Interplay between depth and width for interpolation in neural ODEs

X Partial differential equations, optimal design and numerics

Antonio Álvarez-López

Joint work with Arselane Hadj Slimane and Enrique Zuazua

Department of Mathematics,
Universidad Autónoma de Madrid

August 22, 2024

ELSEVIER

Neural Networks
Available online 19 August 2024, 106640
In Press, Journal Pre-proof ? What's this?

Full Length Article

Interplay between depth and width for interpolation in neural ODEs

Antonio Álvarez-López [a], Arselane Hadj Slimane [b], Enrique Zuazua [a c d]

Table of contents

Supervised Learning

Goal

Input space $(\mathcal{X}, \mu^*) \subset \mathbb{R}^d \xrightarrow{\quad F^* \quad} $ Output space $\mathcal{Y} \subset \mathbb{R}^m$

Approximate (*learn*) F^* from a dataset $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$:

$$\mathbf{x}_n \sim \mu^*, \qquad \mathbf{y}_n = F^*(\mathbf{x}_n), \qquad n = 1, \dots N.$$

$x_1 \in \mathbb{R}^{3N_{pixels}}$

$x_2 \in \mathbb{R}^{3N_{pixels}}$

$\xrightarrow{\quad F^* \quad}$

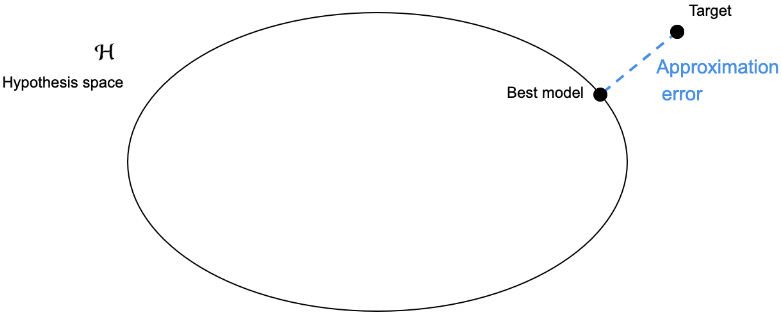
$y_1 = (1, 0) \equiv$ Cat

$y_2 = (0, 1) \equiv$ Dog

Main paradigms I: Approximation

Fix a hypothesis space $\mathcal{H} = \mathcal{H}_\theta$.

How close is $\mathcal{H}$ to the target F^* given a specified bound on θ?

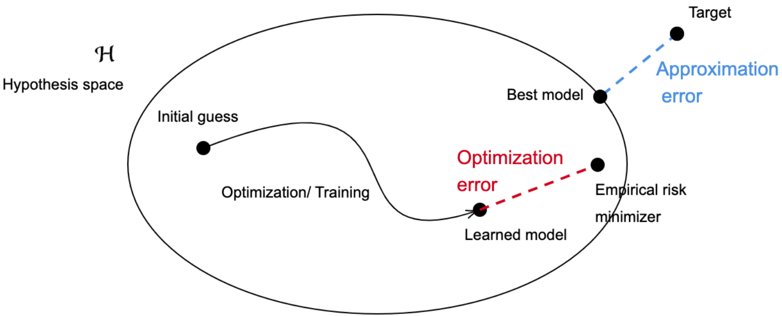


Origin: Expressivity vs overfitting.

Main paradigms II: Optimization

Fix an objective function $\mathcal{J}(\theta) := \frac{1}{N}\sum_{n=1}^{N} L(F_\theta(\mathbf{x}_n), \mathbf{y}_n) + R(\theta)$.

$$\text{How can we find } \hat{F} := \underset{F_\theta \in \mathcal{H}}{\operatorname{argmin}}\ \mathcal{J}(\theta)?$$

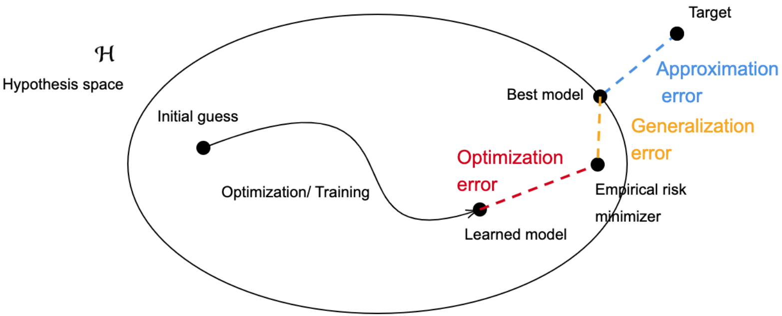


Origin: Non-convexity of L with respect to θ.

Main paradigms III: Generalization

Unknown population μ^*.

Can $\hat{F}$ correctly predict the value of F^* in any new point $\mathbf{x} \in \mathcal{X} \setminus \mathcal{D}$?



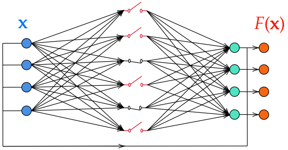
Origin: Gap $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^*} L(F_\theta(\mathbf{x}), \mathbf{y})$ vs $\frac{1}{N}\sum_{n=1}^{N} L(F_\theta(\mathbf{x}_n), \mathbf{y}_n)$.

The hypothesis space of ResNets[1]

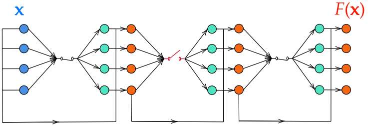
$$\begin{cases} \mathbf{x_{k+1}} & = \mathbf{x_k} + \sum_{i=1}^{p} \mathbf{w_{k,i}} \sigma\left(\mathbf{a_{k,i}} \cdot \mathbf{x_k} + b_{k,i}\right), \qquad k = 0, \dots, L-1, \\ \mathbf{x_0} & \in \mathbb{R}^d. \end{cases}$$

Depth $L \geq 1$ (number of hidden layers);
Parameters $(\mathbf{w_{k,i}}, \mathbf{a_{k,i}}, b_{k,i}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$;

Width $p \geq 1$;
Activation $\sigma : \mathbb{R} \to \mathbb{R}$.



(a) Limiting case 1: $p \gg 1, L = 1$



(b) Limiting case 2: $p = 1, L \gg 1$

[1] [1] K. He, X Zhang, S. Ren, J Sun, "Deep residual learning for image recognition" (2016)

Neural ODEs (continuous-time limit)

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right), \qquad t \in (0, T). \tag{1}$$

- **Control**: $\theta := (\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^{p}, \quad \theta(t) \in L^\infty\left((0,T); (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p\right).$
- **ReLU activation**: $\sigma(z) = (z)_+$ Lipschitz, nonlinear.
- **Flow map** in time T generated by (1) is well defined:

$$\begin{aligned} \Phi_T(\cdot;\theta) : \mathbb{R}^d &\to \mathbb{R}^d \\ \mathbf{x_0} &\mapsto \mathbf{x}(T; \mathbf{x_0}). \end{aligned}$$

Assume θ piecewise constant in $(0, T)$, $\underbrace{L \text{ discontinuities}}_{\sim \text{ Transitions between layers}}$

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p}\sum_{j=1}^{L} \mathbf{w_{i,j}}\sigma\left(\mathbf{a_{i,j}}\cdot\mathbf{x} + b_{i,j}\right)\mathbf{1}_{(t_{j-1},t_j)}(t), \qquad t \in (0, T).$$

Neural ODEs (continuous-time limit)

$$\boxed{\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right), \qquad t\in(0,T). \qquad (1)}$$

- **Control**: $\theta := (\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^{p}, \quad \theta(t)\in L^\infty\left((0,T);\left(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}\right)^p\right).$
- **ReLU activation**: $\sigma(z) = (z)_+$ Lipschitz, nonlinear.
- **Flow map** in time T generated by (1) is well defined:

$$\begin{aligned} \Phi_T(\cdot;\theta) : \mathbb{R}^d &\to \mathbb{R}^d \\ \mathbf{x_0} &\mapsto \mathbf{x}(T;\mathbf{x_0}). \end{aligned}$$

Assume θ piecewise constant in $(0,T)$, $\underbrace{L \text{ discontinuities}}_{\sim \text{ Transitions between layers}}$

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p}\sum_{j=1}^{L} \mathbf{w_{i,j}}\sigma\left(\mathbf{a_{i,j}}\cdot\mathbf{x} + b_{i,j}\right)\mathbf{1}_{(t_{j-1},t_j)}(t), \qquad t\in(0,T).$$

Problem statement

Dataset $\mathcal{D} := \{(\mathbf{x_n}, \mathbf{y_n})\} \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbf{x_n} \neq \mathbf{x_m}, \quad \mathbf{y_n} \neq \mathbf{y_m}, \quad$ if $n \neq m$.

$$\mathcal{J}(\theta) := \frac{1}{N}\sum_{n=1}^{N} |\Phi_T(\mathbf{x}_n, \theta) - \mathbf{y}_n|^2 + R(\theta).$$

Problem

- For any $T > 0$, find a control θ s.t. $\Phi_T(\mathbf{x_n}; \theta) = \mathbf{y_n}$ for all n, with **minimal complexity** (number of switches $L \times$ width p).
- How can L and p **interact** with each other to achieve the goal?

Motivation

- Theoretical: Understanding dynamics and architecture, measure of **expressivity** (the complexity required to interpolate).
- Practical: New methods to attack generalization, **optimal design** of neural ODEs, **initialization** of parameters for optimization.

Problem statement

Dataset $\mathcal{D} := \{(\mathbf{x_n}, \mathbf{y_n})\} \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\quad \mathbf{x_n} \neq \mathbf{x_m}, \quad \mathbf{y_n} \neq \mathbf{y_m}, \quad$ if $n \neq m$.

$$\mathcal{J}(\theta) := \frac{1}{N} \sum_{n=1}^{N} |\Phi_T(\mathbf{x}_n, \theta) - \mathbf{y}_n|^2 + R(\theta).$$

Problem

- For any $T > 0$, find a control θ s.t. $\Phi_T(\mathbf{x_n}; \theta) = \mathbf{y_n}$ for all n, with **minimal complexity** (number of switches $L \times$ width p).
- How can L and p **interact** with each other to achieve the goal?

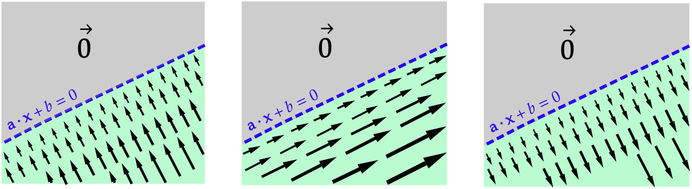
Motivation

- Theoretical: Understanding dynamics and architecture, measure of **expressivity** (the complexity required to interpolate).
- Practical: New methods to attack generalization, **optimal design** of neural ODEs, **initialization** of parameters for optimization.

Basic interpretation of the dynamics

$$p = 1: \qquad \dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t))$$

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



From left to right: Compression, laminar motion, expansion.

Exact control (L vs p)

Theorem (A. Á-L, A. Hadj-Slimane, E. Zuazua)

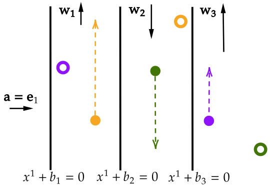
For any $T > 0$, there exists a control

$$\theta \in L^{\infty}\left((0,T); \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^{p}\right)$$

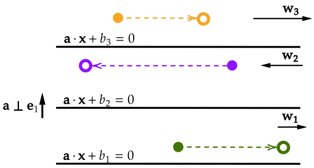
such that

$$\Phi_T(\mathbf{x_n};\theta) = \mathbf{y_n}, \quad \textit{for all } n = 1,\dots,N.$$

Moreover, θ is piecewise constant with $L = 2\lceil N/p \rceil - 1$ discontinuities.



(a) Step 1: Simultaneous control of $d-1$ coordinates $x^{(2)},\dots,x^{(d)}$.



(b) Step 2: Simultaneous control of the remaining coordinate $x^{(1)}$.

More sparsity: Semi-autonomous system

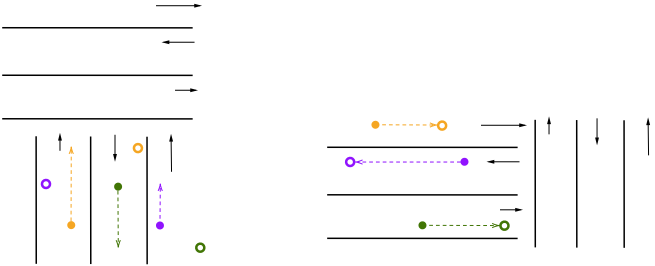
For any $T > 0$, there exists a control

$$\theta = (\mathbf{w_i}, \mathbf{a_i}, b_i)_{i=1}^{p} \in \left(\mathbb{R}^d \times \mathbb{R}^d \times L^\infty((0,T);\mathbb{R})\right)^p$$

such that

$$\Phi_T(\mathbf{x_n};\theta) = \mathbf{y_n}, \quad \text{for all } n = 1, \dots, N.$$

Moreover, $(b_1, \dots, b_p)$ is piecewise constant with $L = 2\lceil N/p \rceil - 1$ discontinuities.



(a) Step 1. (b) Step 2.

For width $\geq$ number of data: $\quad L = 2\lceil N/p \rceil - 1 = 2 - 1 = 1.$

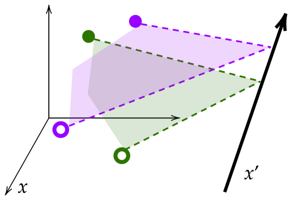
Is it possible to achieve exact control using $\mathbf{L = 0}$ discontinuities?

Autonomous system ($L = 0$): Approach I

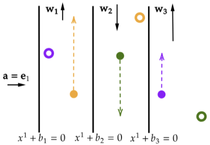
High-dimensional setting

In the conditions of the previous theorem, if $d > N$ then we can improve to

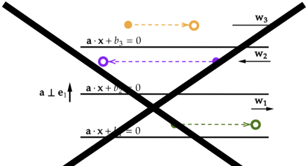
$$L = 2\lceil N/p \rceil - 2 \quad \text{discontinuities.}$$



Change axis $x \mapsto x'$ s.t. $x_n^{(1)} = y_n^{(1)}$ for all n in the new vector basis.



(a) Step 1: Simultaneous control of $d - 1$ coordinates $x^{(2)}, \ldots, x^{(d)}$.



(b) Step 2: Simultaneous control of the remaining coordinate $x^{(1)}$.

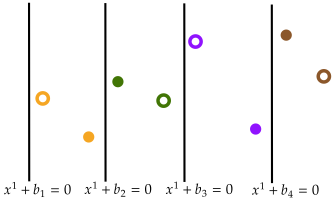
Autonomous system ($L = 0$): Approach II

Probabilistic control

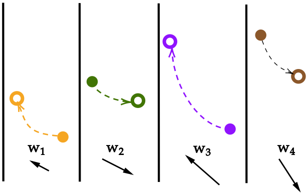
Assume that $\mathbf{x}_n, \mathbf{y}_n \sim U([0,1]^d)$ for all n. Then, with probability P bounded as

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

there exists $\theta \in \mathbb{R}^{d\times N} \times \mathbb{R}^{N\times d} \times \mathbb{R}^N$ such that $\Phi_T(\cdot, \theta)$ interpolates the dataset.



(a) Step 1: Separation.



(b) Step 2: Transversal velocity fields.

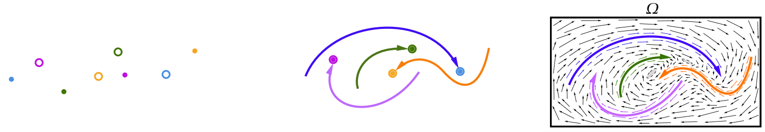
Autonomous system: Approach III

Relaxation to approximate control

For any $T > 0$, there exists a constant control $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^{p}$ such that

$$\sup_{n\in\{1,\ldots,N\}} |\mathbf{y_n} - \Phi_T(\mathbf{x_n};\theta)| \leq C\, \frac{\log_2(\kappa)}{\kappa^{1/d}},$$

where $\kappa = (d+2)dp$ and $C > 0$ is independent of κ.



Lemma (F. Bach, 2014)

Let $\Omega := [-R, R]^d$ *and* $f \in \mathrm{Lip}(\Omega, \mathbb{R})$. *There exists a shallow network* F_p *of width* p *s.t.*

$$\sup_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - F_p(\mathbf{x})| \leq C_{d,R}\, \mathrm{Lip}(f)\, \frac{\log_2 \kappa}{\kappa^{1/d}}, \qquad \textit{where } \kappa = (d+2)p.$$

Neural transport equation

$$\begin{cases} \dot{\mathbf{x}}(t) = \displaystyle\sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right), & t \in (0,T), \\ \mathbf{x}(0) = \mathbf{x}_n \sim \mu_0 \in \mathcal{P}(\mathbb{R}^d), & n = 1,\dots,N. \end{cases}$$

$$\downarrow$$

$$\begin{cases} \partial_t \mu + \mathrm{div}_x \left(\mu \sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right)\right) = 0 \\ \mu(0) = \mu_0. \end{cases} \tag{2}$$

Neural transport equation

$$\begin{cases} \dot{\mathbf{x}}(t) = \displaystyle\sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right), & t \in (0,T), \\ \mathbf{x}(0) = \mathbf{x}_n \sim \mu_0 \in \mathcal{P}(\mathbb{R}^d), & n = 1,\ldots,N. \end{cases}$$

$$\downarrow$$

$$\begin{cases} \partial_t \mu + \mathrm{div}_x \left(\mu \sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right)\right) = 0 \\ \mu(0) = \mu_0. \end{cases} \tag{2}$$

Interpolation of measures

- **Space:** $\mathcal{P}_{ac}^{c}(\mathbb{R}^d)$.
- **Metric:** $W_q(\mu,\nu) := \left(\min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\mathbf{x}-\mathbf{y}|^q d\gamma(x,y)\right)^{1/q}$,
  where $\Pi(\mu,\nu) \subset \mathcal{P}_{ac}^{c}(\mathbb{R}^d\times\mathbb{R}^d)$ is the set of all couplings of μ and ν.
- The curve in $\mathcal{P}_{ac}^{c}(\mathbb{R}^d)$ defined by the **push-forward measure**

$$\mu(t)(\cdot) := \Phi_t(\cdot;\theta)\#\mu_0, \qquad t\in(0,T),$$

solves

$$\partial_t\mu + \mathrm{div}_x\Bigg(\mu \underbrace{\sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x}+b_i(t)\right)}_{\text{Lipschitz in } \mathbf{x}}\Bigg) = 0, \qquad \mu(0)=\mu_0.$$

Problem

Fix $\mu_* := U([0,1]^d)$. For any $\mu_0 \in \mathcal{P}_{ac}^{c}(\mathbb{R}^d)$, find a control $\theta := (\mathbf{w_i},\mathbf{a_i},b_i)_{i=1}^{p}$ s.t.

$$W_q\left(\mu(T),\,\mu_*(\cdot)\right)\approx 0.$$

Interpolation of measures

- **Space:** $\mathcal{P}_{ac}^{c}(\mathbb{R}^d)$.
- **Metric:** $W_q(\mu,\nu) := \left(\min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\mathbf{x}-\mathbf{y}|^q d\gamma(x,y)\right)^{1/q}$, where $\Pi(\mu,\nu) \subset \mathcal{P}_{ac}^{c}(\mathbb{R}^d\times\mathbb{R}^d)$ is the set of all couplings of μ and ν.
- The curve in $\mathcal{P}_{ac}^{c}(\mathbb{R}^d)$ defined by the **push-forward measure**

$$\mu(t)(\cdot) := \Phi_t(\cdot;\theta)\#\mu_0, \qquad t\in(0,T),$$

solves

$$\partial_t\mu + \mathrm{div}_x\left(\mu \underbrace{\sum_{i=1}^{p} \mathbf{w_i}(t)\sigma\left(\mathbf{a_i}(t)\cdot\mathbf{x} + b_i(t)\right)}_{\text{Lipschitz in } \mathbf{x}}\right) = 0, \qquad \mu(0)=\mu_0.$$

Problem

Fix $\mu_* := U([0,1]^d)$. For any $\mu_0 \in \mathcal{P}_{ac}^{c}(\mathbb{R}^d)$, find a control $\theta := (\mathbf{w_i}, \mathbf{a_i}, b_i)_{i=1}^{p}$ s.t.

$$W_q\left(\mu(T), \mu_*(\cdot)\right) \approx 0.$$

Interpolation of measures

Theorem (A. Á-L, A. Hadj-Slimane, E. Zuazua)

For any $d, p \geq 1$, $T, \varepsilon > 0$ *and* $q \in [1, \frac{d}{d-1})$, *there exists a piecewise constant control*

$$\theta \in L^\infty\left((0, T); \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p\right)$$

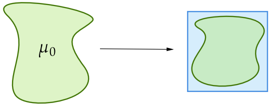
such that the solution $\mu(t)$ *of* (2), *taking* μ_0 *as initial condition, satisfies*

$$W_q(\mu(T), \mu_*) < \varepsilon,$$

and the number of switches of θ *is* $L = \lceil 2d/p \rceil + \left\lceil \frac{1}{p-d+1} \left(\frac{3^{1+d/q}\sqrt{d}}{\varepsilon} \right)^{\frac{d}{1+d/q-d}} \right\rceil - 1.$

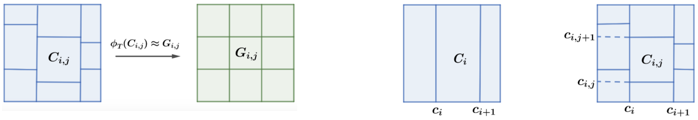
In particular, if $q = 1$ *then* $L = \lceil 2d/p \rceil + \left\lceil \frac{1}{p-d+1} \left(\frac{3^{1+d}\sqrt{d}}{\varepsilon} \right)^{d} \right\rceil - 1.$

Idea of the proof:

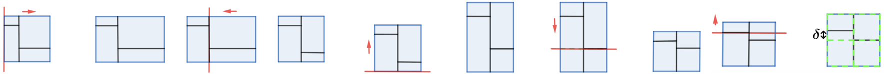


Step 1. We compress μ_0 into $[0, 1]^d$.

Neural transport equation: Interpolation of measures



Step 2. We define two partitions of $[0,1]^d$ into rectangles $C_{i,j}$ and $G_{i,j}$ which contain the same small mass as distributed by μ_0 and μ_*, respectively.



Step 3. Transformation of each rectangle $C_{i,j}$ into the corresponding rectangle $G_{i,j}$ through a sequence of compressions and expansions (from left to right).

Conclusions

- Exact interpolation of data and measures can be constructively attained, showing a trade-off between depth and width.
- Error decay for autonomous, wide enough models via universal approximation.
- In high dimensions, the required width scales with the size of the dataset.

Open problems

- Minimize the number of switches. Is it sharp?
- Explicit control algorithm for the autonomous regime?
- Same for the semi-autonomous model with continuous (linear?) bias $\mathbf{b}(t)$.
- Other activation functions? Which is the optimal one?
- Extension to infinite width as the mean-field limit?

$$\dot{\mathbf{x}}(t) = \int_{\mathbb{R}^{2d+1}} \mathbf{w}\sigma(\mathbf{a}\cdot\mathbf{x}(t) + b)d\mu(t).$$

- Interpolation of measures supported in $\mathbb{R}^d$?

Antonio Álvarez-López, Rafael Orive-Illera, and Enrique Zuazua.
Optimized classification with neural ODEs via separability.
arXiv preprint arXiv:2312.13807, 2023.

Antonio Álvarez-López, Arselane H. Slimane, and Enrique Zuazua.
Interplay between depth and width for interpolation in neural ODEs.
arXiv preprint arXiv:2401.09902, 2024.

Domènec Ruiz-Balet and Enrique Zuazua.
Neural ODE Control for Classification, Approximation, and Transport.
SIAM Review, 65(3):735--773, 2023.

Thank you for your attention!