Controllability of Neural ODEs for classification

Antonio Álvarez-López[1,2], Enrique Zuazua[1,2]

Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

We explore the capacity of neural ODEs for supervised learning from the perspective of simultaneous control. We consider the parameters as piecewise constant functions in time and construct them explicitly ($\Rightarrow$ suboptimal controls).

First, we focus on data classification by controlling clusters of points belonging to two classes. We estimate the number of neurons that neural ODEs require to classify a **generic** pair of point sets.

Secondly, we analyze the interaction and exchangeability of depth and width for simultaneous control. We then focus on the case of constant parameters, where the model is autonomous.

Model

Residual networks: $\quad \mathbf{x_{k+1}} = \mathbf{x_k} + hW_k\sigma\left(A_k\mathbf{x_k} + \mathbf{b_k}\right), \quad k = 0, \dots, \mathrm{N_{layers}} - 1.$

$$\downarrow \quad (h \to 0)$$

Neural ordinary differential equations (neural ODEs, [5])

$$(1) \quad \begin{cases} \dot{\mathbf{x}}(t) &= \sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right), \qquad t \in (0,T), \\ \mathbf{x}(0) &= \mathbf{x_0} \in \mathbb{R}^d, \end{cases}$$

where $d \geq 2$ and

$\theta := (\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p : (0,T) \to (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$ piecewise constant (PC) controls.

- Predictive model: Flow map
  $$\Phi^t(\cdot;\theta) : \mathbb{R}^d \longrightarrow \mathbb{R}^d, \quad \mathbf{x_0} \longmapsto \mathbf{x}(t) \quad \text{solution of (1).}$$
- Complexity = Number of switches L × constant width p.
- Finite dataset $\mathcal{D} = \{(\mathbf{x_n}, \mathbf{y_n})\} \subset \mathbb{R}^d \times \mathcal{Y}$
  $$\begin{cases} \text{Binary classification:} & \mathcal{Y} = \{1, 0\} \longleftrightarrow \underbrace{\{x^{(j)} > 1\}}_{\Omega_1}, \ \underbrace{\{x^{(j)} < 1\}}_{\Omega_0}. \\ \text{Interpolation:} & \mathcal{Y} = \mathbb{R}^d. \end{cases}$$
- Worst-case scenario (W-CS): Random $(\mathbf{x_n}, \mathbf{y_n})$, indep. and uniformly distributed.

Basic dynamics:

- $\mathbf{a}(t), b(t)$ define a hyperplane $H(\mathbf{x}) = \mathbf{a}(t)\cdot\mathbf{x}(t) + b(t) = 0$ in $\mathbb{R}^d$.
- $\sigma(z) = (z)_+$ "activates" the half-space $H(\mathbf{x}) > 0$ and "freezes" $H(\mathbf{x}) \leq 0$.
- $\mathbf{w}(t)$ determines the direction of the field in $H(\mathbf{x}) > 0$.

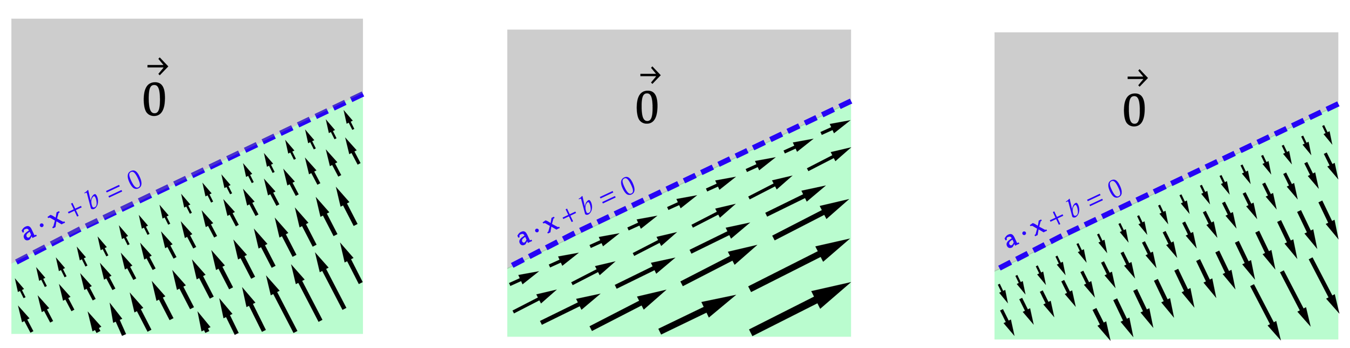


Figure 1. Contraction (left), translation (center), expansion (right).

Problem 1: Classification ("The Rubik's cube")

Statement

For any $T > 0$, find θ such that $\Phi^T(\mathbf{x_n};\theta) \in \Omega_{y_n}$ for all n with minimal complexity L (having fixed $p = 1$) .

Theorem 1 (Probabilistic bound on complexity, [1])

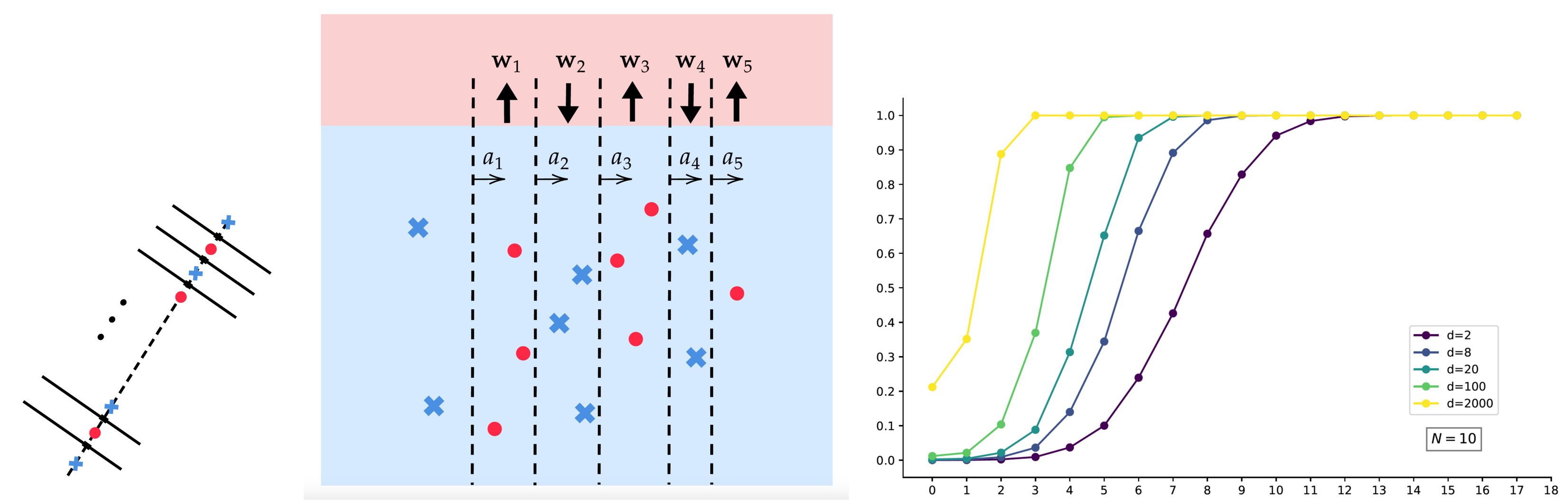
Assume that $\#\{\mathbf{x_n}\} = \#\{\mathbf{x_m}\} = N$ *and* $\mathbf{x_n}, \mathbf{x_m} \sim U([0,1]^d)$. *For any* $T > 0$, *there exist* $j \in \{1, \dots, d\}$ *and* $\theta : (0,T) \to \mathbb{R}^{2d+1}$ *PC such that for all* n, m

$$\Phi^T(\mathbf{x_n};\theta)^{(j)} > 1 \quad \text{and} \quad \Phi^T(\mathbf{x_m};\theta)^{(j)} < 1,$$

and for $k = 0, \dots, 2N-2$, *the number of switches* L *follows*

$$\mathbb{P}(L \leq k) = 1 - \left(\sum_{p=\lceil \frac{k+3}{2} \rceil}^{N} \binom{N-1}{p-1}^2 + \sum_{p=\lceil \frac{k+1}{2} \rceil}^{N-1} \binom{N-1}{p}\binom{N-1}{p-1} \right)^d \left(\frac{2(N!)^2}{(2N)!}\right)^d.$$

- Linear separability/Constant controls ($k = 0$):
  $$\mathbb{P}(L = 0) \geq 1 - \left(\frac{2(N!)^2}{(2N)!}\right)^d \sim 1 - \exp\left(-\frac{\sqrt{\pi N}d}{2^{2N-1}}\right).$$



(a) Case $k = 2N-2$ (b) Algorithm for theorem 1. (c) $\mathbb{P}(L \leq k)$ vs k for $N = 10$ and several d

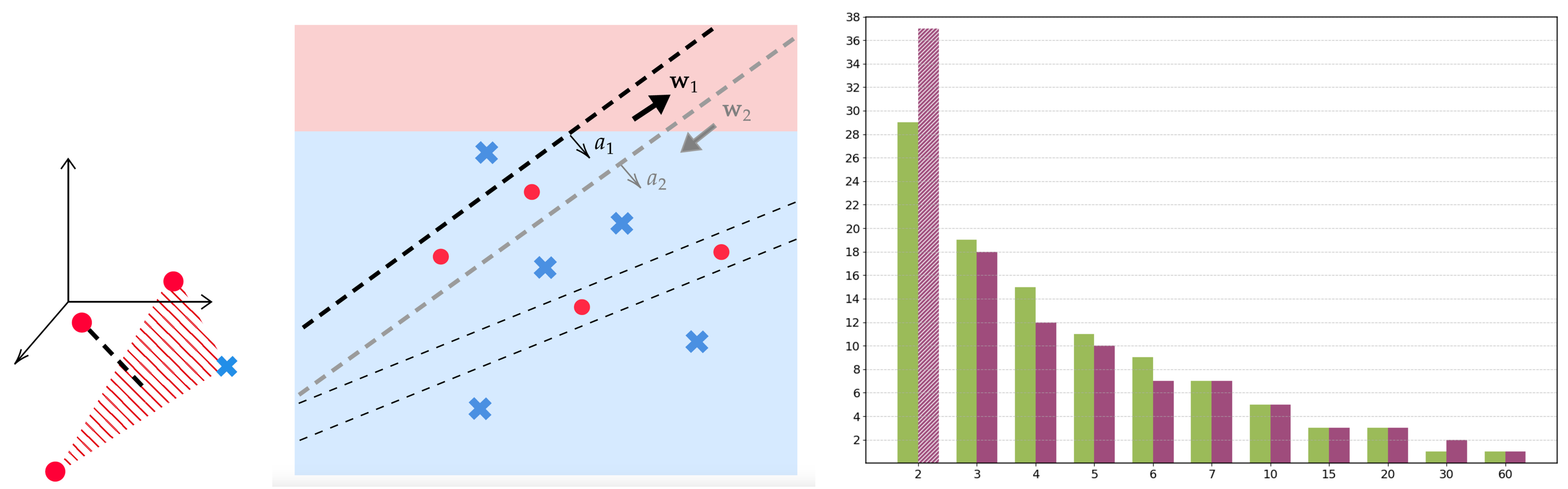
Theorem 2 (Cluster-based classification in W-CS, [1])

Let $\{\mathbf{x_n}\}, \{\mathbf{x_m}\} \subset \mathbb{R}^d$ *be in general position*[a]. *For any* $T > 0$ *and* $j \in \{1, \dots, d\}$, *there exists* $\theta : (0,T) \to \mathbb{R}^{2d+1}$ *PC such that for all* n, m

$$\Phi^T(\mathbf{x_n};\theta)^{(j)} > 1 \quad \text{and} \quad \Phi^T(\mathbf{x_m};\theta)^{(j)} < 1.$$

Furthermore, the number of switches is

$$L = 4\left\lceil \frac{\min\{\#\{\mathbf{x_n}\}, \#\{\mathbf{x_m}\}\}}{d} \right\rceil - 1.$$



(d) GP in $\mathbb{R}^3$. (e) Algorithm for theorem 2 (one iteration). (f) L (green) and $L_{\mathrm{experimental}}$ (purple) vs d.

[a] *No* $d+1$ *points lie on the same hyperplane.*

Problem 2: Interpolation ("Depth vs width")

Statement

For any $T > 0$, study the relation between p, L that guarantees the existence of θ such that $\Phi^T(\mathbf{x_n};\theta) = \mathbf{y_n}$ for all n.

Theorem 3 (Simultaneous control, [2])

Let $\{(\mathbf{x_n}, \mathbf{y_n})\}_{n=1}^N \subset (\mathbb{R}^d)^2$ *and* $T > 0$. *There exists* $\theta : (0,T) \to \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ *PC such that*

$$\Phi^T(\mathbf{x_n};\theta) = \mathbf{y_n}, \qquad \text{for all } n = 1, \dots, N.$$

Furthermore, the number of switches is $L = 2\left\lceil \frac{N}{p} \right\rceil - 1.$

(g) Step 1: Control $d-1$ coordinates. (h) Step 2: Control the remaining coordinate.

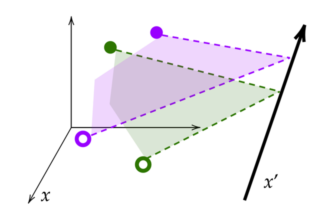
Is it possible to achieve $L = 0$?

Special case 1: High dimensions

If $d > N$ then it can improved to $L = 2\left\lceil \frac{N}{p} \right\rceil - 2.$

Special case 2: Semi-autonomous

In Thm 3 we can take constant $\mathbf{w}$, $\mathbf{a}$ and $b = b(t)$.



Build new basis to eliminate Step 1.

Theorem 4 (Relaxation: Approximate control, [2])

Let $\{(\mathbf{x_n}, \mathbf{y_n})\}_{n=1}^N \subset (\mathbb{R}^d)^2$ *and* $T > 0$. *There exists* $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ *and* $C > 0$ *independent of* p *such that*

$$\sup_{n\in\{1,\dots,N\}} |\mathbf{y_n} - \Phi^T(\mathbf{x_n};\theta)| \leq C\frac{\log_2(m)}{m^{1/d}}, \qquad \text{for } m = (d+2)dp.$$

Figure 2. Handmade vector field that interpolates $\mathcal{D}$, later approximated with system (1).

Conclusions

- The complexity required to classify any generic dataset is $1 + O(N/d)$.
- Increasing p allows reducing L for interpolation as $1 + O(N/p)$.
- An autonomous, wide enough neural ODE can achieve approx. control.

References

[1] A. Álvarez-López, R. Orive-Illera, and E. Zuazua. Optimized classification with neural odes via separability. *arXiv:2312.13807*, 2023.

[2] A. Álvarez-López, A. Hadj Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural odes. *Neural Networks*, 180:106640, 2024.

[3] F. Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18(19), 2017.

[4] D. Ruiz-Balet and E. Zuazua. Neural ODE control for classification, approximation, and transport. *SIAM Rev.*, 65(3):735–773, 2023.

[1]Universidad Autónoma de Madrid, Spain. [2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

antonio.alvarezl@uam.es