Deusto

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Universidad Autónoma de Madrid

Control, Cibernética e Inteligencia Artificial

Enrique Zuazua

Centro de Ciencias de Benasque, 23.08.2022

European Research Council
Established by the European Commission

Unterstützt von / Supported by
Alexander von Humboldt
Stiftung / Foundation

Algunas cuestiones habituales

¿Las Matemáticas para qué sirven?

¿Matemática o Matemáticas?

¿Las Matemáticas no están ya todas hechas, acabadas? ¿Para qué seguir en la investigación Matemática?

¿Qué aportan las Matemáticas?

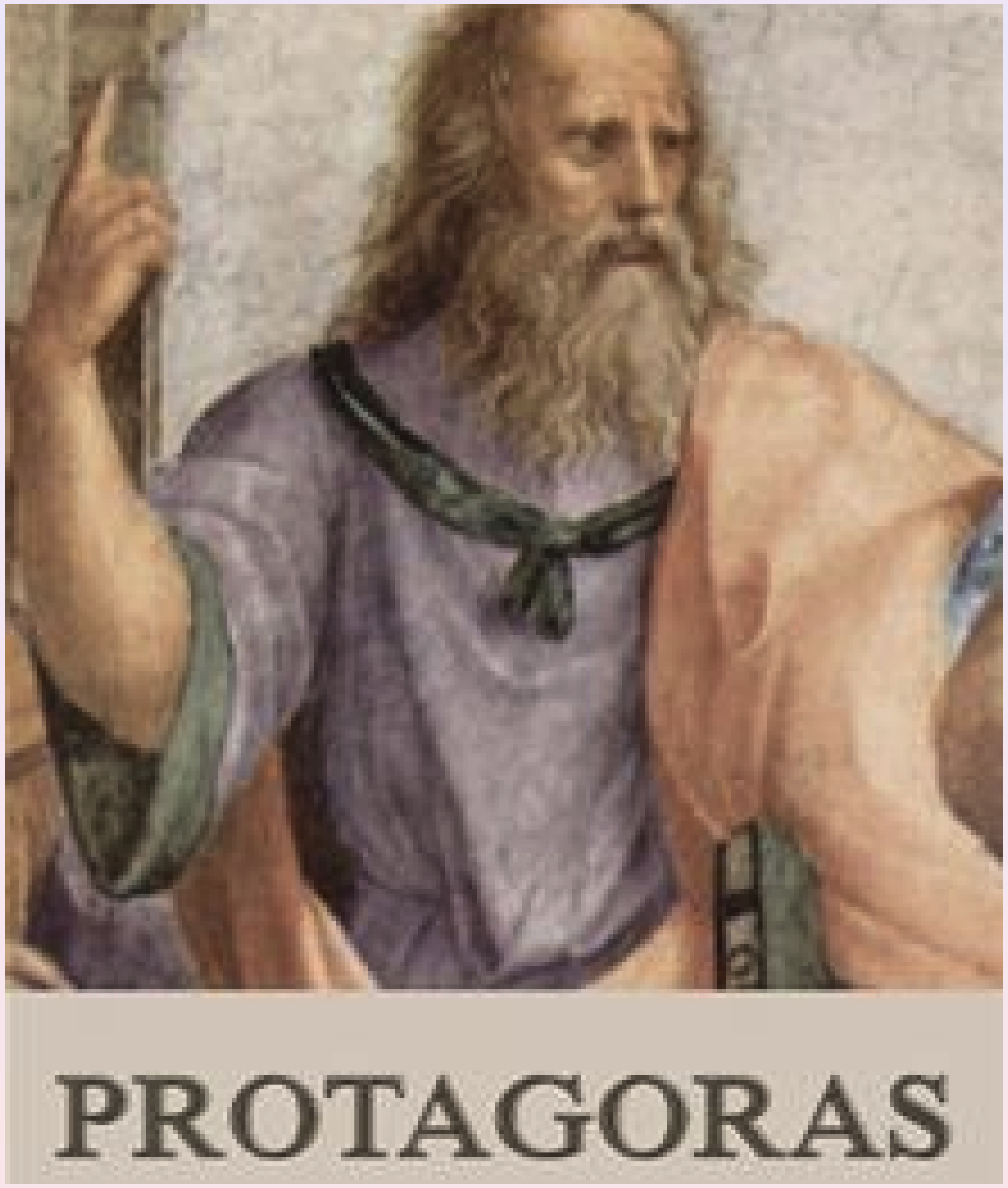
¿Qué salidas laborales ofrecen las Matemáticas?

¿Son necesarias las Matemáticas en el sistema educativo?

Table of Contents

Hombre/Mujer = unidad de medida

El ser humano es la medida de todo



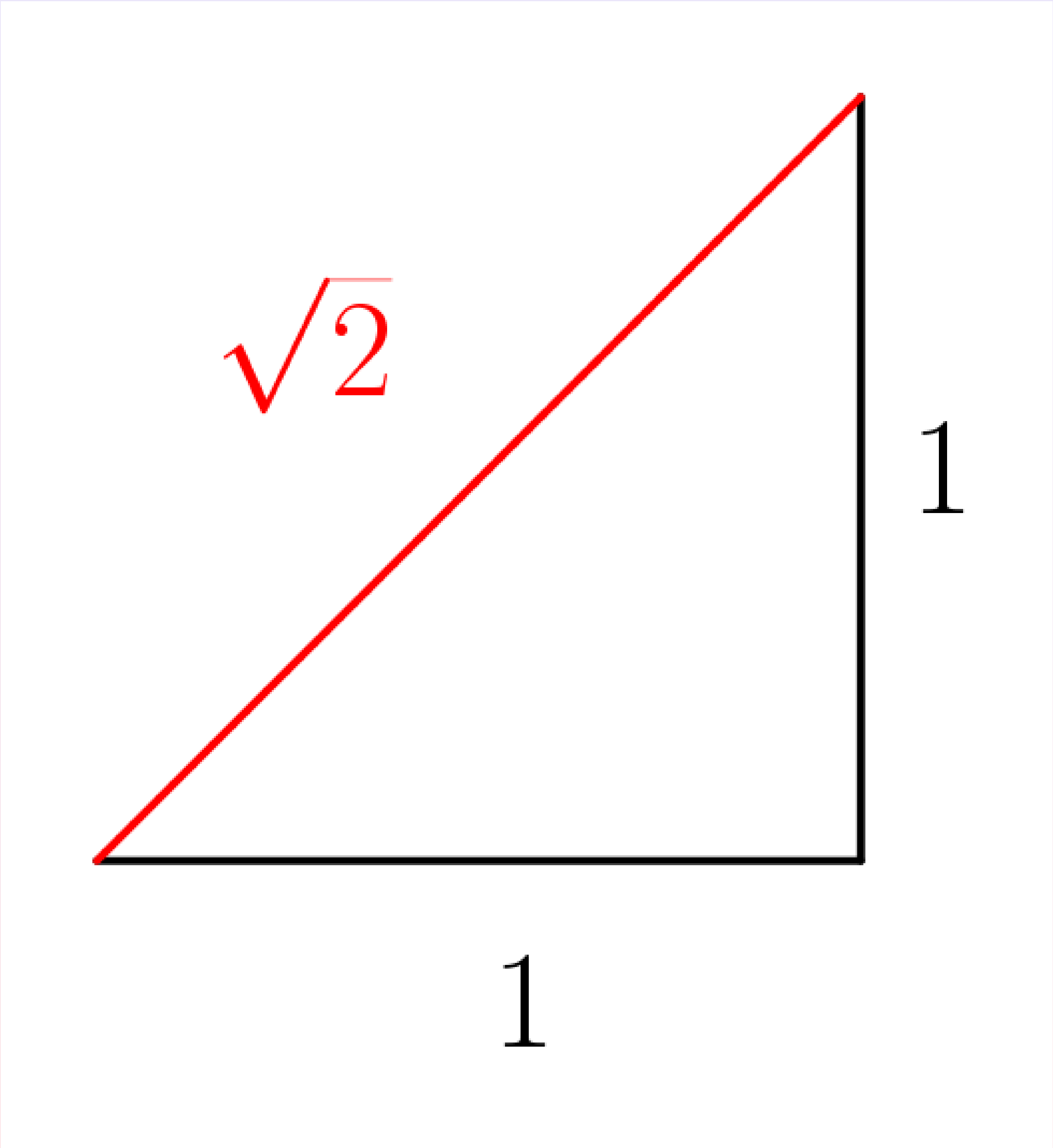
Protágoras d'Abdera (485 a. C. - 411 a. C.)

Pensador nómada, útil allí donde fuera

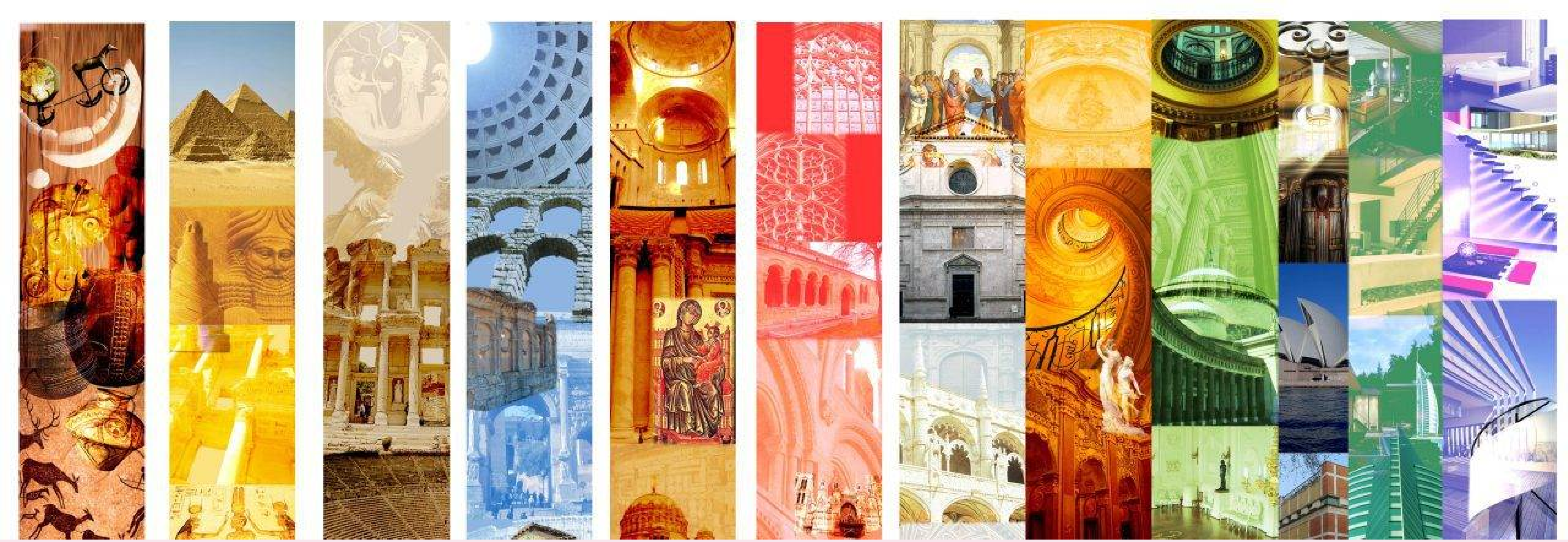
Pitagóricos

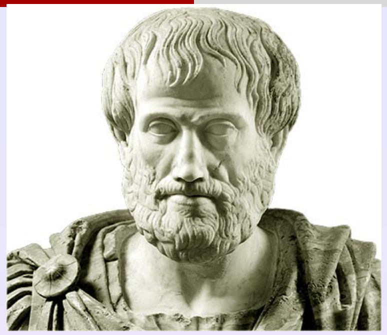
Escuela pitagórica: formada por astrólogos, médicos, matemáticos y filósofos, cuya creencia más importante era que todas las cosas son, en esencia, números (siglo VI a.C.).

Matemáticas: Exactas ?



De civilización en civilización





*"... 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."*

Chapter 3, Book 1, of the monograph "Politics", 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.

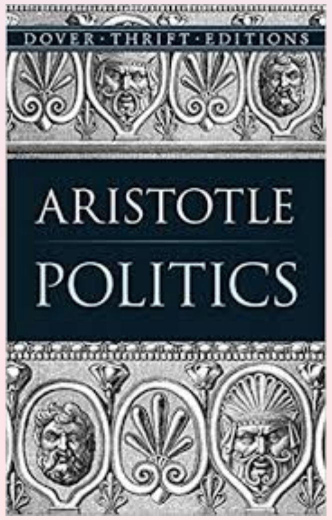
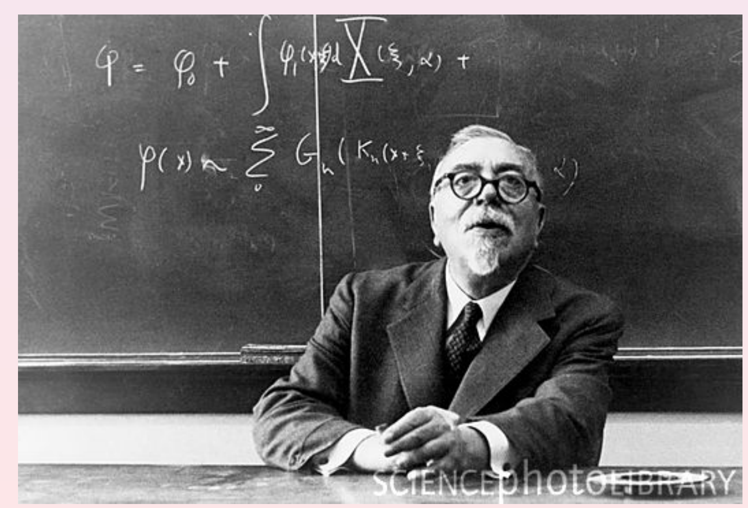


Table of Contents

"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".

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.



Robotic arm

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

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

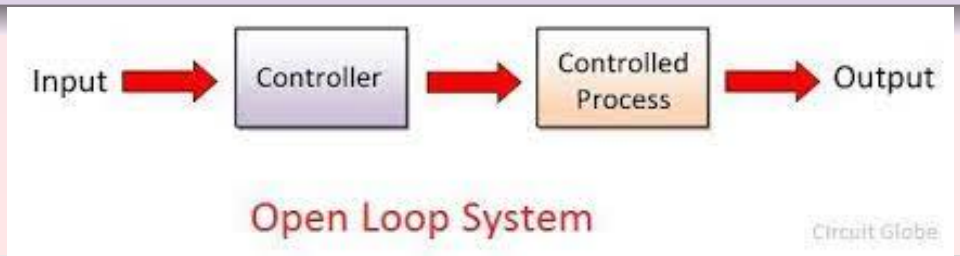
In (1), A is a $n \times n$ real matrix, B is of dimensions $n \times m$ and x^0 is the initial sate of the sytem in $\mathbb{R}^n$. The function $x : [0, T] \longrightarrow \mathbb{R}^n$ represents the*state* and $u : [0, T] \longrightarrow \mathbb{R}^m$ the *control*.

¿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*

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



Proof:

From the variation of constants formula:

$$x(t) = e^{At}x^0 + \int_0^t e^{A(t-s)}Bu(s)ds = e^{At}x^0 + \int_0^t \sum_{k\geq 0} \frac{(t-s)^k}{k!} A^k Bu(s)ds.$$

By Cayley[1]-Hamilton's [2] theorem A^k for $k \geq n$ is a linear combination of $I, A, ..., A^{n-1}$.

[1] Arthur Cayley (UK, 1821 - 1895)
[2] William Rowan Hamilton (Ireland, 1805 - 1865)

Proof:

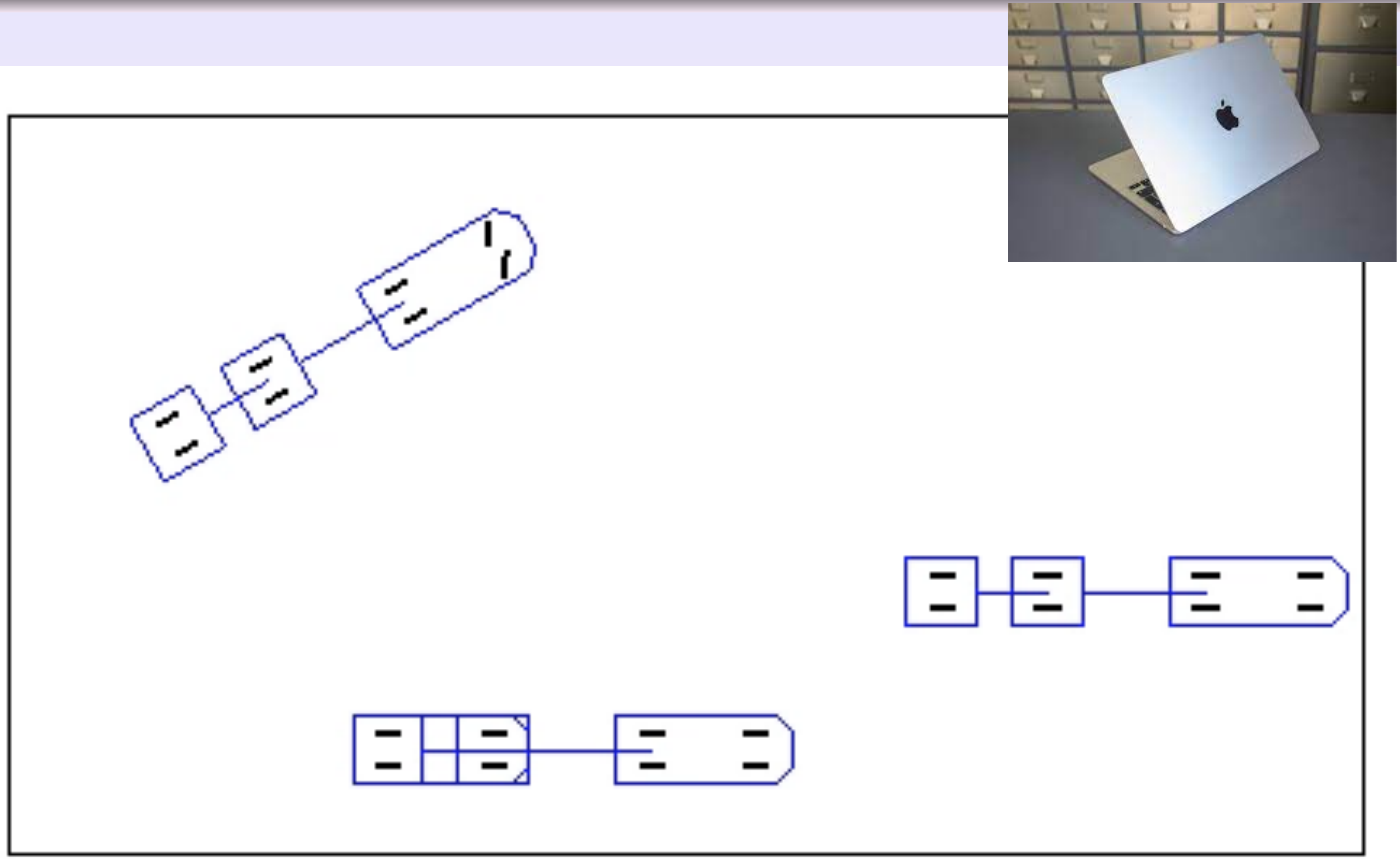
From the variation of constants formula:

$$x(t) = e^{At}x^0 + \int_0^t e^{A(t-s)}Bu(s)ds = e^{At}x^0 + \int_0^t \sum_{k\geq 0} \frac{(t-s)^k}{k!} A^k Bu(s)ds.$$

By Cayley[1]-Hamilton's [2] theorem A^k for $k \geq n$ is a linear combination of $I, A, ..., A^{n-1}$.

[1] Arthur Cayley (UK, 1821 - 1895)
[2] William Rowan Hamilton (Ireland, 1805 - 1865)

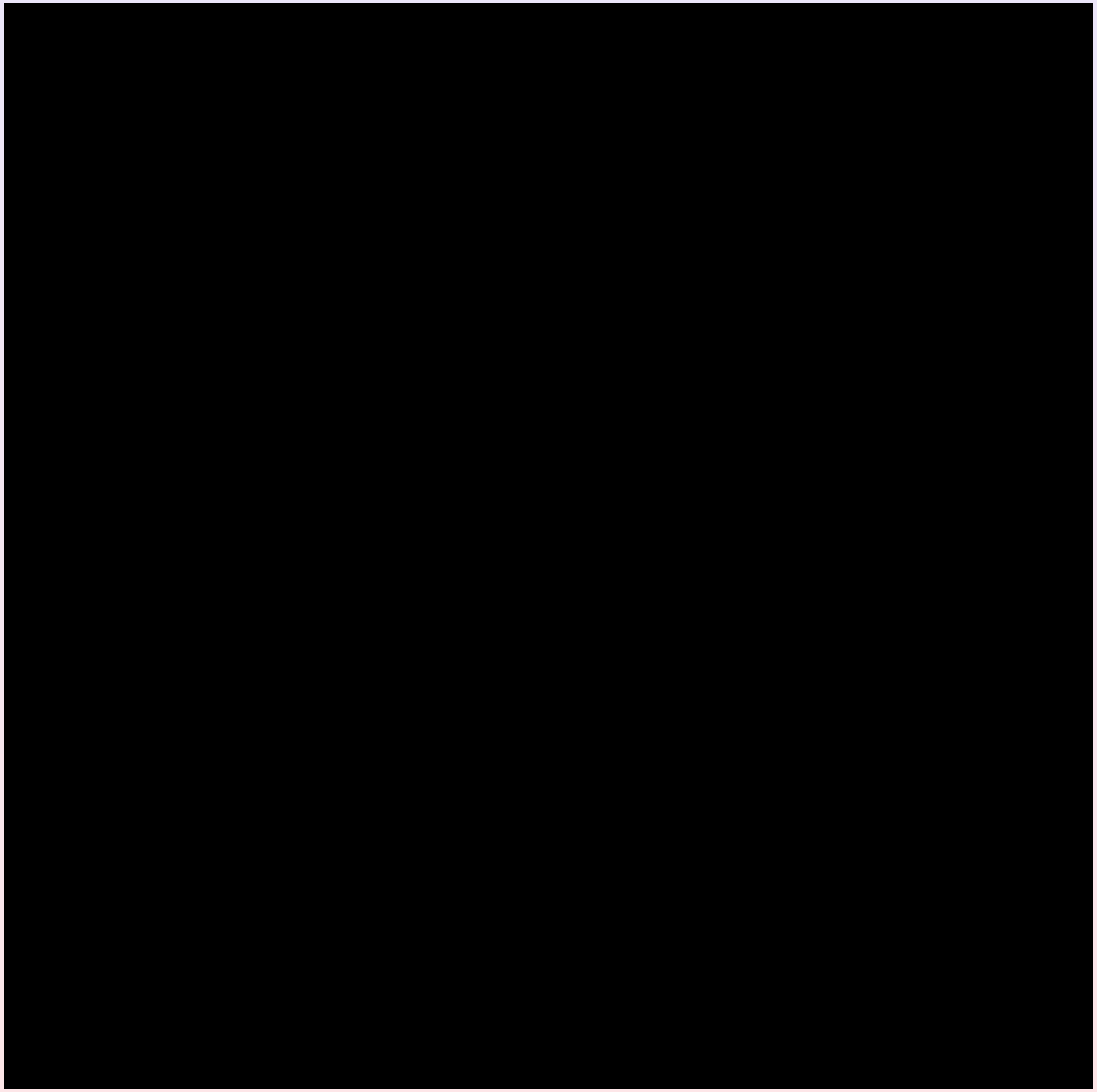
An example: Parking a car



Control in an information rich World, SIAM, R. Murray Ed., 2003.

The mathematical shepherd

R. Escobedo, A. Ibañez and E.Zuazua, Optimal strategies for driving a mobile agent in a “guidance by repulsion” model, Communications in Nonlinear Science and Numerical Simulation, 39 (2016), 58-72.



Duality, J.-L. Lions, SIREV, 1988

Consider the adjoint system

$$\begin{cases} -p' = A^* p, & t \in (0, T) \\ p(T) = p_T \end{cases}$$

and minimize

$$J(p_T) = \frac{1}{2} \int_0^T |\, B^* p \,|^2 \, dt + \langle x^0, p(0) \rangle$$

Then

$$u = B^* \widehat{p}$$

is the control of minimal L^2-norm.[4]
And the functional J is coercive iff the Kalman rank condition is satisfied.

The Kalman condition is equivalent to the Unique Continuation property

$$B^* p \equiv 0 \Rightarrow p^T \equiv 0.$$

The observability inequality plays a key role

$$||p^T||^2 \leq C \int_0^T |\, B^* p \,|^2 \, dt$$

[4] This confirms Wiener's vision "..control and communication..."

Table of Contents

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

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°26(2003), 79–140

Control theory: History, mathematical achievements and perspectives*

E. Fernández-Cara[1] and E. Zuazua[2]

Universal approximation theorem I

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*

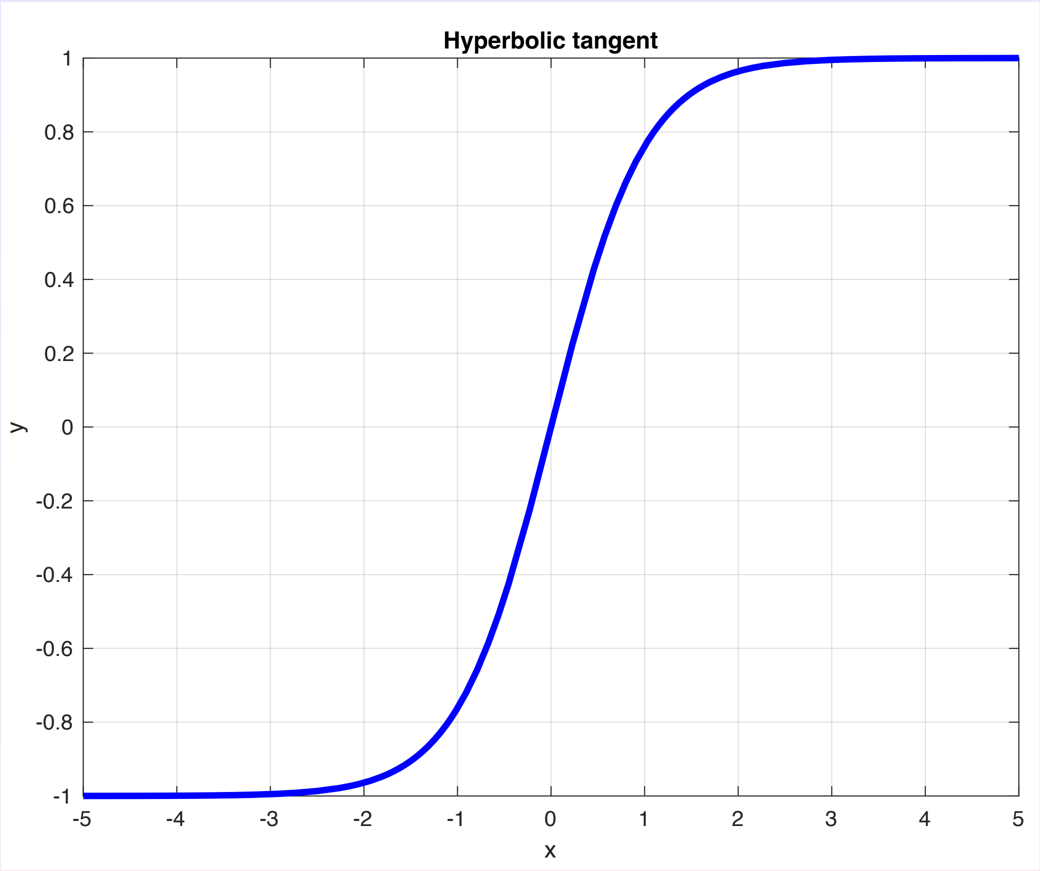
G. Cybenko†

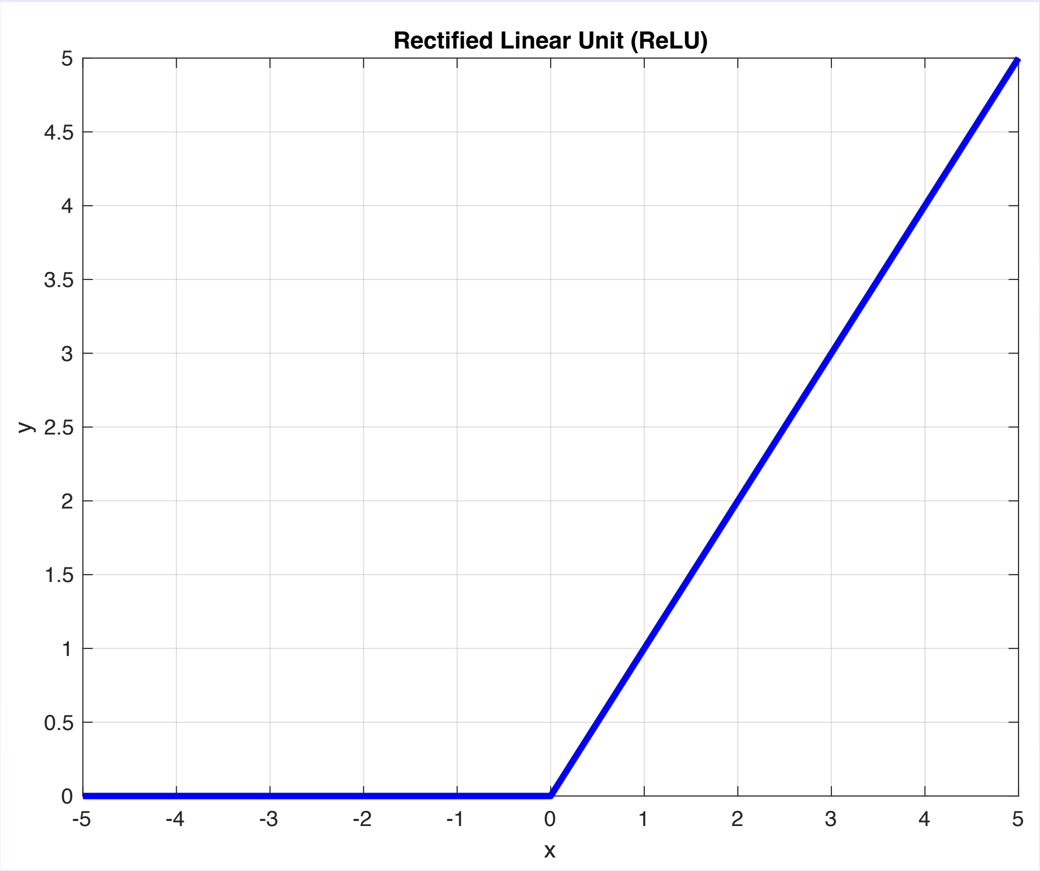
$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j), \tag{1}$$

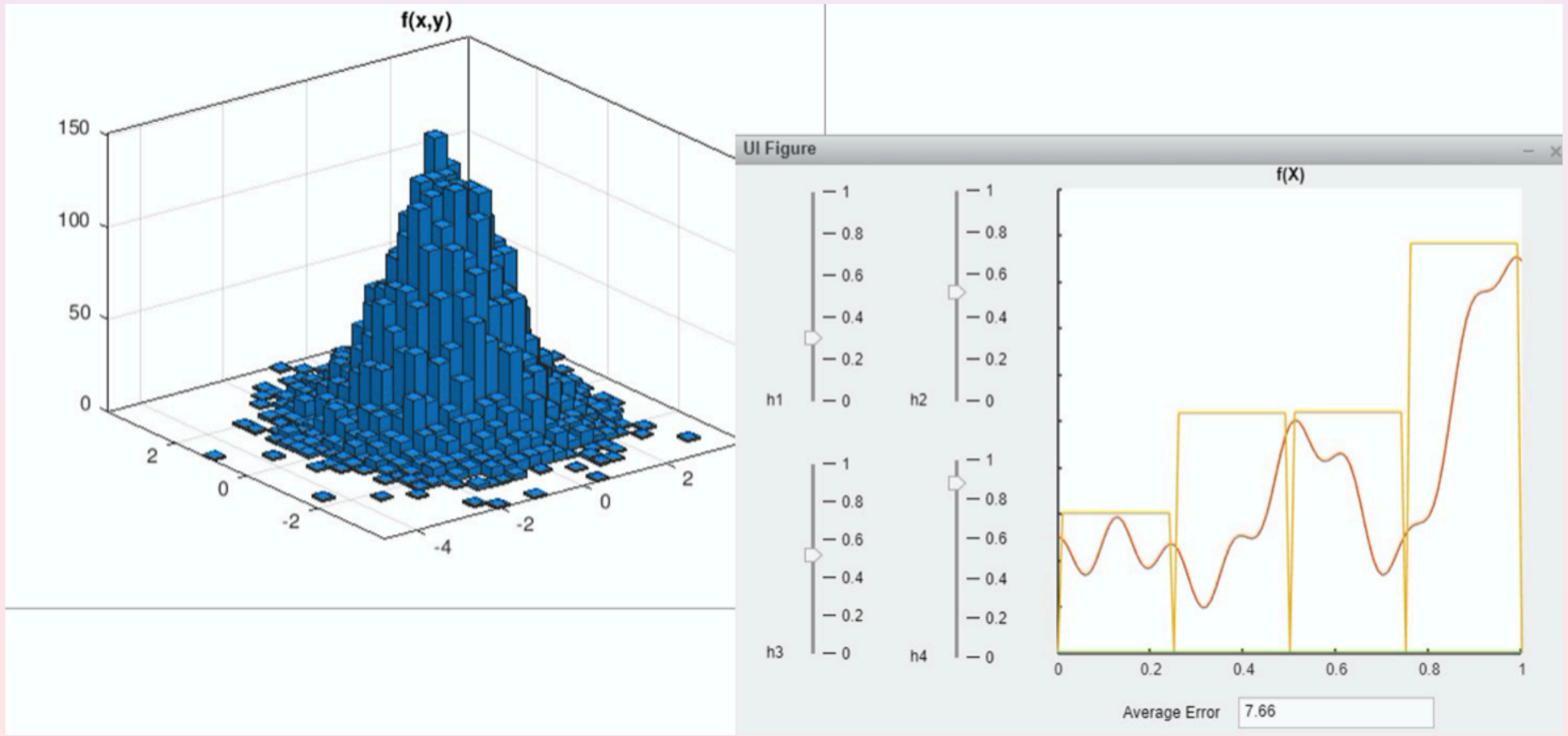
where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^{T} is the transpose of y so that $y^{\mathrm{T}}x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

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

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

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

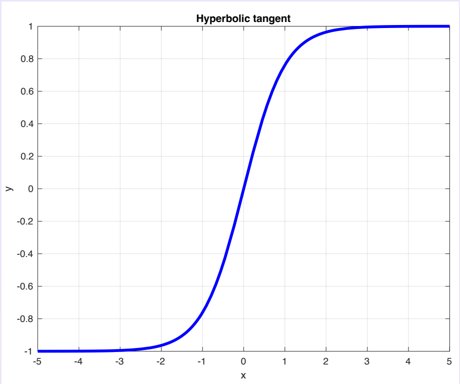
drawn from an unknown probability measure ρ on $\mathbb{R}^d \times \mathbb{R}^m$.

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

Residual neural networks

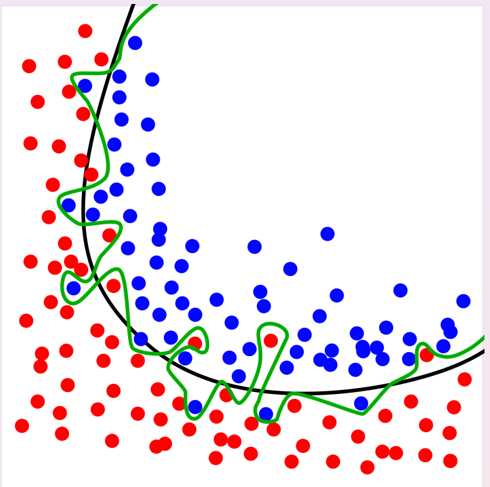
[1] K He, X Zhang, S Ren, J Sun, 2016: Deep residual learning for image recognition
[2] E. Weinan, 2017. A proposal on machine learning via dynamical systems.
[3] R. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, 2018. Neural ordinary differential equations.
[4] E. Sontag, H. Sussmann, 1997, Complete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

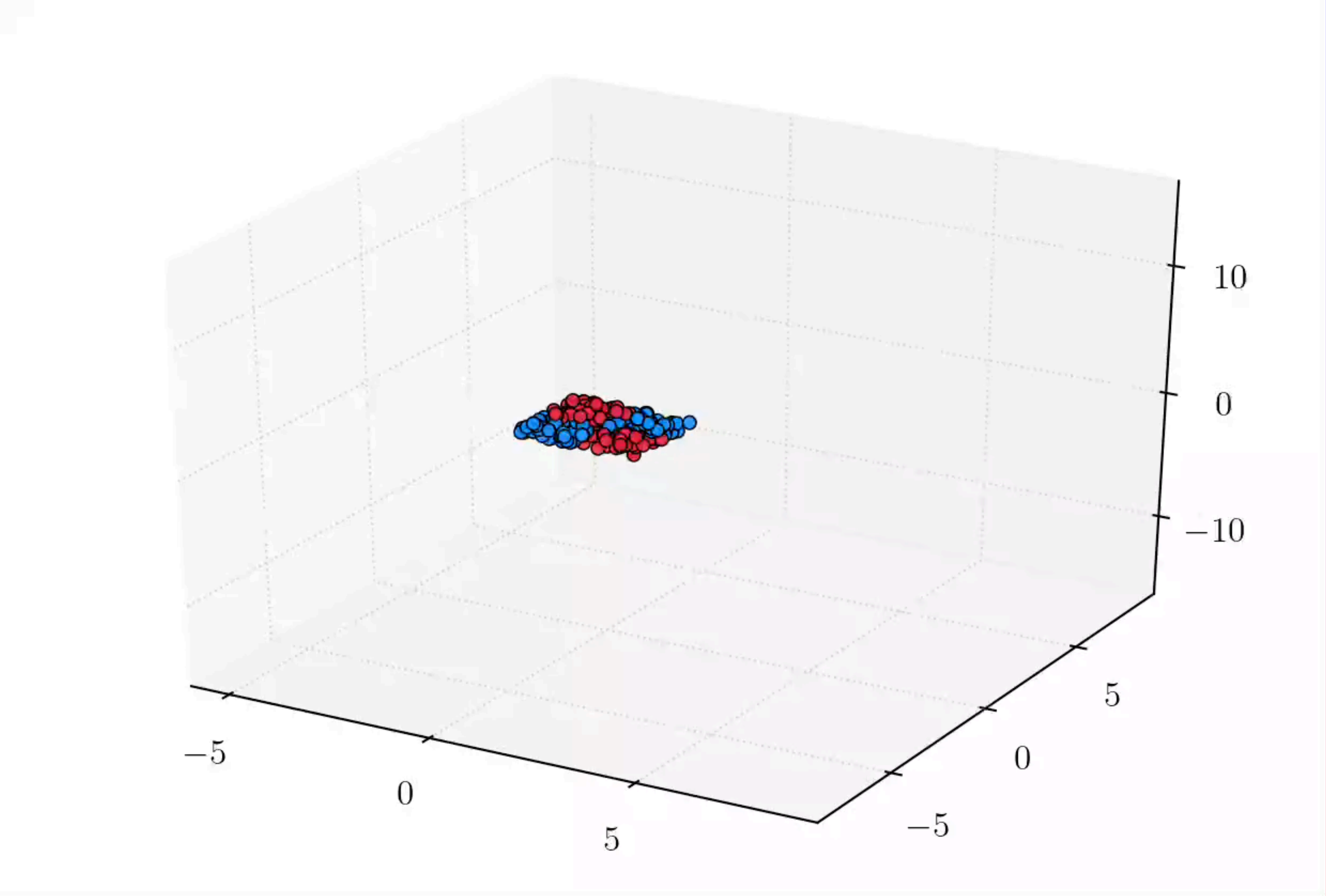
[9] D. Ruiz-Balet & Zuazua, Neural ODE control for classification, approximation and transport, arXiv:2104.05278

Special features of the control of ResNets

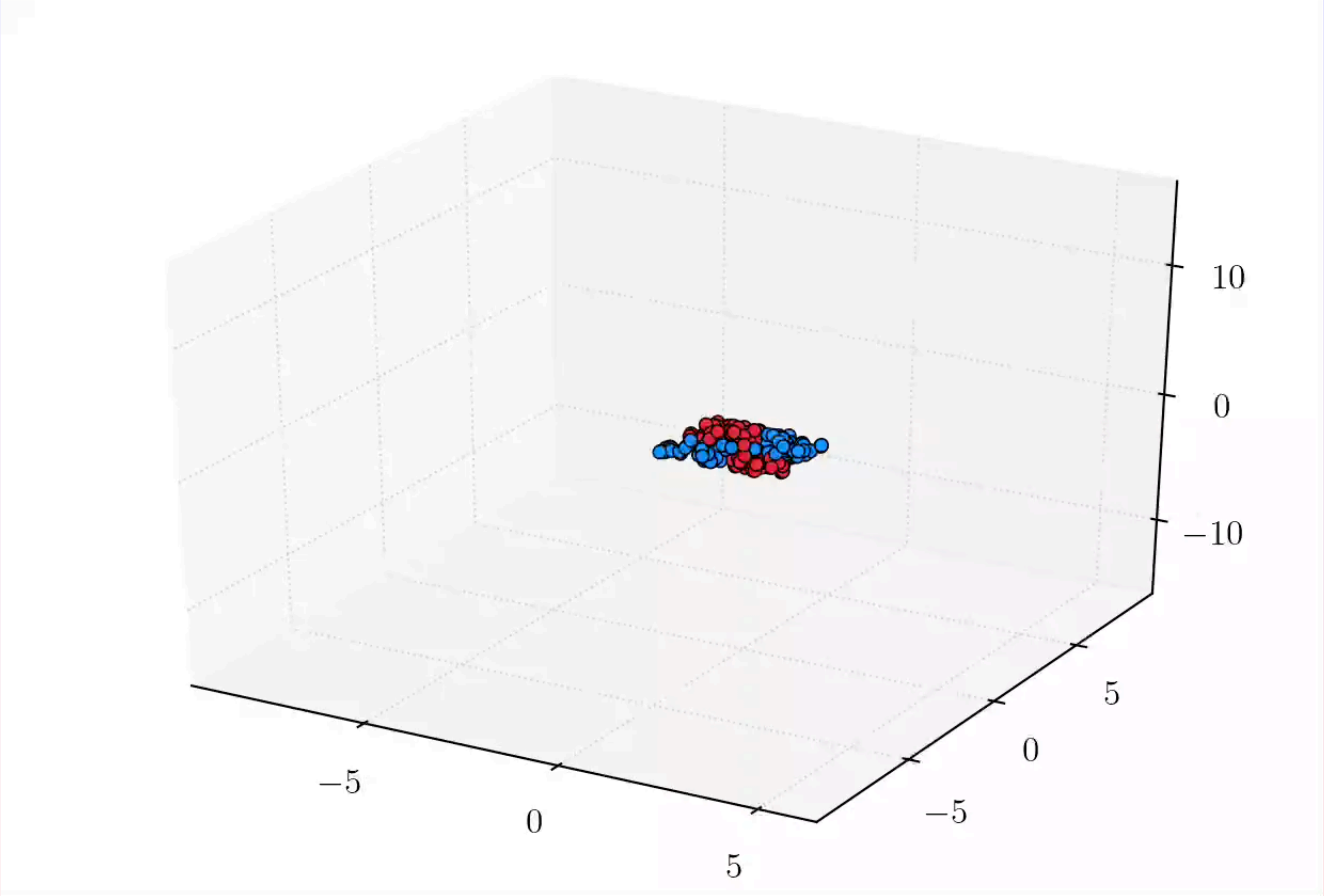
- Nonlinearities are unusual in Mechanics: σ is flat in half of the phase space.
- We need to control many trajectories (one per item to be classified) with the same control! [10]

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.

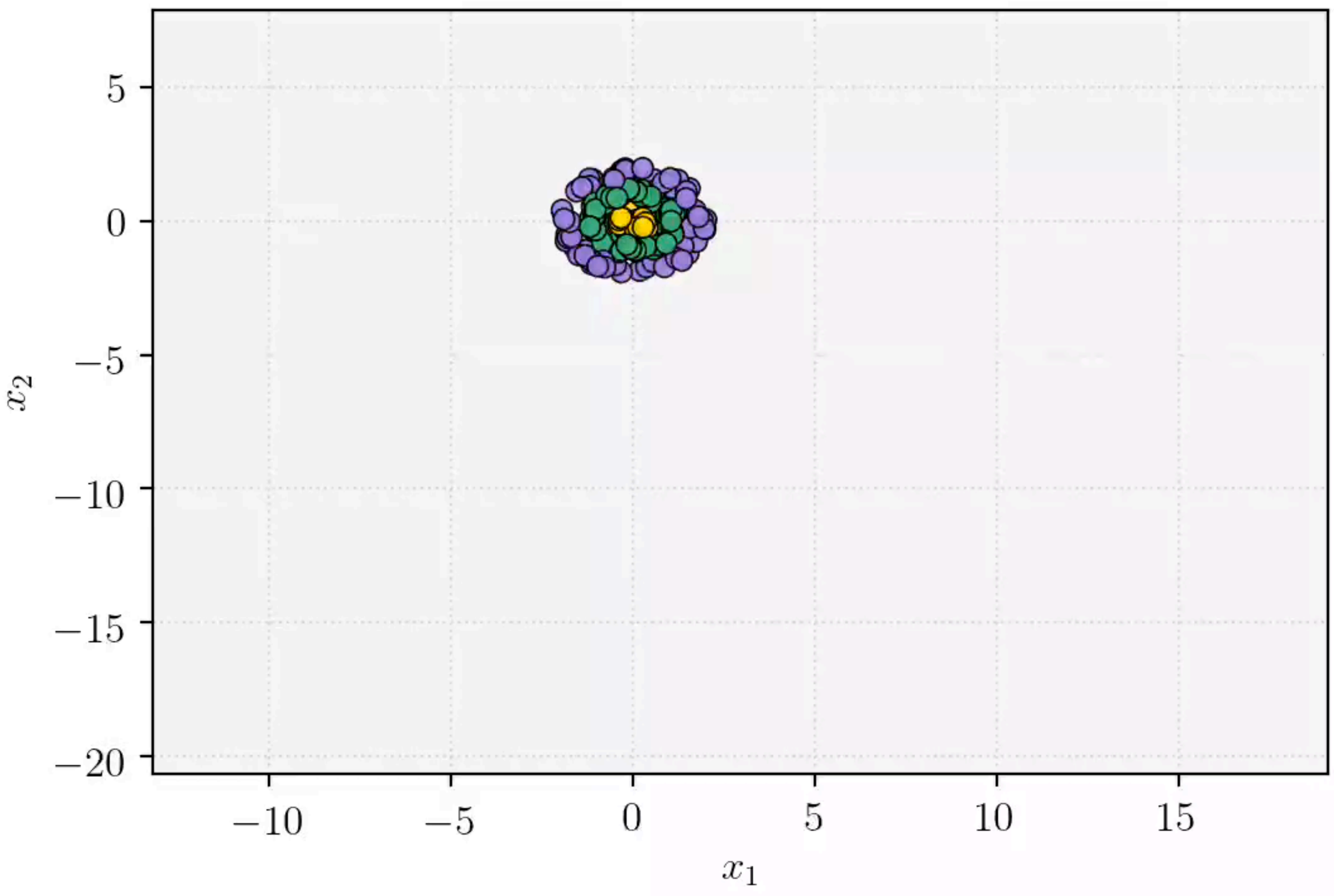
[10] This would be impossible for instance, for the standard linear system $x' = Ax + Bu$.



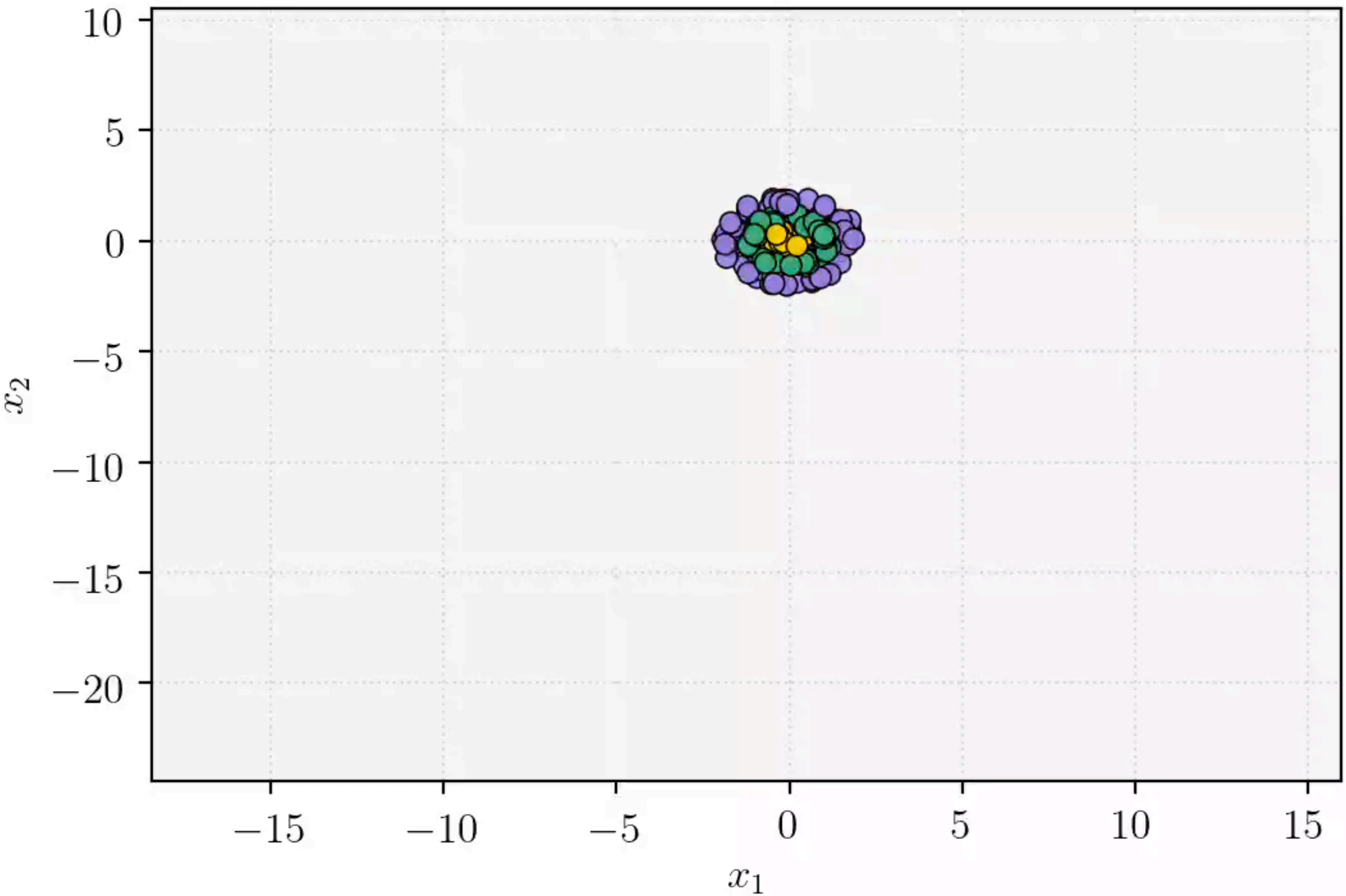
B. Geshkovski, MIT



B. Geshkovski, MIT



B. Geshkovski, MIT



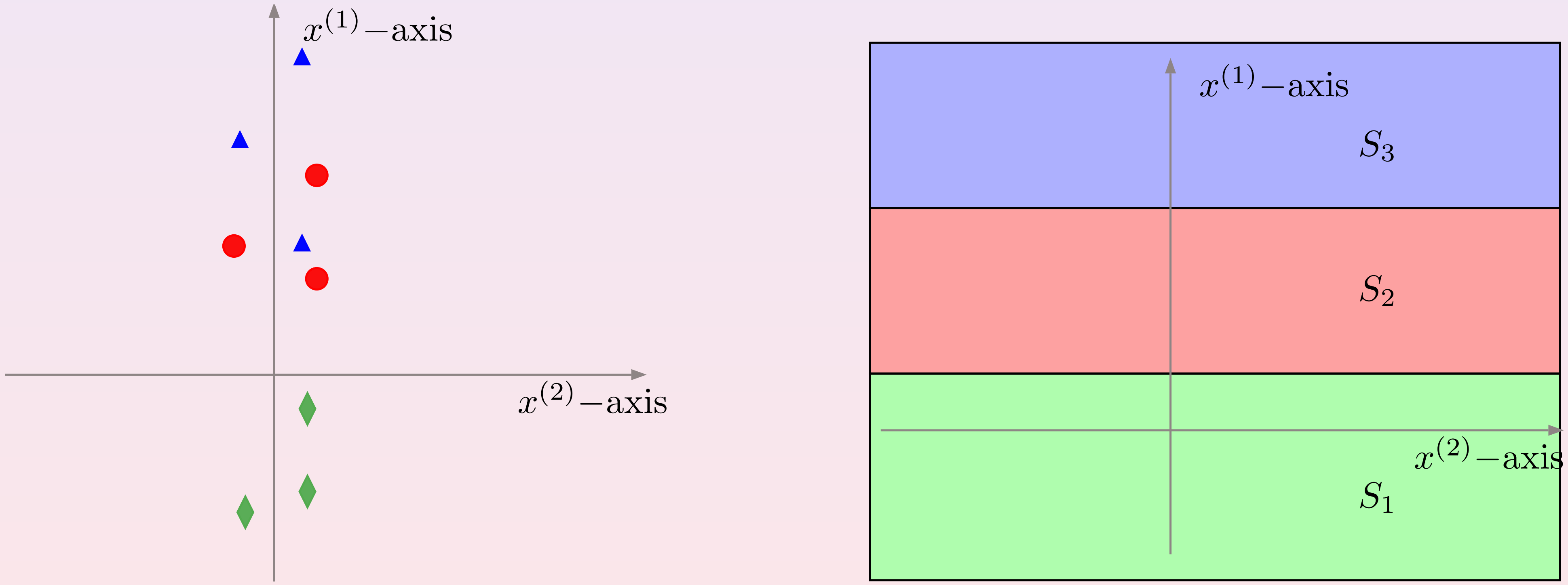
B. Geshkovski, MIT

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 $y_i \in \{1, \ldots, M\}$.

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.

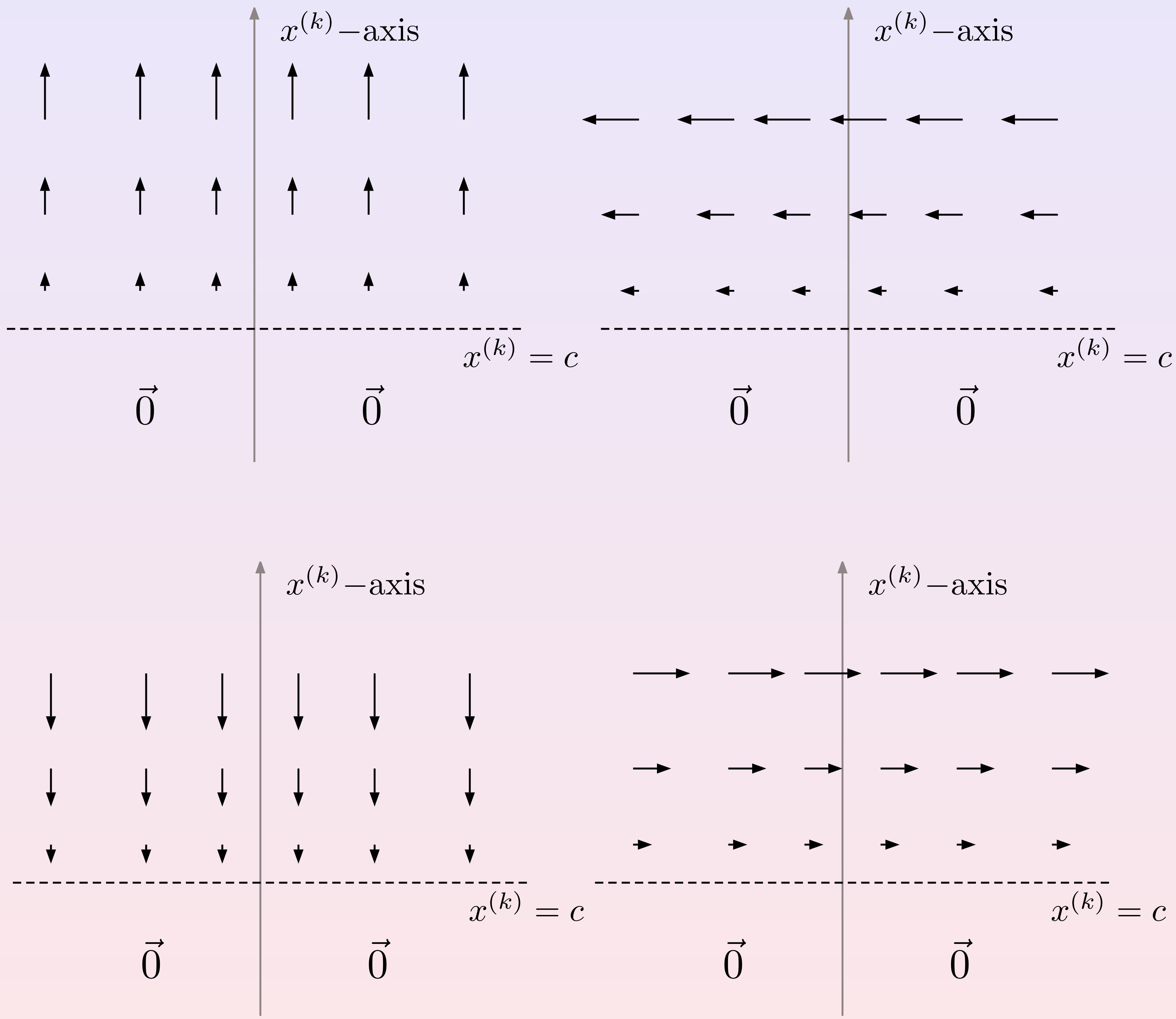


Basic control actions

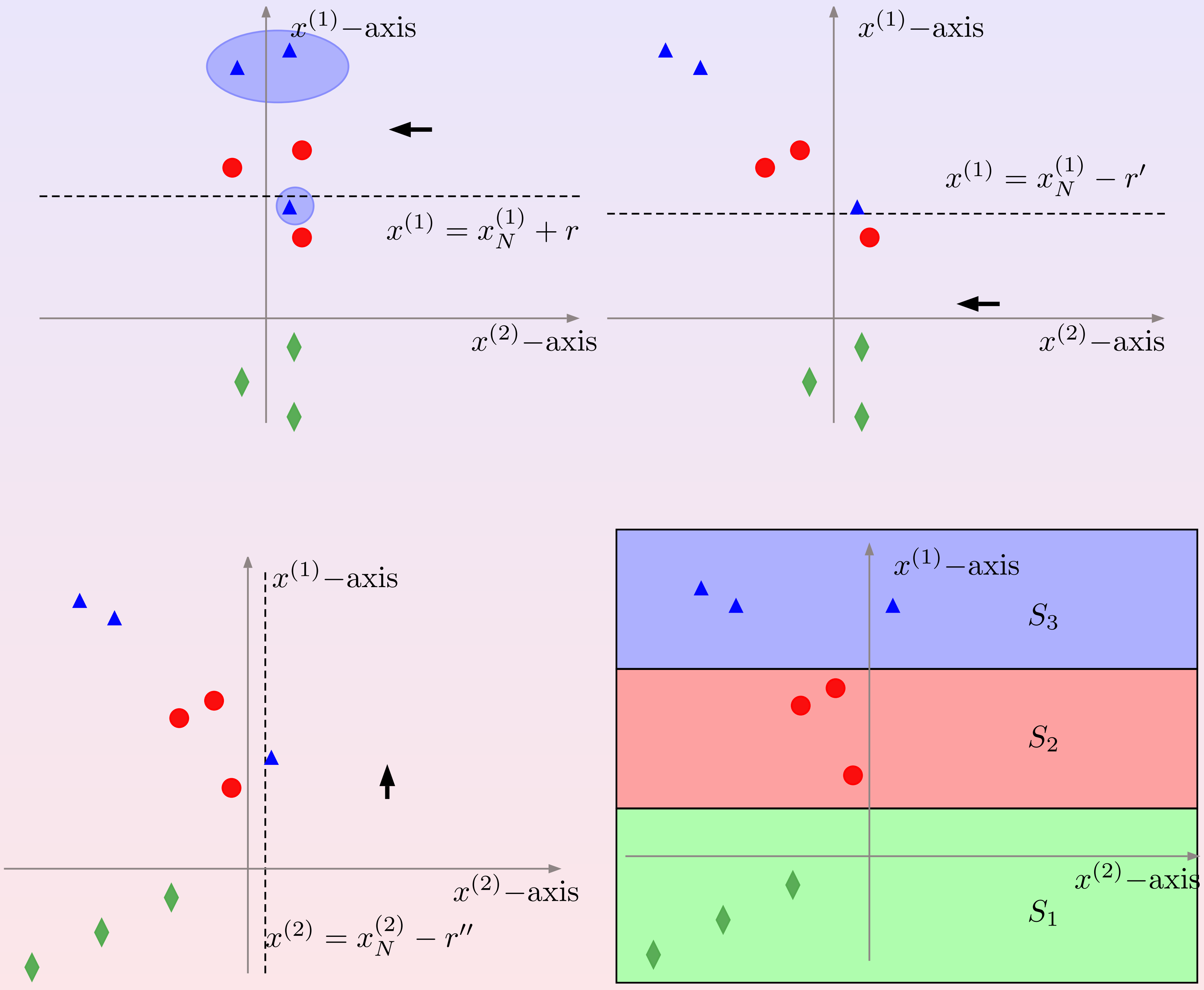
$$\dot{\mathbf{x}}(t) = W(t)\sigma(A(t)\mathbf{x}(t) + b(t))$$

- $b(t)$ 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



Neural transport equations

Note that the differential equation

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

corresponds to the characteristics of the transport equation:

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

The results above can therefore be understood in terms of the controllability of the transport equation: "Atomic initial data can be driven to atomic final targets". This also allow for a more general interpretation in terms of approximation control in Wasserstein-1 distance. Or for systems of transport equations, so that each scalar component corresponds to the density within one of the classes of data.

This establishes a link to the Theory of Optimal Transport: Neural Transport? [12]

12

$$\mathcal{W}_1(\mu, \nu) = \sup_{Lip(g) \leq 1} \left\{ \int_{\mathbb{R}^d} g d\mu - \int_{\mathbb{R}^d} g d\nu \right\}$$

where $Lip(g) \leq 1$ stands for the class of Lipschitz functions with Lipschitz constant less or equal than 1.

In Mathematics and Economy: Optimal transport of resources. Formulated any the French mathematicians **Gaspard Monge** in 1781 ("Sur la théorie des déblais et des remblais" (Mém. de l'Acad. de Paris, 1781))

In the ancient Egypt the "harpenodaptai" (string stretchers), were specialized in stretching very long strings leading to long straight segments to help in large constructions, with the understanding that (a) "The shortest distance between two points is the straight line" and (b) "Among all the paths of a given length the one that produces the longest distance between its extremes is the straight line as well."

Neural transport equations

The simultaneous control of the nODE

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

to arbitrary terminal states

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

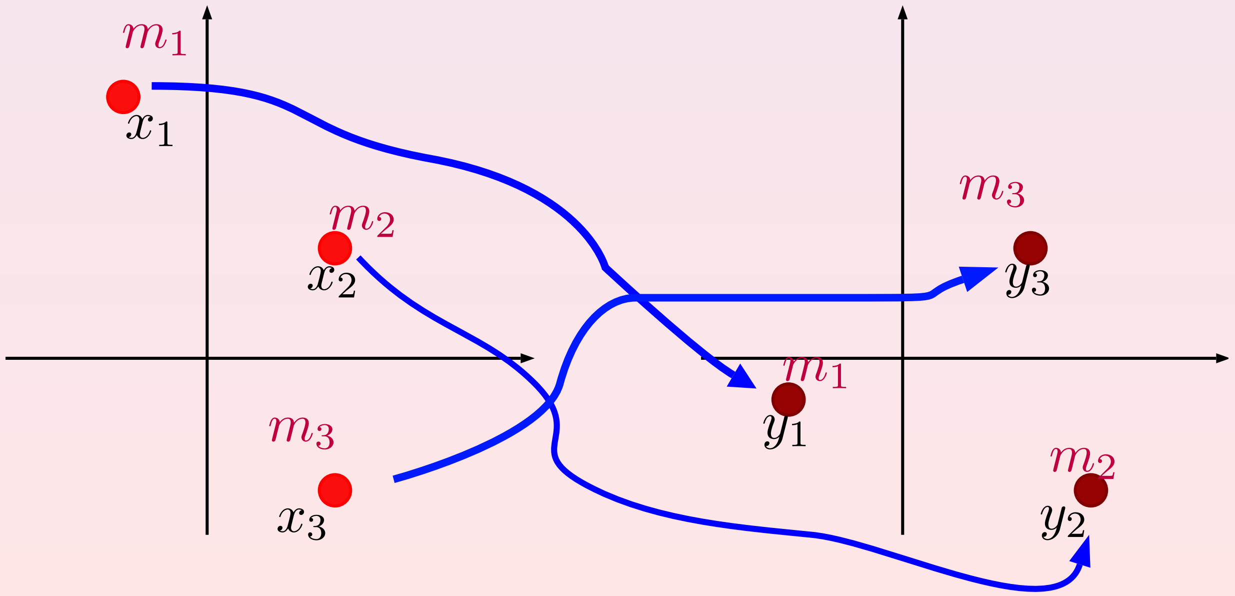
in terms of the transport equation, leads to the control of an atomic initial datum from

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

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.

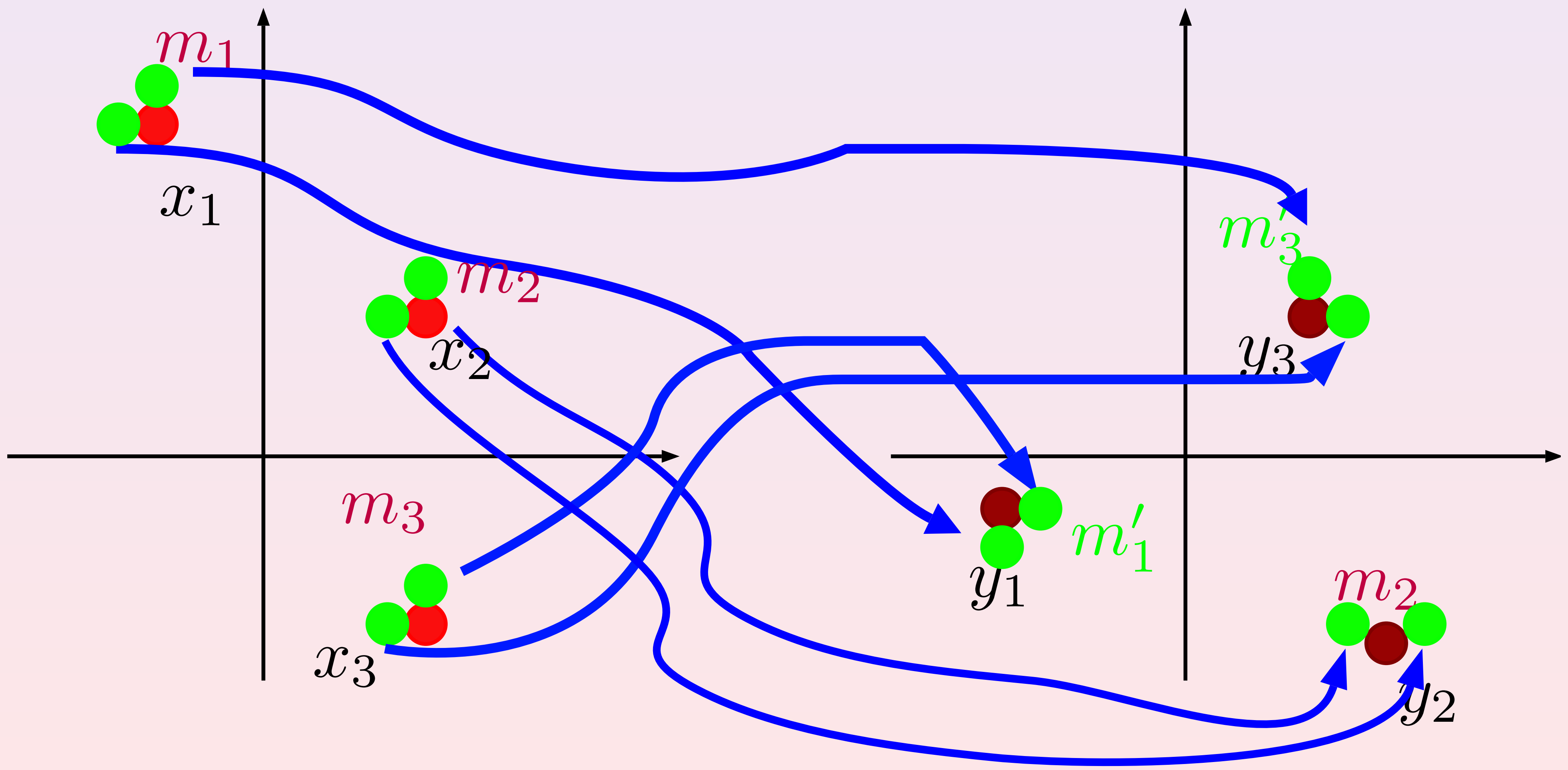


Neural transport equations

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

Let us approximate a piece-weise constant function taking two different values P and Q on the sets represented by colors blue and red.
We aim to build a nODE so that the solution of

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

is such that

$$\varphi(T, x) = P, \quad \text{when } x \in \text{Blue Set}$$

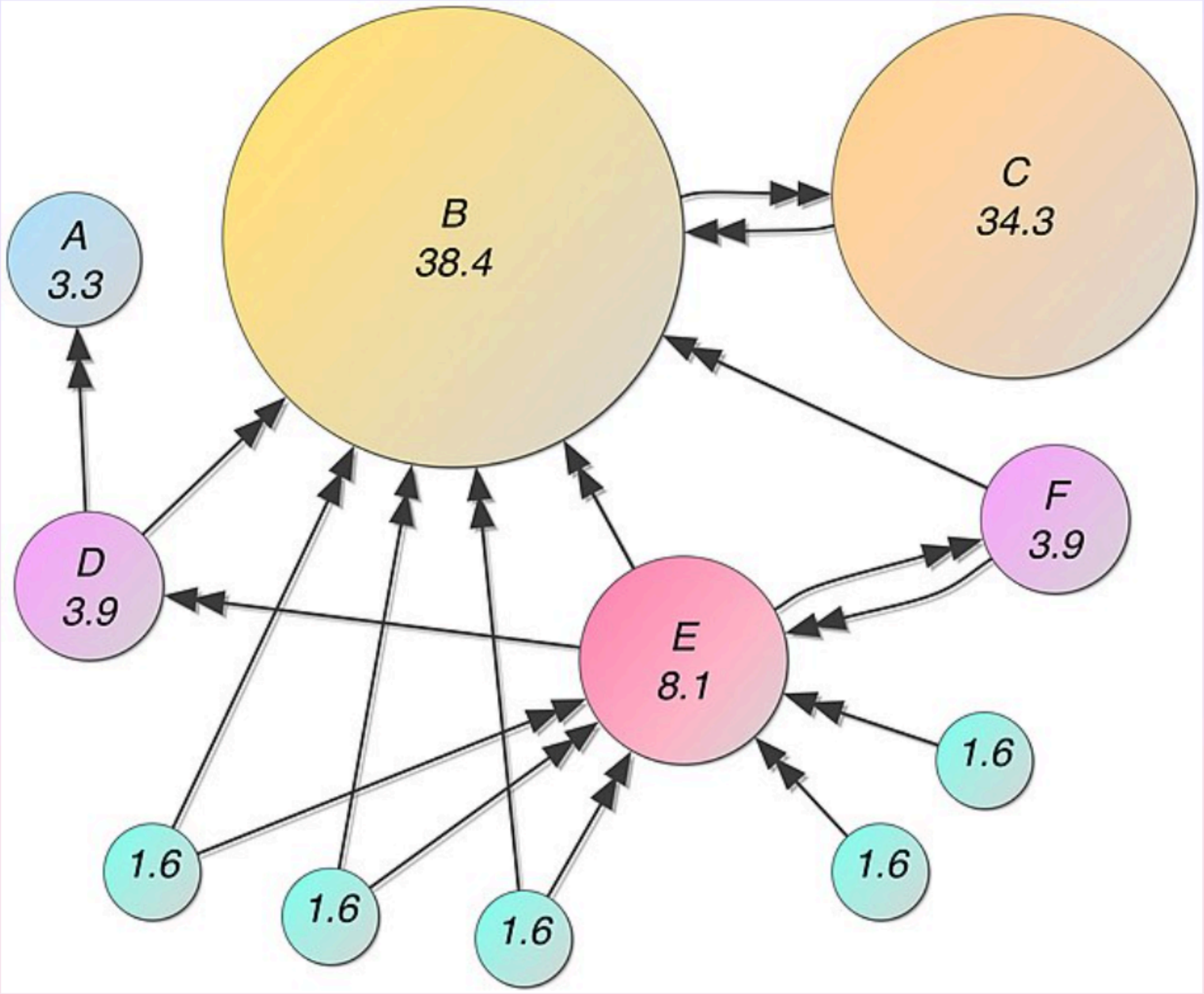
and

$$\varphi(T, x) = Q, \quad \text{when } x \in \text{Red Set}.$$

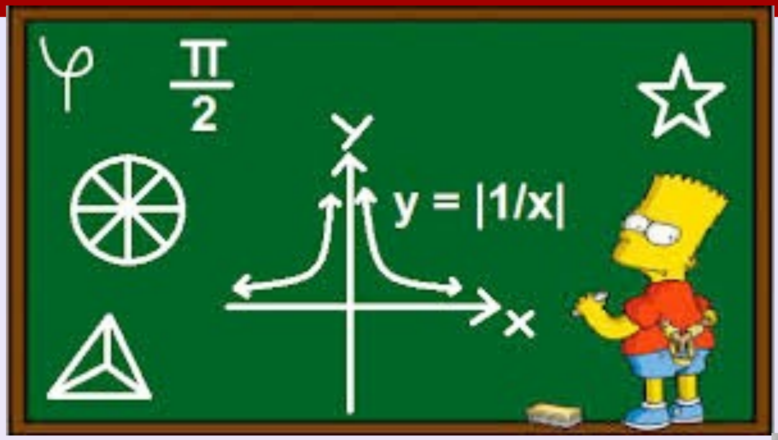
The same control inspired strategies allow to achieve the result up to an ε- error.

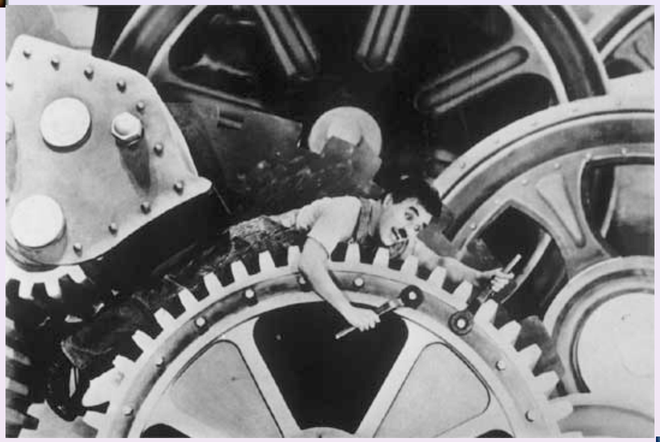
Table of Contents

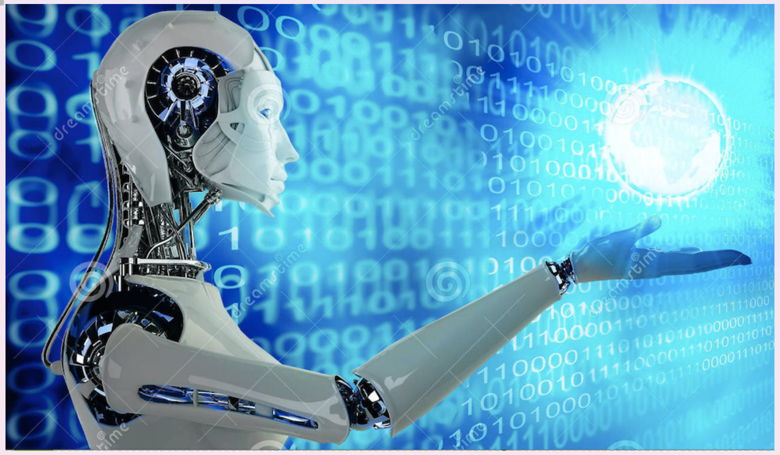
El nuevo paradigma: Matemáticas + Inteligencia Artificial
PageRank: Larry Page and Sergey Brin (Google, 1998)



Inspirados por el Teorema de Perron-Frobenius (Oskar Perron (1907) y Georg Frobenius (1912))


π/2
Y
y = |1/x|
X





$$i = \sqrt{-1}$$

Enrique Zuazua

https://dcn.nat.fau.eu