FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Neural ODEs, Control and Machine Learning

Enrique Zuazua

Chair of Dynamics, Control and Numerics
Department of Data Science
FAU
Erlangen, Germany
https://caa-avh.nat.fau.eu

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*

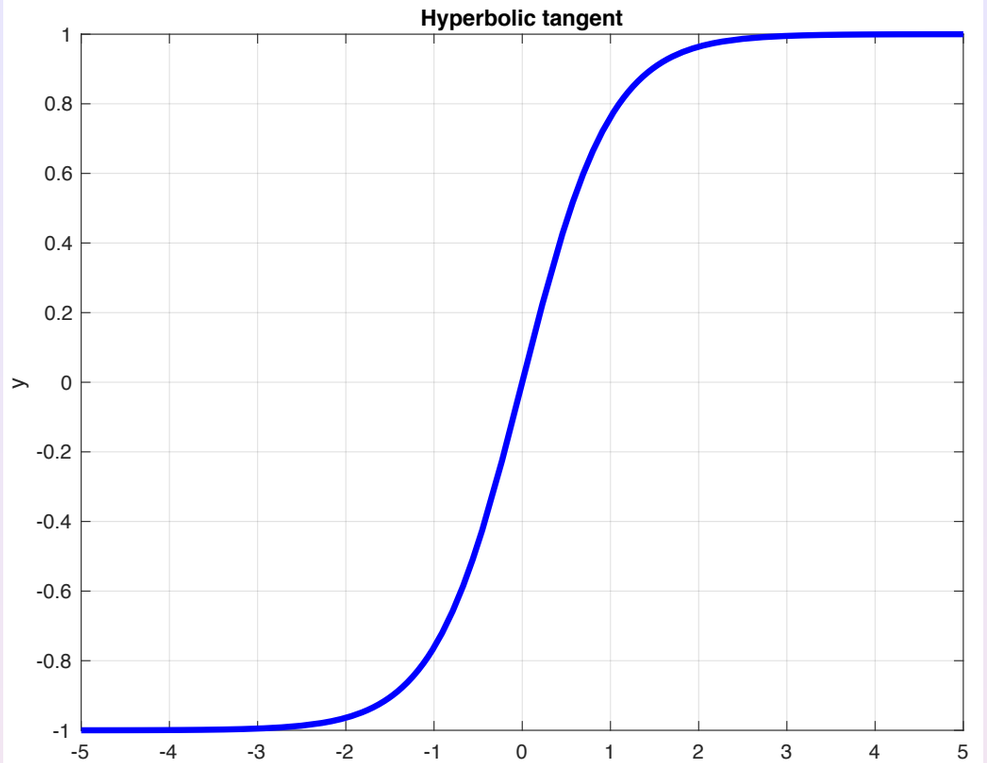
G. Cybenko†

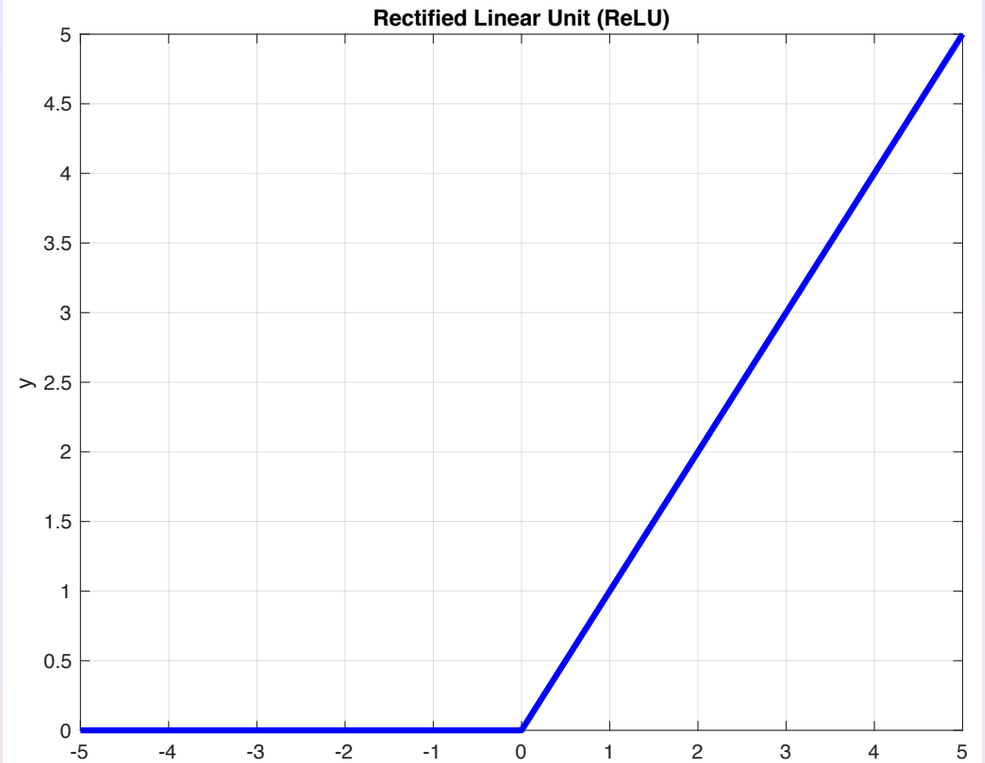
$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j), \tag{1}$$

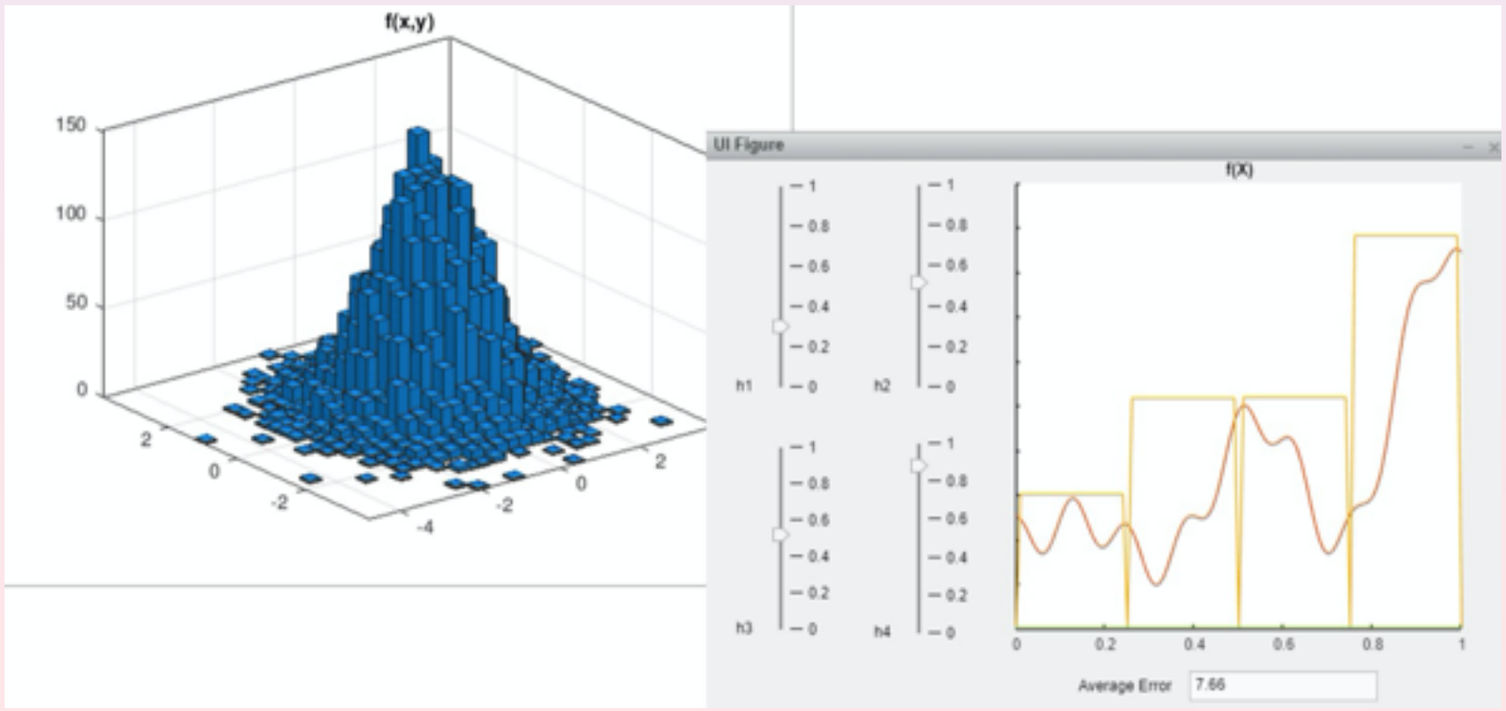
where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^T is the transpose of y so that $y^T x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

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

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**.

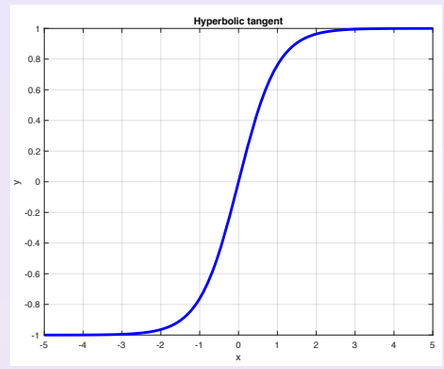
Residual neural networks

[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

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

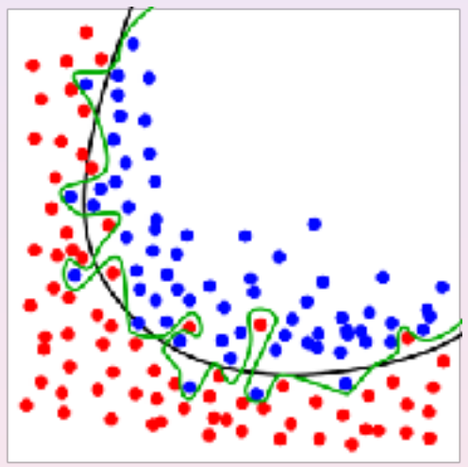
where $h = 1$, σ globally Lipschitz $\sigma(0) = 0$.



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?[1] [2]

[1] C. Esteve, B. Geshkovski, D. Pighin, E. Zuazua, Large-time asymptotics in deep learning, arXiv:2008.02491
[2] 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! [3]

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.

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

Turnpike for ResNets

$T \to \infty \quad \sim \quad N_{\text{layers}} \to \infty.$

Turnpike refers to the fact that, in long time-horizons, optimal controls and trajectories are exponentially close to the optimal steady-state control and state in most of the time-horizon.

Supervised Learning* $\Longleftrightarrow$ **minimize**[4]

$$\frac{1}{N}\int_0^T \|P\mathbf{x}(t)-\overline{y}\|^2 dt + \alpha \|u\|^2_{H^1(0,T;\mathbb{R}^{d_u})}. \tag{SL*}$$

$\overline{y} := [\vec{y}_1, \ldots, \vec{y}_N]$, $u := [A, W, b]$ in (5) and $P : \mathbb{R}^d \to \mathbb{R}^m$.

Theorem (Turnpike): Under controllability assumptions, for any sufficiently large T, an optimal solution $(u_T, \mathbf{x}_T)$ to (SL*)–(5) satisfies

$$\|u_T\|_{H^1(0,T;\mathbb{R}^{d_u})} \leq C_1$$

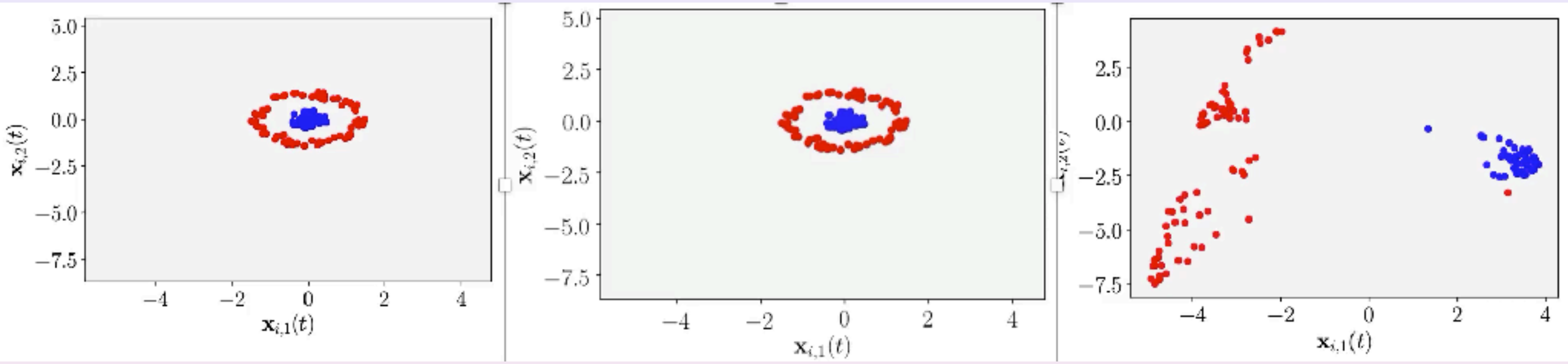
and

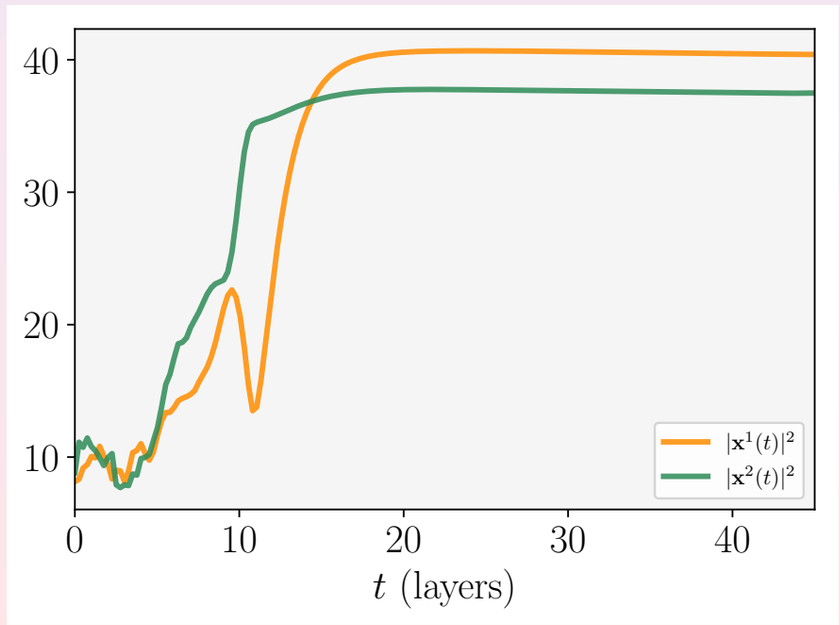
$$\|P\mathbf{x}_T(t)-\overline{y}\| \leq C_2 e^{-\mu t} \qquad \forall t \in [0,T]$$

for some $C_1 > 0$, $C_2 > 0$ and $\mu > 0$, all independent of T.

[4] Note that in this context we do not impose a perfect classification. We just expect that it will occur with high probability as an outcome of the optimar control problem

$$T \to \infty \quad \sim \quad N_{\text{layers}} \to \infty.$$

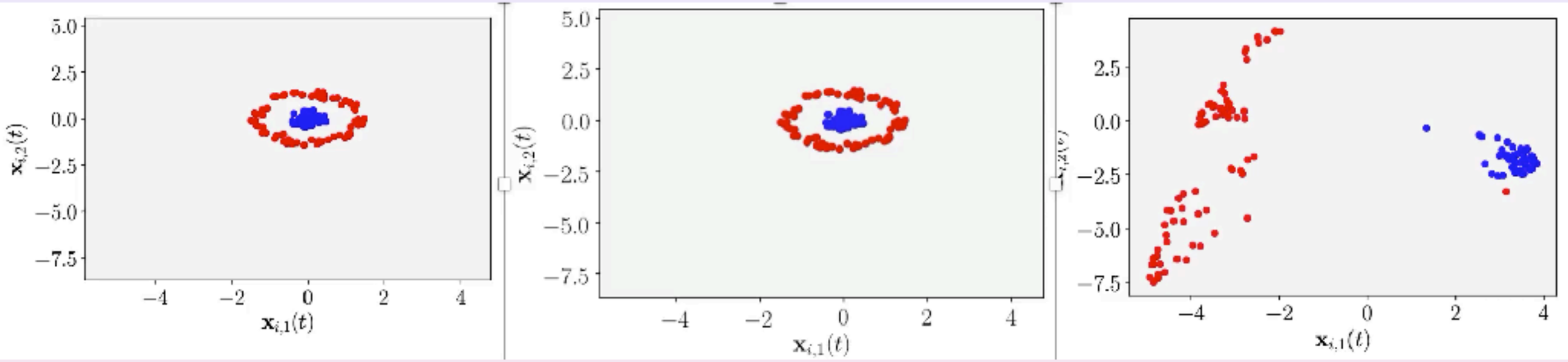


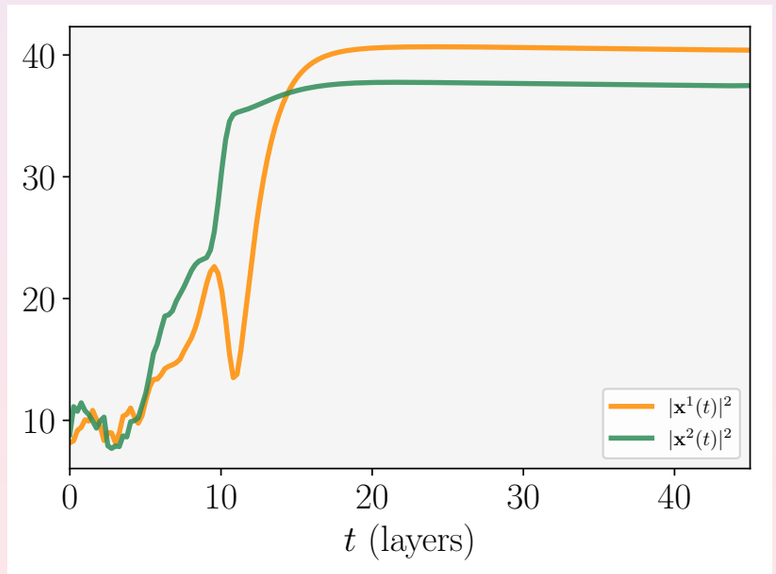


Classical SL problem?

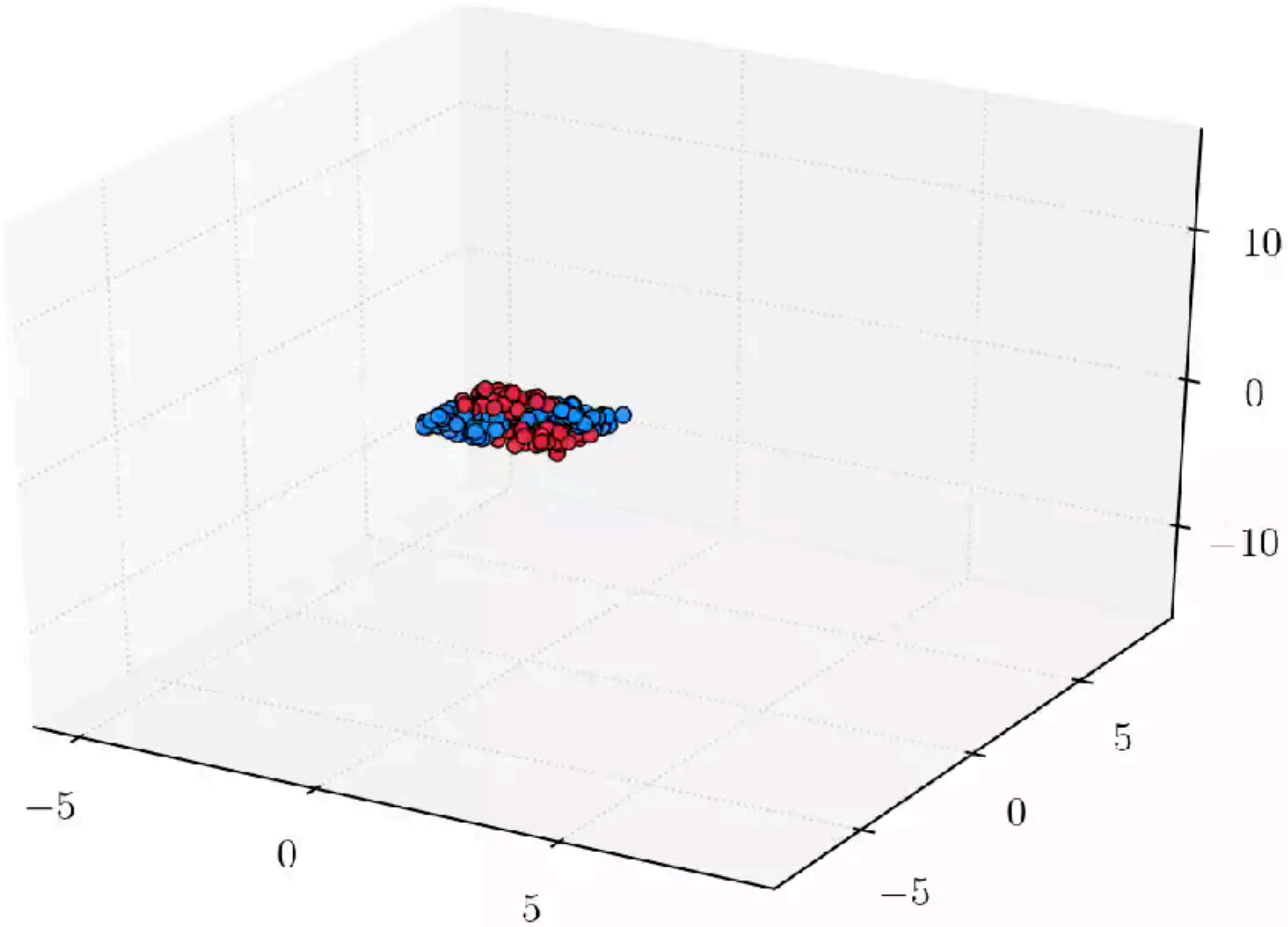
$P : \mathbb{R}^d \to \mathbb{R}^m$, **minimize**

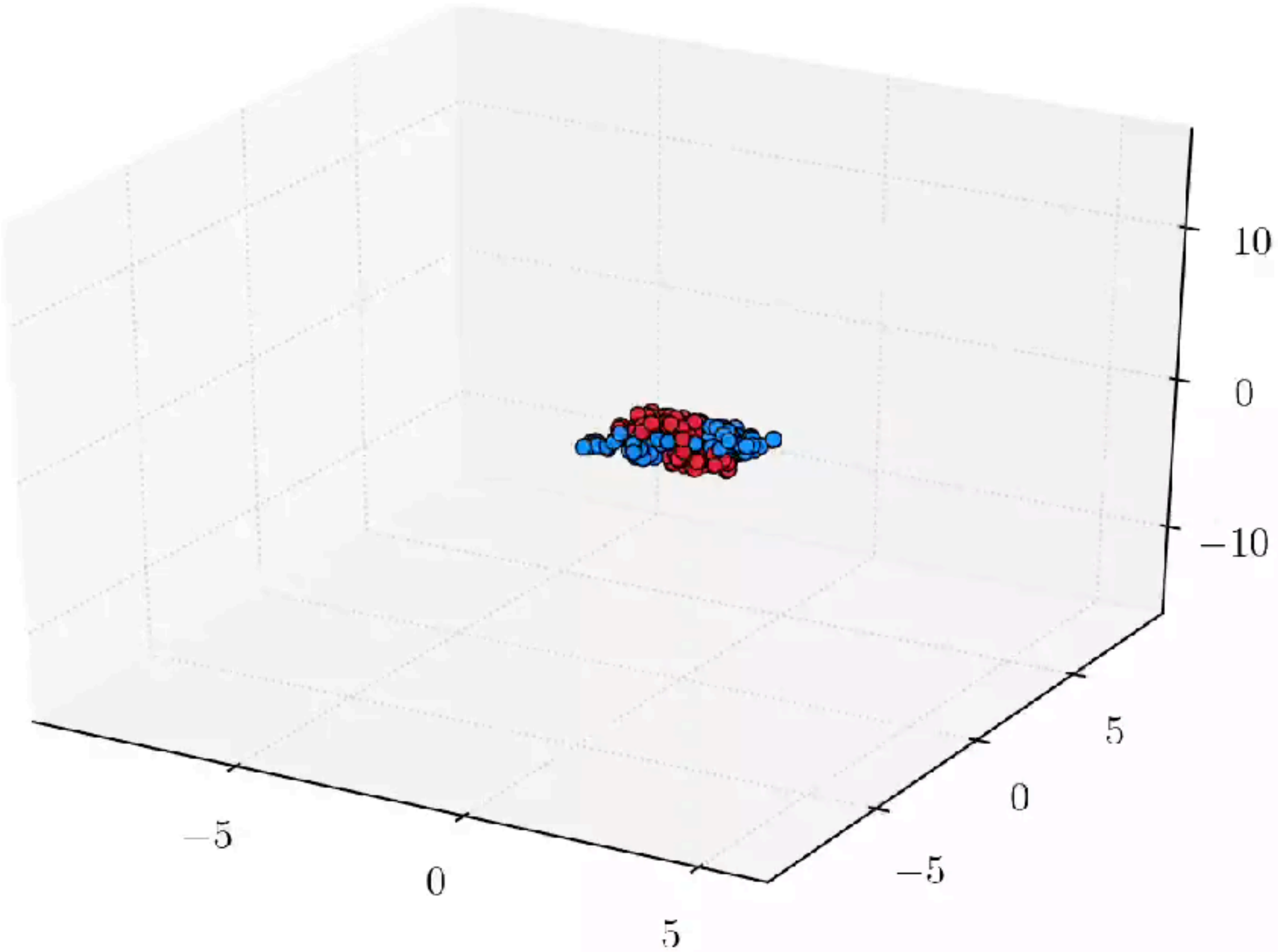
$$\frac{1}{N}\|P\mathbf{x}(T) - \overline{y}\|^2 + \alpha \|u\|^2_{H^1(0,T;\mathbb{R}^{d_u})}. \tag{SL}$$

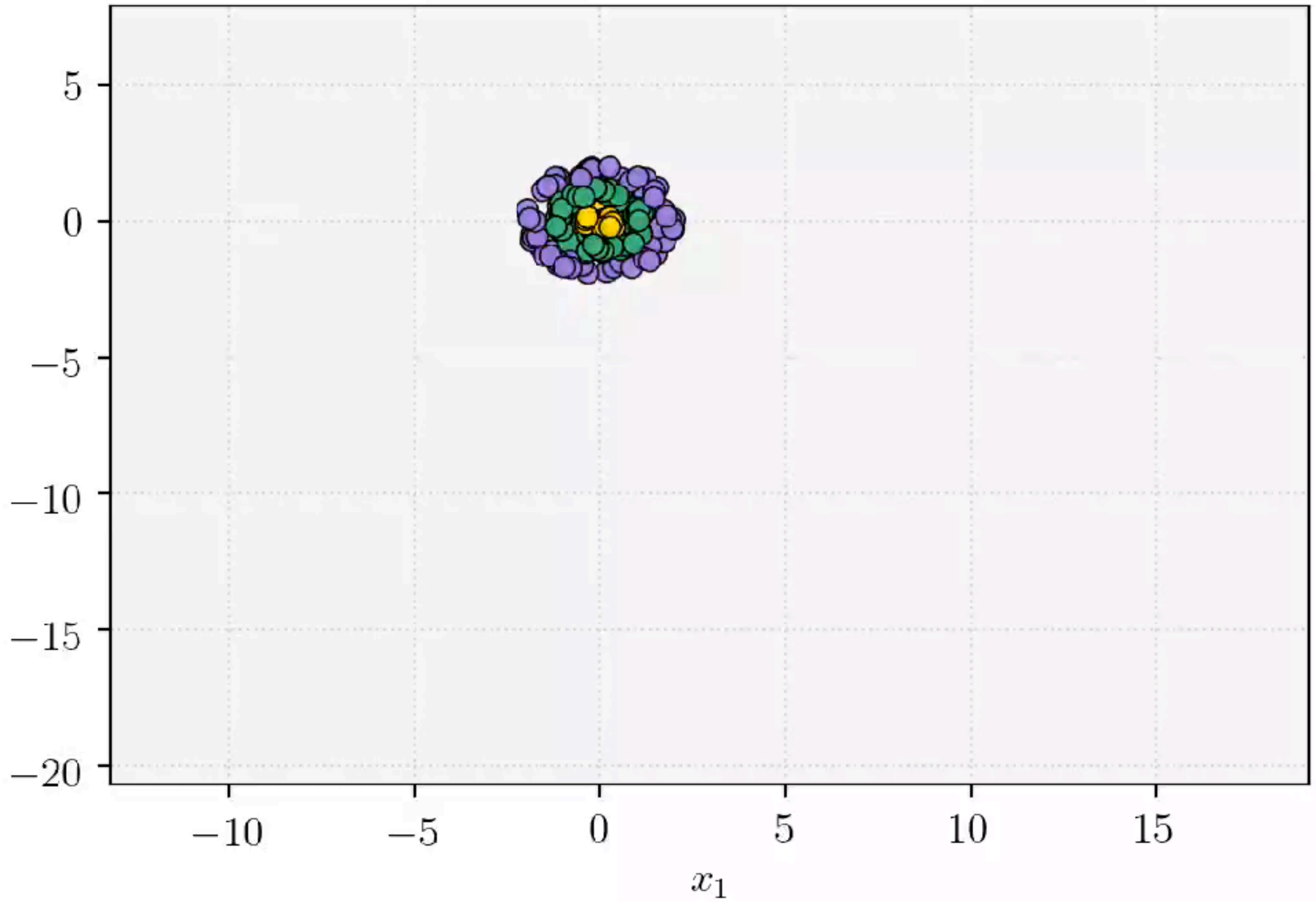


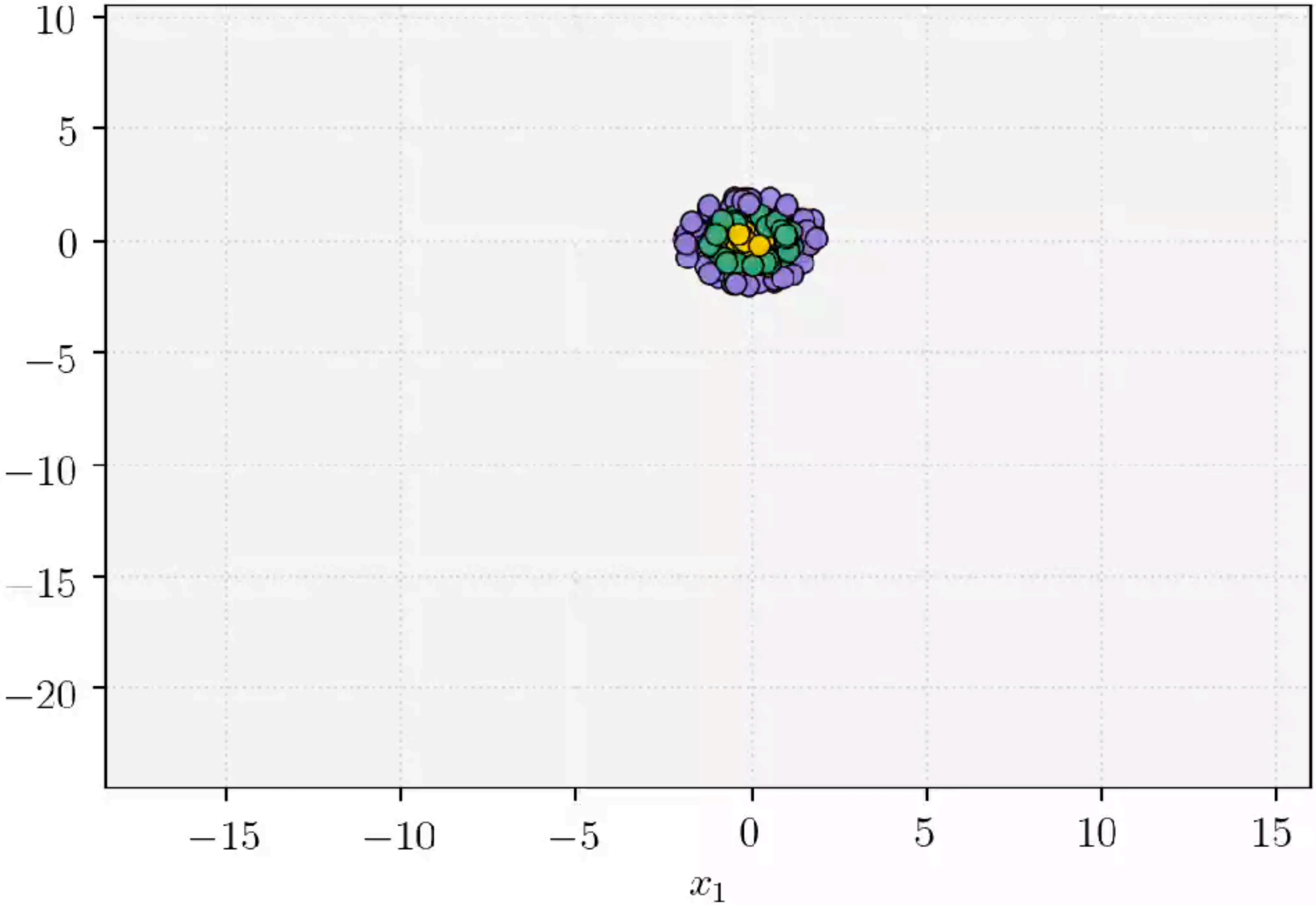


Convergence of $\mathbf{x}(T)$ to $P^{-1}(\{\overline{y}\})$ when $T \to \infty$, but slow (no turnpike).


10
0
−10
−5
0
5
5
0
−5


10
0
−10
−5
0
5
5
0
−5


5
0
−5
−10
−15
−20
x_2
−10
−5
0
5
10
15
x_1

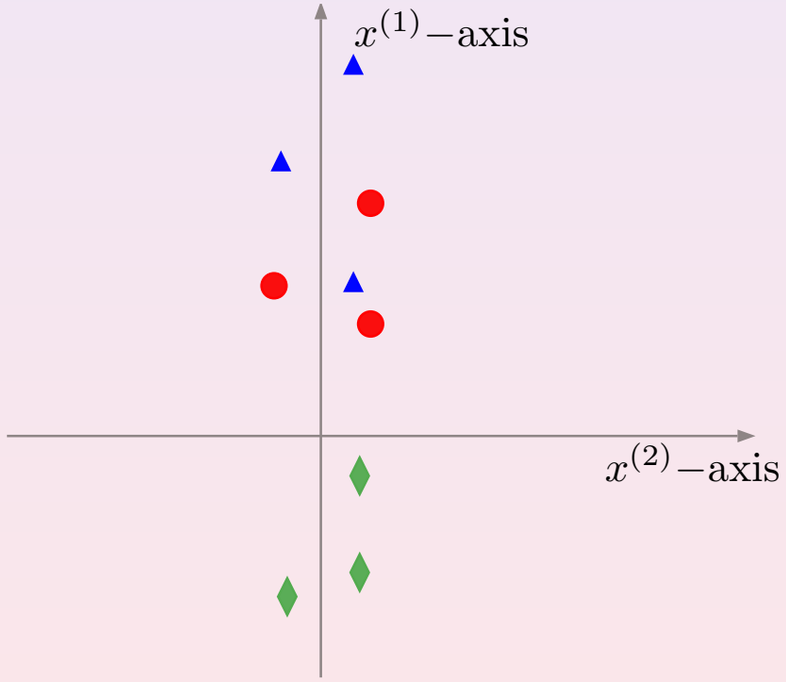

10
5
0
−5
−10
−15
−20
x_2
−15
−10
−5
0
5
10
15
x_1

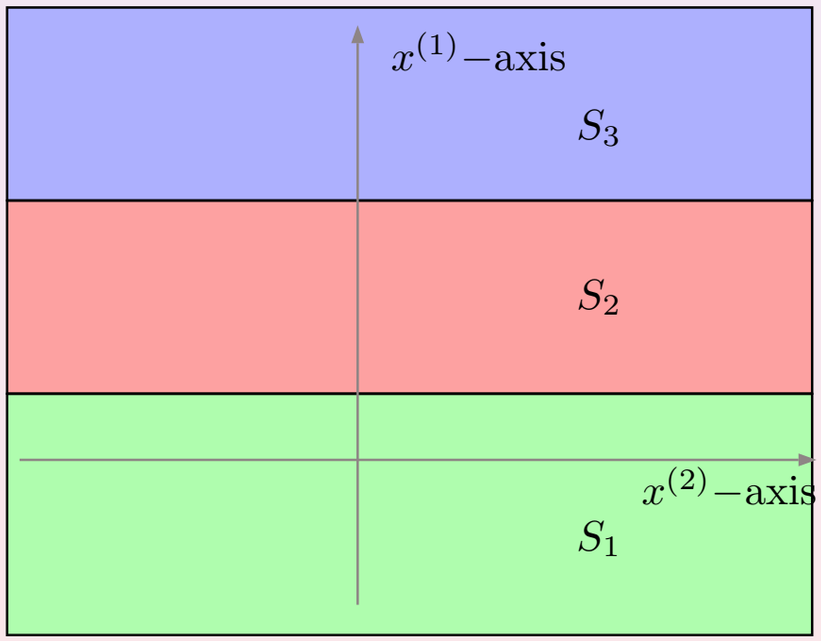
What is actually the ResNet doing?

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.



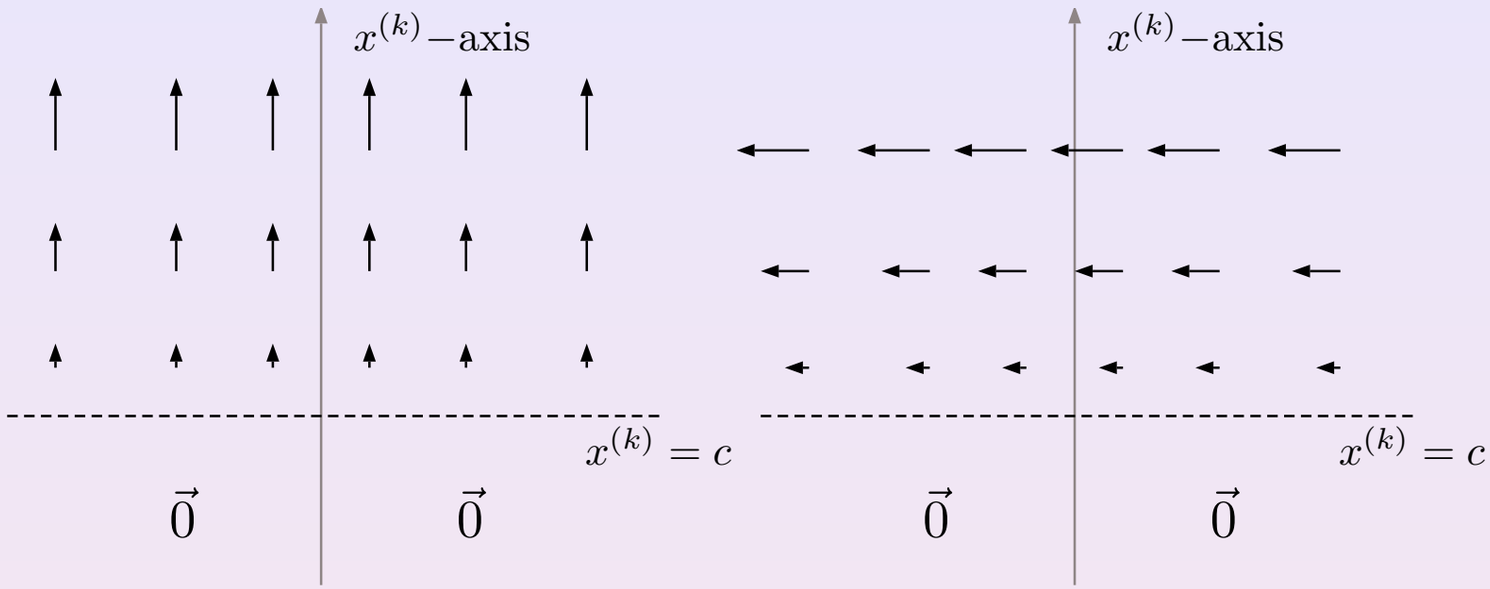


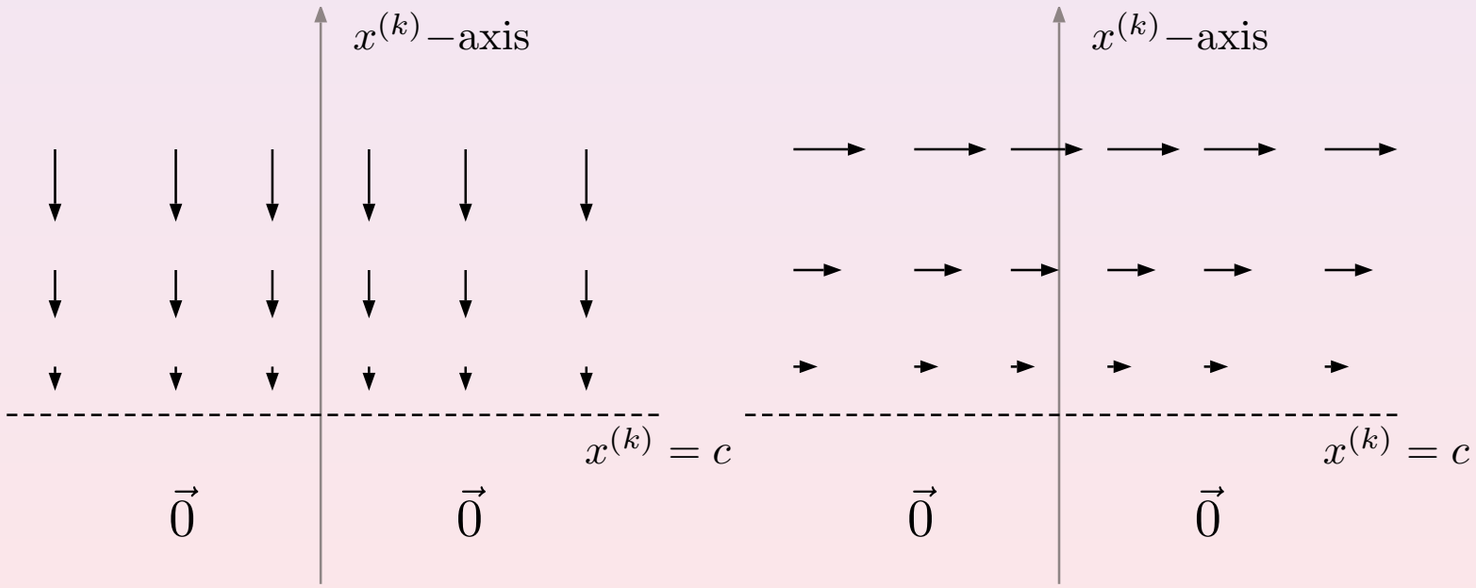
Basic control actions

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

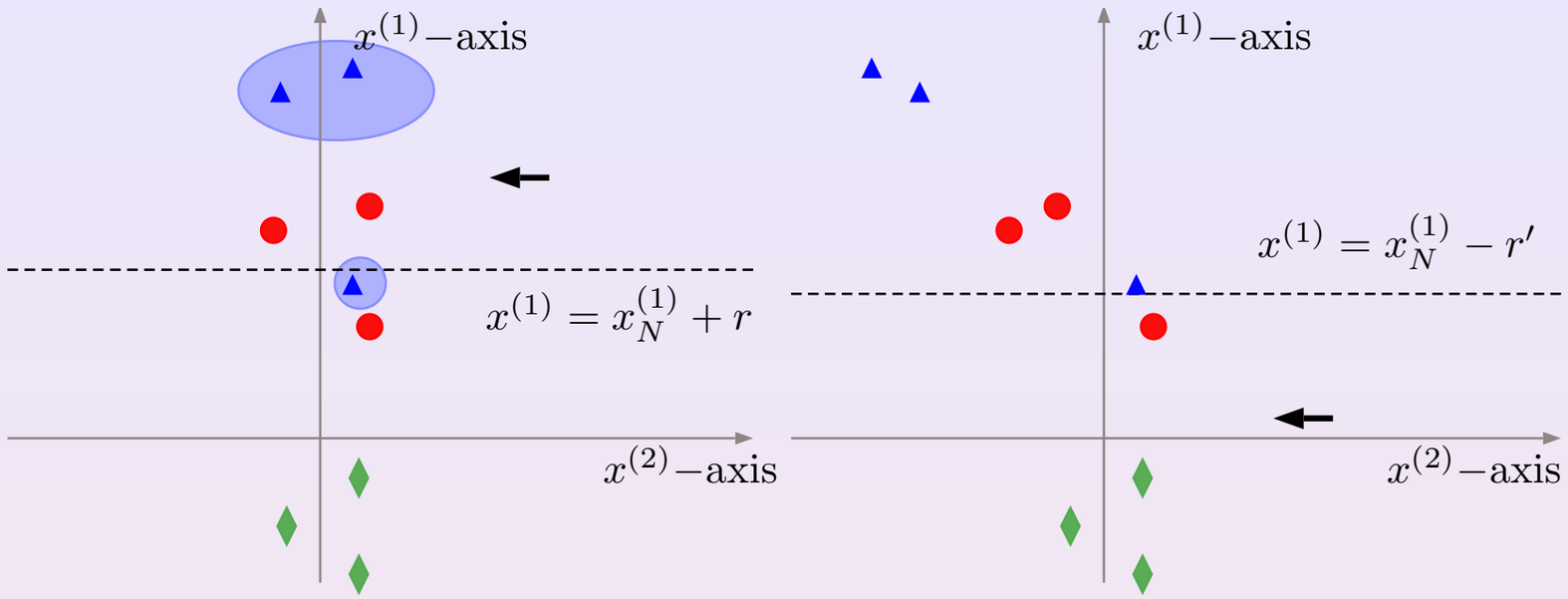
- $b(t)$ induces a time-dependent translation of the Euclidean space. It plays an important role to place the center of the action of the sigmoid.
- $A(t)$ compresses, expands, and induces rotations in the euclidean space with different objectives:
  - Compression can help gathering data into clusters 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.

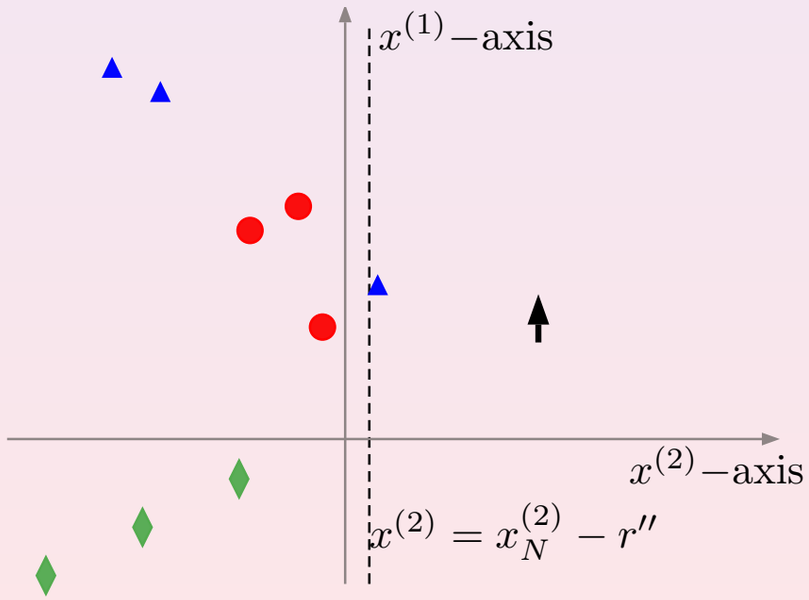
Some canonical flows induced by nODE

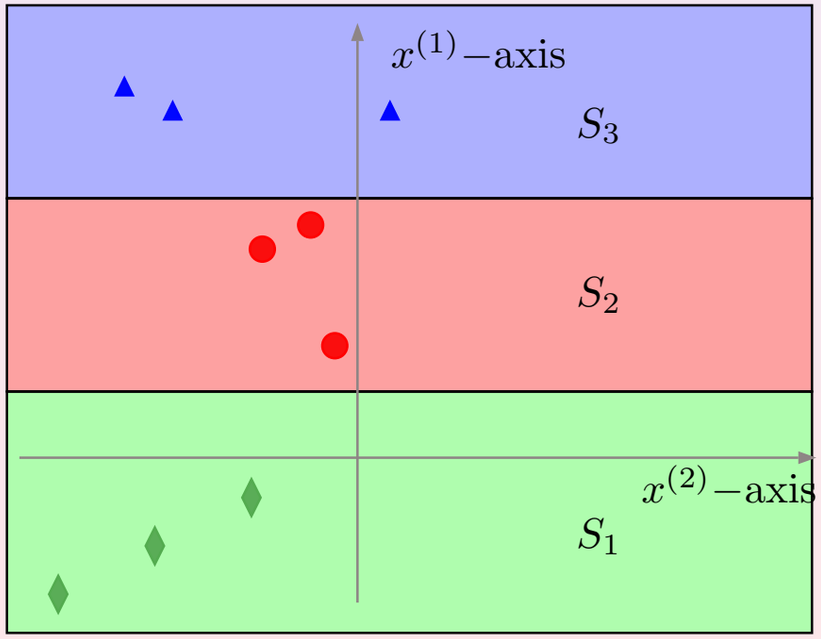




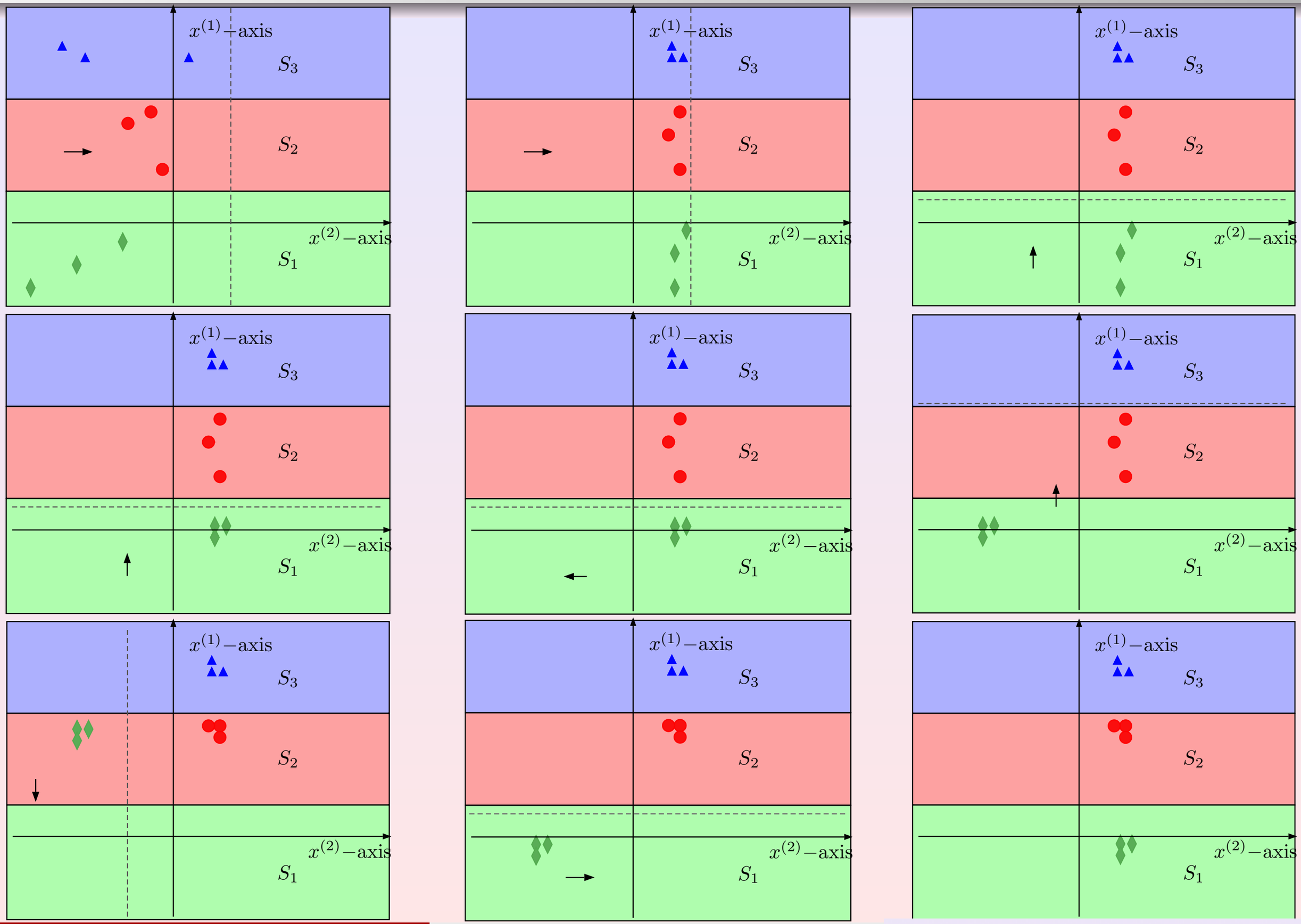
Controlling one datum

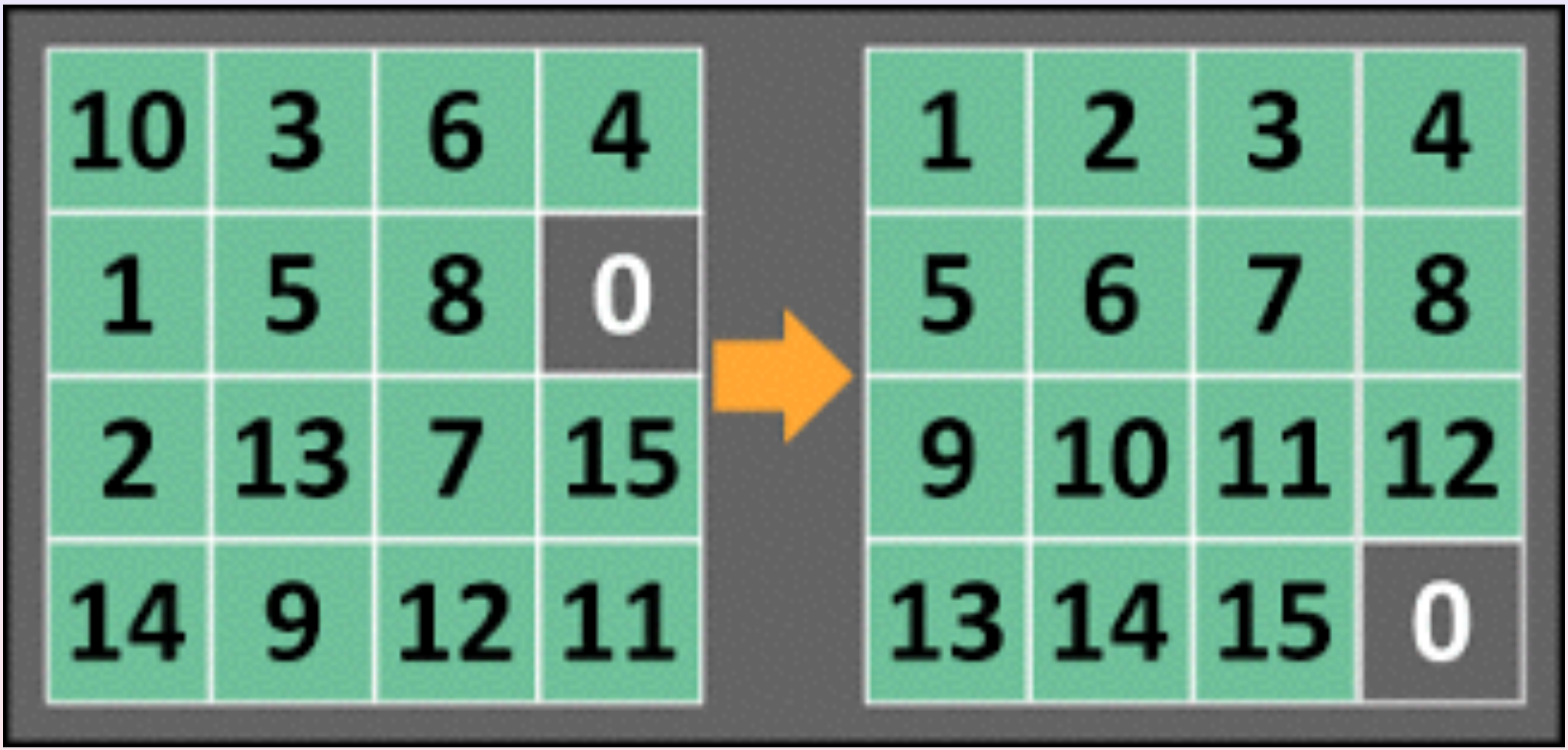






Compression after classification




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.

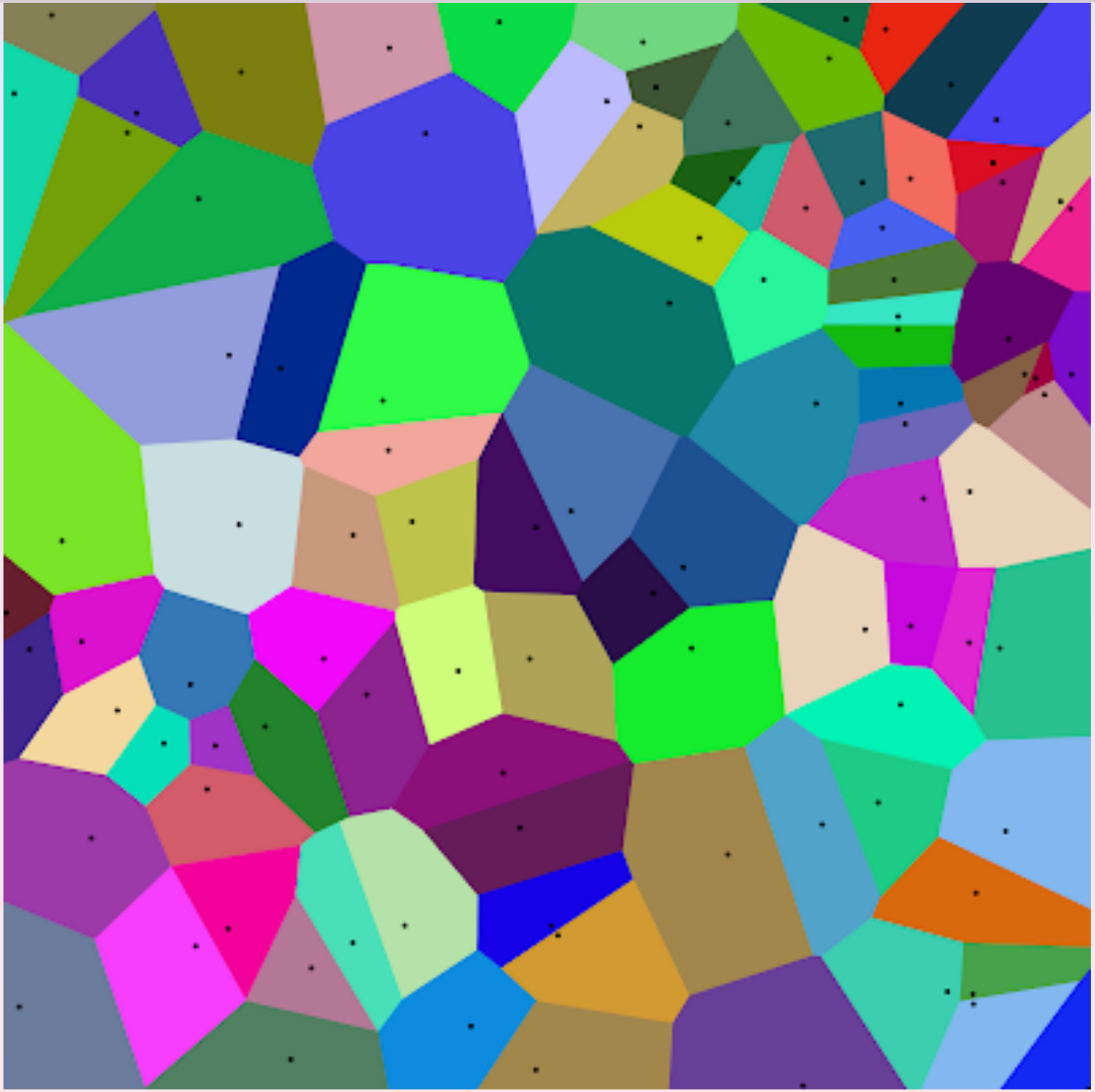
$$\phi_T(x_i; A, W, b) \in S_{m_i}, \quad \forall i = 1, ..., N.$$

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).

Complexity of the control strategy



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? [5]

[5]

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

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

Neural transport equations

The simultaneous control of the nODE

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

to arbitrary terminal states

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

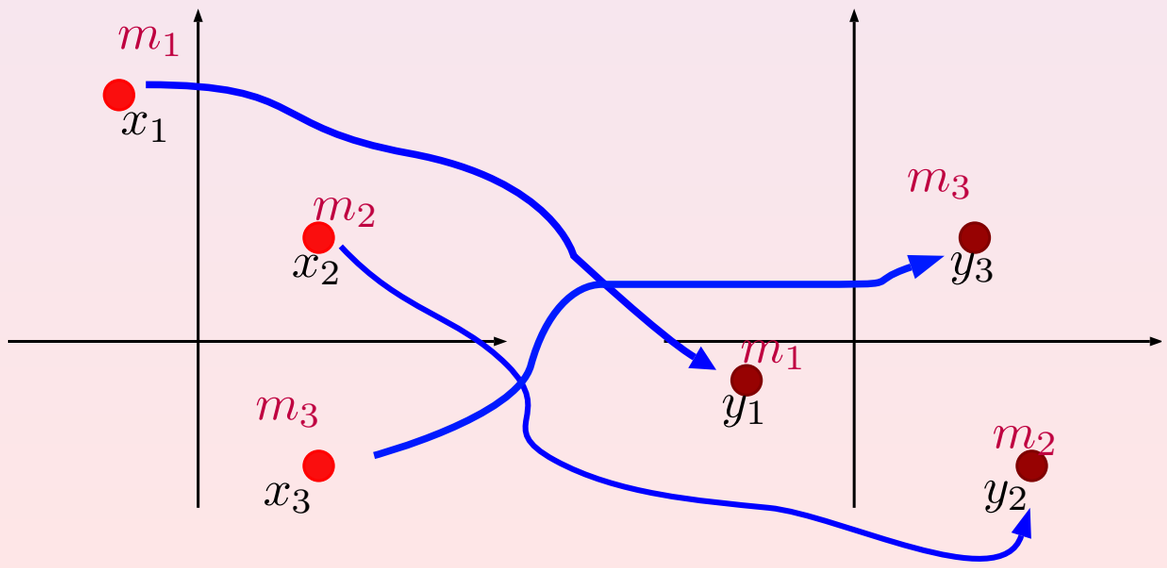
in terms of the transport equation, leads to the control of an atomic initial datum from

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

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.

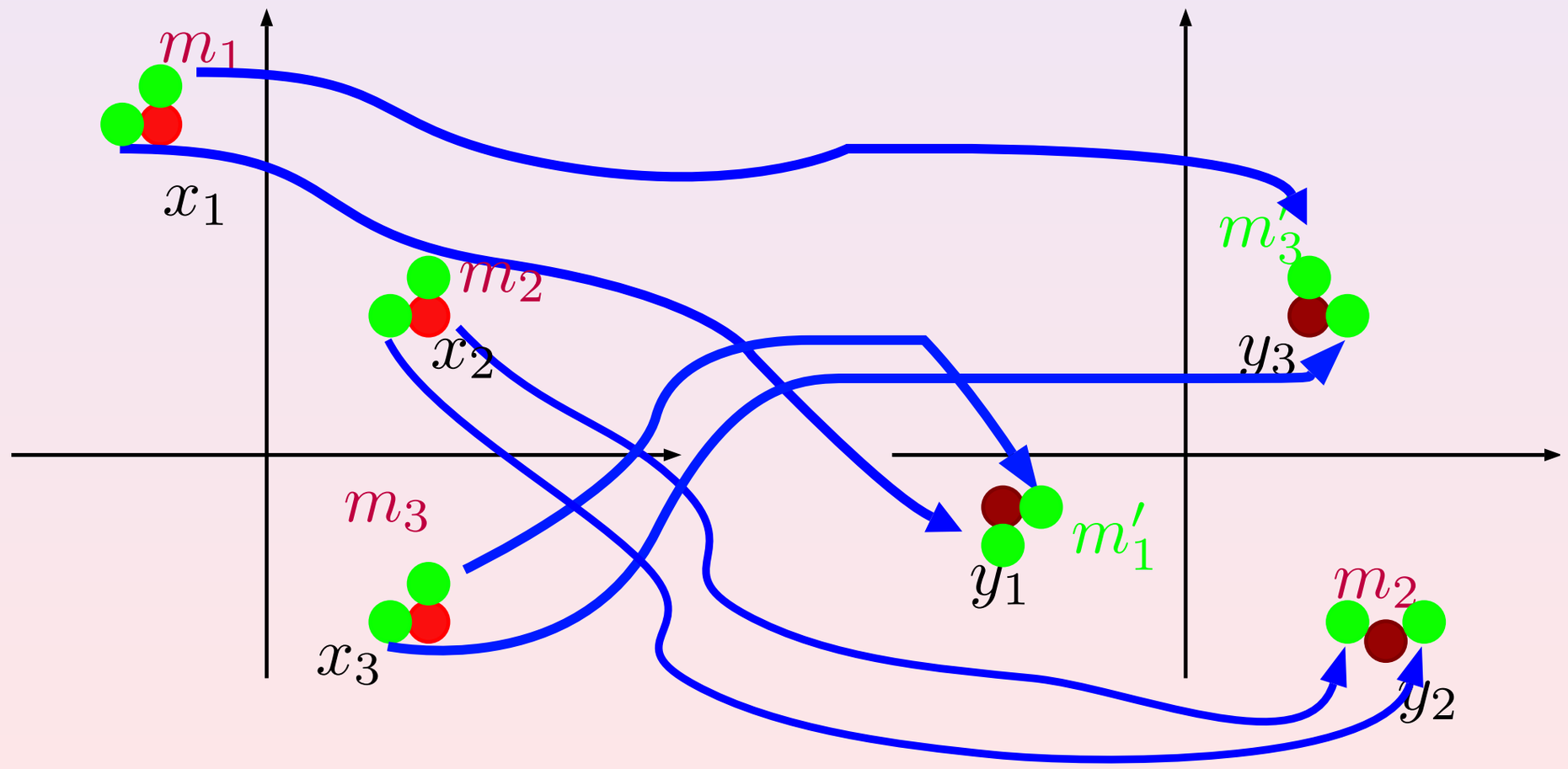


Neural transport equations

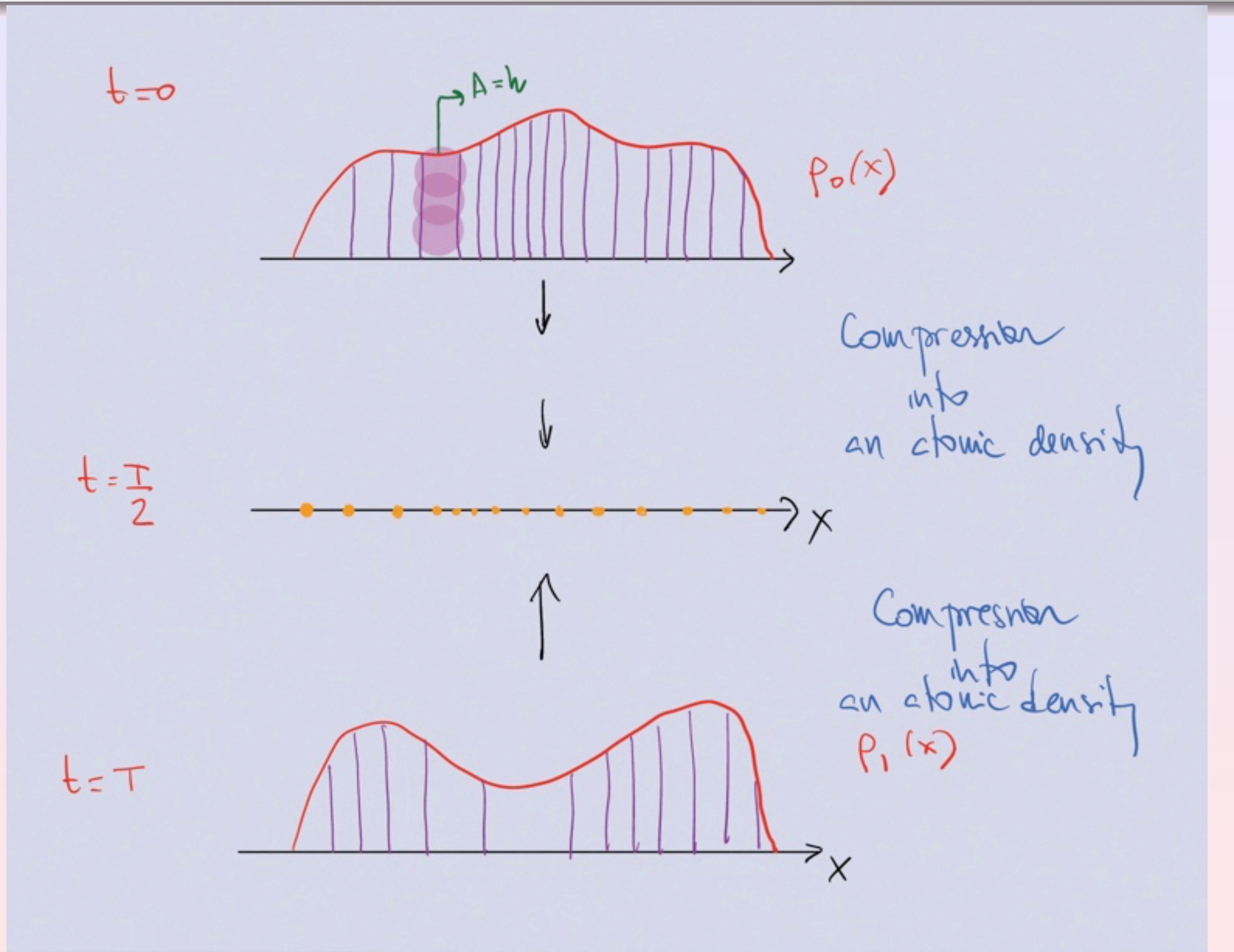
We can enrich the strategy above to also regulate the amplitude of the masses. But this requires to relax the control statement into an ε-approximate one.
For that to be done we need to split initial masses so that

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

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

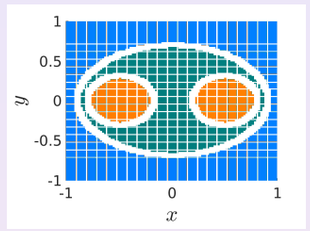


Neural Transport

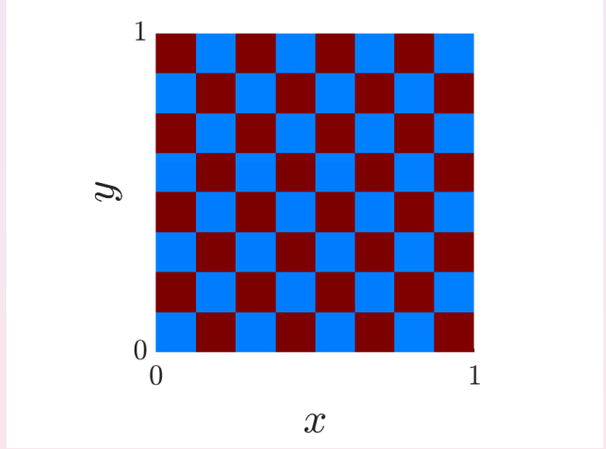


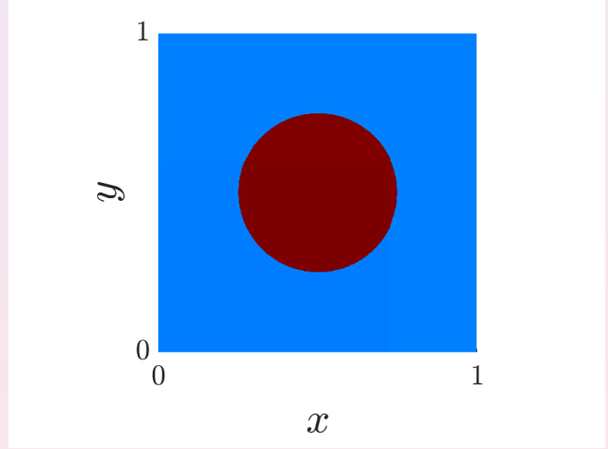
Universal approximation

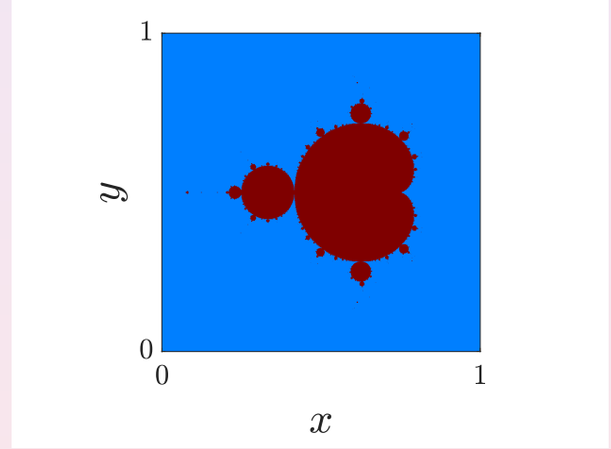
- As a corollary we can achieve Universal Approximation.
- By density, it is sufficient to consider targets that are simple piecewise constant functions.
- We can proceed making a partition of the departure and arrival spaces so that the problem becomes a countable version of simultaneous control of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



- The complexity of the needed controls depends on the nature of functions one aims to approximate.[6]







---

[6] $N_\Gamma(h)$ 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 $\epsilon^{-\frac{2dD}{d-D}}$.

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

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

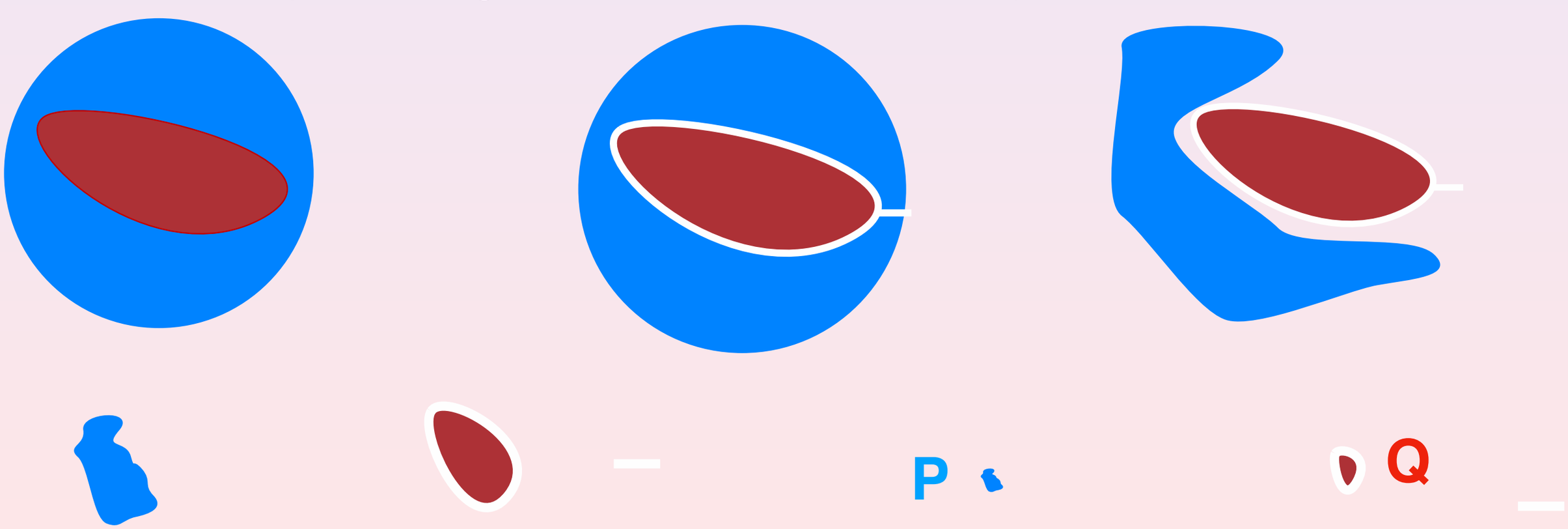
is such that

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

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.



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 (Innovalia), 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 & Deusto).
- Funded by the ERC Advanced Grant DyCon and an Alexander von Humboldt Professorship and Marie-Sklodowska Curie ITN "ConFlex"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 765579.