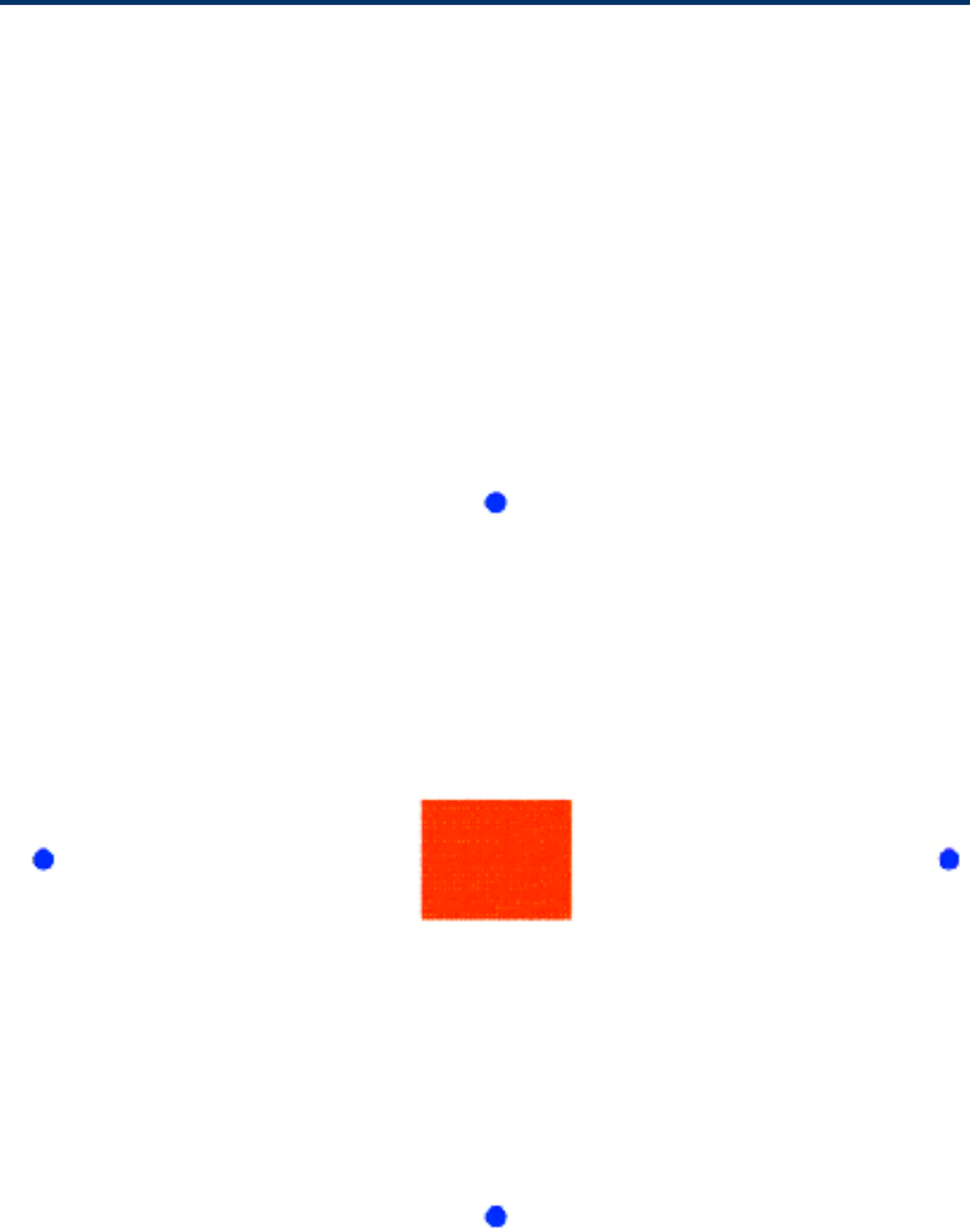
Control and Machine Learning

Enrique Zuazua

FAU & AvH, Erlangen, Germany

Outline

Two neighbouring fields



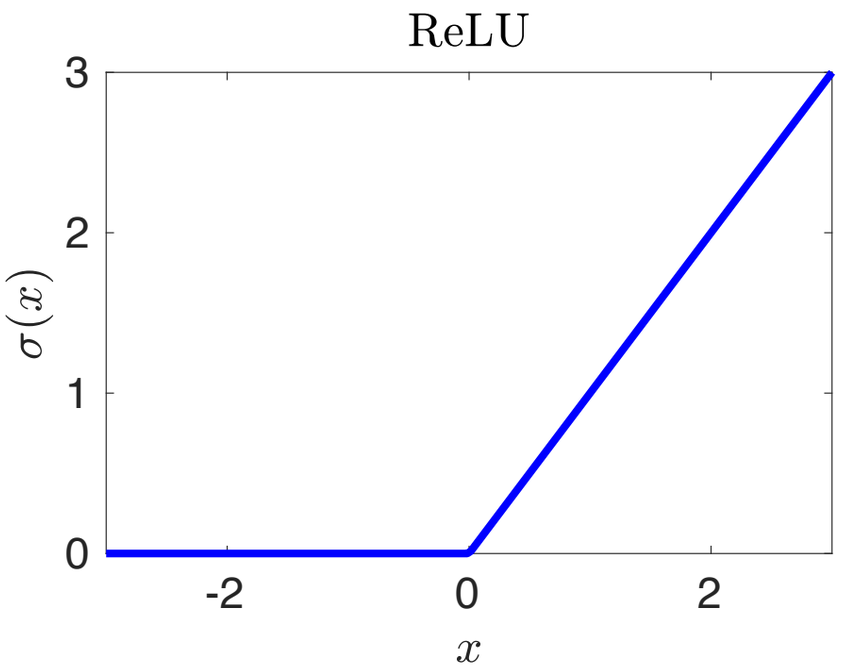
Control: Dogs-Sheep

Supervised Learning

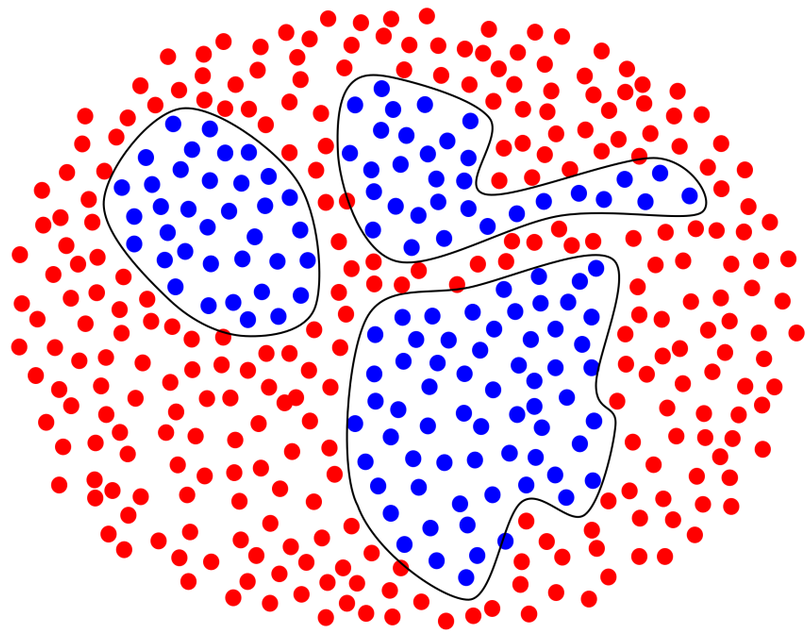
Neural differential equations

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

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



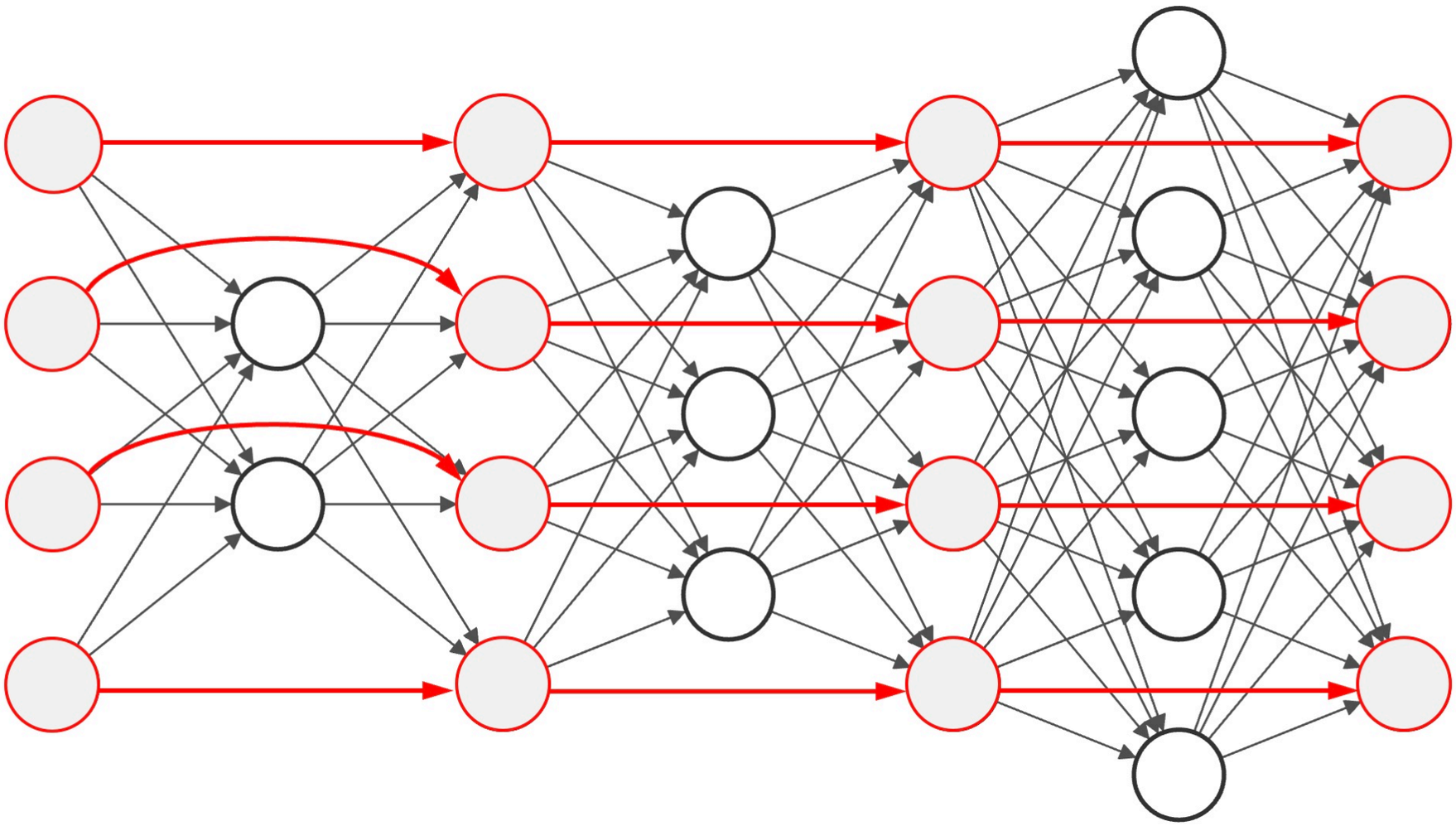
$$f(x) \sim \sum_{j=1}^{K} \mathbf{w}_j\ \sigma(\mathbf{a}_j\cdot x + b_j)$$



Outline

Standard computational practice

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\int_0^T \text{loss}\left(x_i(t), y^{(i)}\right)\,dt}_{\text{empirical risk} := E(x(\cdot))} + \alpha \int_0^T \|(\mathbf{a}(t), \mathbf{w}(t), b(t))\|^2\,dt$$

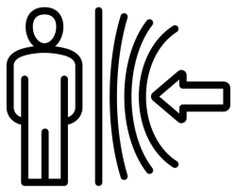


Objectives

To open the black-box of Machine Learning with Control theoretical tools

Explainability

Generalization



Robustness

Complexity

Computational cost

New ML methods

Our recent contributions

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

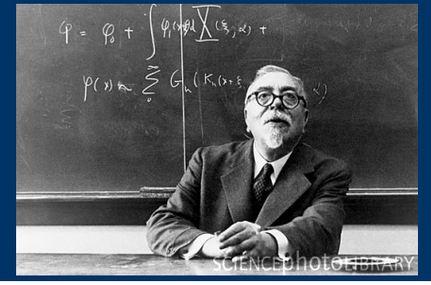
B. Geshkovski, E. Zuazua, *Turnpike in optimal control of PDEs, ResNets, and beyond*, Acta Numer. 31 (2022), 135–263

D. Ruiz-Balet, E. Zuazua, *Neural ODE control for classification, approximation and transport*, SIAM Review, to appear.

Outline

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.

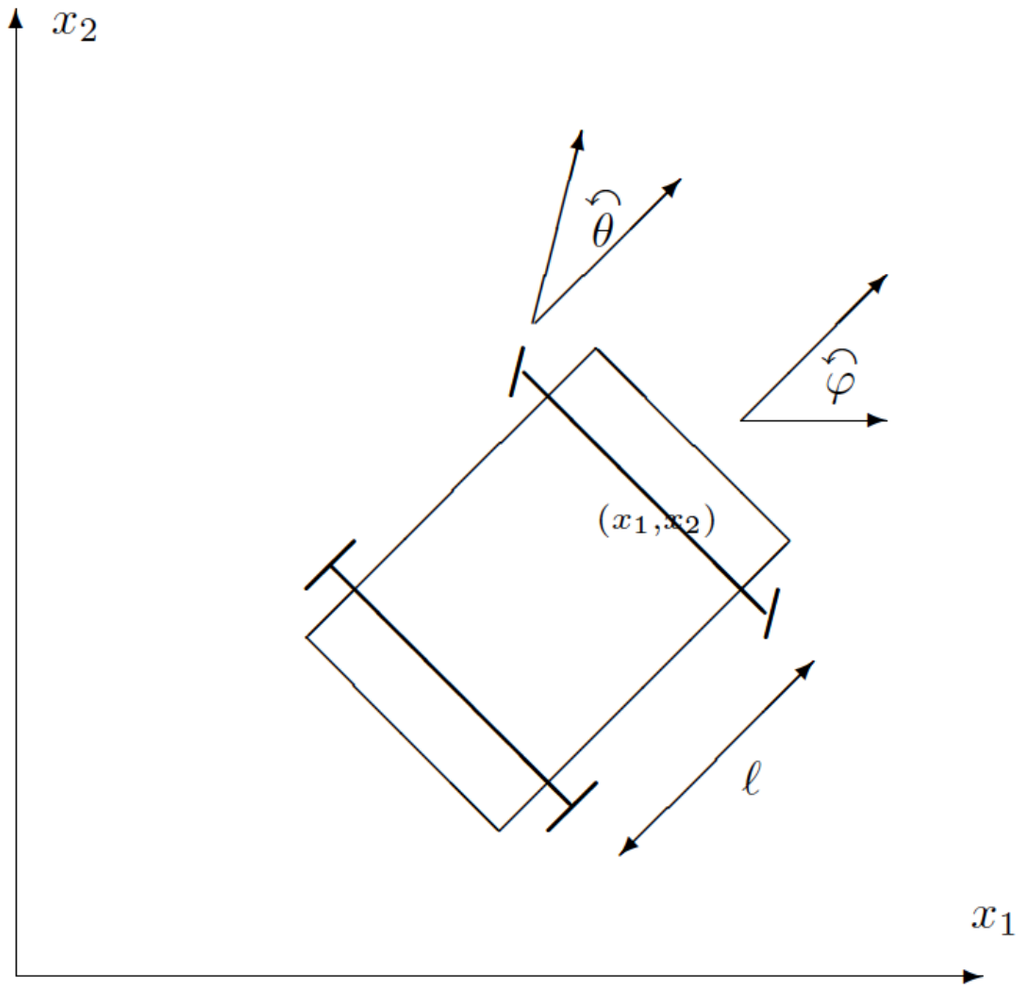
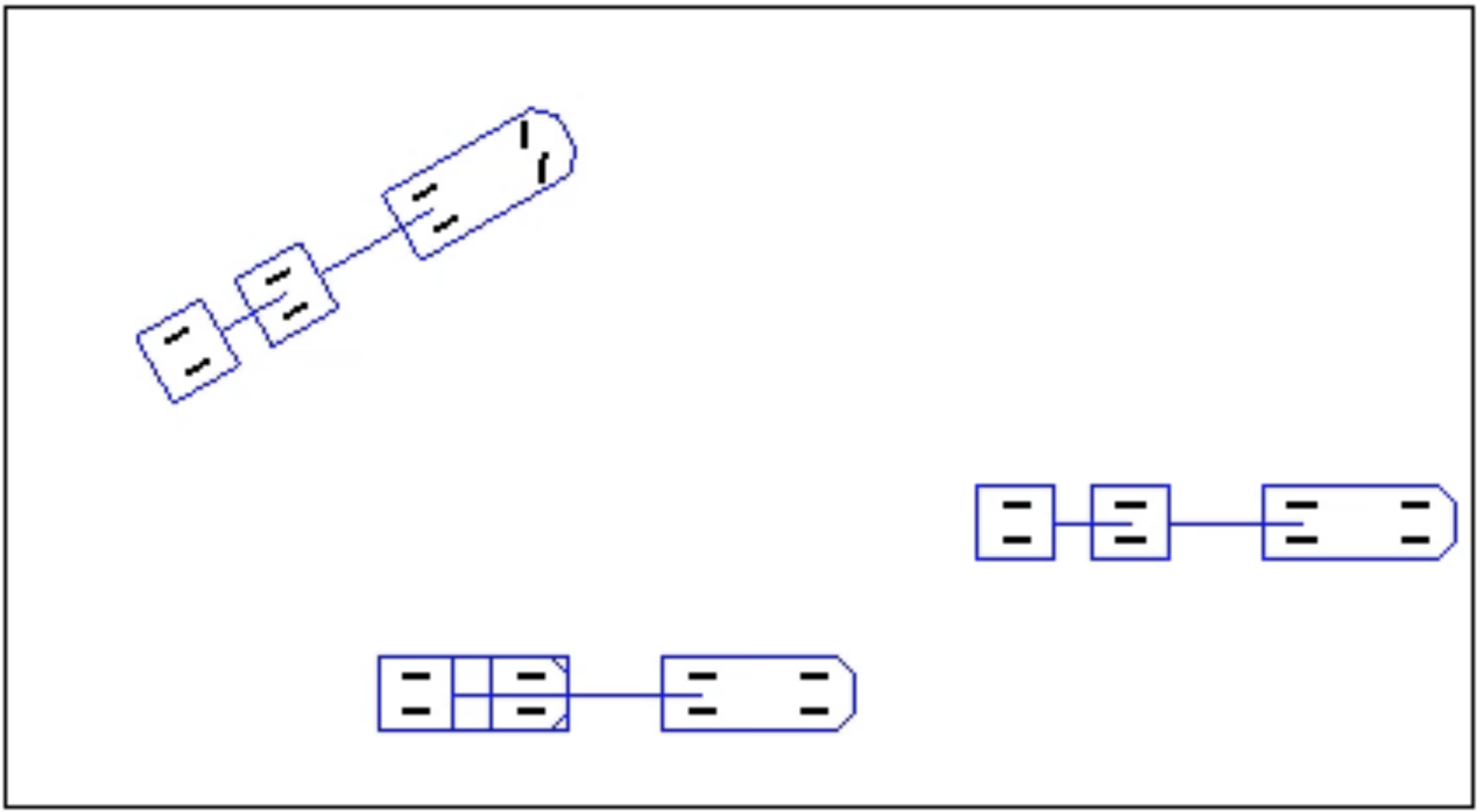


Figure 4.1: *4-dimensional car model.*

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

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.

Outline

Supervised learning

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

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

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

Classification: match points (images) to respective labels (cat, dog).
This is typically done **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.
[4] E. Sontag, H. Sussmann, 1997.

For each item $i = 1, ..., N$:

ResNets: Residual Neural Networks

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + hA^k\sigma(w^k\mathbf{x}_i^k + b^k)$$
$$\text{for} \quad k \in \{0, \dots, N_{layers} - 1\}$$

NODEs: Neural Ordinary Differential Equations

$$\dot{\mathbf{x}}_i(t) = A(t)\sigma(w(t)\mathbf{x}_i(t) + b(t))$$
$$\text{for} \quad t \in (0, T)$$

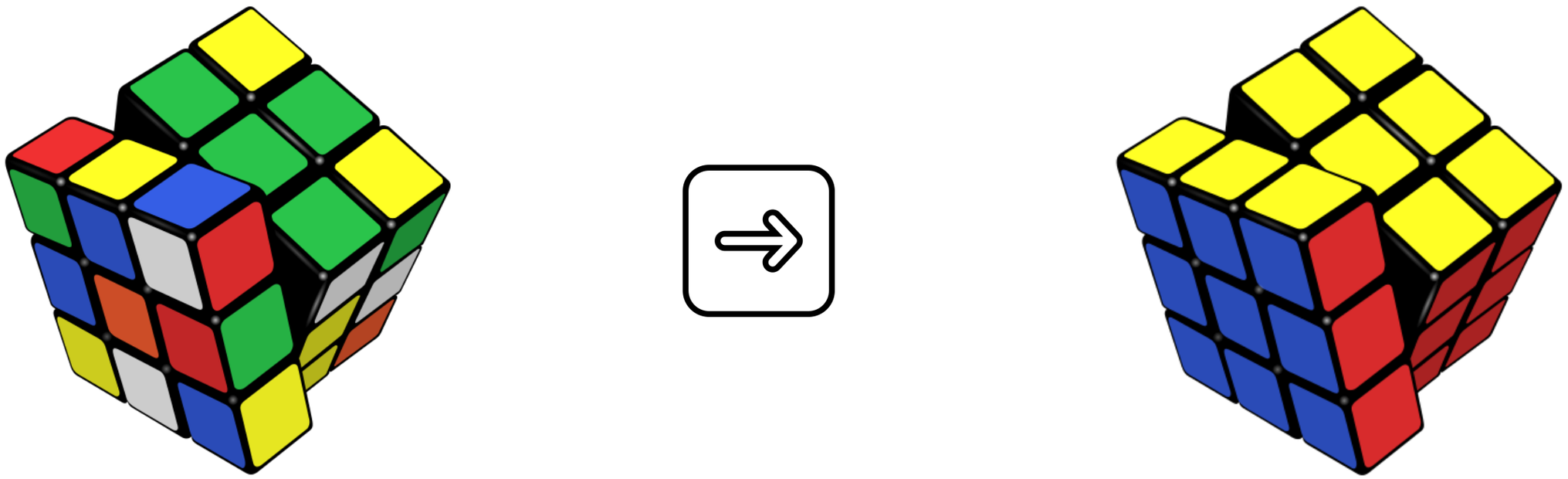
This constitutes a huge ensemble or simultaneous control problem.

ResNets in action (Borjan Geshkovski)

ResNets and the Rubik Cube

- 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 achieving this giant simultaneous control goal. The fact that σ leaves half of the phase space invariant while deforming the other one allows for dynamics not encountered in mechanics, for which such kind of simultaneous control property is unlikely.



Classification by control

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

a

Consider the NODE with the ReLU as activation function. Then, in any time horizon $[0, T]$*, a finite arbitrary number of items can be driven to an arbitrary number of open subsets of the Euclidean space corresponding to its labels.*

- *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.*

[a] 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.

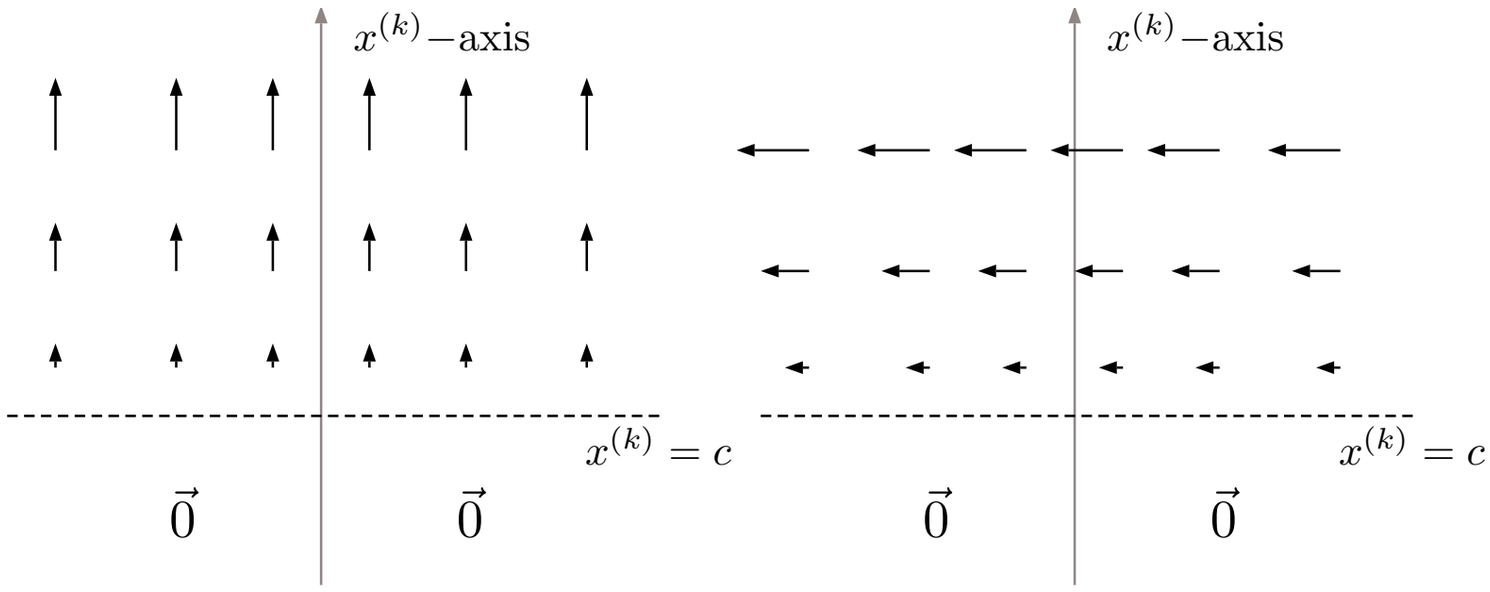
What is the ResNet doing? Basic control actions

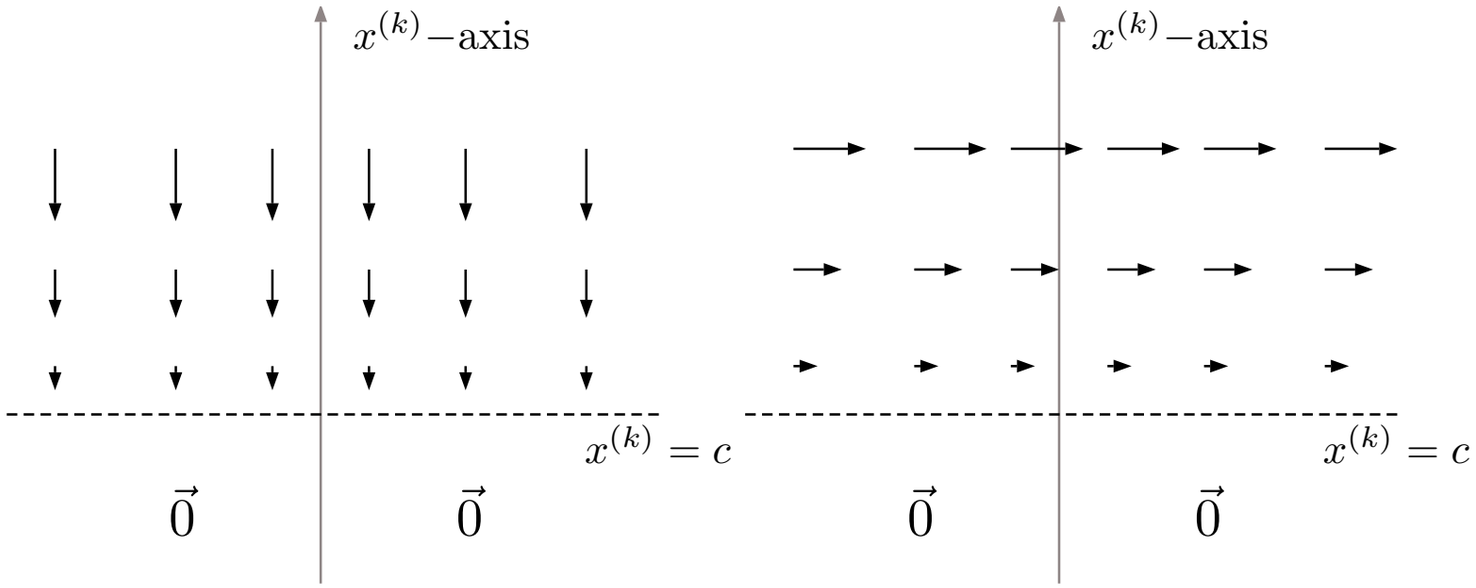
$$\dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t))$$

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

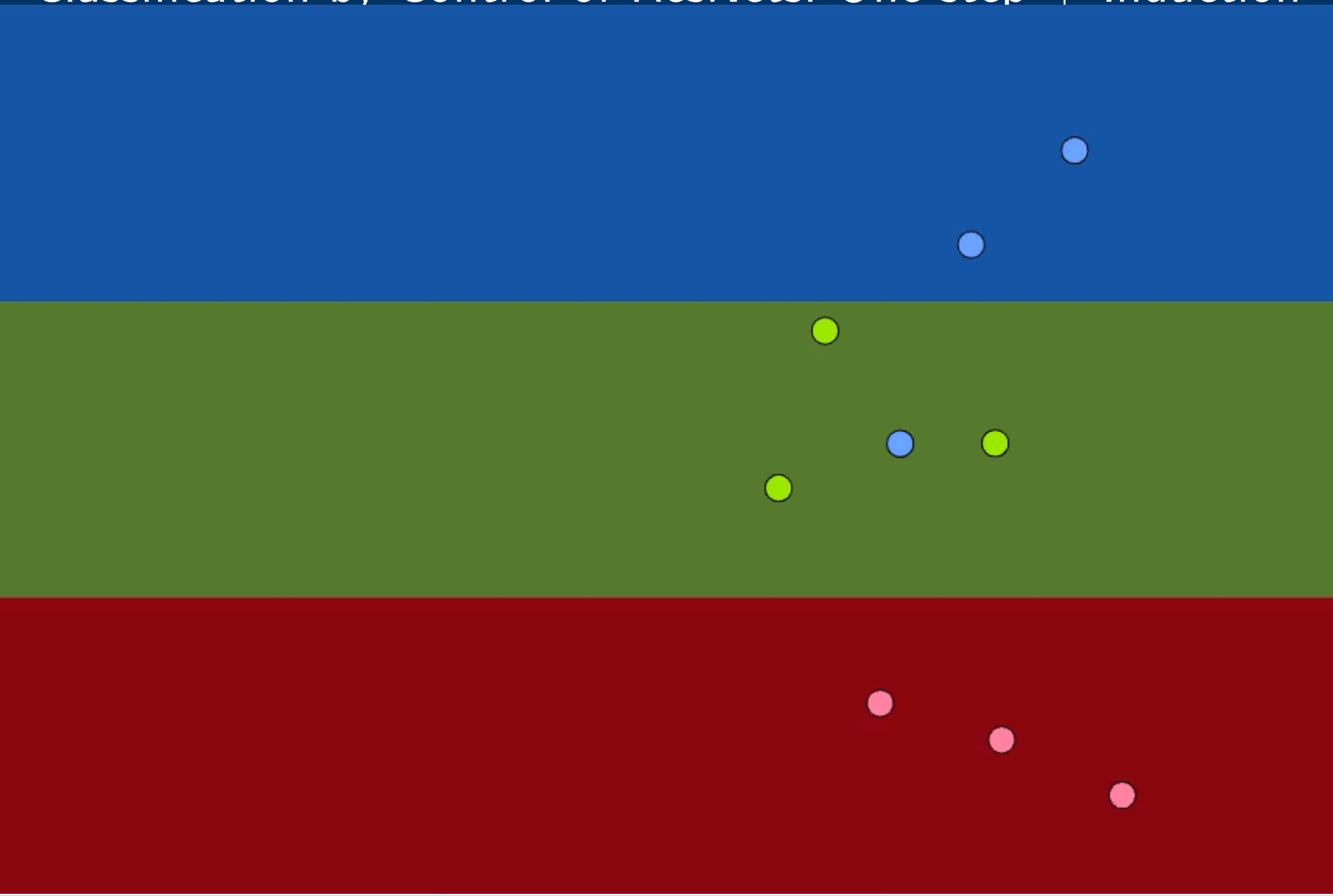
An intelligent piecewise constant choice of controls, by induction, assures the needed simultaneous control property.

Some canonical flows induced by nODE



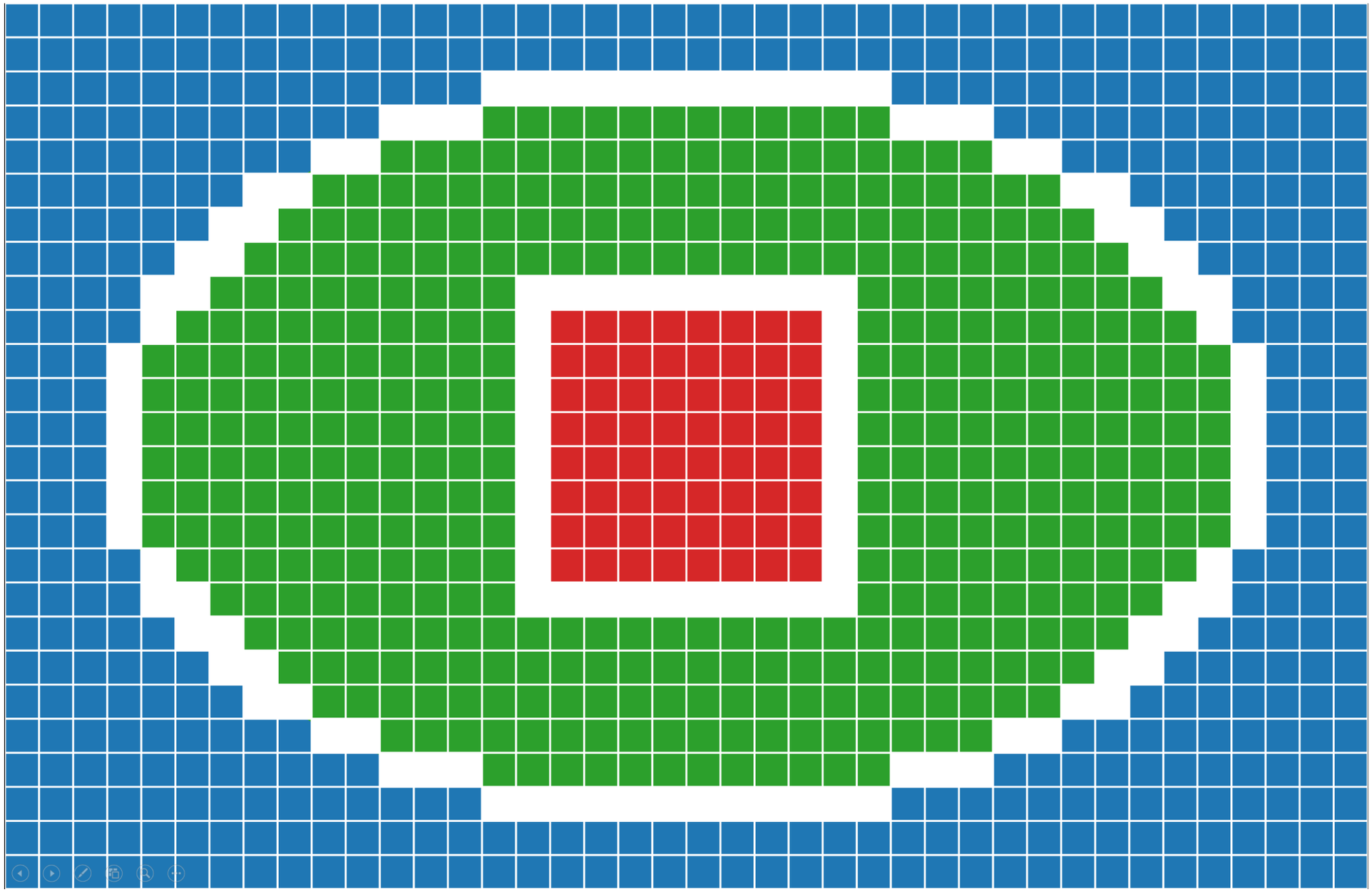


Classification by Control of ResNets: One step + Induction



Complexity

$$||w_\epsilon||_\infty \leq Per(\Gamma)\epsilon^{-2d^2}$$



Outline

Deep neuronal network (non-residual case)

Let $L \geq 1$ and the parameters $\{\mathbf{a}^k, b^k\}_{k=1}^{L} \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$ with $d_k \geq 1$ for every $k \in \{0, \dots, L-1\}$.

Consider the discrete dynamics

$$\mathbf{x}^{k+1} = \sigma(\mathbf{a}^k \cdot \mathbf{x}^k + b^k), \quad k \in \{0, \dots, L-1\}.$$

Here σ corresponds to the ReLU activation function, possibly interpreted in a vector-valued form,

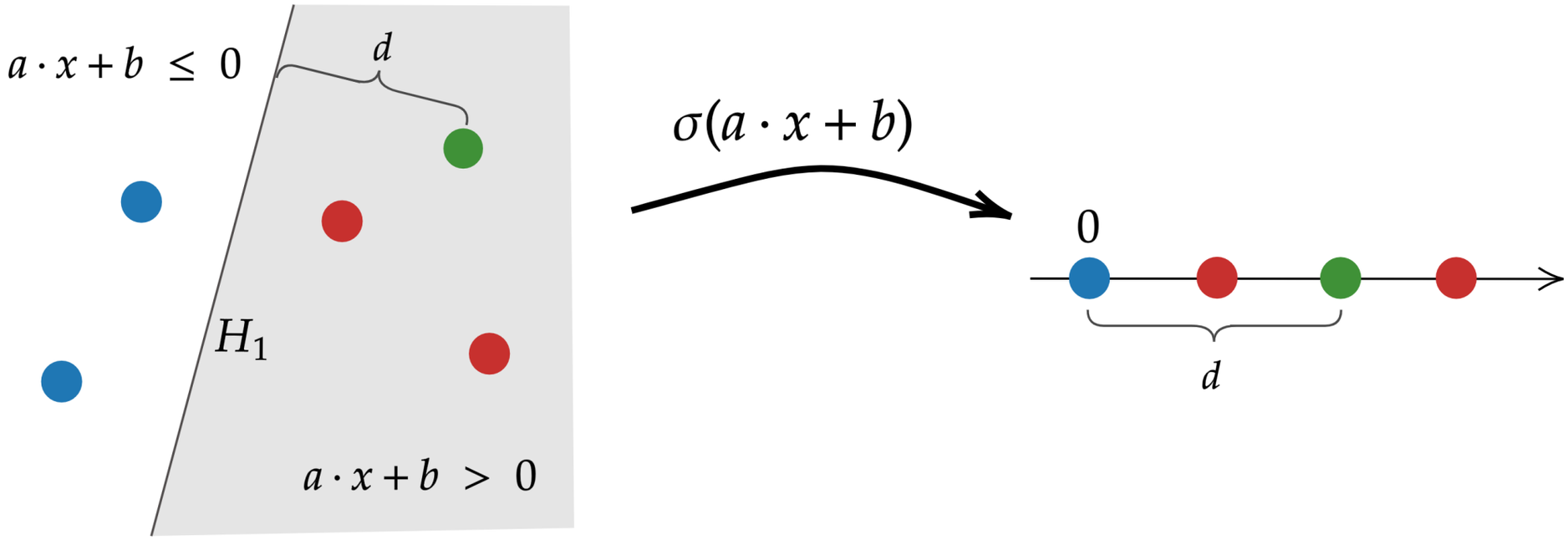
$$\sigma \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_d) \end{pmatrix}.$$

Geometric analysis of dynamics I

If $a \in \mathbb{R}^{1\times 2}$ and $b \in \mathbb{R}$ then

$$H(a,b) = \{x \in \mathbb{R}^2 : a \cdot x + b = 0\},$$

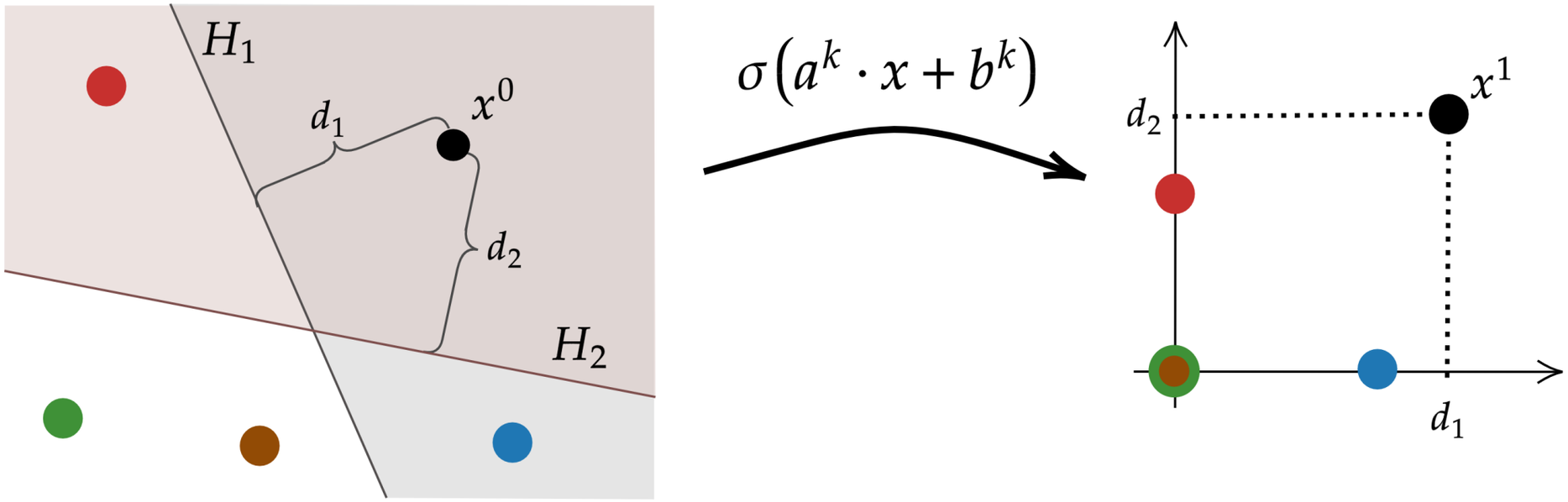
define a hyperplane.



All points at the left of the hyperplane H_1 collapse to zero.

Geometric analysis of dynamics II

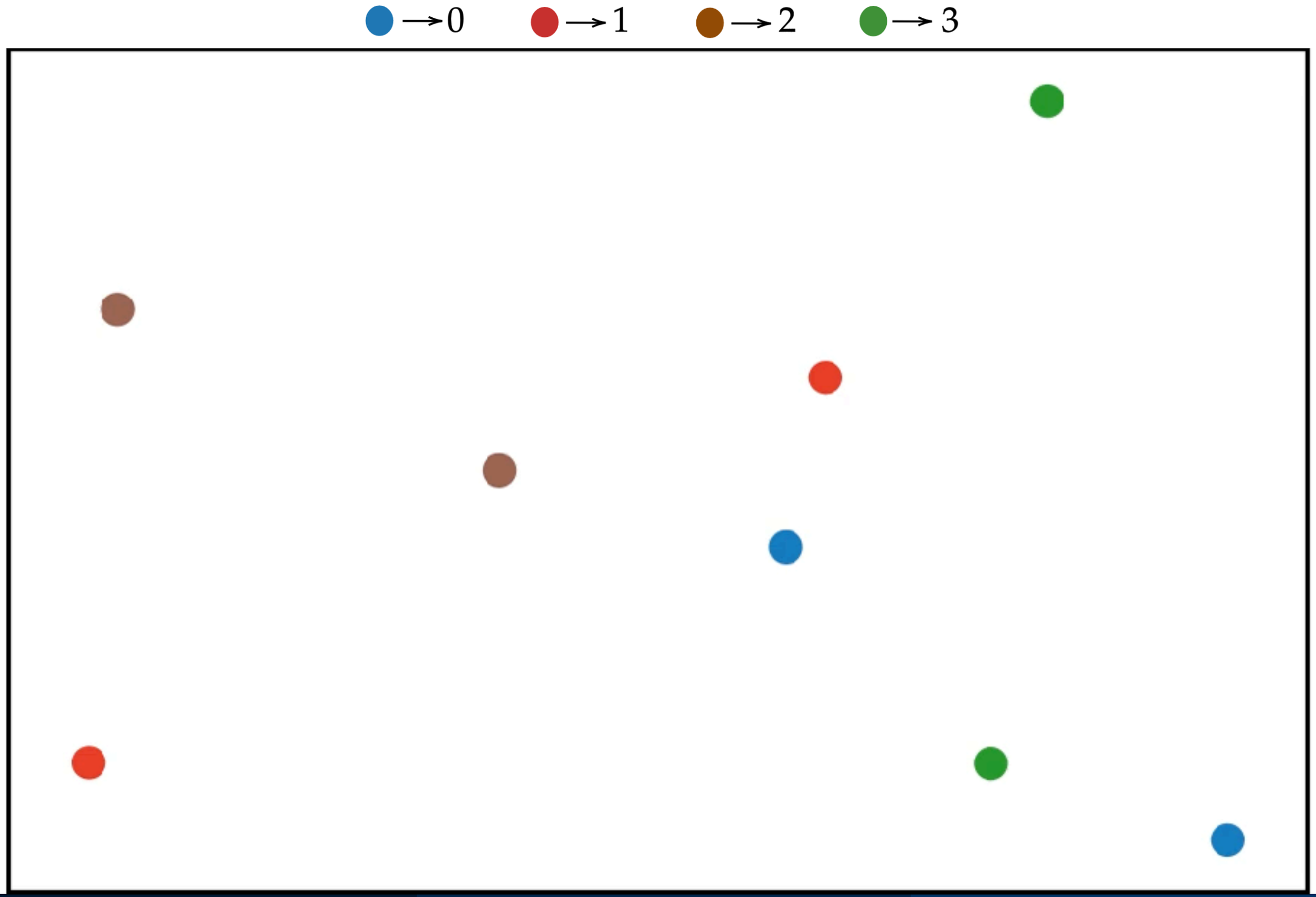
When $(a_1, a_2)^T = a \in \mathbb{R}^{2\times 2}$ and $(b_1, b_2)^T = b \in \mathbb{R}^2$ they define two hyperplanes $H_1(a_1, b_1)$ and $H_2(a_2, b_2)$.



Different regions are mapped to different locations. All points in the white region are mapped to the same position $(0, 0)$.

Idea: Construct the parameters $\{a^k, b^k\}_k$ such that in each iteration, points of the same color collapse in the same point.

Deep neural network in action

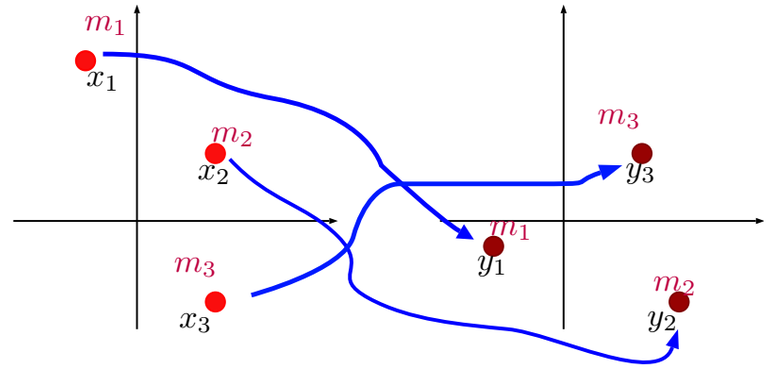


Outline

Neural transport equations

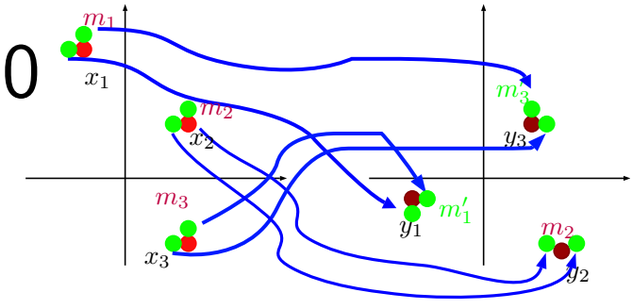
Note that the differential equation

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



corresponds to the characteristics of the transport equation:

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



Atomic initial data can be driven to atomic final targets.

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

Outline

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.

Thank you for the invitation and attention