Some control aspects in deep learning

Domènec Ruiz-Balet[1]

[1]Imperial College London

25th August 2022

Benasque, IX Partial differential equations, optimal design and numerics, Special session for Enrique Zuazua's 60th birthday

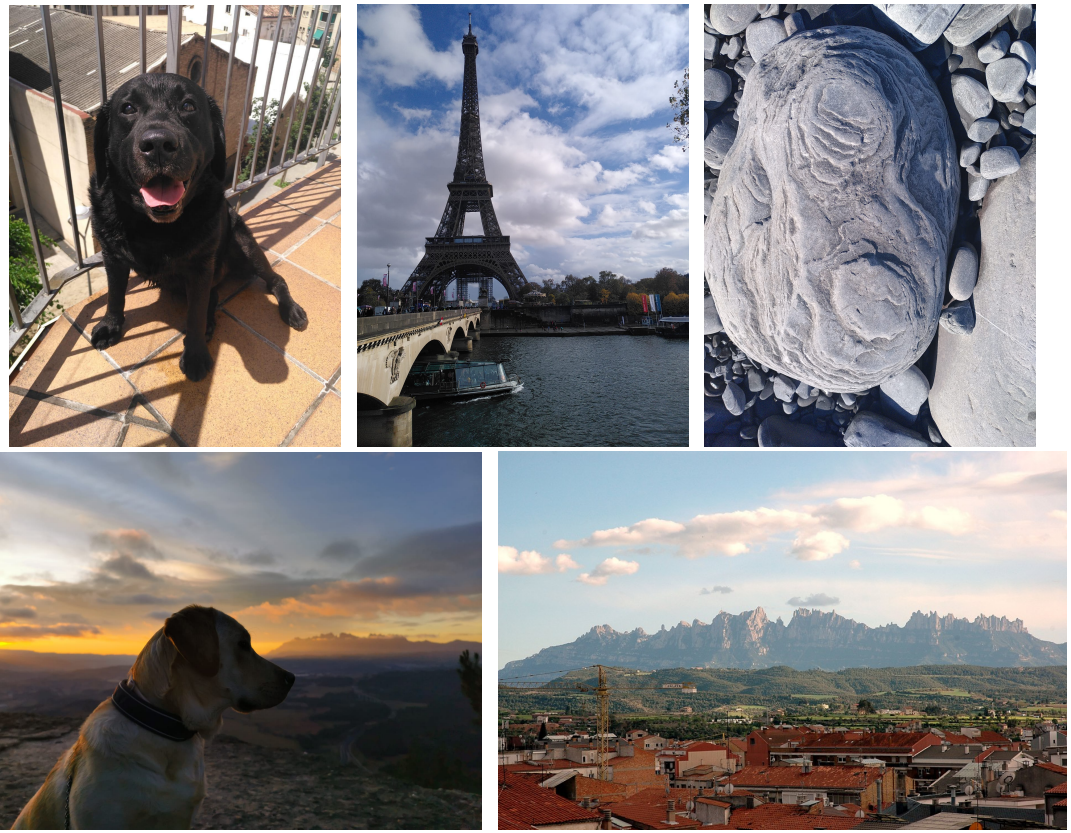
Joint works with Enrique Zuazua

Imperial College London

Elastic Manufacturing Systems

Classification

We have a large data set of pictures



How can we separate the dog pictures from the other types of pictures?

Learning a function through samples

Let $A \subset \mathbb{R}^d$ be the subset of dog pictures.

Classification

Can we recover the function $\mathbb{1}_A$ given a finite amount of samples of $\mathbb{1}_A$?

$$\{(x_i, y_i = \mathbb{1}_A(x_i))\}_{i=1}^N \subset \mathbb{R}^d \times \{0, 1\}$$

Approach

The approach would be to find an approximation of $\mathbb{1}_A$ in a "large" family of functions

The family we will consider are Neural Networks

- Deep Residual networks (ResNets)[1] were implemented to ease the training of deep neural networks.

$$x^{k+1} = x^k + W^k \sigma(A^k x^k + b^k)$$

- This formulation reminds an Euler discretization of an ODE[2][3].

$$x' = W(t)\sigma(A(t)x + b(t))$$

[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.

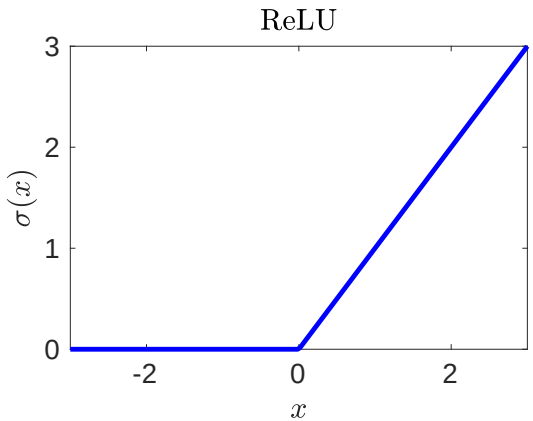
[3] R. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud 2018. Neural ordinary differential equations

A Neural ODE has the form of:

$$x' = W(t)\sigma(A(t)x + b(t))$$

where
Consider $\sigma : \mathbb{R} \to \mathbb{R}$ being the ReLU

$$\sigma(x) = \max(x, 0)$$



Control problem

Let $f : \mathbb{R}^d \times \mathbb{R}^{d_u} \to \mathbb{R}^d$ be a Lipschitz function.
Controllability problem: Let $x_0, x_T \in \mathbb{R}^d$, $\exists u \in L^\infty((0,T);\mathbb{R}^{d_u})$ s.t.

$$\begin{cases} x' = f(x,u) \\ x(0) = x_0, x(T) = x_T \end{cases}$$

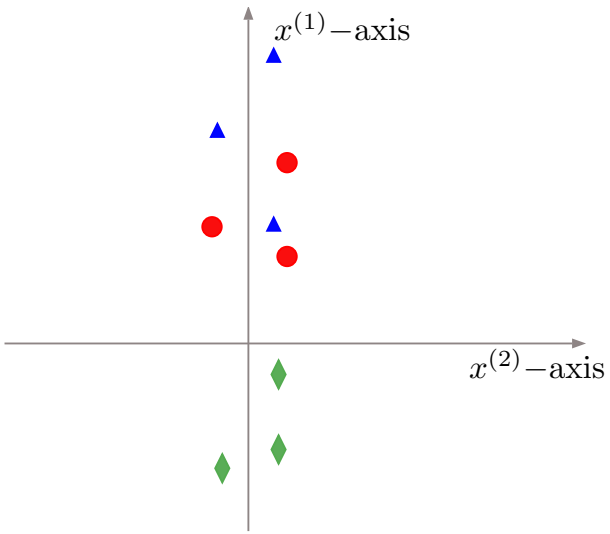
is satisfied?
Simultaneous controllability problem: Does $\boldsymbol{u} \in L^\infty((0,T);\mathbb{R}^{d_u})$ exist such that

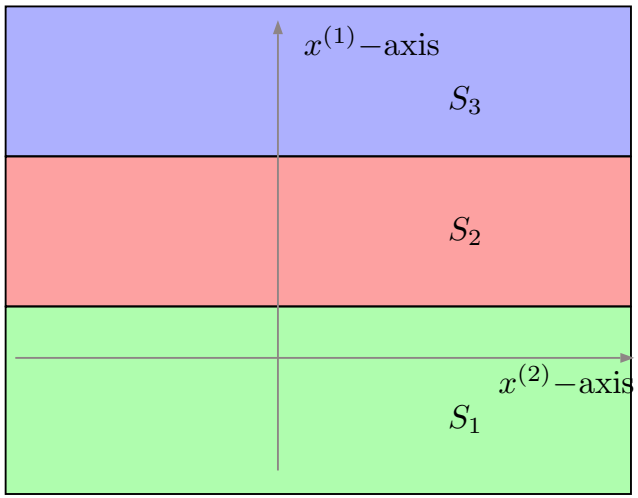
$$\begin{cases} x' = f(x,\boldsymbol{u}) \\ x(0) = x_{0,1}, x(T) = x_{T,1} \end{cases} \qquad \begin{cases} x' = f(x,\boldsymbol{u}) \\ x(0) = x_{0,2}, x(T) = x_{T,2} \end{cases}$$

are satisfied?

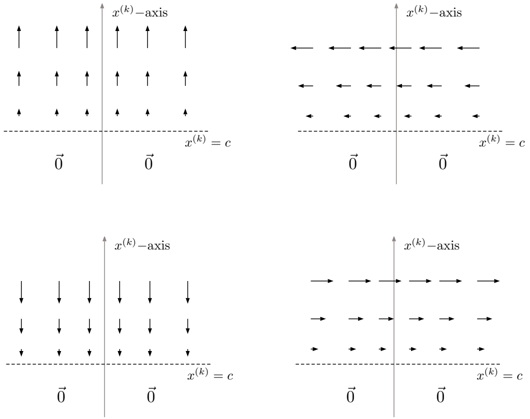
Consider N samples $\{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^d \times \{1, ..., M\}$ where x_i is the data associated with a class $y_i \in \{1, ..., M\}$.

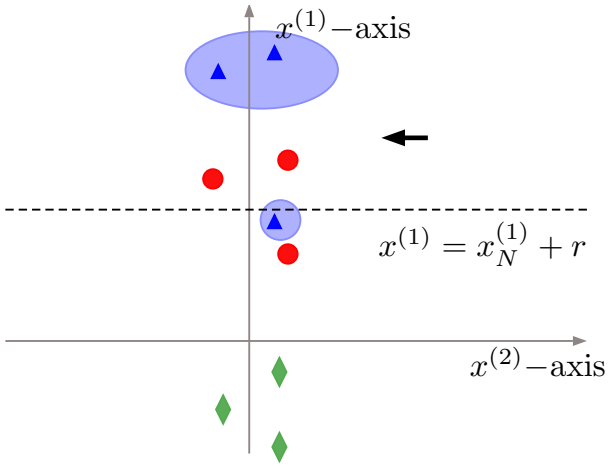
Find a **control** strategy that brings **simultaneously** all points to their **prefixed locations**

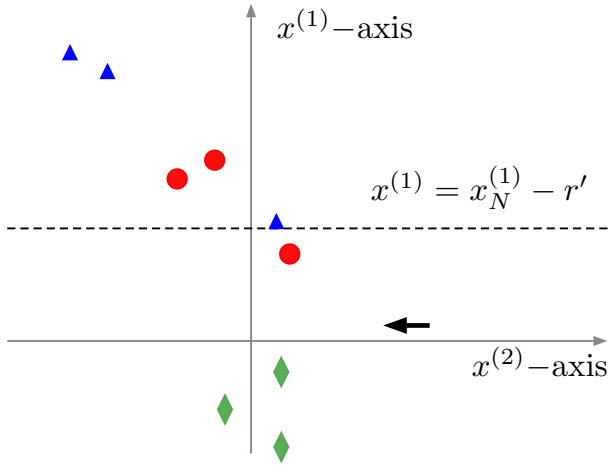


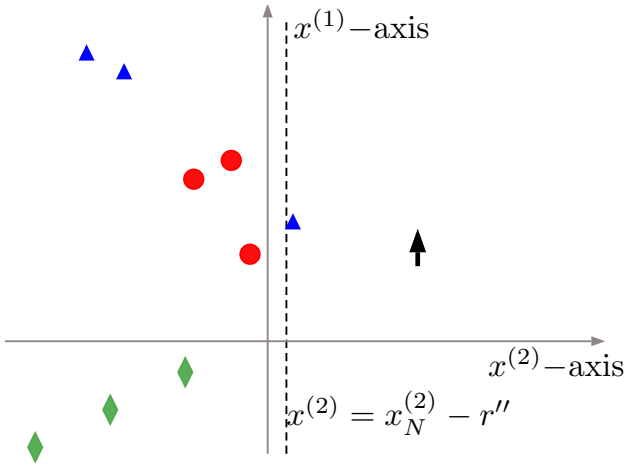


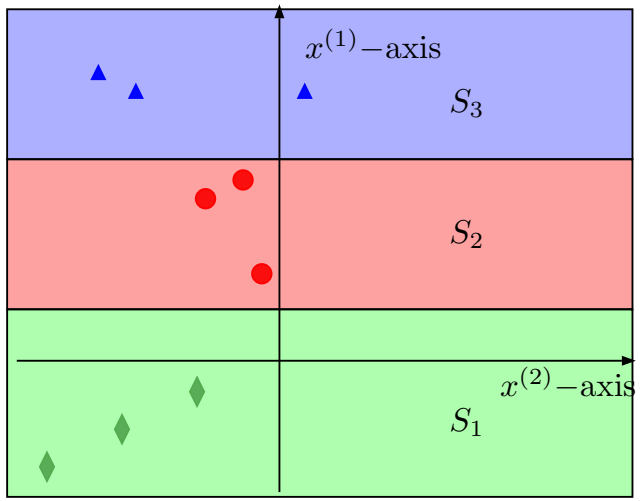
$$x' = W(t)\sigma(A(t)x + b(t))$$




$x^{(1)}-$axis
$x^{(1)} = x_N^{(1)} + r$
$x^{(2)}-$axis


$x^{(1)}-$axis
$x^{(1)} = x_N^{(1)} - r'$
$x^{(2)}-$axis


$x^{(1)}-$axis
$x^{(2)}-$axis
$x^{(2)} = x_N^{(2)} - r''$


$x^{(1)}-$axis
S_3
S_2
S_1
$x^{(2)}-$axis

Classification

Let $d \geq 2$ and $M \geq 2$. Consider $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \{1, ..., M\}$. Assume $x_i \neq x_j$ if $i \neq j$. Then, for every $T > 0$, $\exists A, W \in L^\infty\left((0,T); \mathbb{R}^{d\times d}\right)$ and $b \in L^\infty\left((0,T), \mathbb{R}^d\right)$ such that

$$\phi_T(x_i; A, W, b) \in S_{m_i},$$

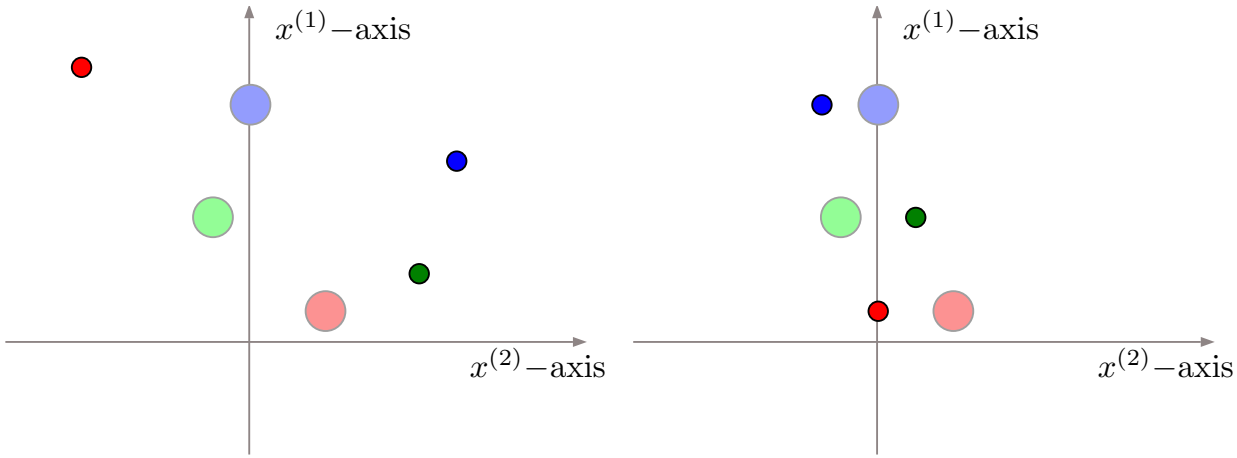
S_{m_i} being the subset corresponding to the label y_i.

Theorem RB-Zuazua (Simultaneous Control/Interpolation)

For $d \geq 2$ the system is approximately simultaneously controllable.

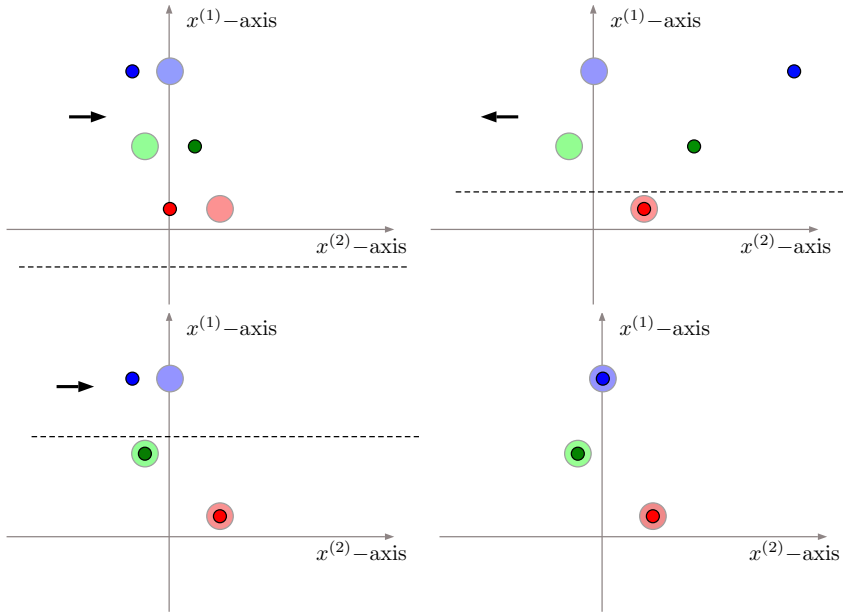
Simultaneous Control

Step 1: Apply the same arguments than in the classification



Simultaneous Control

Step 2: build, in an iterative manner piecewise controls that steer one by one each point to their target.



Universal approximation

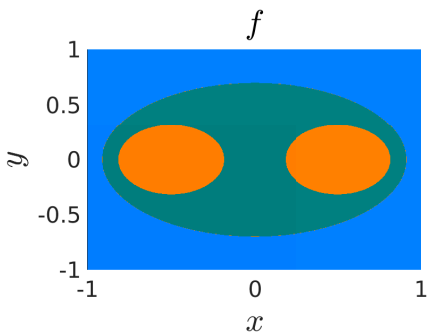
Universal Approximation

Let $d \geq 2$ and $T > 0$. Let $\Omega \subset \mathbb{R}^d$ be a bounded set, and σ be the ReLU. Then, for any $f \in L^2(\Omega; \mathbb{R}^d)$ and $\epsilon > 0$ there exist $A, W \in L^\infty((0,T); \mathbb{R}^d)$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that the flow generated by the Neural ODE, $\phi_T(\cdot; A, W, b)$, satisfies:

$$\|\phi_T(\cdot; A, W, b) - f(\cdot)\|_{L^2(\Omega)} < \epsilon.$$

In the NODE formulation, the result translates into the following each input $x \in \mathbb{R}^d$ has a target $f(x) \in \mathbb{R}^d$. The actual value of $f(x)$. Ideally, we would like to find controls A, W and b such that for every $x \in \Omega$ we have that

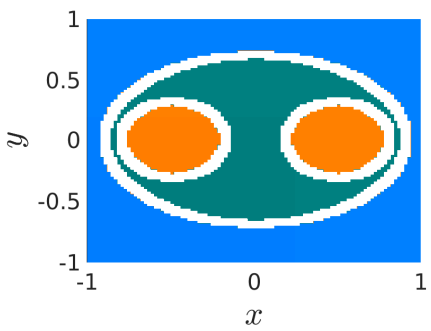
$$f = \sum_{m=1}^{M} \alpha_m \chi_{\Omega_m},$$

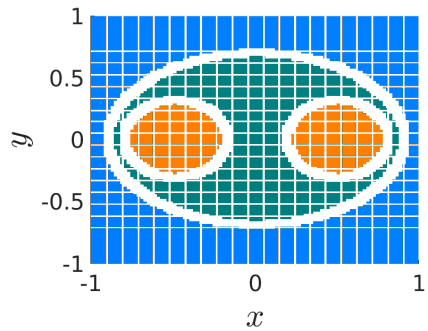


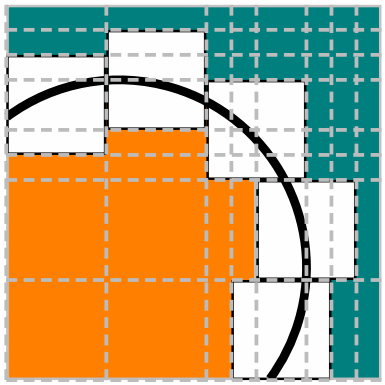
$$\begin{cases} \dot{x} = W(t)\sigma(A(t)x(t) + b(t)) \\ x(0) = x \\ \phi_T(x; A, W, b) = f(x). \end{cases}$$

The proof is based on approximating the target function by a **simple function** and then trying to reduce the problem to a simultaneous control one.

1. **Cover the boundary** of the characteristic sets with small hypercubes and generate a mesh with them.
2. **Remove a small strip around the mesh**
3. The white areas in the figure above belong to Ω_h and the function will **not be very well approximated** there.

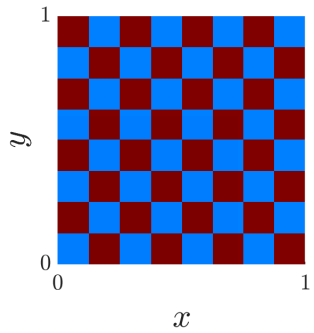


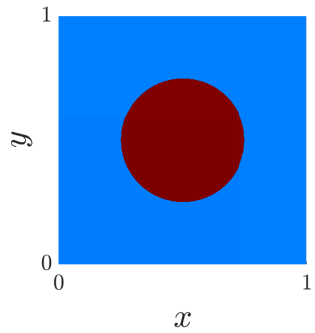


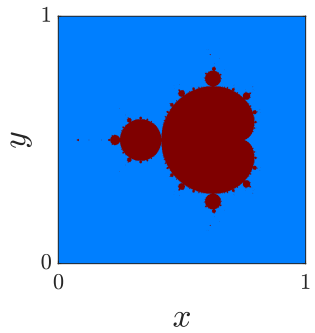


”Badly mixed data“

- The number of hyperrectangles depend on how many hypercubes one needs.
- If the interface between the characteristic sets is very irregular, one will need more hypercubes and therefore the control cost will be higher







”Badly mixed data“

Let D be the **box-counting dimension** of Γ[4]. Let $N_\Gamma(h)$ be the number of hypercubes of side h needed to cover the boundary Γ, the box-counting dimension is defined as

$$D := \lim_{h \to 0} \frac{\log N_\Gamma(h)}{\log\left(\frac{1}{h}\right)}.$$

The bounds on the control cost are able to capture such complexity.

$$\|W\|_{L^\infty} \lesssim \epsilon^{-2(D+1)d}, \qquad \|b\|_{L^\infty} \lesssim \epsilon^{-2dD} \qquad \text{as } \epsilon \to 0$$

The number of switches of A, W, b will be of the order of ϵ^{-2dD}.

[4] K. Falconer 06: Fractal geometry: mathematical foundations and applications.

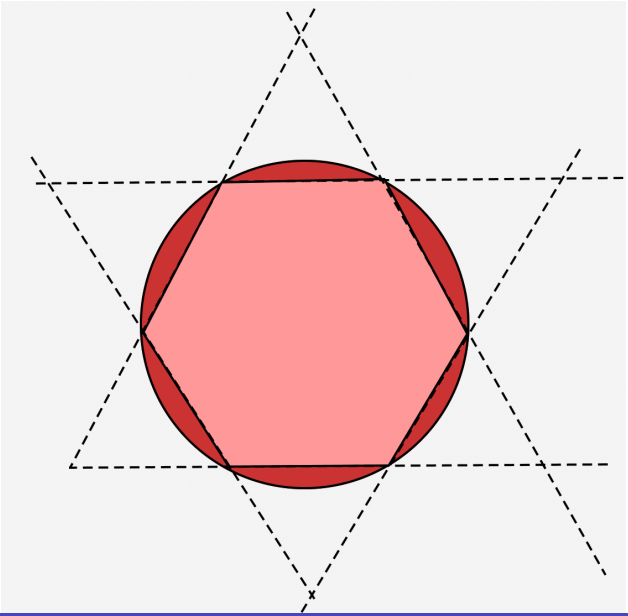
Curse of dimensionality

Constructive procedure: Meshing the boundary produces a high number of components and hence a high control cost

The number of switches of A, W, b will be, at most, of the order of ϵ^{-2dD}.

ONE CANNOT AVOID THE DEPENDENCE ON THE DIMENSION

For $\mathbb{B}_d$, the number of switches of A, W, b will be of the order of $\epsilon^{-\frac{d-1}{2}}$.



$$\min_{p_N \in P_N} |\mathbb{B}_d \triangle p_N| \sim \frac{c(d)}{N^{2/(d-1)}}$$

K. Böröczky Jr, Polytopal approximation bounding the number of k-faces, Journal of Approximation Theory, (2000)

Transport

Note that the set of differential equations

$$\begin{cases} x' = W(t)\sigma(A(t)x + b(t)) \\ x(0) = z \in \mathbb{R}^d \end{cases}$$

correspond to the projected characteristics of the transport equation:

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

Definition

Let $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$ be probability measures. The Wasserstein-1 distance $\mathcal{W}_1(\mu, \nu)$ is defined by as:

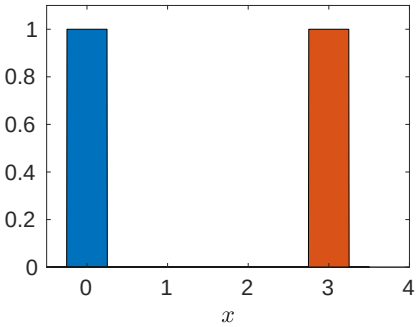
$$\mathcal{W}_1(\mu, \nu) = \sup_{Lip(g) \leq 1} \left\{ \int_{\mathbb{R}^d} g d\mu - \int_{\mathbb{R}^d} g d\nu \right\}$$

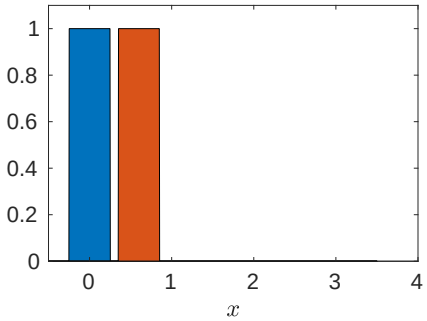
where $Lip(g) \leq 1$ stands for the class of Lipschitz functions with Lipschitz constant less or equal than 1.

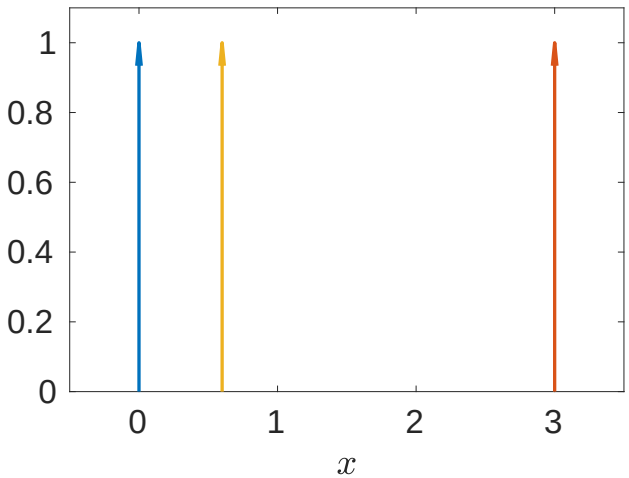
56.

[5] C. Villani 2008: Optimal transport: old and new

[6] L.V. Kantorovich 1942, On the transfer of masses.







The L^p-norm does not "see" the Euclidean distance between the supports

$$\|v_1 - v_2\|_{L^p} = 1$$

$$\|v_1 - \tilde{v}_2\|_{L^p} = 1$$

However, the Wasserstein distance does,

$$\mathcal{W}_1(\delta_{x_1}, \delta_{x_2}) = |x_1 - x_2|$$

Consider target measures ρ^* in the form

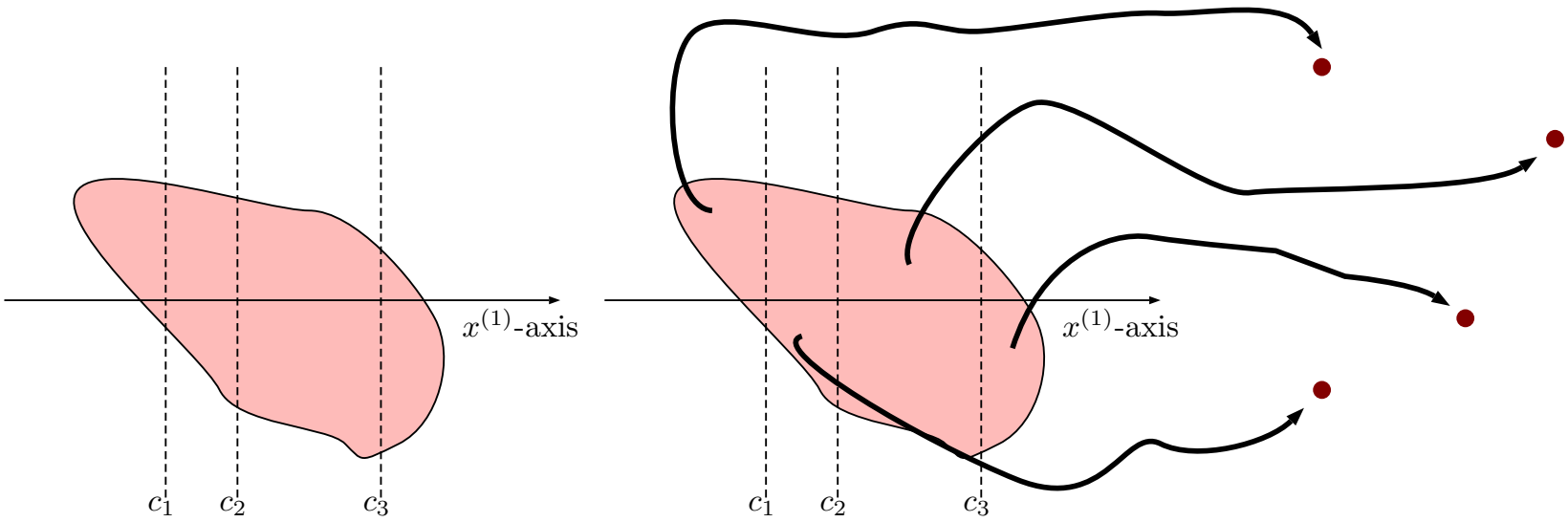
$$\rho^* = \sum_{m=1}^{M} \beta_m \delta_{\alpha_m}, \qquad \sum_{m=1}^{M} \beta_m = \int_{\mathbb{R}^d} \rho^0 dx$$

where δ_{α_m} is the Dirac delta located at $\alpha_m \in \mathbb{R}^d$ and $\beta_m > 0$.

Theorem (RB-Zuazua 2021 Control transport)

Let $T > 0$, $d \geq 2$. Then, for every $\epsilon > 0$, $\exists W, A \in L^\infty((0,T);\mathbb{R}^{d\times d})$ and $b \in L^\infty((0,T);\mathbb{R}^d)$ s.t.

$$\mathcal{W}_1(\rho(T), \rho^*) < \epsilon.$$


$x^{(1)}$-axis
c_1
c_2
c_3
$x^{(1)}$-axis
c_1
c_2
c_3

Simultaneous Control of NTEs

Remark

The scalar transport equation does not allow distinguishing among labels. One should take a vectorial structure in order to have a transport formulation for classification.

- Let us consider M classes, and M compactly supported probability densities ρ_m for $m = 1, ..., M$.
- Assume that for every x there exists, at most, a unique label y. This implies that the supports of the probability measures ρ_m are disjoint:

$$\text{supp}(\rho_m) \cap \text{supp}(\rho_{m'}) = \varnothing, \qquad \text{if } m \neq m'.$$

The system reads:

$$\begin{cases} \partial_t \rho_m + \mathrm{div}_x \left[(W(t)\sigma(A(t)x + b(t))\rho_m) \right] = 0, & m \in \{1, ..., M\} \\ \rho_m(0) = \rho_m^0 \in C_c(\mathbb{R}^d), & m \in \{1, ..., M\} \\ \mathrm{supp}(\rho_m^0) \cap \mathrm{supp}(\rho_{m'}^0) = \varnothing \quad \text{if } m \neq m'. \end{cases}$$

Assume that the target functions satisfy:

$$\mathrm{supp}(\rho_m^*) \cap \mathrm{supp}(\rho_{m'}^*) = \varnothing \qquad \text{if } m \neq m', \qquad \int \rho_m^* = \int \rho_m^0 dx = 1 \qquad m$$

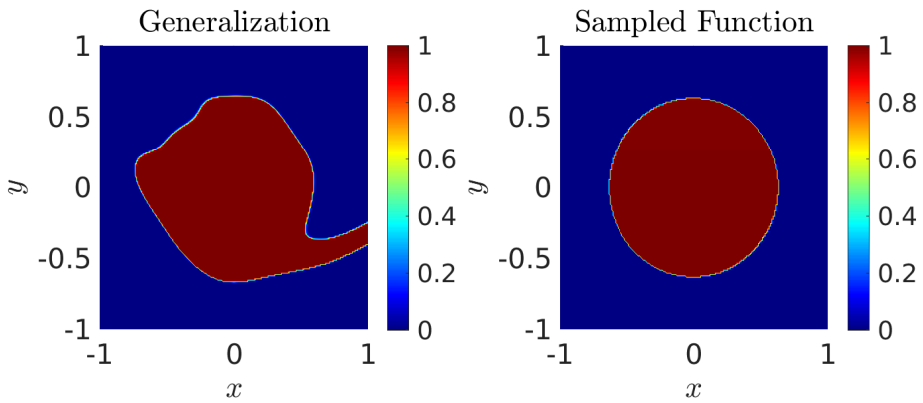
Theorem RB-Zuazua 2021: Simultaneous control transport

Let $T > 0$. Then, for any $\epsilon > 0$, $\exists W, A \in L^\infty((0, T), \mathbb{R}^{d \times d})$ and $\exists b \in L^\infty((0, T), \mathbb{R}^d)$ such the solution satisfies:

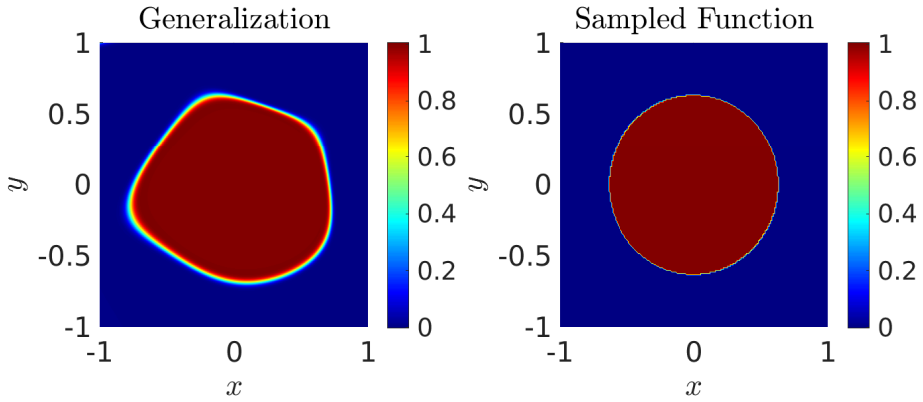
$$\mathcal{W}_1(\rho_m(T), \rho_m^*) < \epsilon \quad m \in \{1, ..., M\}$$

Other type of dynamics: Momentum ResNet

$$x'' + x' = W\sigma(Ax + b)$$



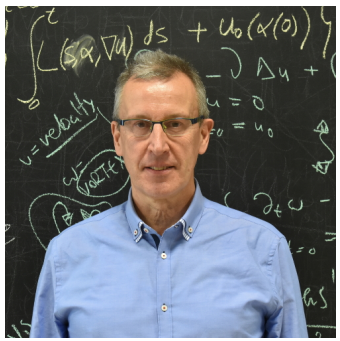
Neural ODE



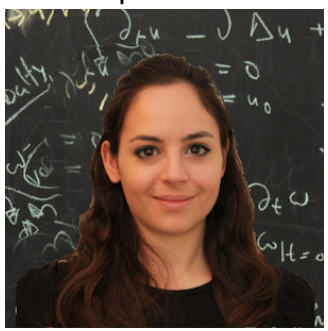
Momentum ResNet

Zorionak!

- D Ruiz-Balet, E Zuazua, Neural ODE control for classification, approximation and transport,arXiv preprint arXiv:2104.05278 (Accepted in SIAM Review)
- D Ruiz-Balet, E Affili, E Zuazua, Interpolation and approximation via momentum ResNets and neural ODEs, Systems & Control Letters 162, 2022



Enrique Zuazua



Elisa Affili

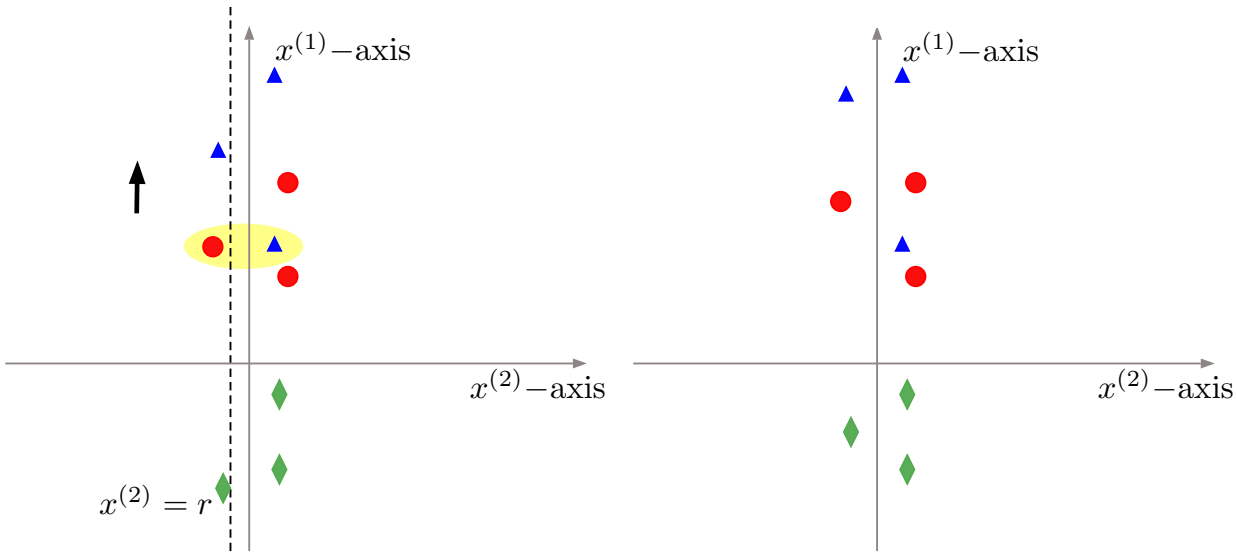
1. **Freeze**. The fact that the vector field can be zero in half-space, implies that any vector field generated by a hyperplane leaves half being critical points for the dynamics.
2. **Allocate**. One can allocate the hyperplane at any place in $\mathbb{R}^d$.
3. Setting an appropiate w one can generate
   - an **expansion**.
   - a **compression**.
   - a **translation**.

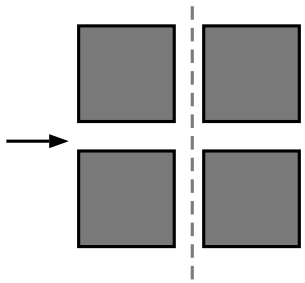
This allows us to consider fundamental vector fields that will be key for controlling the dynamics.

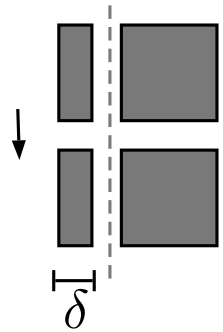
Classification Problem

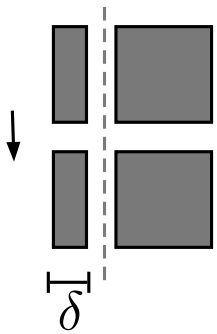
First of all, before applying the main argument, we will have to prepeare our data set. We need that to find controls such that can justify without loss of generality the following assumption

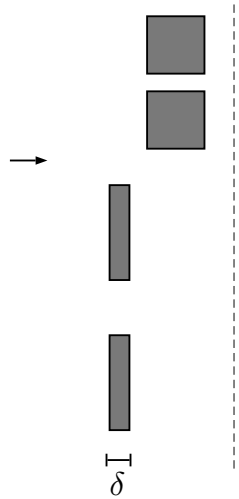
$$x_j^{(1)} \neq x_i^{(1)} \quad \text{if } i \neq j$$

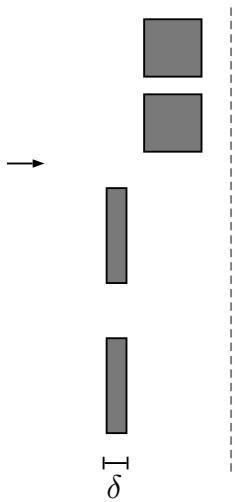


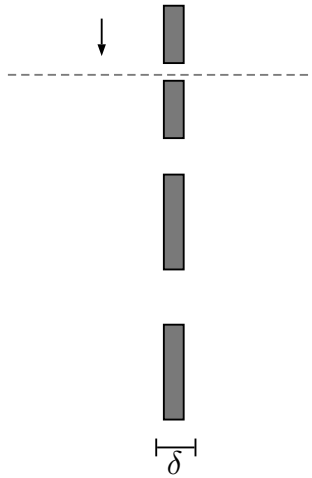


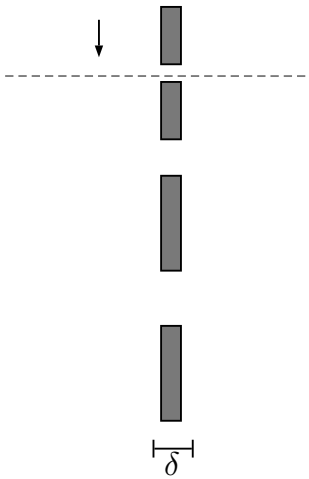

δ

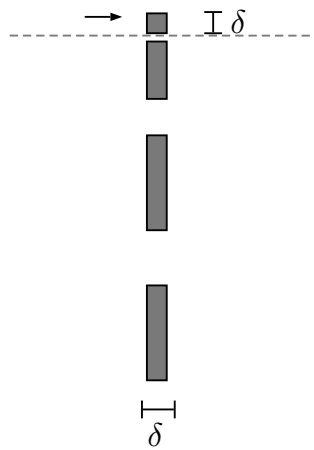

δ

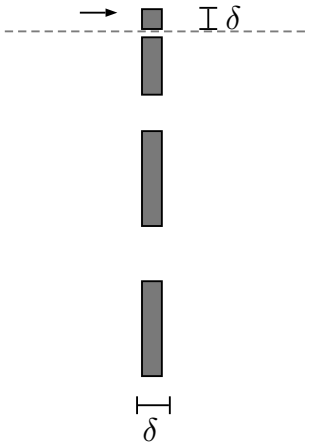

δ

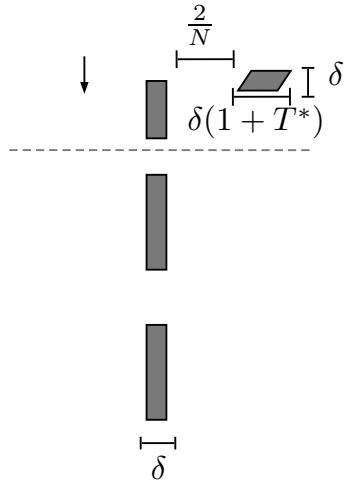

δ

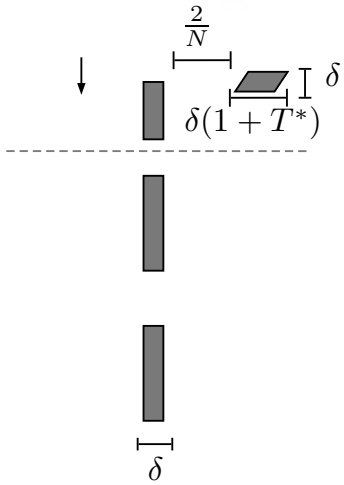

δ

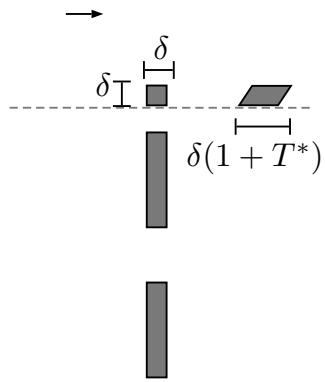

δ

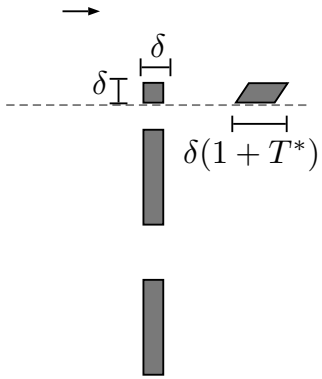

δ
δ

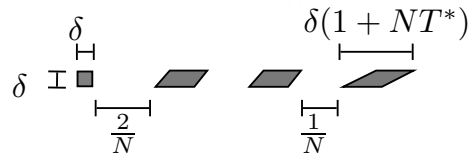

δ
δ

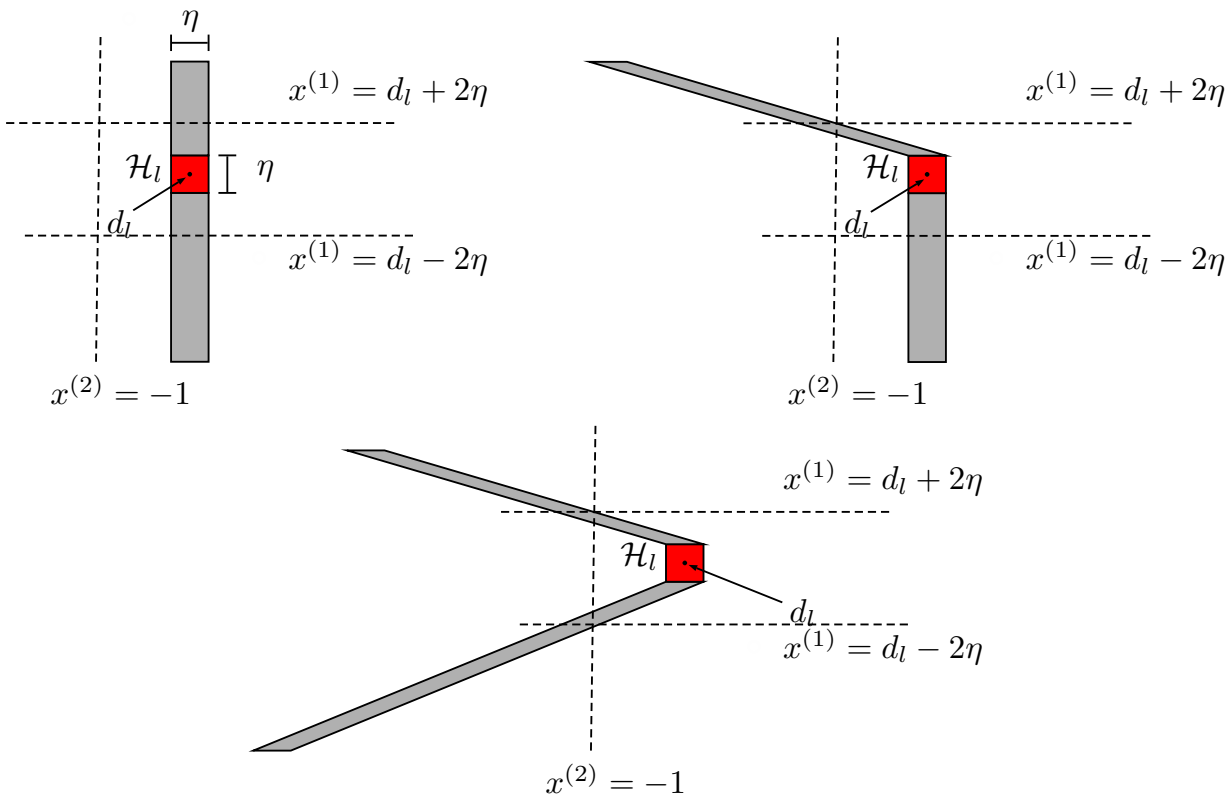

$\frac{2}{N}$
δ
$\delta(1+T^*)$
δ


$\frac{2}{N}$
δ
$\delta(1+T^*)$
δ


δ
δ
$\delta(1+T^*)$


δ
δ
$\delta(1+T^*)$


δ
δ
$\frac{2}{N}$
$\delta(1+NT^*)$
$\frac{1}{N}$


η
$x^{(1)} = d_l + 2\eta$
$\mathcal{H}_l$
η
d_l
$x^{(1)} = d_l - 2\eta$
$x^{(2)} = -1$
$x^{(1)} = d_l + 2\eta$
$\mathcal{H}_l$
d_l
$x^{(1)} = d_l - 2\eta$
$x^{(2)} = -1$
$x^{(1)} = d_l + 2\eta$
$\mathcal{H}_l$
d_l
$x^{(1)} = d_l - 2\eta$
$x^{(2)} = -1$

- The restriction of $d \geq 2$ comes from the limitation already pointed out before. However, the transport equation

$$\partial_t \rho + \mathrm{div}_x[V(x,t)\rho] = 0$$

  with V as a control, can be approximately controlled in the one-dimensional case.
- It would be enough to design appropriate locations of attractors and repulsors depending on the mass distribution of ρ^0.
- This would allow concentrating the mass in a finite number of points.
- Later, one needs to design a dynamical system that brings each mass approximately on the approximation of the target. Neural Transport Equations can achieve this in dimension $d \geq 2$ just with the controls A, W and b.