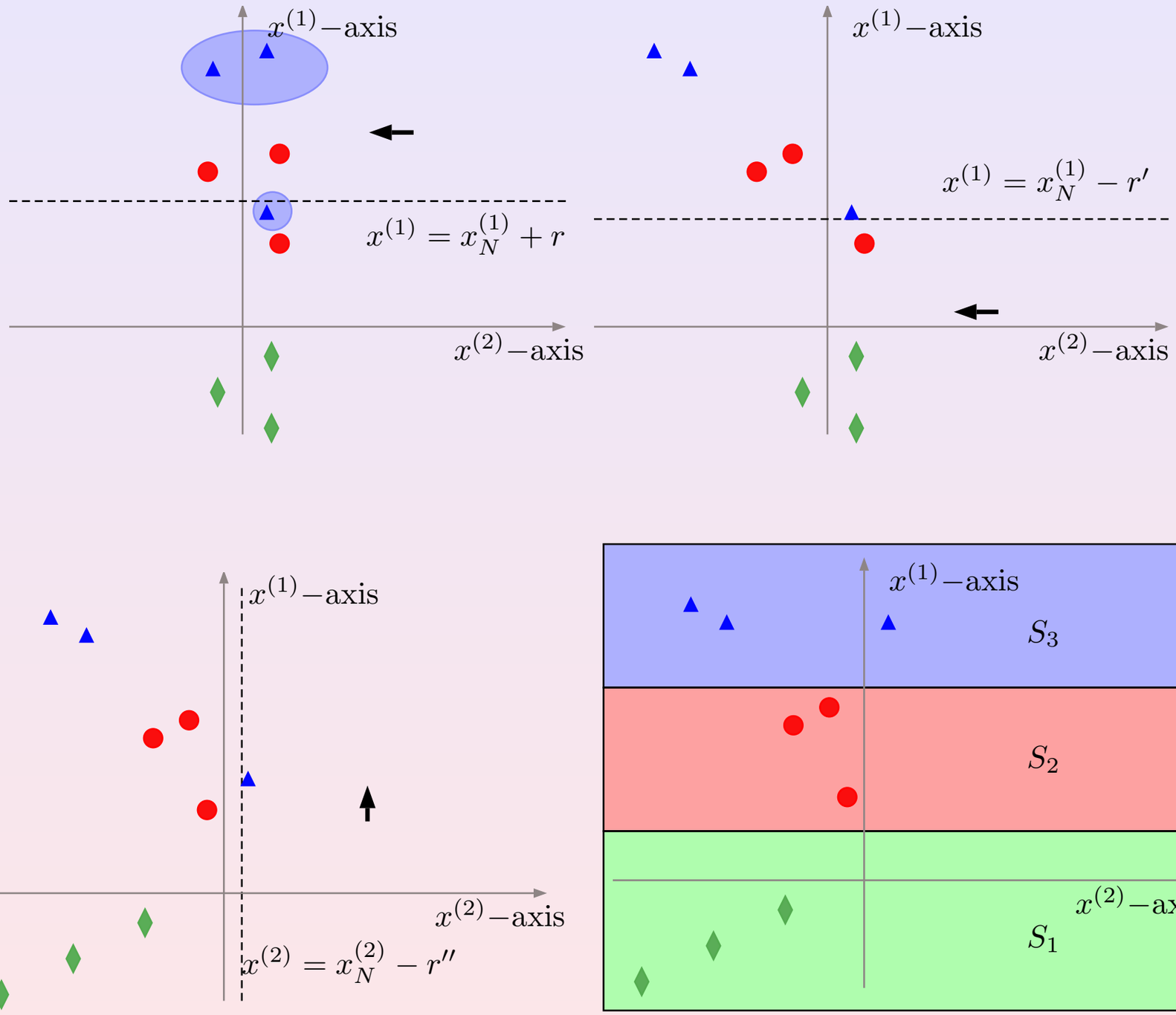
$$\dot{\mathbf{x}}(t) = W(t)\sigma(A(t)\mathbf{x}(t) + b(t)).$$

- $b(t)$ induces a time-dependent translation of the Euclidean space. It plays an important role to place the center of the action of the sigmoid.
- $A(t)$ compresses, expands, and induces rotations in the euclidean space with different objectives:
  - Compression can help gathering data into clusters so that they might be manipulated simultaneously.
  - Expansion allows to separate data of different classes to better focus the action of the control on just one of them.
  - Rotations allow to better choose the hemisphere where the action will be focused.
- $W(t)$ determines the direction and intensity with which the flow will evolve in the active hemisphere.

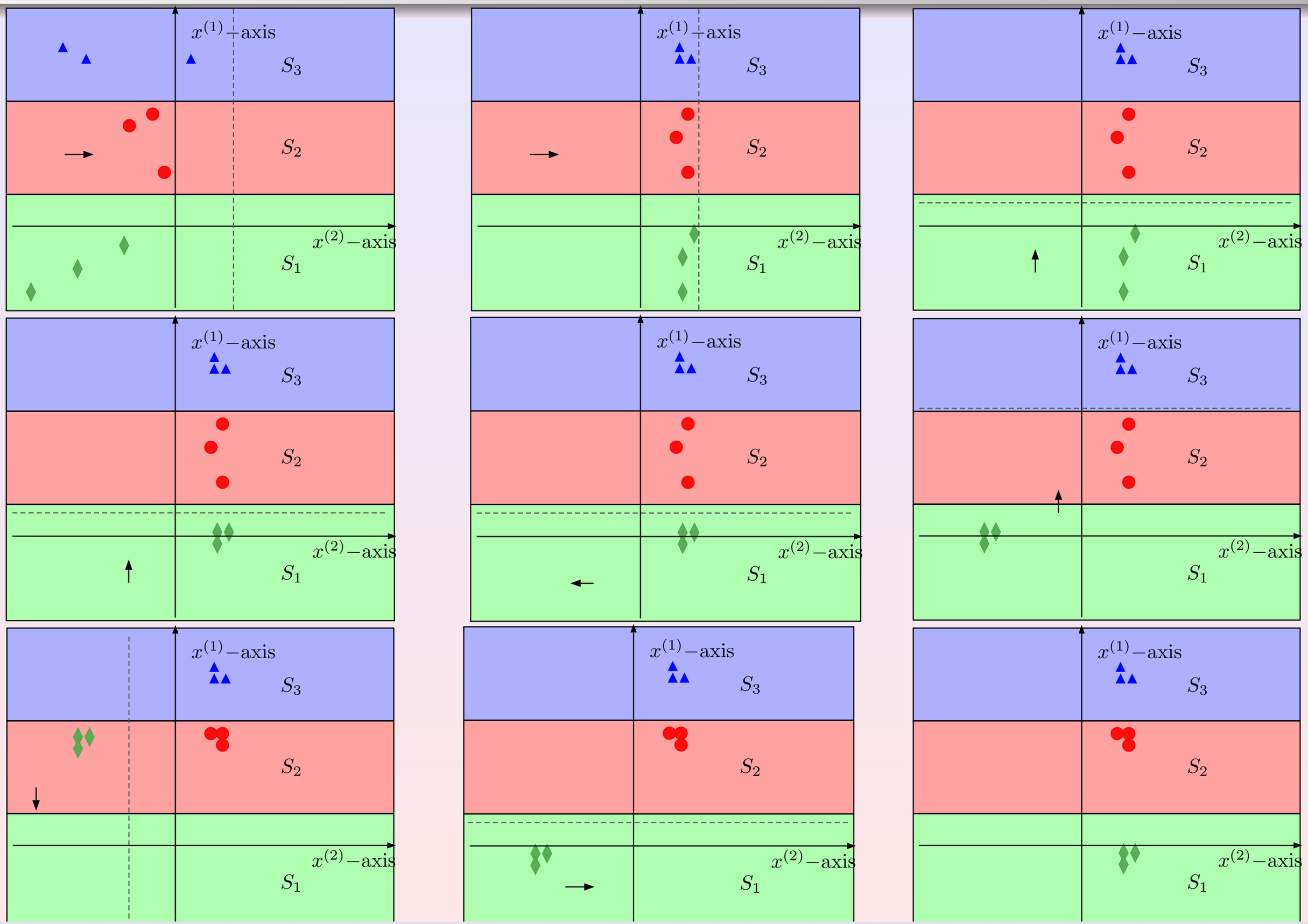
Some canonical flows induced by nODE

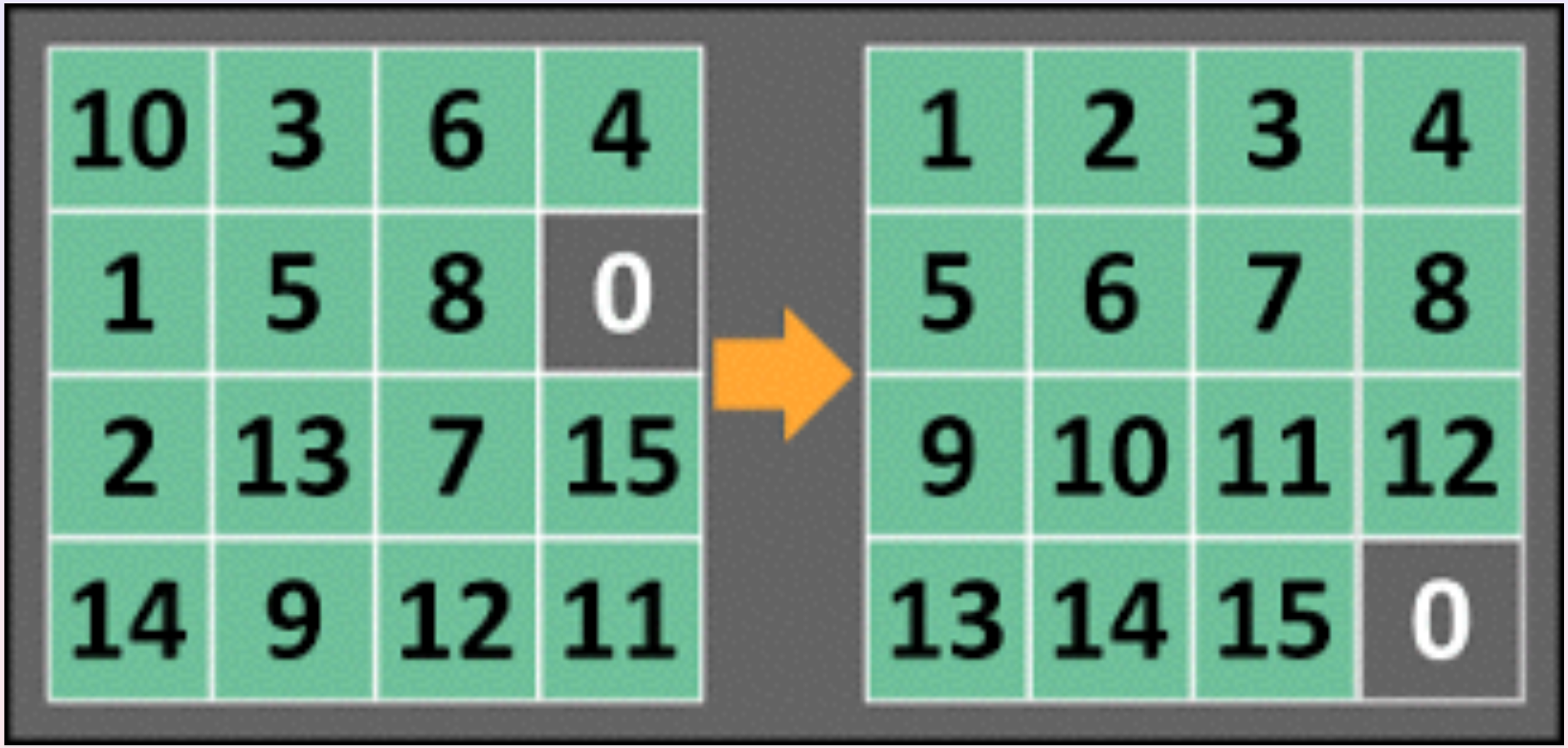


Controlling one datum



Compression after classification




10 3 6 4
1 5 8 0
2 13 7 15
14 9 12 11
1 2 3 4
5 6 7 8
9 10 11 12
13 14 15 0

Classification by control

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

[a] *Let σ be the ReLU.*
Let $d \geq 2$, 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 $A, W \in L^\infty\left((0,T); \mathbb{R}^{d\times d}\right)$ and $b \in L^\infty\left((0,T), \mathbb{R}^d\right)$ 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.*
- *The complexity of controls also diminishes when the control requirement is relaxed so that not all data need to be classified.*
- *The targets S_m can be just N distinct points in the euclidean space.*

[a] Related results for smooth sigmoids using Lie bracket control techniques: A. Agrachev and A. Sarychev, arXiv:2008.12702, (2020).

Neural transport

The simultaneous control of the nODE

$$\begin{cases} \dot{x} = W(t)\boldsymbol{\sigma}(A(t)x + b(t)) \\ x(0) = x_i, \quad i = 1, ..., N \end{cases}$$

to arbitrary terminal states

$$x(T) = y_i, \quad i = 1, ..., N$$

in terms of the transport equation, leads to the control of an atomic initial datum from

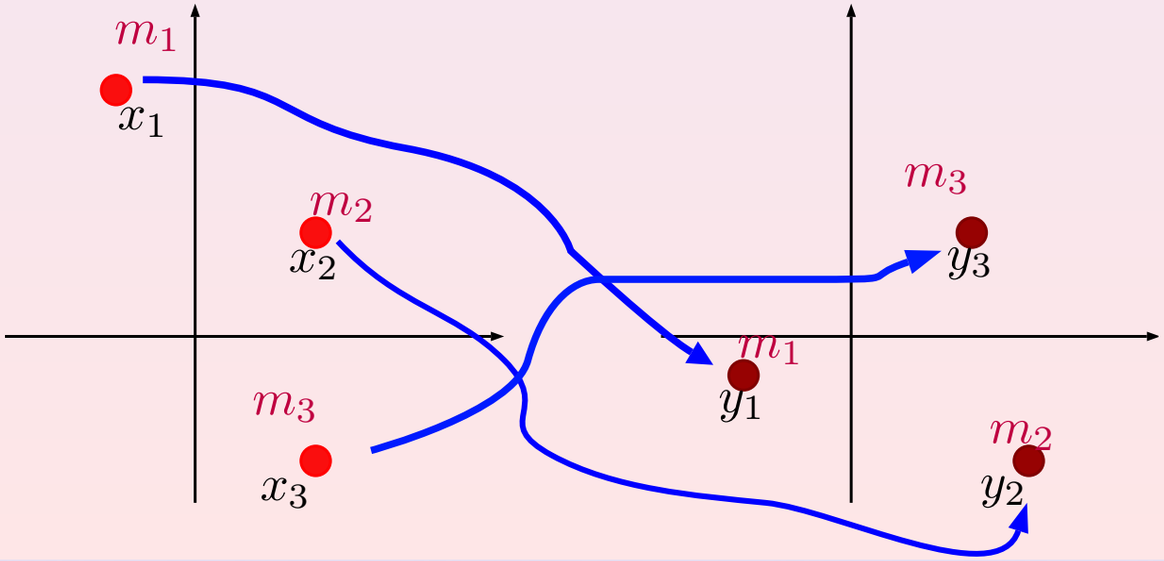
$$\rho(x, 0) = \sum_{i=1}^{N} m_i \delta_{x_i}$$

to the terminal one

$$\rho(x, T) = \sum_{i=1}^{N} m_i \delta_{y_i}.$$

But note that, even if the locations of the masses are transported, the amplitude of the masses do not vary.

Optimal Transport
Monge-Kantorovich

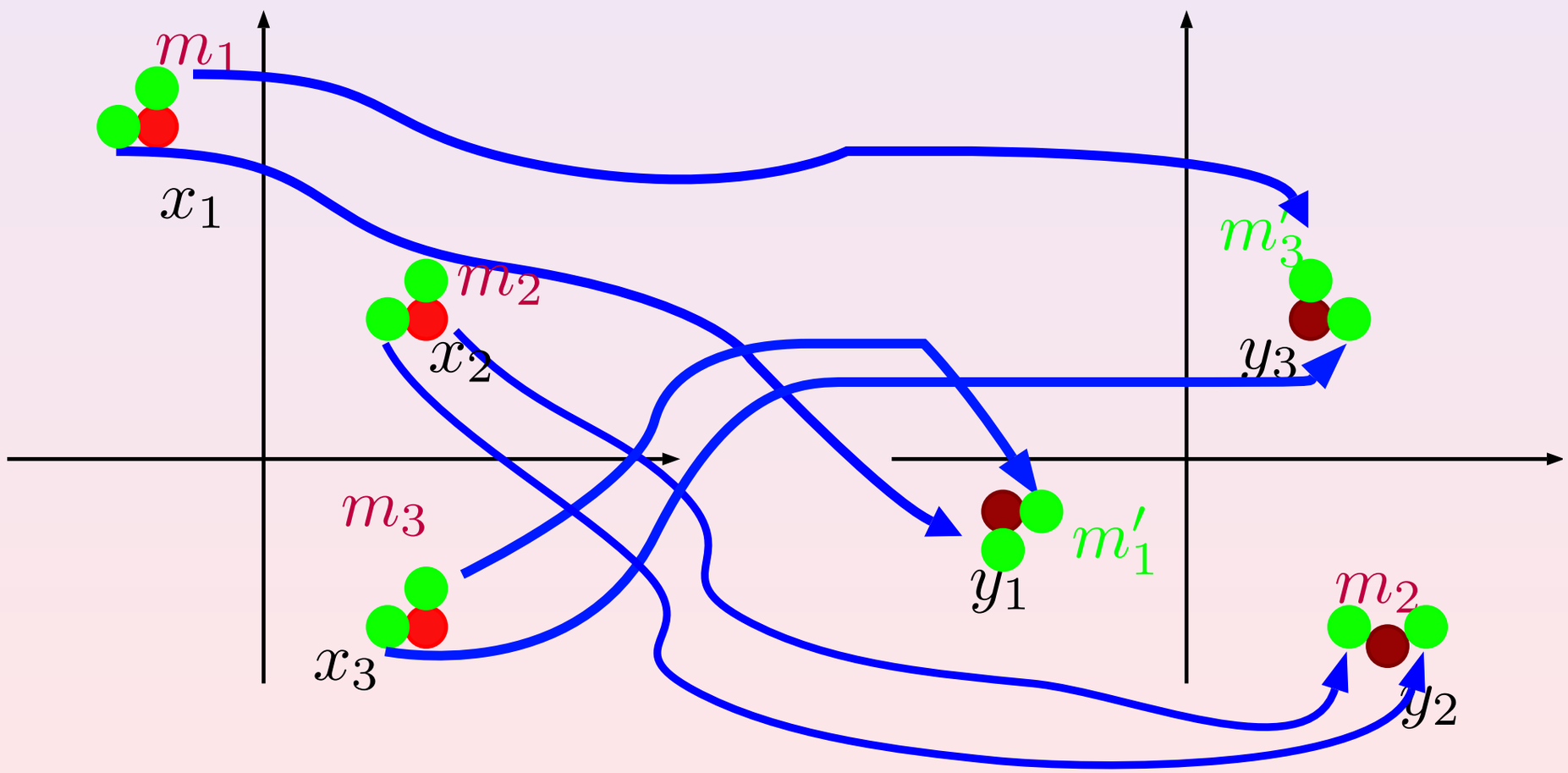


Neural transport

We can enrich the strategy above to also regulate the amplitude of the masses. But this requires to relax the control statement into an ε-approximate one.
For that to be done we need to split initial masses so that

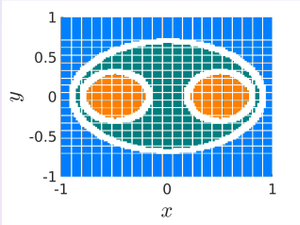
$$m_i = \sum_{j=1}^{J_i} m_{i,j}, \quad i = 1, ..., N$$

they are dispersed from the center x_i into the neighboring points $x_{i,j}$. This allows to enrich the transport diagram.

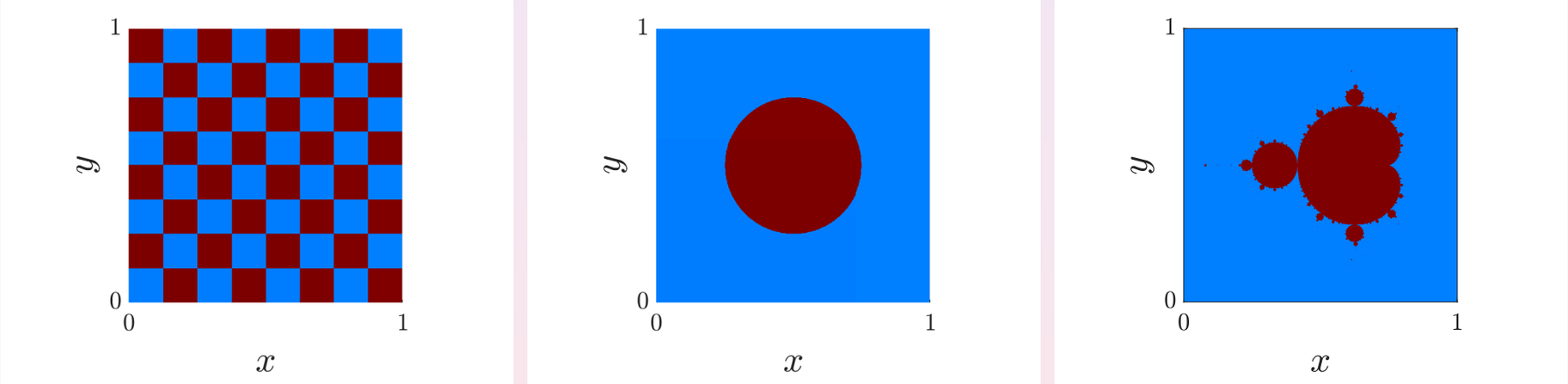


Universal approximation

- As a corollary we can achieve Universal Approximation.
- By density, it is sufficient to consider targets that are simple piecewise constant functions.
- We can proceed making a partition of the departure and arrival spaces so that the problem becomes a countable version of simultaneous control of the nODE.



- The complexity of the needed controls depends on the nature of functions one aims to approximate.[13]



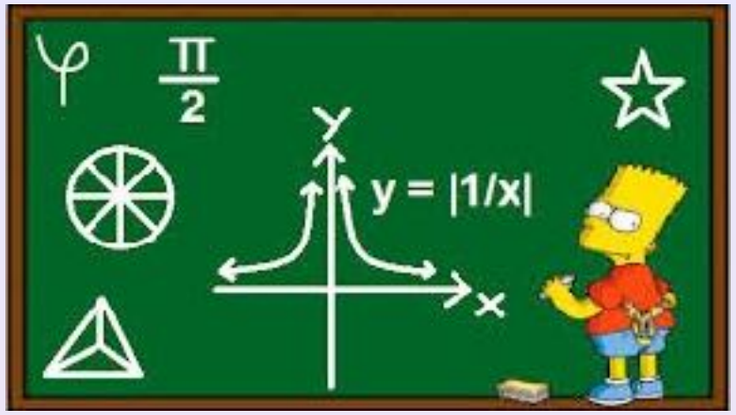
[13] $N_\Gamma(h)$ being the number of hypercubes of side h needed to cover the boundary Γ, the box-counting dimension is

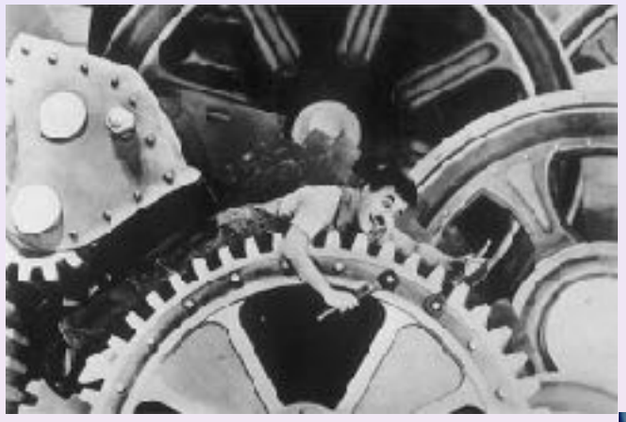
$$D := \lim_{h \to 0} \left[\log N_\Gamma(h) / \log \left(h^{-1} \right) \right].$$

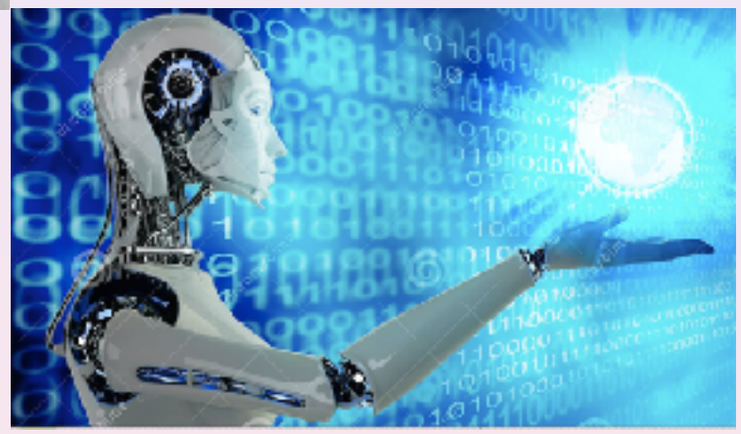
Then

$$\|W\|_{L^\infty} \lesssim \epsilon^{-\frac{4Dd}{d-D}}, \qquad \|b\|_{L^\infty} \lesssim \epsilon^{-\frac{-2dD}{d-D}} \qquad \text{as } \epsilon \to 0$$

and the number of switches of A, W, b will be of the order of $\epsilon^{-\frac{2dD}{d-D}}$.


y = |1/x|





Concluding remarks

An extraordinary and fertile field in the interplay between Dynamical Systems, Control, Machine Learning and applications

- Control and dynamical systems tools allow to explain the amazing efficiency of Neural Networks (NN) in some specific applications.
- Long-time / Turnpike control arise naturally in Deep Learning
- Interesting open questions:
  - How to deal with Neural ODEs that switch in dimension of the Euclidean phase space.
  - Are there results explaining how the clustering of data (number of separating interfaces needed) diminishes in higher dimensions?
  - How close is our piecewise constant control strategy from the optimal one (in the Pontryagin sense?)
  - How does our control strategy compare to those obtained in a purely NN setting?
  - How does the complexity of the controls diminish when we relax the classification criteria?
  - Links with Optimal Transport.
  - Other objectives: Unsupervised learning?

- My sincere thanks to collaborators: A. Porretta (Roma 2), E. Trélat (Paris Sorbonne), M. Gugat (FAU), D. Pighin (Sherpa.ai), B. Geskhovski (UAM & Deusto), C. Esteve (UAM & Deusto), M. Schuster (FAU), M. Lazar (Dubrovnik), V. Hernández-Santamaria (Mx), N. Sakamoto (Nanzan), J. Heiland (Magdeburg), H. Kouhkouh (Padova), D. Ruiz-Balet (UAM

THANK YOU FOR THE INVITATION AND YOUR ATTENTION!