

NEURAL ODES FOR INTERPOLATION AND TRANSPORT: FROM TIGHT TO SHALLOW NEURAL NETWORKS

Antonio Álvarez López, Arselane Hadj Slimane, Enrique Zuazua

July 27, 2023

INTRODUCTION

RESNETS AND NEURAL ODES

- Resnets are a special architecture of neural networks (NNs):

$$\begin{cases} \mathbf{x}_{k+1} &= \mathbf{x}_k + \sum_{i=1}^p \mathbf{w}_k^i \sigma(\mathbf{a}_k^i \mathbf{x}_k + b_k^i), & k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_0 &= \mathbf{x} \in \mathbb{R}^d. \end{cases} \quad (1)$$

N_{layers} is the *depth* of the NN; p is the *width* of the NN;
 $\mathbf{w}_k^i, \mathbf{a}_k^i \in \mathbb{R}^d, b_k^i \in \mathbb{R}$ are optimizable weights; $\mathbf{x}_k \in \mathbb{R}^d$ are the unknowns.

- They are the discretization of neural odes (nodes):

$$\begin{cases} \dot{\mathbf{x}}(t) &= \sum_{i=1}^p \mathbf{w}_i(t) \sigma(\mathbf{a}_i(t) \cdot \mathbf{x}(t) + b_i(t)) \\ \mathbf{x}(0) &= \mathbf{x}_0 \in \mathbb{R}^d, \end{cases} \quad (2)$$

where $(\mathbf{w}_i, \mathbf{a}_i, b_i) \in L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ are control variables.

- We assume that the controls are piecewise-constant with M time discontinuities and take the *ReLU* activation function $\sigma(z) = \max\{z, 0\}, \forall z \in \mathbb{R}$.

INTRODUCTION

GOALS

- ▶ Two antipodal cases can be reached:
 - *Tight neural odes* are obtained when $p = 1$

$$\dot{\mathbf{x}}(t) = \mathbf{w}(t)\sigma(\mathbf{a}(t) \cdot \mathbf{x}(t) + b(t)). \quad (3)$$

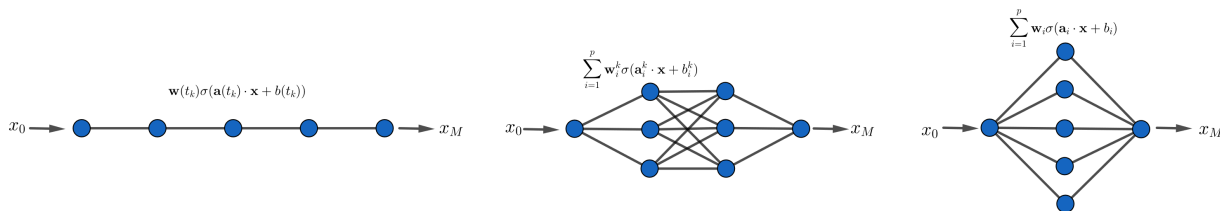
- *Deep neural odes* are obtained when $p \geq 2$ and $M \geq 1$

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^p \mathbf{w}_i(t)\sigma(\mathbf{a}_i(t) \cdot \mathbf{x}(t) + b_i(t))$$

- *Shallow neural odes* are obtained when $M = 0$:

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^p \mathbf{w}_i\sigma(\mathbf{a}_i \cdot \mathbf{x}(t) + b_i). \quad (4)$$

- ▶ Goal: To build a constructive theory encompassing and linking the tight, deep and shallow models.



From left to right: tight, deep and shallow ResNet

INTERPOLATION/SIMULTANEOUS CONTROL

SETTING

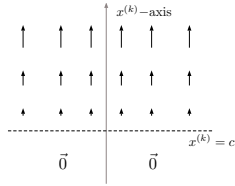
- Consider a dataset

$$\mathcal{S} = \{(\mathbf{x}_n, \mathbf{y}_n)\} \subset \mathbb{R}^d \times \mathbb{R}^d.$$

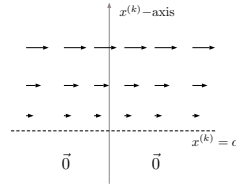
We want to know if neural odes can *interpolate* \mathcal{S} , i.e, if there exist controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ such that

$$\phi_T(\mathbf{x}_n; W, A, \mathbf{b}) = \mathbf{y}_n, \quad \forall n = 1, \dots, N,$$

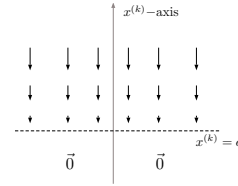
where $\phi_T(\cdot; W, A, \mathbf{b}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the flow associated to (2), taking
$$\begin{cases} W &= (\mathbf{w}_1, \dots, \mathbf{w}_p), \\ A &= (\mathbf{a}_1, \dots, \mathbf{a}_p), \\ \mathbf{b} &= (b_1, \dots, b_p). \end{cases}$$



(a) Dilatation



(b) Translation



(c) Compression

Basic operations allowed by one neuron $(\mathbf{w}_i, \mathbf{a}_i, b_i)$, extracted from [Ruiz-Balet and Zuazua 2021]

INTERPOLATION/SIMULTANEOUS CONTROL

FROM TIGHT TO DEEP NEURAL ODES

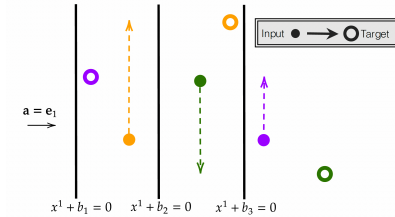
Theorem 1 (Simultaneous control with p neurons)

Consider the neural ode (2) with $d \geq 2$ and the dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbf{y}_n \neq \mathbf{y}_m$, for all $n \neq m$. Let $T > 0$ be fixed. There exist controls, $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$, such that the flow $\phi_T(\cdot; W, A, \mathbf{b})$ generated by the neural ODE satisfies:

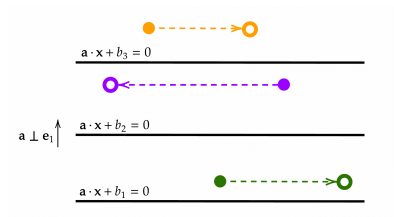
$$\phi_T(\mathbf{x}_n; W, A, \mathbf{b}) = y_n, \quad \forall n = 1, \dots, N$$

Moreover, the controls are piecewise constant and the number of time switches is

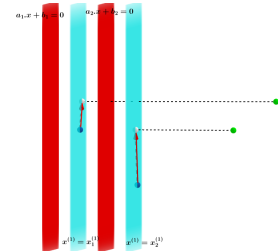
$$M = 2 \left\lfloor \frac{N}{p} \right\rfloor + 1.$$



(a) Step 1 of the steering



(b) Step 2 of the steering



(c) Step 1 in $d = 3$

INTERPOLATION/SIMULTANEOUS CONTROL

FROM TIGHT TO DEEP NEURAL ODES

Corollary 1 (Approximate control with p neurons)

Consider the neural ode (2) with $d \geq 2$ and any dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$.

Let $T > 0$ be fixed. For any $\epsilon > 0$, there exist controls, $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \subset L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$, such that the flow generated by the neural ODE satisfies

$$|\phi_T(\mathbf{x}_n; W, A, \mathbf{b}) - \mathbf{y}_n| < \epsilon, \quad \forall n = 1, \dots, N$$

Moreover, the controls are piecewise constant and the number of time switches is

$$M = 2 \left\lfloor \frac{N}{p} \right\rfloor + 1.$$

Remark: As the number of neurons p increases, the required number of time switches M decreases proportionally. In terms of neural networks, this is interpreted as the *exchangeability* between width and depth because they play the same role in the steering.

When $p > N$, the selected controls exhibit a single discontinuity, indicating a transition to a 2-hidden layer neural network, instead of the shallow neural ode (4).

How can we reach the autonomous ansatz?

INTERPOLATION/SIMULTANEOUS CONTROL

THE SHALLOW CASE

Lemma 1 (Exact controllability with Lipschitz fields)

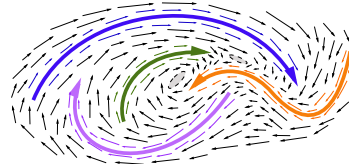
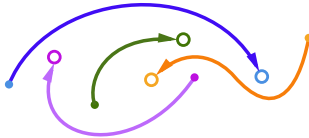
Consider any dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ for $d \geq 2$, and let $T > 0$ be fixed. There exists a vector field $\mathbf{V} \in \text{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ such that the flow ψ_T of the equation $\dot{\mathbf{x}} = \mathbf{V}(\mathbf{x})$ interpolates the dataset \mathcal{S} .

Theorem 2 (Approximate controllability with shallow nodes)

Consider any dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ for $d \geq 2$, and let $T > 0$ be fixed. For any interpolating field $\mathbf{V} \in \text{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ with Lipschitz constant L_V , there exist constant controls $W, A \in \mathbb{R}^{p \times d}$, $\mathbf{b} \in \mathbb{R}^p$ such that the flow ϕ_T generated by the shallow neural ode (4) satisfies

$$|\mathbf{y}_n - \phi_T(\mathbf{x}_n; W, A, \mathbf{b})| \leq C_{d,L_V} \frac{\log_2 m}{m^{1/d}} T \exp \{ \min\{L_V, L_{NN}\} T \}, \quad \forall n = 1, \dots, N, \quad (5)$$

where $m = (d^2 + 2d)p$ is the total number of parameters in the network field and L_{NN} denotes its Lipschitz constant.



INTERPOLATION/SIMULTANEOUS CONTROL

THE SHALLOW CASE

Corollary 2 (Case $d > N$)

Consider a dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ with distinct targets and let $T > 0$. Suppose that $d > N$. Then, there exist piecewise constant controls such that the flow ϕ_T generated by the neural ode (2) satisfies

$$\phi_T(\mathbf{x}_n; W, A, \mathbf{b}) = y_n, \quad \forall n = 1, \dots, N,$$

Moreover, the number of time discontinuities is

$$M = \left\lfloor \frac{N}{p} \right\rfloor$$

.

INTERPOLATION/SIMULTANEOUS CONTROL

THE SHALLOW CASE

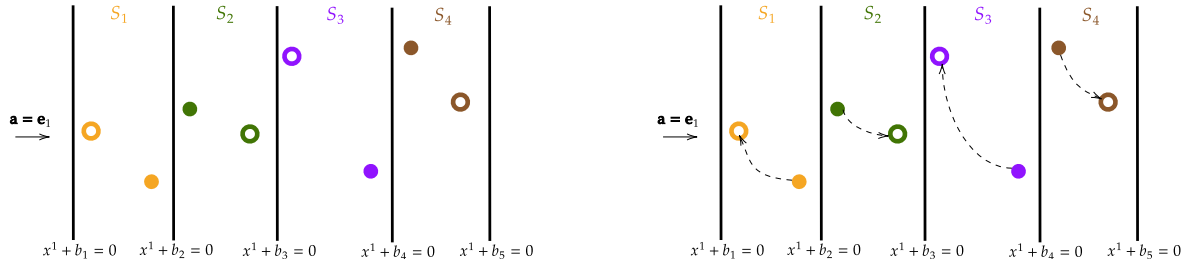
Assumption 1

Given the dataset \mathcal{S} , there exists a vector $\mathbf{a} \in \mathbb{S}^{d-1}$, a permutation τ of N elements and a sequence $-\infty < b_{N+1} < b_N < \dots < b_1 < \infty$ such that

$$-b_n < \mathbf{a} \cdot \mathbf{x}_{\tau(n)} < -b_{n+1} \quad \text{and} \quad -b_n < \mathbf{a} \cdot \mathbf{y}_{\tau(n)} < -b_{n+1}, \quad \text{for all } n = 1, \dots, N-1.$$

Theorem 3 (Exact controllability with shallow nodes)

Consider any dataset $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ for $d \geq 2$, under Assumption 1 and take $p = N$. For any fixed time horizon $T > 0$, there exist constant controls, $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^N \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^N$, such that the flow ϕ_T associated to (4) interpolates \mathcal{S} .



INTERPOLATION/SIMULTANEOUS CONTROL

THE SHALLOW CASE

Proposition 1

Let ρ be an absolutely continuous probability measure defined on the hypercube $[-L, L]^d$ for some $L > 0$ such that all the marginal measures $\rho_i : [-L, L] \rightarrow \mathbb{R}_{\geq 0}$ for each coordinate are independent and identically distributed. If every point \mathbf{x}_n and \mathbf{y}_n of the dataset \mathcal{S} is sampled from ρ , the probability P that Assumption 1 is fulfilled is bounded as

$$1 - \left[1 - \frac{1}{\sqrt{2}} \left(\frac{e}{2N} \right)^N \right]^d \leq P \leq 1.$$

Remark: The uniform probability measure in $[-L, L]^d$ or any isotropic Gaussian distribution centered in the origin fulfills the hypothesis of Proposition 1.

NEURAL TRANSPORT EQUATION

MATHEMATICAL SETTING

- Consider that the input for the neural ODE is a probability measure $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ instead of a dataset. Now the question is whether or not the system is able to transform ρ_0 into any given target probability measure ρ_* . More precisely, we want to construct controls W, A, \mathbf{b} such that the flow ϕ_T generated by the neural ODE satisfies

$$\phi_T(\cdot, W, A, \mathbf{b}) \# \rho_0 \approx \rho_*(\cdot).$$

- The curve in the space of measures defined by $\rho(t)(\cdot) = \phi_t(\cdot; W, A, \mathbf{b}) \# \rho_0$ solves the equation

$$\begin{cases} \partial_t \rho + \operatorname{div}_{\mathbf{x}} (\sum_{i=1}^p \mathbf{w}_i \sigma(\mathbf{a}_i \cdot \mathbf{x} + b_i)) &= 0 \\ \rho(0) &= \rho_0, \end{cases} \quad (6)$$

- To compute the difference between two measures, we consider:

1. The space of probability measures

$$\mathcal{P}_{ac}^c(\mathbb{R}^d) = \left\{ \mu : \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1] : \operatorname{supp}(\mu) \text{ is compact, } \mu \ll \mathcal{L}_d \right\}.$$

2. The *Wasserstein-1 distance*, defined for $\mu, \nu \in \mathcal{P}_{ac}^c(\mathbb{R}^d)$ as

$$W_1(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |\mathbf{x} - \mathbf{y}| d\gamma(dx, dy) \right\},$$

where $\Pi(\mu, \nu)$ denotes the set of measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ s.t. $\operatorname{proj}_x(\gamma) = \mu$ and $\operatorname{proj}_y(\gamma) = \nu$.

NEURAL TRANSPORT EQUATION

MAIN RESULT

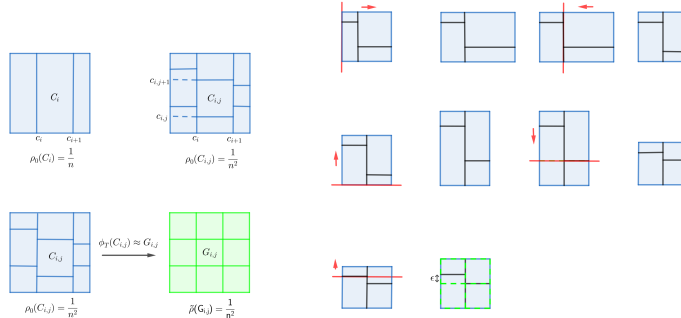
Theorem 4

Let $d, p \in \mathbb{N}^*$, and let $\rho_0 \in \mathcal{P}_{ac}^c(\mathbb{R}^d)$, $\tilde{\rho}$ be the uniform probability over $[0, 1]^d$ and $T > 0$ be fixed. Consider the neural transport equation.

For any $\epsilon > 0$, there exist piece-wise constant controls $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \in L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$, such that the solution of the transport equation with initial condition ρ_0 , $\rho(T)$, satisfies:

$$W_1(\rho(T), \rho_*) \leq \epsilon$$

Moreover, the number of discontinuities of the controls is: $M \sim_{d, \rho_0} \frac{1}{p \epsilon^d}$



(a) Step 1: partition

(b) Step 2: control

NEURAL TRANSPORT EQUATION

EXPLICIT COMPUTATION OF THE NUMBER OF DISCONTINUITIES

There exists a constant $C = C_{d,\rho_0} > 0$ such that, for any $n \in \mathbb{N}^*$, there exist controls, $(\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p \in L^\infty((0, T); \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$, such that the associated solution of the neural transport equation satisfies

$$W_1(\rho(T), \rho_*) \leq C \frac{1}{n}.$$

Moreover, the number M of values taken by the controls is:

- In the case $p = 1$:

$$M = (2d - 1) + (n + \dots + n^d) = (2d - 1) + \frac{n^{d+1} - n}{n - 1}.$$

- In the case $p \geq 1$:

$$M = \left\lfloor \frac{2d}{p} \right\rfloor + \left(\left\lfloor \frac{n}{p} \right\rfloor + 1 + \dots + \left\lfloor \frac{n^d}{p} \right\rfloor + 1 \right)$$


- In the case $p = p_1 + \dots + p_d$, we can simultaneously put p_i velocities on the i th coordinate and we have:

$$M = \left\lfloor \frac{2d}{p} \right\rfloor + \max \left\{ \left\lfloor \frac{n}{p_1} \right\rfloor + 1, \dots, \left\lfloor \frac{n^d}{p_d} \right\rfloor + 1 \right\}.$$

OPEN PROBLEMS

- ▶ **Better algorithm for the general case:** the question of approaching the autonomous regime as $p \rightarrow +\infty$, in the general case or with less restrictive hypothesis, remains open.
- ▶ **Universal approximation:** We proved that we can interpolate any function in any finite set of points. The natural question after this is, how can we now approximate function on the whole space \mathbb{R}^d ? and, what regularity hypothesis must make on the target function?
- ▶ **Control of the transport equation:** The question of controlling the neural transport equation to any target probability, and what hypothesis has to be made on it, remains open.
- ▶ **Optimal Lipschitz constant:** The question of finding the optimal Lipschitz constant for the autonomous field that achieves the simultaneous control, is an interesting question that may help us with the other open questions.
- ▶ **Switching dimensions:** The equation describing a resnet imposes that the dimension stays the same as we go from one layer to the other. It is interesting to think what we can gain from allowing the dimension to switch at strategic times, both shrinking the dynamics to reduce complexity or augmenting the dimensionality of the states to create space. How to carry out the projections or produce new coordinates (maybe through nonlinear functions applied on the data?) are interesting questions to work on.

REFERENCES I

-  Ruiz-Balet, Domènec and Enrique Zuazua (2021). *Neural ODE control for classification, approximation and transport*. DOI: 10.48550/ARXIV.2104.05278. URL: <https://arxiv.org/abs/2104.05278>.