Controllability of Neural ODEs for classification

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

Friedrich-Alexander-Universität Erlangen-Nürnberg

Motivation

Neural ODEs offer a solid framework for efficient continuous-time modeling. Specifically, they hold significant potential in the functional development of digital twins, which require real-time tracking and forecasting of the physical twin. Neural ODEs naturally integrate data from irregular time series and show high compatibility in mirroring the physical principles guiding real-world systems, such as fluid dynamics in engineering or biological processes in healthcare.
Two key tasks:

1. Optimize the complexity they require for data classification. This facilitates fast, adaptive decision-making and timely responses to evolving conditions, essential in risk situations.
2. Understand their expressivity through data control. This demands a detailed analysis of the role played by the architecture (depth, width of the model), and contributes to high fidelity in the digital replication of physical assets. It allows to predict future states based on data trends or to capture intrinsic dynamics.

Model

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

$\downarrow$ (Continuous limit $h \to 0$)

Neural ordinary differential equations (neural ODEs, [4])

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

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

- Predictive model: Flow in time $t = T$ generated by (1), $\Phi^T(\cdot;\theta) : \mathbb{R}^d \to \mathbb{R}^d$, which maps $\mathbf{x_0} \in \mathbb{R}^d \longmapsto \mathbf{x}(T)$ solution of (1) evaluated in $t = T$.
- Complexity = Number of time switches (L) $\times$ constant width (p).
- Dataset $\mathcal{D} = \{(\mathbf{x_n}, \mathbf{y_n})\}_{n=1}^N \subset \mathbb{R}^d \times y$. $\begin{cases} \text{Binary classification:} & y = \{1, 0\}. \\ \text{Simultaneous control:} & y = \mathbb{R}^d. \end{cases}$
- Worst-case scenario: (W-CS) Random $(\mathbf{x_n}, \mathbf{y_n})$, indep. and uniformly distributed.
  Balanced classes: $\#\{(\mathbf{x_n}, 1)\} = \#\{(\mathbf{x_n}, 0)\}$.

Basic dynamics:

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

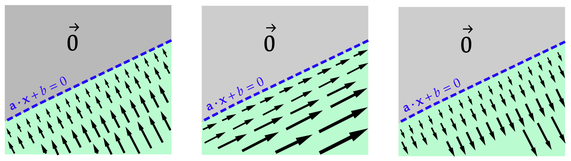


Figure 1. Contraction (left), translation (center), expansion (right).

Binary classification

Problem statement

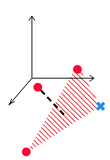
Define a pair of disjoint target regions: $\Omega_1 = \{x^{(j)} > 1\}$ and $\Omega_0 = \{x^{(j)} < 1\}$.
For any given $T > 0$, find a control θ s.t. $\Phi^T(\mathbf{x_n};\theta) \in \Omega_{y_n}$ for all $n = 1, \dots, N$.

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

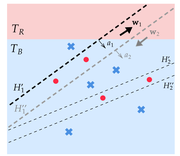
Let $2 \le d < 2N$ and $\{\mathbf{x_n}\}_{n=1}^N$, $\{\mathbf{x_{N+n}}\}_{n=1}^N \subset \mathbb{R}^d$ be in general position[a] (GP). Consider the neural ODE (1) with $p = 1$. For any time $T > 0$ and $j \in \{1, \dots, d\}$, there exists a piecewise constant control $\theta : (0,T) \to \mathbb{R}^{2d+1}$ such that

$$\Phi^T(\mathbf{x_n};\theta)^{(j)} > 1 \quad \text{and} \quad \Phi^T(\mathbf{x_{N+n}};\theta)^{(j)} < 1, \qquad \text{for all } n = 1, \dots, N.$$

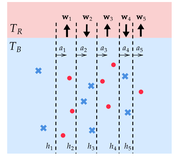
Furthermore, the number of time switches is $L = 1 + O(N/d)$.



(a) 4 points in $\mathbb{R}^3$ in GP.



(b) Classification method for theorem 1 (one iteration).



(c) Classification method for theorem 2.

Theorem 2 (Probabilistic bound on complexity, [1])

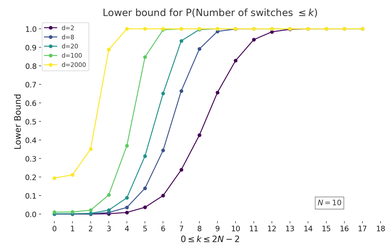
Let $d \ge 2$ and consider the neural ODE (1) with $p = 1$. Assume that $\mathbf{x_n}, \mathbf{x_{N+n}} \sim U([0,1]^d)$, for all $n = 1, \dots, N$. For any time $T > 0$, there exist $j \in \{1, \dots, d\}$ and a piecewise constant control $\theta : (0,T) \to \mathbb{R}^{2d+1}$ such that

$$\Phi^T(\mathbf{x_n};\theta)^{(j)} > 1 \quad \text{and} \quad \Phi^T(\mathbf{x_{N+n}};\theta)^{(j)} < 1, \qquad \text{for all } n = 1, \dots, N,$$

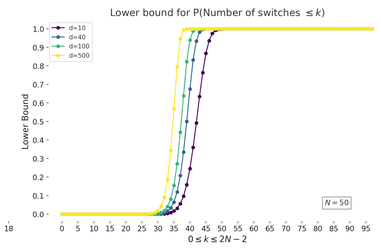
and the number of switches L satisfies the probabilistic bound, for $k = 0, \dots, 2N-2$:

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

- Linear separability: $P(L=0) \ge 1 - (2(N!)^2/(2N)!)^d$.
- Asymptotics: For $d, N \gg 1$, $P(L=0) \sim 1 - \exp\left\{-d\sqrt{N}/2^N\right\}$.



(d)



(e)

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

Interpolation/Simultaneous control

Problem statement

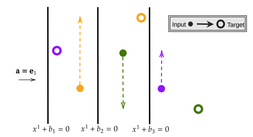
For any given $T > 0$, find a control θ s.t. $\Phi^T(\mathbf{x_n};\theta) = \mathbf{y_n}$ for all $n = 1, \dots, N$.

Theorem 3 (Architecture: depth vs width, [2])

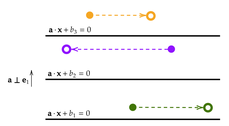
Let $d \ge 2$. Consider a dataset $\{(\mathbf{x_n}, \mathbf{y_n})\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$. For any time $T > 0$, there exists a piecewise constant control $\theta : (0,T) \to \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ such that

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

Furthermore, the number of time switches is $L = 1 + O(N/p)$.



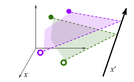
(f) Step 1: Control $d-1$ coordinates of p points simultaneously.



(g) Step 2: Control the remaining coordinate of the p points simultaneously.

Special case: High dimensions

If $d > N$, then L can be improved to $L = O(N/p)$.

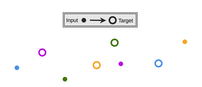


Build new basis by $\mathbf{x} \mapsto \mathbf{x}'$ to eliminate Step 1.

Theorem 4 (Approximate control with autonomous model, [2])

Let $d \ge 2$. Consider a dataset $\{(\mathbf{x_n}, \mathbf{y_n})\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$. For any time $T > 0$, there exists a constant control $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ such that

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



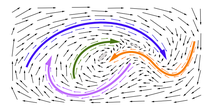


Figure 2. Handmade vector field that interpolates $\mathcal{D}$, later approximated with system (1).

Conclusions

- Clustering of data enables a reduction of complexity with high probability.
- Increasing d diminishes the complexity as $O(N) \to 1 + O(N/d)$.
- Increasing the width p allows reducing depth L as $1 + O(N/p)$.
- An autonomous, sufficiently wide neural field can achieve approx. control.

References

[1] Antonio Álvarez-López, Rafael Orive-Illera, and Enrique Zuazua. Optimized classification with neural odes via separability. *arXiv preprint arXiv:2312.13807*, 2023.

[2] Antonio Álvarez-López, Arselane Hadj Slimane, and Enrique Zuazua. Interplay between depth and width for interpolation in neural odes. *arXiv preprint arXiv:2401.09902*, 2024.

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

[4] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11.

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

antonio.alvarezl@uam.es