Departamento de Matemáticas,
Universidad Autónoma de Madrid

Deep Learning with Controlled Flows

Expressivity, Generalization and Generation

Antonio Álvarez López

Ph.D. Thesis Defense | July 1, 2026

PART I



Supervised Learning

Supervised Learning

Input space $\mathcal{X} \subset \mathbb{R}^d$ $\xrightarrow{\exists y(\cdot)}$ Output space $\mathcal{Y} \subset \mathbb{R}^m$.

Goal: Build a model that approximates (*learns*) $y(\cdot)$ from a finite dataset

$$\mathcal{D} := \{(x_i, y_i = y(x_i))\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y} \qquad (N < \infty).$$

Basic assumptions: Deterministic setting with no noise, and $x_i \neq x_j$ for all $i \neq j$.

Supervised Learning

Input space $\mathcal{X} \subset \mathbb{R}^d$ $\xrightarrow{\exists y(\cdot)}$ Output space $\mathcal{Y} \subset \mathbb{R}^m$.

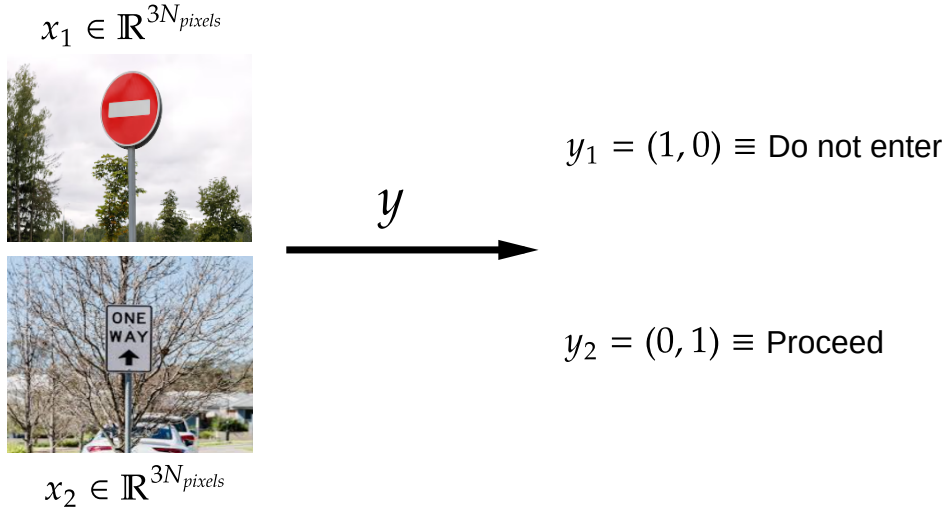
Goal: Build a model that approximates (*learns*) $y(\cdot)$ from a finite dataset

$$\mathcal{D} := \{(x_i, y_i = y(x_i))\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y} \qquad (N < \infty).$$

Basic assumptions: Deterministic setting with no noise, and $x_i \neq x_j$ for all $i \neq j$.

Classification

range(y) is finite.



Supervised Learning

Input space $\mathcal{X} \subset \mathbb{R}^d$ $\xrightarrow{\exists y(\cdot)}$ Output space $\mathcal{Y} \subset \mathbb{R}^m$.

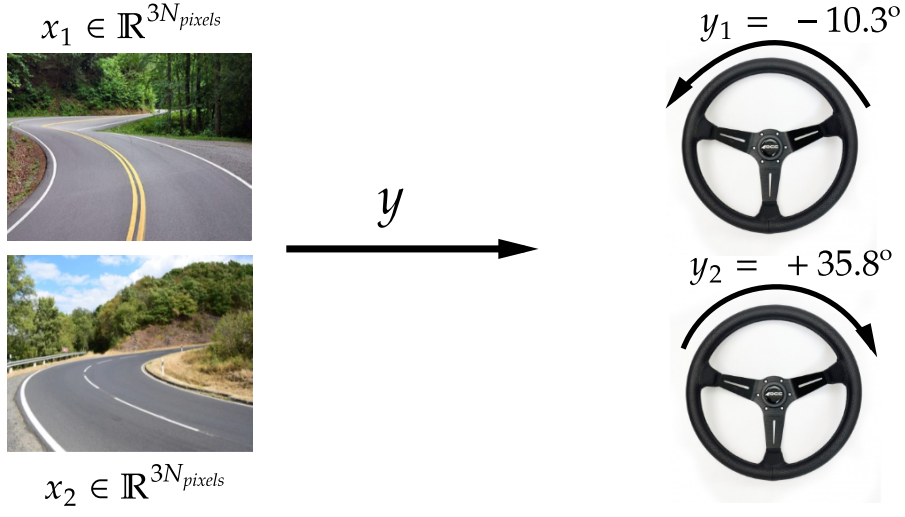
Goal: Build a model that approximates (*learns*) $y(\cdot)$ from a finite dataset

$$\mathcal{D} := \{(x_i, y_i = y(x_i))\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y} \qquad (N < \infty).$$

Basic assumptions: Deterministic setting with no noise, and $x_i \neq x_j$ for all $i \neq j$.

Regression

range(y) is continuous.



Main paradigms I: Approximation

Fix a (parameterized) model class $\mathcal{H}_{\Theta_p} = \{f_\theta : \mathcal{X} \to \mathcal{Y} \ : \ \theta \in \Theta_p\}$.

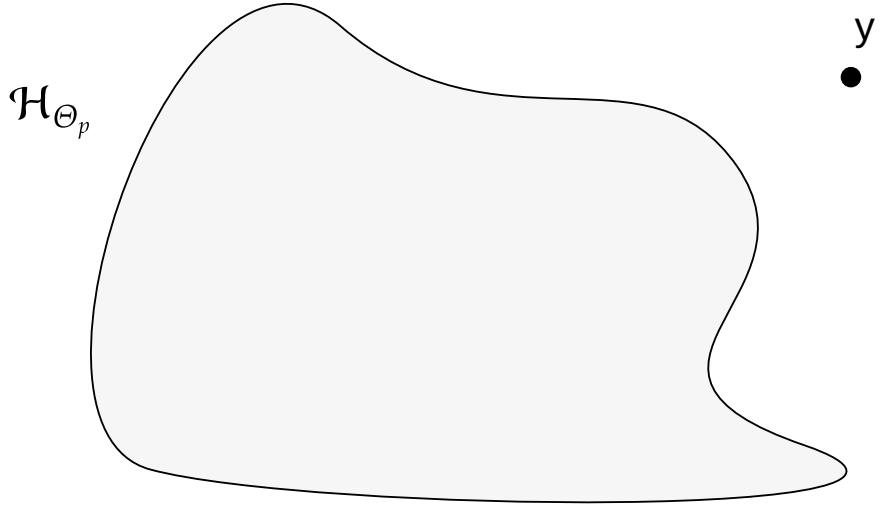
How close is $\mathcal{H}_{\Theta_p}$ to the target $y(\cdot)$ for a given parameter budget p?

y

Main paradigms I: Approximation

Fix a (parameterized) model class $\mathcal{H}_{\Theta_p} = \{f_\theta : \mathcal{X} \to \mathcal{Y} \ : \ \theta \in \Theta_p\}$.

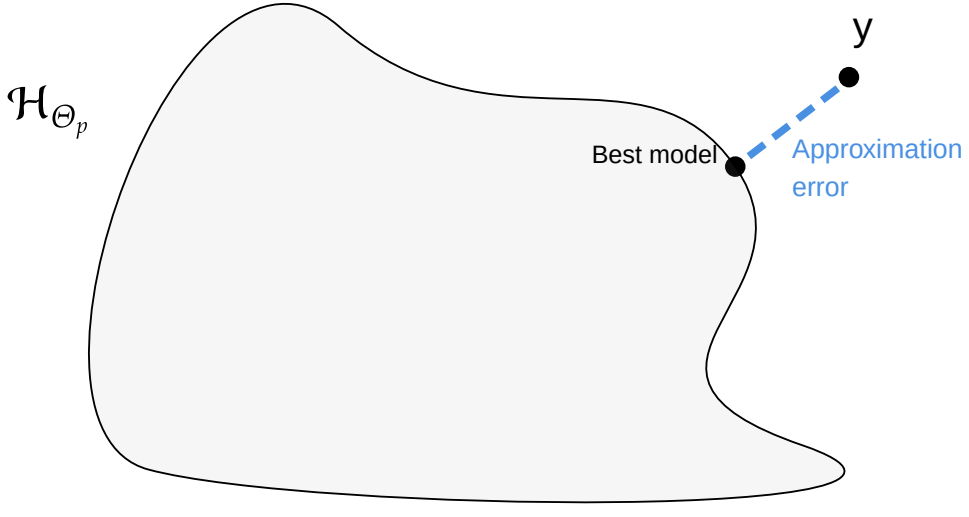
How close is $\mathcal{H}_{\Theta_p}$ to the target $y(\cdot)$ for a given parameter budget p?



Main paradigms I: Approximation

Fix a (parameterized) model class $\mathcal{H}_{\Theta_p} = \{f_\theta : \mathcal{X} \to \mathcal{Y} \ : \ \theta \in \Theta_p\}$.

How close is $\mathcal{H}_{\Theta_p}$ to the target $y(\cdot)$ for a given parameter budget p?



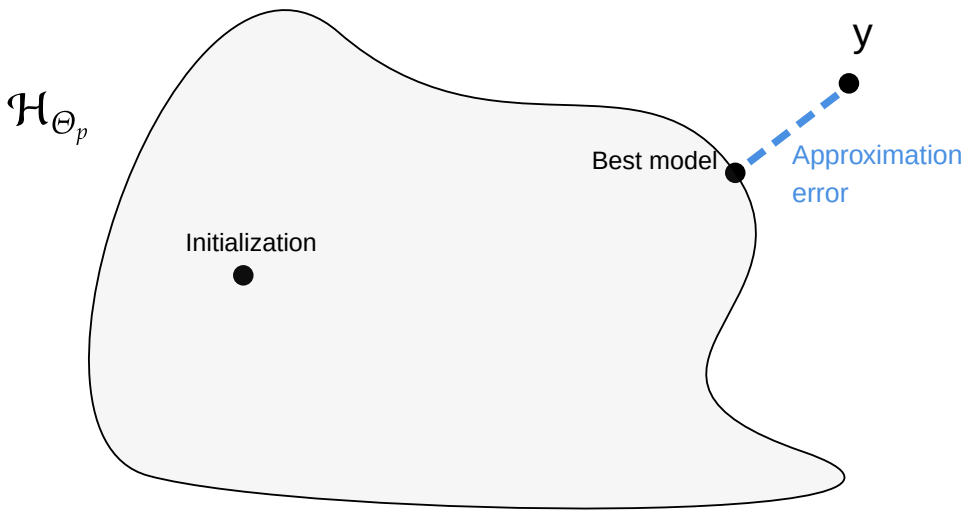
Main paradigms II: Optimization

The standard approach is to minimize the empirical risk:

$$\min_{\theta \in \Theta} \mathcal{R}_N(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(x_i), y_i\big)$$

(ℓ *is a chosen loss function.*)

How to navigate the resulting optimization landscape, which may be non-convex?



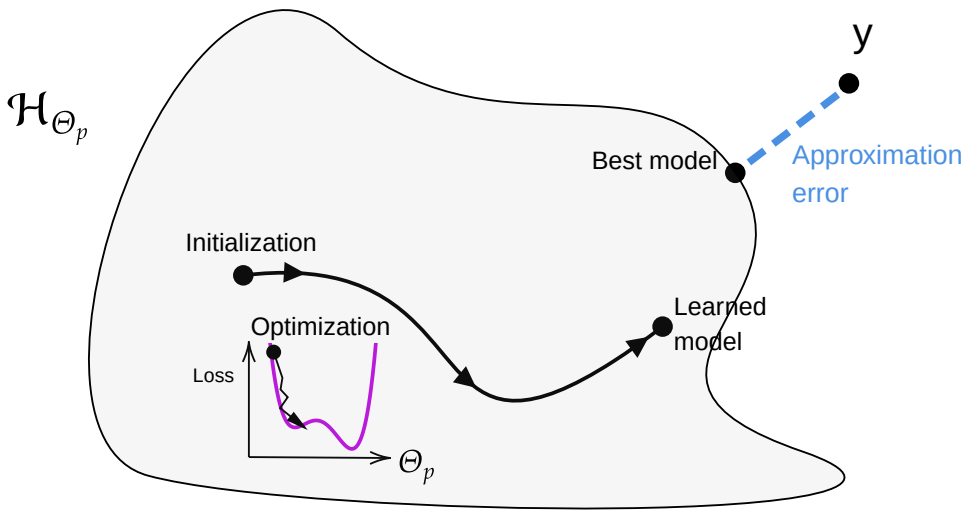
Main paradigms II: Optimization

The standard approach is to minimize the empirical risk:

$$\min_{\theta \in \Theta} \mathcal{R}_N(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(x_i), y_i\big)$$

(ℓ *is a chosen loss function.*)

How to navigate the resulting optimization landscape, which may be non-convex?



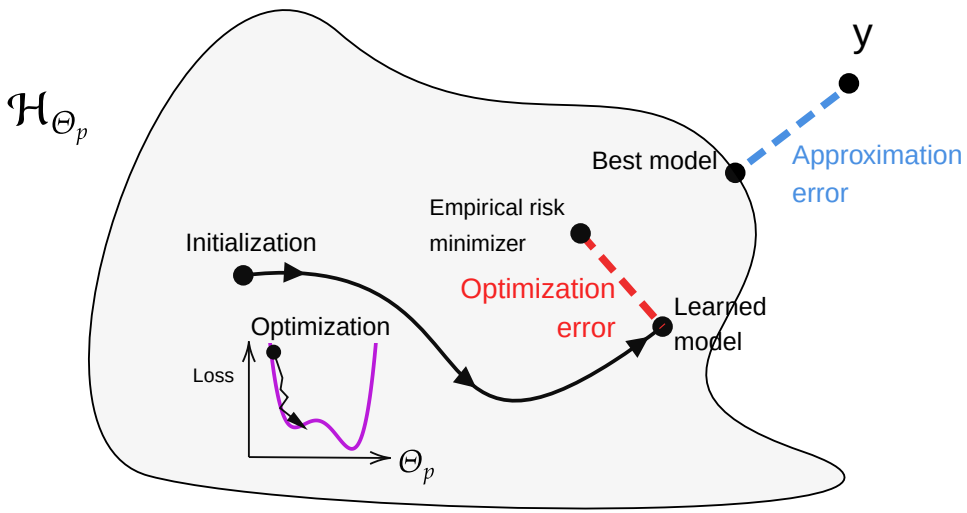
Main paradigms II: Optimization

The standard approach is to minimize the empirical risk:

$$\min_{\theta \in \Theta} \mathcal{R}_N(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(x_i), y_i\big)$$

(ℓ *is a chosen loss function.*)

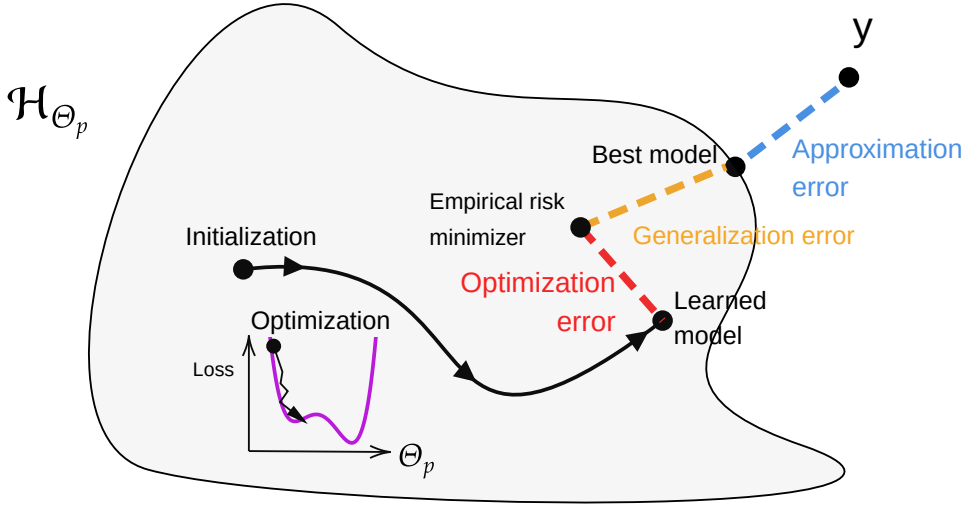
How to navigate the resulting optimization landscape, which may be non-convex?



Main paradigms III: Generalization

But the ultimate theoretical goal is to minimize the population risk:

$$\mathcal{R}(\theta) := \int_{\mathcal{X}} \ell\big(f_\theta(x), y(x)\big)\, \mathrm{d}\mu(x)$$

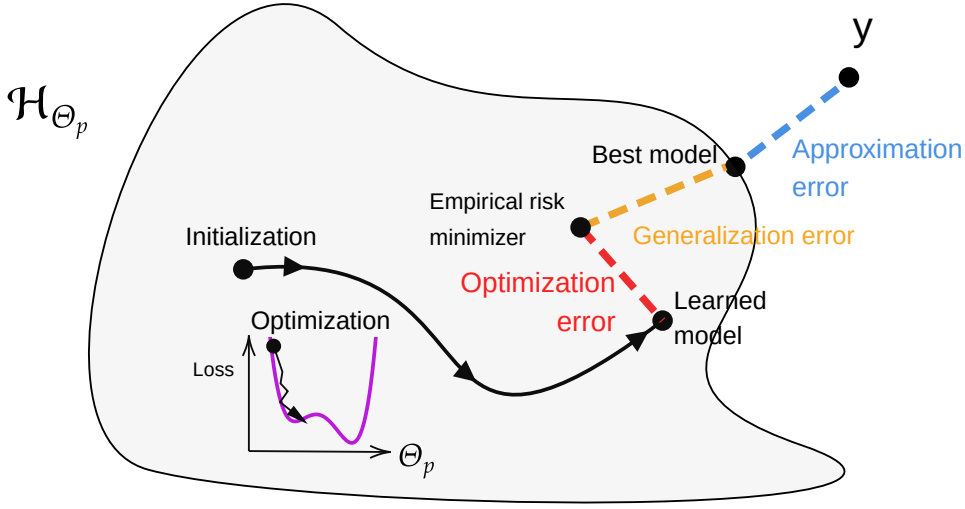


Main paradigms III: Generalization

But the ultimate theoretical goal is to minimize the population risk:

$$\mathcal{R}(\theta) := \int_{\mathcal{X}} \ell\big(f_\theta(x), y(x)\big)\, \mathrm{d}\mu(x) \quad \leq \quad \mathcal{R}_N(\theta) \quad + \quad \sup_{\vartheta \in \Theta} \big|\mathcal{R}(\vartheta) - \mathcal{R}_N(\vartheta)\big|$$

How to balance expressivity and complexity and obtain performance guarantees?



Model classes I: linear models

Family of linear combinations of a collection of fixed basis functions (features):

$$f_\theta(x) = \sum_{j=1}^{p} \theta_j \phi_j(x) \qquad (x \in \mathbb{R}^d).$$

Parameters: $\theta = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$. **Features:** $\phi_1, \ldots, \phi_p : \mathbb{R}^d \to \mathbb{R}$.

Classical examples (in 1D):

- Polynomials: $\phi_j(x) = x^{j-1}$.
- Trigonometric functions: $\phi_j(x) = \sin(jx)$ or $\cos(jx)$.
- Finite elements: Piecewise polynomials on a fixed grid.

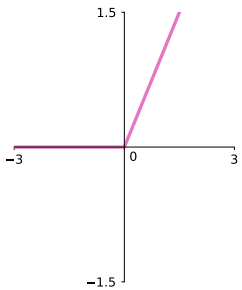
Pros & Cons: Convex optimization for convex losses, but expressivity limited by the fixed features.

Model classes II: shallow neural networks

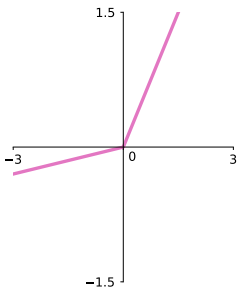
We let each feature adapt to the data by taking the form of a neuron[1]/ridge function:

$$f_\theta(x) = \sum_{j=1}^{p} w_j \sigma(a_j^\top x + b_j) \qquad (x \in \mathbb{R}^d).$$

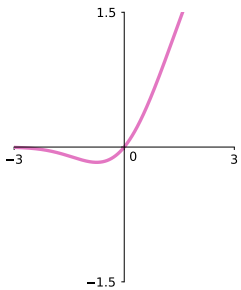
Parameters: $\theta = (w_j, a_j, b_j)_{j=1}^p \in \left(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}\right)^p$. **Activation:** $\sigma : \mathbb{R} \to \mathbb{R}$.



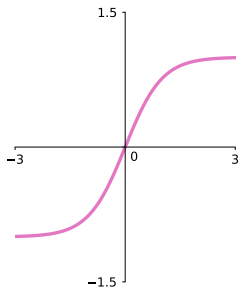
ReLU
$\sigma(s) = s_+$



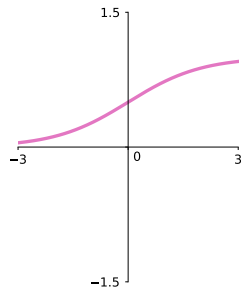
Leaky ReLU
$\sigma(s) = s \vee (0.1s)$



GELU
$\sigma(s) = \frac{s}{2}\left(1 + \mathrm{erf}(\frac{s}{\sqrt{2}})\right)$



Tanh
$\sigma(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$



Sigmoid
$\sigma(s) = \frac{1}{1+e^{-s}}$

Pros & Cons: Higher representation power[2], but the loss landscape is generally non-convex.

---

[1] Inspired by biological nervous systems, artificial neurons originated as static threshold logic gates W. S. McCulloch & W. Pitts. Bull. Math. Biophys. (1943) and evolved into the adaptive perceptron capable of learning from data F. Rosenblatt. Psychological Review (1958).

[2] Three classical references are: G. Cybenko. Math. Control Signals Syst. (1989); K. Hornik. Neural Networks (1991); A. Pinkus. Acta Numerica (1999)

Model classes III: multilayer perceptrons

Instead of adding more neurons to one layer (width), we compose several layers (depth):

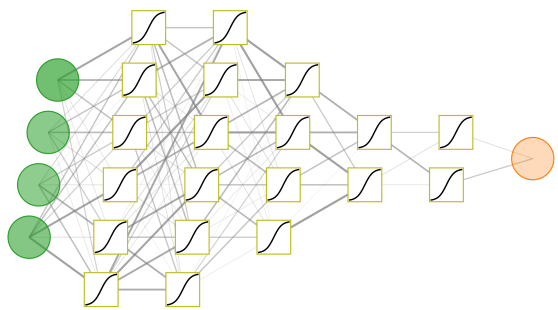
$$\begin{cases} x_0 = x \in \mathbb{R}^d, \\ x_{\ell+1} = \sigma(A_\ell x_\ell + b_\ell) & (\ell = 0, \dots, L-1), \\ f_\theta(x) = A_L x_L + b_L. \end{cases}$$

Parameters: $\theta = (A_\ell, b_\ell)_{\ell=0}^{L} \in \prod_{\ell=0}^{L} \left(\mathbb{R}^{d_{\ell+1} \times d_\ell} \times \mathbb{R}^{d_{\ell+1}}\right)$

Pros & Cons: Hierarchical representations boost expressivity, and efficient gradient computation via the backpropagation algorithm[3]:

$$\frac{\partial f_\theta(x)}{\partial \theta_\ell} = \frac{\partial f_\theta(x)}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_{L-1}} \cdots \frac{\partial x_{\ell+2}}{\partial x_{\ell+1}} \cdot \frac{\partial x_{\ell+1}}{\partial \theta_\ell}.$$

But large L can cause gradients to vanish or explode!



Graph representation of MLP.

[3] Introduced in S. Linnainmaa, Master's thesis (1970) and P. J. Werbos, PhD thesis (1974); popularized by D. E. Rumelhart et al. Nature (1986).

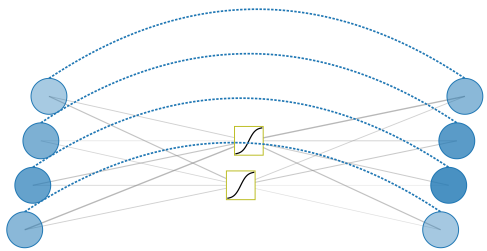
Model classes IV: residual networks

Residual networks[4] (ResNets) help mitigate instabilities by including skip connections.

$$\begin{cases} x_0 = x \in \mathbb{R}^d, \\ x_{\ell+1} = x_\ell + W_\ell \sigma(A_\ell x_\ell + b_\ell) \qquad (\ell = 0, \dots, L-1), \\ f_\theta(x) = x_L \end{cases}$$

Parameters: $\theta = (W_\ell, A_\ell, b_\ell)_{\ell=0}^{L-1} \in \left(\mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p\right)^L$. (here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$)

ResNets are now a standard baseline, but mathematical analysis calls for a simpler framework.



Graph representation of ResNet block.

[4] First proposed in the context of image recognition in K. He, X. Zhang, S. Ren, and J. Sun. CVPR (2016).

Model classes IV: residual networks

Residual networks[4] (ResNets) help mitigate instabilities by including skip connections.

$$\begin{cases} x_0 = x \in \mathbb{R}^d, \\ x_{\ell+1} = x_\ell + W_\ell \sigma(A_\ell x_\ell + b_\ell) \qquad (\ell = 0, \dots, L-1), \\ f_\theta(x) = x_L \end{cases}$$

Parameters: $\theta = (W_\ell, A_\ell, b_\ell)_{\ell=0}^{L-1} \in \left(\mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p\right)^L$. (here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$)

This structure looks familiar...

It takes the exact form of a forward Euler discretization with unit step size.

[4] First proposed in the context of image recognition in K. He, X. Zhang, S. Ren, and J. Sun. CVPR (2016).

Neural ODEs: The continuous-depth limit

The layer index becomes time $t \in [0, T]$ for a fixed $T > 0$.

$$\begin{cases} \dot{x}(t) = W(t)\,\sigma\left(A(t)x(t) + b(t)\right), \quad t \in [0, T], \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases}$$

Neural ODEs: The continuous-depth limit

The layer index becomes time $t \in [0, T]$ for a fixed $T > 0$.

$$\begin{cases} \dot{x}(t) = \sum_{j=1}^{p} w_j(t)\, \sigma\left(a_j(t)^\top x(t) + b_j(t)\right), & t \in [0, T], \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases}$$

Neural ODEs: The continuous-depth limit

The layer index becomes time $t \in [0, T]$ for a fixed $T > 0$.

$$\begin{cases} \dot{x}(t) = \sum_{j=1}^{p} w_j(t)\, \sigma\left(a_j(t)^\top x(t) + b_j(t)\right), & t \in [0, T], \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases}$$

- **Parameter**: $\theta := (w_j, a_j, b_j)_{j=1}^p \in L^\infty\left(0, T; \left(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\right)^p\right)$, piecewise constant. ($L$ discontinuities)
- **Model**: Flow map at time $T > 0$ is well-defined (Cauchy-Lipschitz/Carathéodory):

$$f_\theta = \Phi_T^\theta(\cdot) : x_0 \mapsto x(T; x_0).$$

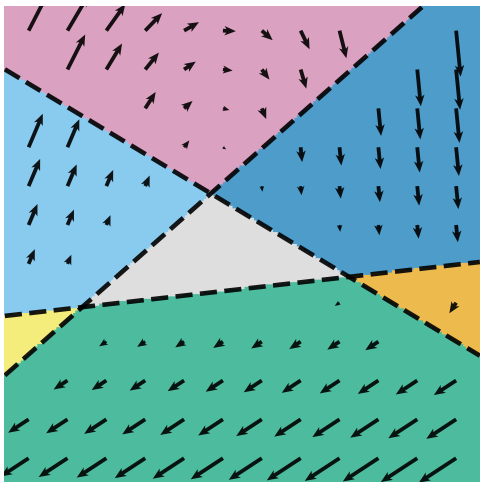
- **Training**: Backpropagation $\rightarrow$ Adjoint method (backward-in-time ODE).

The dynamical-systems viewpoint on deep learning was developed in Weinan E, Commun. Math. Stat. (2017) and E. Haber and L. Ruthotto, Inverse Problems (2018). Neural ODEs and adjoint-based training were introduced in R. T. Q. Chen et al., NeurIPS (2018). See M. E. Sander et al., NeurIPS (2022) and P. Marion et al., ICLR (2024) for quantified ResNet--ODE correspondences, particularly during training.

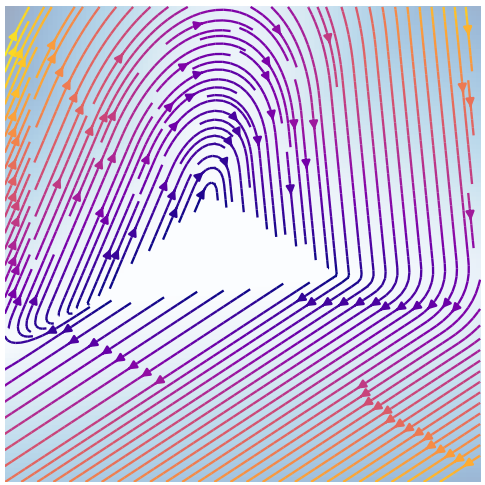
Neural ODEs: The continuous-depth limit

The layer index becomes time $t \in [0, T]$ for a fixed $T > 0$.

$$\begin{cases} \dot{x}(t) = \sum_{j=1}^{p} w_j(t)\, \sigma\left(a_j(t)^\top x(t) + b_j(t)\right), & t \in [0, T], \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases}$$



(a) Polyhedral partition + field.



(b) Induced streamlines.

The connection with control theory

Control theory *is the branch of mathematics that focuses on the manipulation of dynamical systems by applying external inputs to steer the system towards a desired state or trajectory.*

Supervised Learning		Control Theory
Neural network	$\longleftrightarrow$	Controlled dynamical system
Network parameters (W_ℓ, A_ℓ, b_ℓ)	$\longleftrightarrow$	Control signal $(W(t), A(t), b(t))$
Empirical risk minimization	$\longleftrightarrow$	Controllability
Regularized empirical risk minimization	$\longleftrightarrow$	Optimal control problem

Linear, geometric & nonlinear

- R. E. Kalman. IFAC (1960). State-space controllability and observability.
- V. Jurdjevic, H. J. Sussmann. J. Diff. Eq. (1972). Controllability of systems on Lie groups.
- E. D. Sontag. Springer (1998). Foundations of finite-dimensional control.
- J.-M. Coron. AMS (2007). Controllability and stabilization of nonlinear systems.

Optimal control

- R. Bellman. Princeton Univ. Press (1957). Dynamic programming and the Bellman equation.
- L. S. Pontryagin et al. Interscience (1962). Pontryagin maximum principle.
- W. H. Fleming, R. W. Rishel. Springer (1975). Deterministic and stochastic optimal control.
- E. Trélat. Vuibert (2005). Geometric, analytical and numerical optimal control.

Infinite-dimensional & PDEs

- J.-L. Lions. Dunod (1968). Optimal control of PDE-governed systems.
- D. L. Russell. SIAM Review (1978). Controllability and stabilization of linear PDEs.
- J.-L. Lions. SIAM Review (1988). Exact controllability and the HUM method.
- R. F. Curtain, H. Zwart. Springer (1995). Linear systems in infinite-dimensional spaces.

Our approach

Guiding principle.

Rather than relying on abstract existence,
we construct the explicit controls for ERM and quantify their complexity.

Three supervised learning problems demanding increasingly strict constraints:

1. **Data classification** $\longrightarrow$ **Points to cells**
2. **Data interpolation** $\longrightarrow$ **Points to points**
3. **Generalization** $\longrightarrow$ **Cells to points**

CHAPTER 1



Data Classification

A. ÁLVAREZ-LÓPEZ, R. ORIVE-ILLERA, AND E. ZUAZUA

«Cluster-based classification with neural ODEs via control»

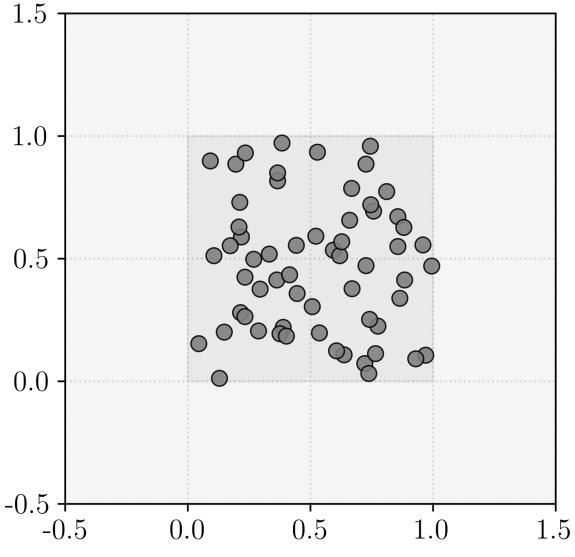
Journal of Machine Learning, 4(2):128--156, 2025



Framework

Worst-case scenario:

Pure noise $\quad x_i \sim U([0,1]^d)$ for all i.



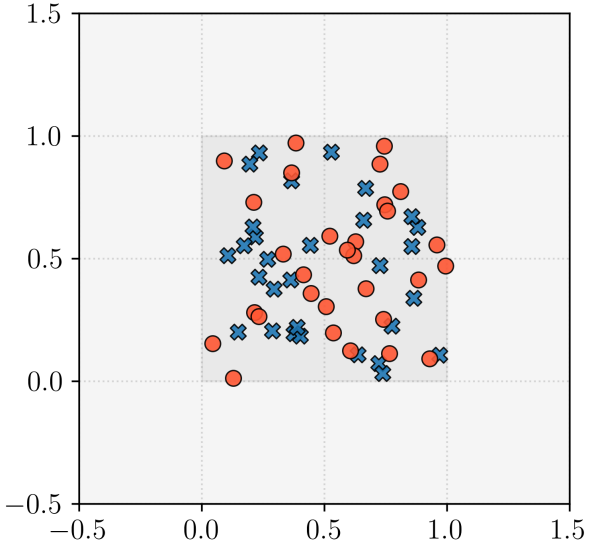
Framework

Worst-case scenario:

Pure noise $\quad x_i \sim U([0,1]^d) \quad$ for all i.

Binary classification:

$$y_i \in \{1, 0\} \quad \text{for all } i \quad \longrightarrow \quad \mathcal{D} = \underbrace{\{(x_i, 1)\}_{i=1}^N}_{R} \cup \underbrace{\{(x_j, 0)\}_{j=1}^N}_{B}.$$



Framework

Worst-case scenario:

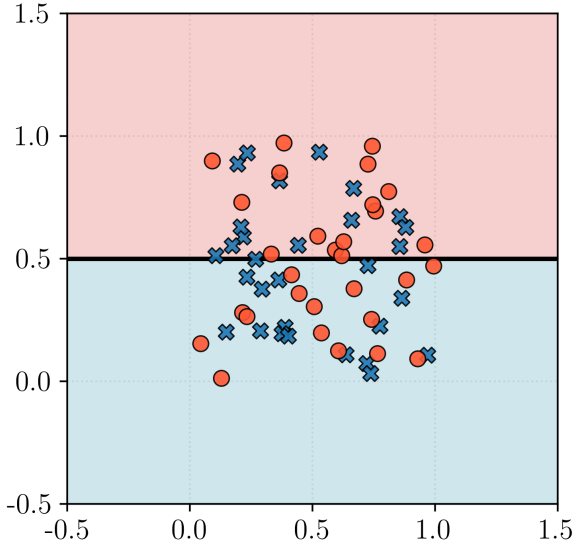
Pure noise $\qquad x_i \sim U([0,1]^d) \quad$ for all i.

Binary classification:

$$y_i \in \{1, 0\} \quad \text{for all } i \qquad \longrightarrow \qquad \mathcal{D} = \underbrace{\{(x_i, 1)\}_{i=1}^N}_{R} \cup \underbrace{\{(x_j, 0)\}_{j=1}^N}_{B}.$$

Target regions:

$$1 \longleftrightarrow \Omega_1 := \{x^{(1)} > 0.5\}, \qquad 0 \longleftrightarrow \Omega_0 := \{x^{(1)} < 0.5\}.$$



Framework

Model: One neuron ($p = 1$) and ReLU activation:

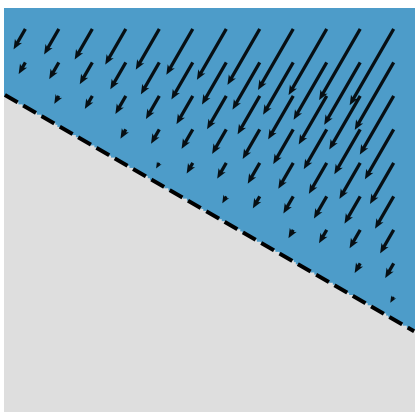
$$\dot{x}(t) = w(t)\,(a(t)^\top x(t) + b(t))_+ \qquad \theta = (w, a, b).$$

Framework

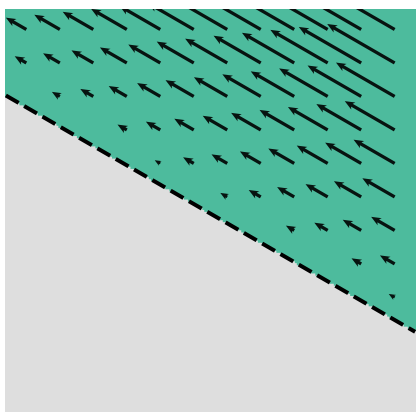
Model: One neuron ($p = 1$) and ReLU activation:

$$\dot{x}(t) = w(t)\,(a(t)^\top x(t) + b(t))_+ \qquad \theta = (w, a, b).$$

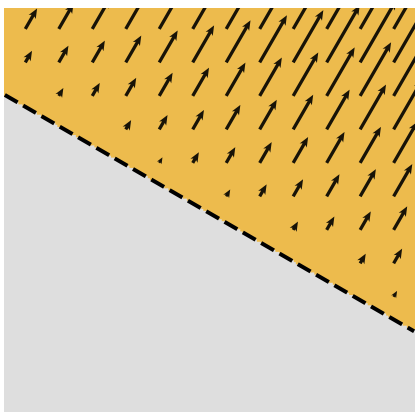
- $a(t), b(t)$ determine the hyperplane in $\mathbb{R}^d$ given by $\mathrm{H}(x) = a(t)^\top x + b(t) = 0$.
- s_+ "activates" the half-space $\mathrm{H}(x) > 0$ and "freezes" $\mathrm{H}(x) \leq 0$.
- $w(t)$ determines the direction of the vector field in the active region.



Compression ($a^\top w < 0$).



One-sided linear shear ($a^\top w = 0$).



Expansion ($a^\top w > 0$).

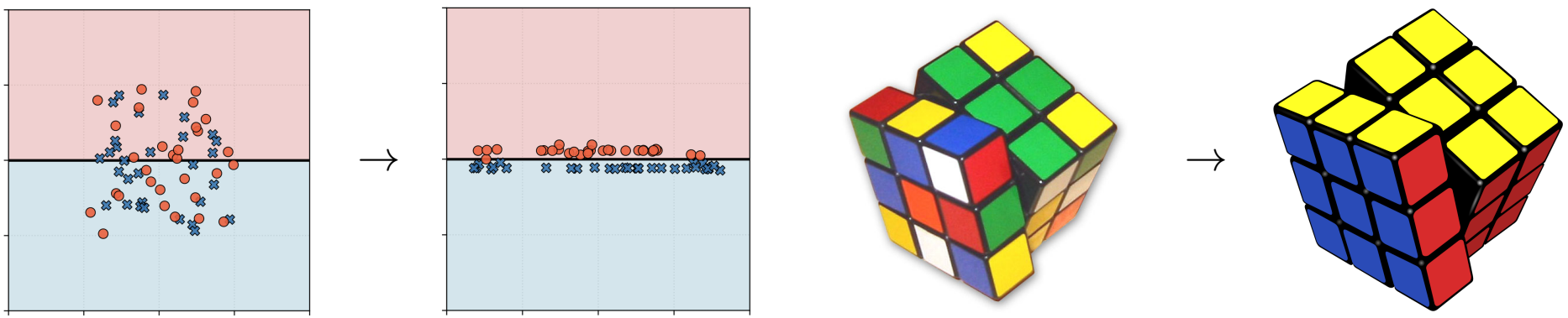
Framework

Model: One neuron ($p = 1$) and ReLU activation:

$$\dot{x}(t) = w(t)\,(a(t)^\top x(t) + b(t))_+ \qquad \theta = (w, a, b).$$

Problem statement

Given $T > 0$, find a piecewise constant control $\theta(\cdot)$ such that $\Phi_T^\theta(x_i) \in \Omega_{y_i}$ for all i.



Prior work

Question 1 (Minimax).
What is the minimum number of switches L that ensures any dataset of a fixed size can be classified?

Prior work uses $L = 6N$ switches via a point-by-point algorithm.

D. Ruiz-Balet and E. Zuazua. SIAM Rev. (2023).

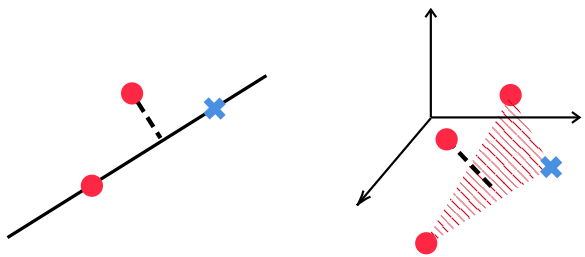
Question 1

Reducing L requires classifying points by clusters.

Theorem (A. Á-L, R. Orive-Illera, E. Zuazua)

Setting $L = 4\lceil N/d \rceil - 1$ switches on the piecewise constant control $\theta : [0, T] \to \mathbb{R}^{2d+1}$ suffices to classify any dataset $\{x_i, x_i\}_{i=1}^N \subset \mathbb{R}^d$ in general position.

A finite set $X \subset \mathbb{R}^d$ is in **general position** if no affine subspace of dimension $m < d$ contains more than $m + 1$ points of X.



Examples in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Question 1

Reducing L requires classifying points by clusters.

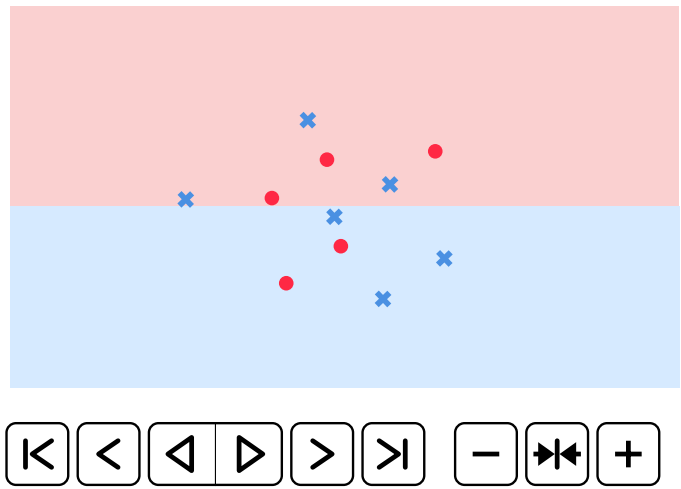
Theorem (A. Á-L, R. Orive-Illera, E. Zuazua)

Setting $L = 4\lceil N/d \rceil - 1$ *switches on the piecewise constant control* $\theta : [0, T] \to \mathbb{R}^{2d+1}$ *suffices to classify any dataset* $\{x_i, x_i\}_{i=1}^N \subset \mathbb{R}^d$ *in general position.*

Proof sketch.

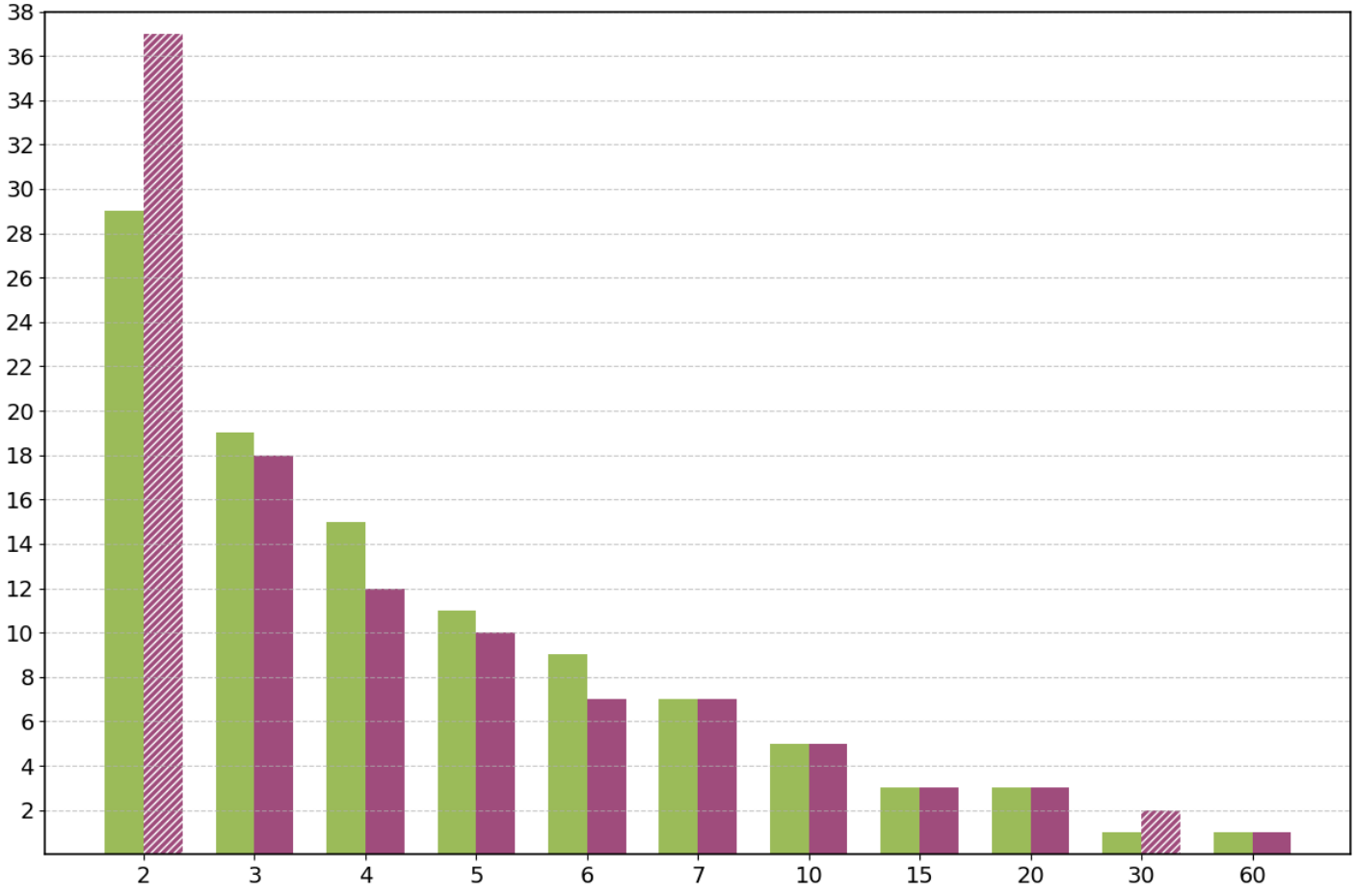
Step 1 (Separation). In general position, any dataset can be partitioned into clusters of d points by pairs of parallel hyperplanes.

Step 2 (Control). Each cluster is classified in at most 2 movements for $\sigma(s) = (s \wedge 1)_+$
$\Longrightarrow$ at most 4 movements for $\sigma(s) = s_+$.



Computational test of optimality

L vs *d*



Our algorithm

Computer training

Question 2

Question 2 (Probabilistic).

How likely is it that a dataset can be classified with at most $L = k$ switches for any $k \geq 0$?

Theorem (A. Á-L, R. Orive-Illera, E. Zuazua)

Assume $x_i, x_i \sim U\left([0,1]^d\right)$ for all i. For any $T > 0$, there exists $\theta = (w, a, b)$ where

$$a \in \mathbb{S}^{d-1} \quad \text{constant}, \qquad (w, b) \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}),$$

that classifies the dataset. Moreover, θ is piecewise constant with L switches such that

$$\mathbb{P}(L = 0) = 1 - \left(1 - \frac{2(N!)^2}{(2N)!}\right)^d \longrightarrow 1 \quad \text{if and only if} \quad d \gtrsim \frac{4^N}{\sqrt{N}}.$$

Question 2

Question 2 (Probabilistic).

How likely is it that a dataset can be classified with at most $L = k$ switches for any $k \geq 0$?

Theorem (A. Á-L, R. Orive-Illera, E. Zuazua)

Assume $x_i, x_i \sim U\left([0,1]^d\right)$ for all i. For any $T > 0$, there exists $\theta = (w, a, b)$ where

$$a \in \mathbb{S}^{d-1} \quad \text{constant}, \qquad (w, b) \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}),$$

that classifies the dataset. Moreover, θ is piecewise constant with L switches such that

$$\mathbb{P}(L = 0) = 1 - \left(1 - \frac{2(N!)^2}{(2N)!}\right)^d \longrightarrow 1 \quad \text{if and only if} \quad d \gtrsim \frac{4^N}{\sqrt{N}}.$$

More generally, for $0 \leq k \leq 2N - 2$:

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

Question 2

Question 2 (Probabilistic).

How likely is it that a dataset can be classified with at most $L = k$ switches for any $k \geq 0$?

Theorem (A. Á-L, R. Orive-Illera, E. Zuazua)

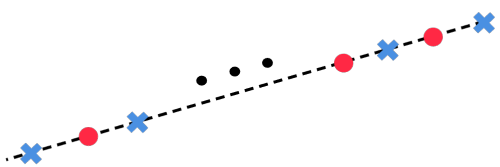
Assume $x_i, x_i \sim U\left([0,1]^d\right)$ for all i. For any $T > 0$, there exists $\theta = (w, a, b)$ where

$$a \in \mathbb{S}^{d-1} \quad \text{constant}, \qquad (w, b) \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}),$$

that classifies the dataset. Moreover, θ is piecewise constant with L switches such that

$$\mathbb{P}(L = 0) = 1 - \left(1 - \frac{2(N!)^2}{(2N)!}\right)^d \longrightarrow 1 \quad \text{if and only if} \quad d \gtrsim \frac{4^N}{\sqrt{N}}.$$

Worst-case ($k = 2N - 2$): Characterized by all points being collinear with alternating labels.



Highly pathological!

CHAPTER 2



Data Interpolation

A. ÁLVAREZ-LÓPEZ, A.H. SLIMANE, AND E. ZUAZUA
«Interplay between depth and width for interpolation in neural ODEs»
Neural Networks, 180:106640, 2024



A. ÁLVAREZ-LÓPEZ, B. GESHKOVSKI, AND D. RUIZ-BALET
«Constructive approximate transport maps with normalizing flows»
Applied Mathematics & Optimization, 92(2):33, 2025



A. ÁLVAREZ-LÓPEZ, L. LIVERANI, AND E. ZUAZUA
«Constructive interpolation and generalization rates for neural ODEs: a control perspective»
Preprint & Submitted, 2026



Problem

Model: We begin by considering

$$\dot{x}(t) = \sum_{j=1}^{p} w_j(t) \left(a_j(t)^\top x(t) + b_j(t)\right)_+, \qquad t \in [0, T].$$

ERM: $\min_{\theta(\cdot)} \sum_{i=1}^{N} \|\Phi_T^\theta(x_i) - y_i\|^2$ $\xrightarrow{\textit{Zero training error}}$ **Simultaneous point controllability**

Problem statement

For any $T > 0$, find a piecewise constant control $\theta = (w_j, a_j, b_j)_{j=1}^p$ such that $\Phi_T^\theta(x_i) = y_i$ for all i, with minimal complexity (number of switches L, width p).

(We assume $y_i \neq y_j$ for all $i \neq j$)

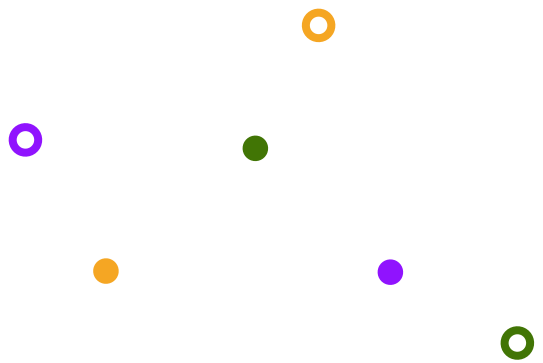
- E. D. Sontag, H. J. Sussmann. Systems Control Lett. (1997). Controllability of continuous-time recurrent neural networks.
- A. Agrachev, Y. Baryshnikov, A. Sarychev. ESAIM COCV (2016). Lie-algebraic controllability criteria for dynamical ensembles.
- C. Cuchiero, M. Larsson, J. Teichmann. SIMODS (2020). Universal interpolation by controlled ODE models.
- A. Agrachev, A. Sarychev. SIAM J. Control Optim. (2020). Simultaneous control of finite and continuous point ensembles.
- D. Ruiz-Balet, E. Zuazua. SIAM Rev. (2023). Constructive control for classification and approximation.
- J. Cheng, Q. Li, T. Lin, Z. Shen. SIAM J. Control Optim. (2025). Links between controllability and approximation for general control families.

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$



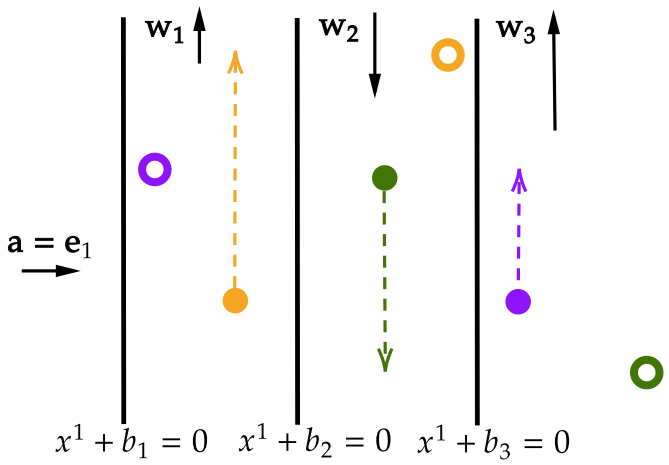
Initial points (solid circles) and target points (empty circles).

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$



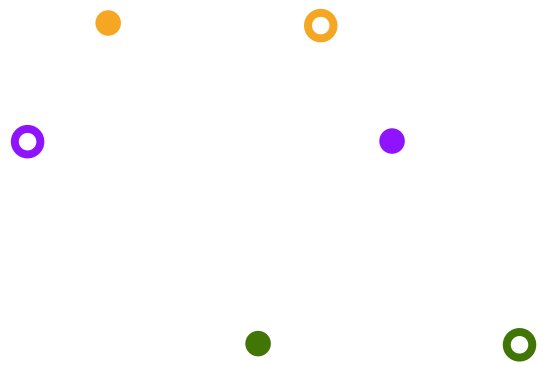
Step 1: Simultaneous control of $d - 1$ coordinates $x^{(2)}, \dots, x^{(d)}$.

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$



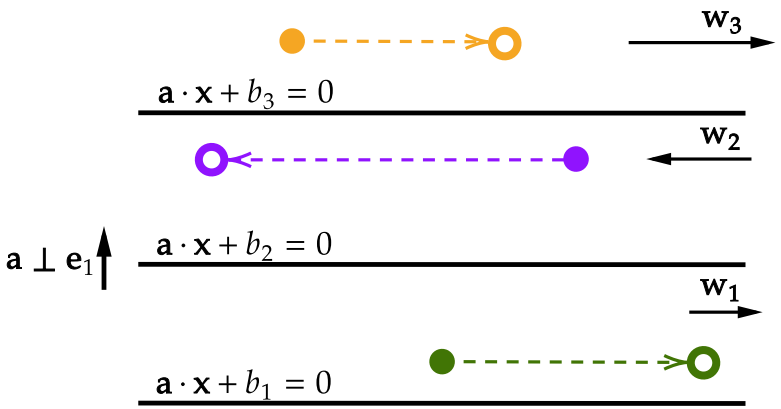
Step 1: Simultaneous control of $d-1$ coordinates $x^{(2)}, \dots, x^{(d)}$.

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$



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

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$



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

Exact control (L vs p)

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

For any $T > 0$ and $p \in \mathbb{N}$, there exists a control $\theta \in L^\infty(0, T; \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p)$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \dots, N$. Moreover, θ is piecewise constant with

$$L = 2\lceil N/p \rceil - 1 \quad \textit{switches}.$$

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

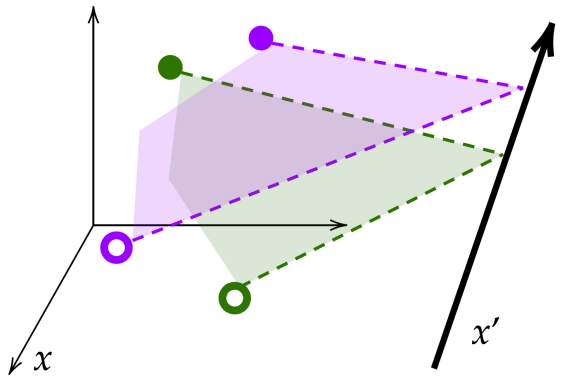
Is it possible to achieve exact control using $L = 0$ switches?

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

High-dimensional setting

Under the assumptions of the previous theorem, if $d > N$ then we can improve it to collinear

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



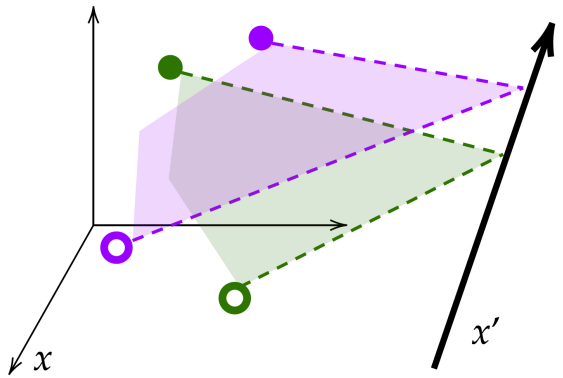
Change of basis so that $x_i^{(1)} = y_i^{(1)}$ for all i.

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

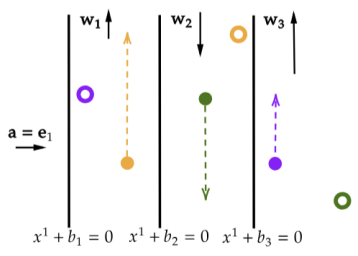
High-dimensional setting

Under the assumptions of the previous theorem, if $d > N$ then we can improve it to collinear

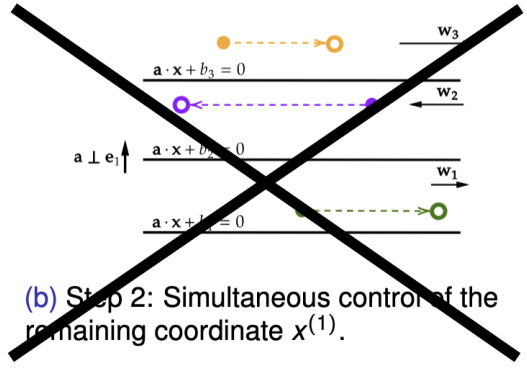
$$L = 2\lceil N/p \rceil - 2 \quad \text{switches.}$$



Change of basis so that $x_i^{(1)} = y_i^{(1)}$ for all i.



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



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

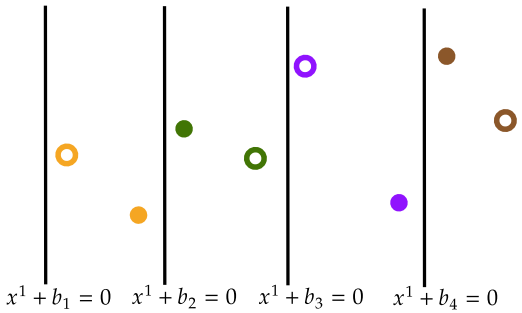
Autonomous system ($L = 0$): Approach II

Probabilistic control via oblique trajectories

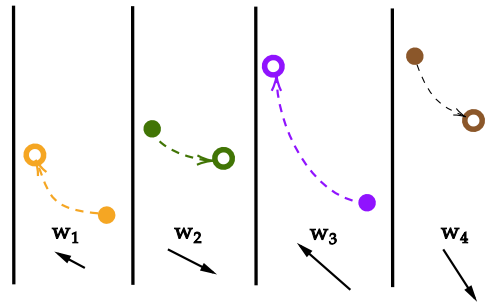
Assume that $x_i, y_i \sim U([0,1]^d)$ for all i. Then, with probability π bounded as

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

there exists $\theta \in \mathbb{R}^{d \times N} \times \mathbb{R}^{N \times d} \times \mathbb{R}^N$ such that $\Phi_T^\theta(x_i) = y_i$ for all $i = 1, \ldots, N$.



(a) Step 1: Separation.



(b) Step 2: Transverse velocity fields.

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

Relaxation to approximate control

For each $p \geq 3$ there exists $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ such that

$$\sup_i \|y_i - \Phi_T^\theta(x_i)\| \lesssim \sqrt{d + \log p}\, p^{-\frac{1}{2}-\frac{1}{d}}.$$

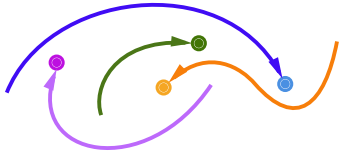
Autonomous system ($L = 0$): Approach III

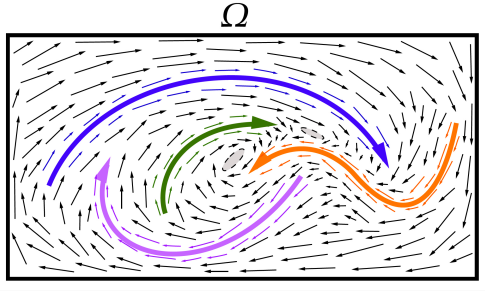
Relaxation to approximate control

For each $p \geq 3$ there exists $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ such that

$$\sup_i \|y_i - \Phi_T^\theta(x_i)\| \lesssim \sqrt{d + \log p}\, p^{-\frac{1}{2}-\frac{1}{d}}.$$







Ingredient 1. Construction of an ideal reference vector field.

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

Relaxation to approximate control

For each $p \geq 3$ there exists $\theta \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p$ such that

$$\sup_i \|y_i - \Phi_T^\theta(x_i)\| \lesssim \sqrt{d + \log p}\, p^{-\frac{1}{2}-\frac{1}{d}}.$$

Ingredient 2. J. M. Klusowski and A. R. Barron. IEEE Trans. Inf. Theory (2018)

Any vector field v with $B_{v,2} = \int \|\omega\|_1^2 |\hat{v}(\omega)| \mathrm{d}\omega < +\infty$ can be uniformly approximated on $[-1,1]^d$ by a shallow ReLU network v_p of width $p \geq 3$ and $\mathrm{Lip}(v_p) \leq 1$ such that:

$$\sup_{x\in[-1,1]^d} \|v(x) - v_p(x)\| \lesssim B_{v,2}\sqrt{d + \log p}\, p^{-\frac{1}{2}-\frac{1}{d}}.$$

Architectural variants

Architecture	**Dynamics** $(\dot{x} =)$	**Cost** (N targets)
Semi-autonomous[1] Constant parameters: $\theta \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^p$	$\sum_{j=1}^{p} w_j \left(a_j^\top x + c_j t + b_j\right)_+$	Width: $p = 2N$ Switches: $L = 0$
2-layer autonomous[1] Constant parameters: $w_k, a_{kj} \in \mathbb{R}^d$, $u_{kj}, b_{kj}, c_k \in \mathbb{R}$	$\sum_{k=1}^{q} w_k \left(\sum_{j=1}^{p} u_{kj} (a_{kj}^\top x + b_{kj})_+ + c_k \right)_+$	Width: $p = 2d, q = N$ Switches: $L = 0$
Linear-in-control[2] Time-varying control: $W(t) \in \mathbb{R}^{d \times d}, b(t) \in \mathbb{R}^d$	$W(t)\, x_+ + b(t)$	Width: $p = d$ Switches: $L = 4N + 2$

[1] **A. Á-L**, L. Liverani, E. Zuazua, Preprint (2026).

[2] **A. Á-L**, B. Geshkovski, D. Ruiz-Balet. Appl. Math. Optim. (2025).

CHAPTER 3



Generalization

A. ÁLVAREZ-LÓPEZ, L. LIVERANI, AND E. ZUAZUA

«Constructive interpolation and generalization rates for neural ODEs: a control perspective»

Preprint & Submitted, 2026



Generalization bounds for continuous-time models

Population risk minimization: $\min_\theta \mathcal{R}(\theta), \qquad \mathcal{R}(\theta) := \int_{\mathcal{X}} \|\Phi_T^\theta(x) - y(x)\|^2 \, \mathrm{d}\mu(x)$

Existing bound. P. Marion. NeurIPS (2023)

For linear-in-control neural ODEs under R_Θ-bounded Lipschitz controls, with probability $1-\delta$:

$$\mathcal{R}(\theta) \lesssim \mathcal{R}_N(\theta) + \underbrace{\sqrt{\frac{(p+1)\log(R_\Theta\, p\, N)}{N}} + \frac{p\sqrt{R_\Theta}}{N^{1/4}} + \sqrt{\frac{\log(1/\delta)}{N}}}_{\sup_{\vartheta\in\Theta} |\mathcal{R}(\vartheta)-\mathcal{R}_N(\vartheta)|}.$$

Tension.

Larger N shrinks the gap but raises the empirical error;
Larger p kills the empirical error but inflates the gap.

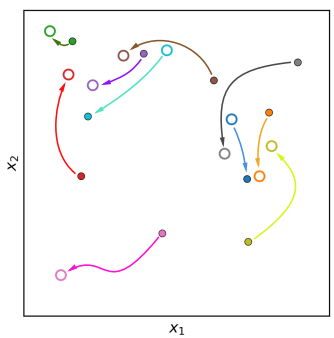
- Verma, Kumar. TMLR (2025).
  Broader nonlinear neural ODE architectures.
- Jia et al. ICLR (2025).
  Feedback mechanisms for generalization.
- Bleistein, Guilloux. ICLR (2024).
  Approximation and generalization of NCDEs.
- Marzouk et al. JMLR (2024).
  Nonparametric statistical theory of distribution learning.

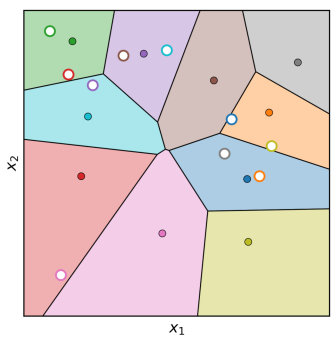
Simultaneous cell controllability

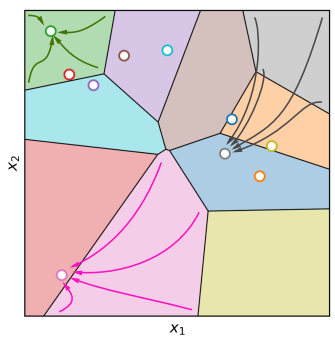
Our goal

Build data-dependent controls θ_N and choose $p = p_N$ so that $\mathcal{R}(\theta_N) \to 0$ at an explicit rate.

Point interpolation is not enough; we need to control regions of the input space.







We say that $\{v_\theta : \theta \in \Theta_p\}$ achieves *simultaneous cell controllability (SCC)* if for any pairwise disjoint compact convex sets $\{A_k\}_{k=1}^K \subset \mathbb{R}^d$, targets $\{r_k\}_{k=1}^K \subset \mathbb{R}^d$, $T > 0$ and $\eta > 0$, there exist a width $p \in \mathbb{N}$ and a control θ such that $\Phi_T^\theta(A_k) \subset B(r_k, \eta)$ for all $k = 1, \ldots, K$.

Semi-autonomous neural ODEs are cell-controllable

Model: semi-autonomous ReLU neural ODE $\dot{x}(t) = \sum_{j=1}^{p} w_j \left(a_j^\top x(t) + c_j\, t + b_j\right)_+$.

Theorem (A. Á-L, L. Liverani, E. Zuazua)

For disjoint compact convex $\{A_k\} \subset [-1,1]^d$ and distinct targets $\{r_k\}$, there exist $\mathfrak{C}, \mathfrak{M}, \eta_0 > 0$ such that, for every $\eta \le \eta_0$ and width

$$p \;\ge\; \mathfrak{C}\, \eta^{-2-\mathfrak{C}}\, (1 - \log \eta)^2,$$

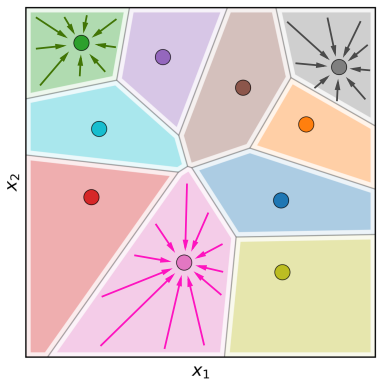
there is a control θ for which $\Phi_T^\theta(A_k) \subset B(r_k, \eta)$ for all k, with $\|\Phi_T^\theta\|_{L^\infty([-1,1]^d)} \le \mathfrak{M}$.

- Width is a *sufficient* (achievability) condition, not sharp. Numerically, smaller widths already match the baselines.
- $\mathfrak{C}$ blows up as the minimal cell separation tends to zero.

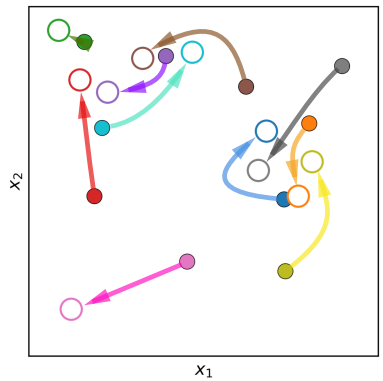
SCC relies on the bias being time-dependent; autonomous fields can still interpolate isolated points, but trajectories cannot cross.

Proof sketch

Build a smooth reference time-dependent vector field, then approximate it.



Step 1. Contract each A_k into a small ball.



Step 2. Transport the balls along disjoint corridors to r_k.

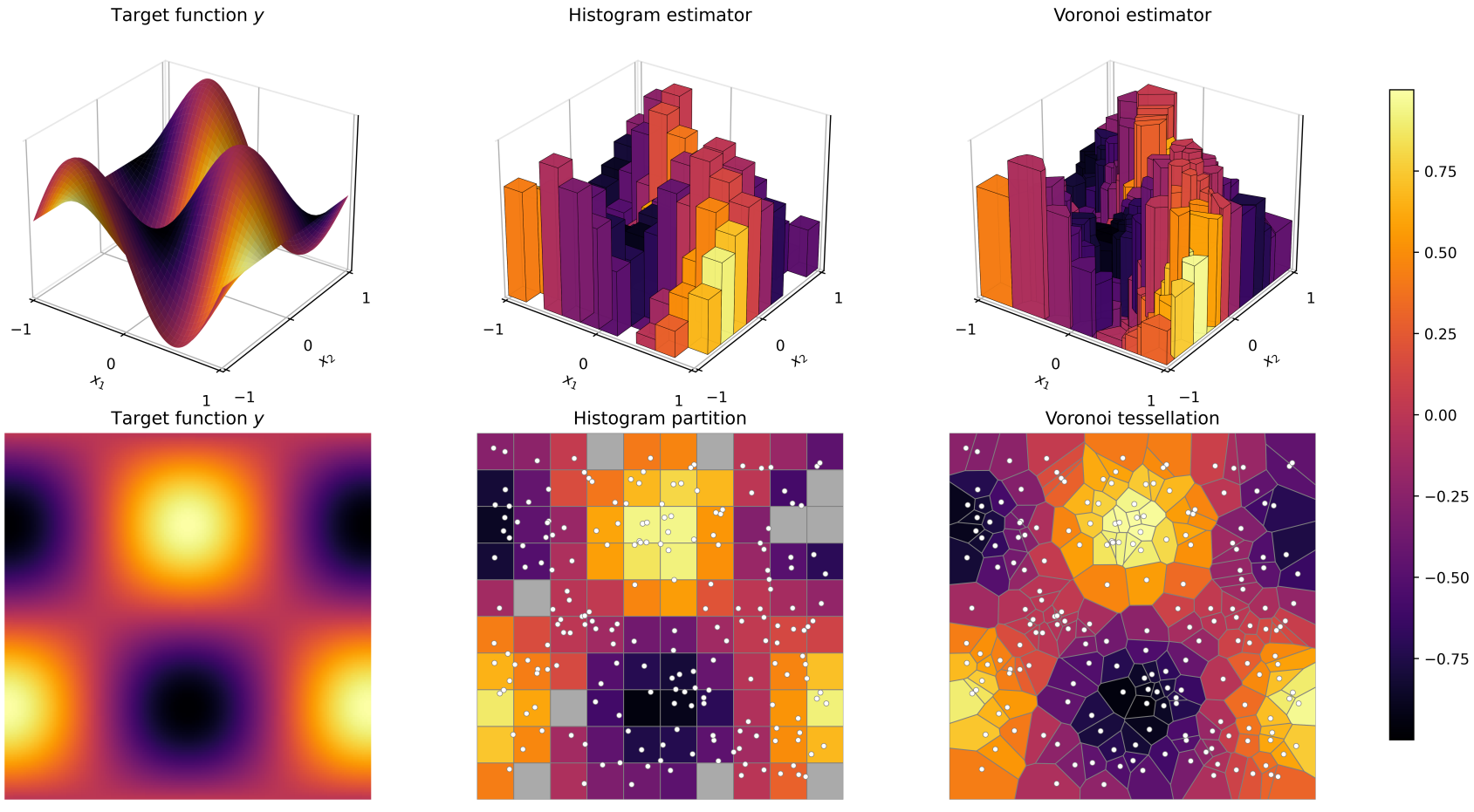
Step 3 (Realization). J. M. Klusowski and A. R. Barron. IEEE Trans. Inf. Theory (2018)

Any vector field v with $B_{v,2} = \int \|\omega\|_1^2 |\hat{v}(\omega)| \mathrm{d}\omega < +\infty$ can be uniformly approximated on $[-1,1]^d$ by a shallow ReLU network v_p of width $p \geq 3$ and $\mathrm{Lip}(v_p) \leq 1$ such that:

$$\sup_{x \in [-1,1]^d} \|v(x) - v_p(x)\| \lesssim B_{v,2} \sqrt{d + \log p}\, p^{-\frac{1}{2} - \frac{1}{d}}.$$

Nonparametric estimators

Cell-wise controllability allows our model to emulate piecewise-constant **nonparametric estimators** by assigning entire cells of the input space to specific labels:



Nonparametric estimators

Cell-wise controllability allows our model to emulate piecewise-constant **nonparametric estimators** by assigning entire cells of the input space to specific labels:

Our population risk bound

Assume $y(\cdot) \in \mathscr{C}^{0,\alpha}([-1,1]^d)$, $x_i \overset{\text{i.i.d.}}{\sim} \mu$ and $p \in \mathbb{N}$ is sufficiently large. Then

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \begin{cases} N^{-\frac{2\alpha}{2\alpha+d}} & \text{for histogram estimators,} \\ \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{d}} & \text{for nearest-neighbor estimators.} \end{cases}$$

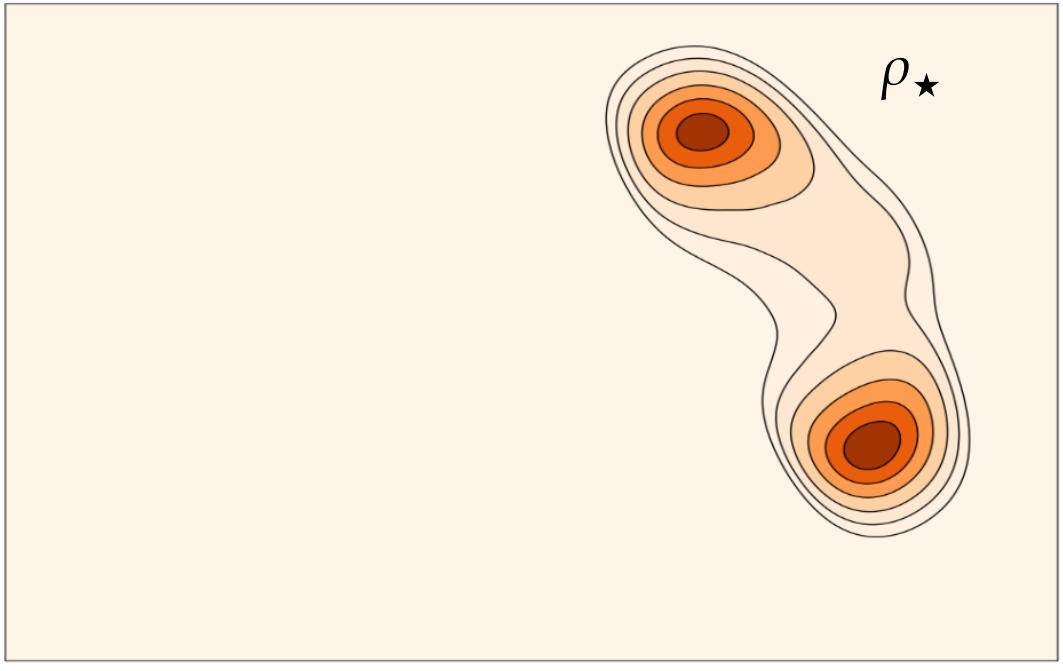
PART II



Generative Modeling

Generative modeling

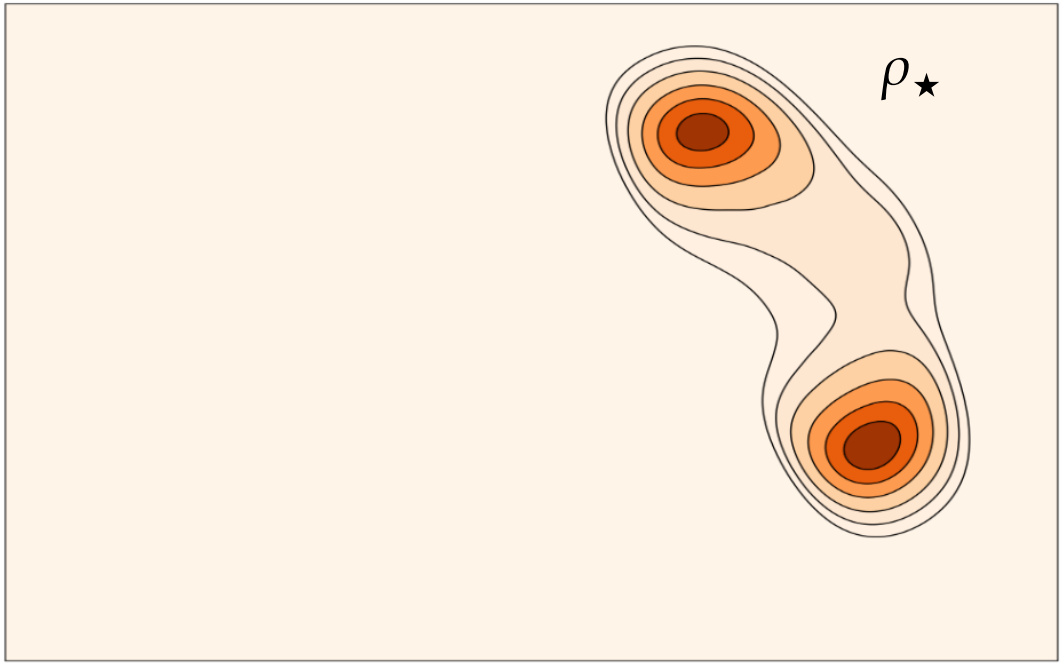
What is flow-based generative modeling?



Generative modeling

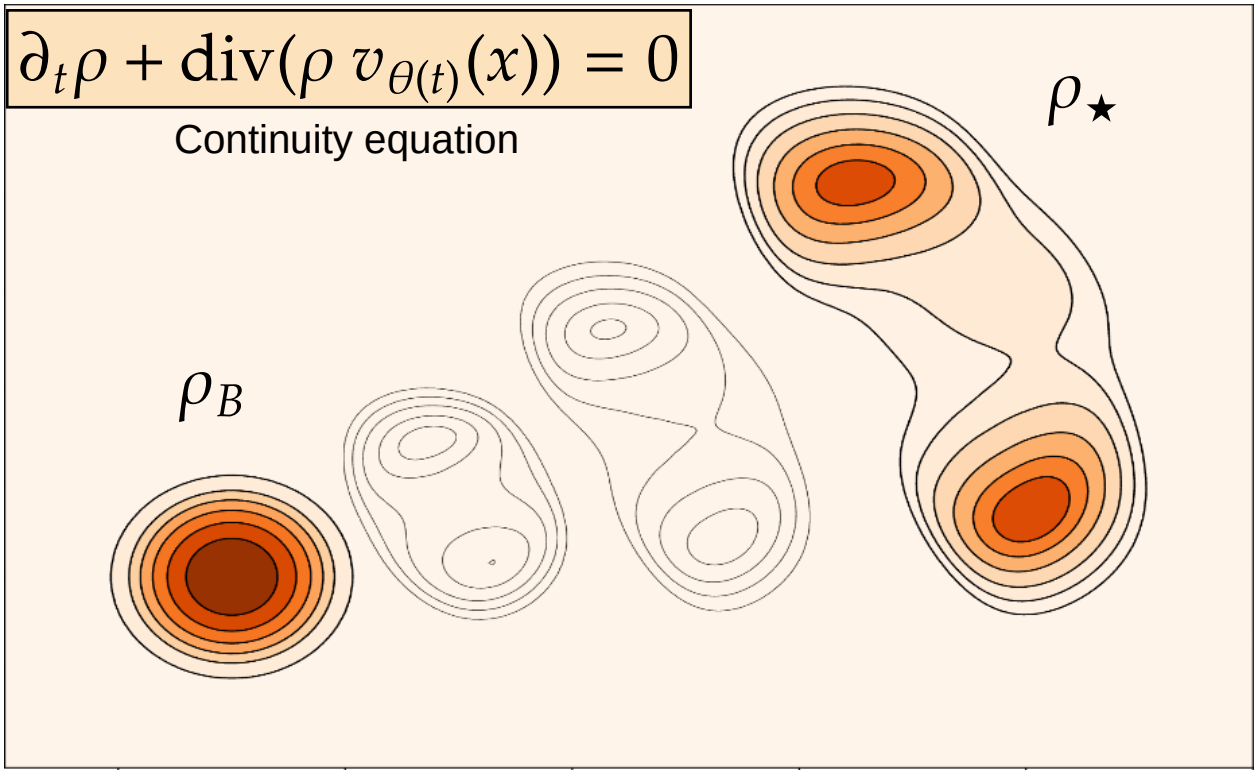
What is flow-based generative modeling?

Generating samples from a target distribution $\mu_\star$ (complex and high-dimensional)
by transforming samples from a simple, base distribution μ_{B} via some learned map $\mathcal{T}$.



Continuous Normalizing Flows

Base $\mu_B \equiv \mathcal{N}(0, \mathrm{Id}) \;\mapsto$ Target $\mu_\star$ accessed through samples $x_1, \ldots, x_N \sim \mu_\star$.



Fix a Lipschitz control family $\{v_\theta(\cdot)\}_\theta$. Find $\theta(\cdot)$ such that $\mu_T^\theta = (\Phi_T^\theta)_{\#}\mu_B \approx \mu_\star$.

A sample $x \sim \mu_B$ then follows the flow: $\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\theta(x) &= v_{\theta(t)}(\Phi_t^\theta(x)) \\ \Phi_0^\theta(x) &= x \end{cases}$

Continuous Normalizing Flows

Many SOTA models share this continuity-equation formulation, differing in their training objectives.

Continuous Normalizing Flows

Many SOTA models share this continuity-equation formulation, differing in their training objectives.

a) Likelihood-trained.[1] Maximize the endpoint (log-)likelihood at the sample, evaluated by solving

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \rho_t^\theta \left(\Phi_t^\theta(x)\right) = -\operatorname{div} v_{\theta(t)} \left(\Phi_t^\theta(x)\right).$$

[1] R. T. Q. Chen et al. NeurIPS (2018); W. Grathwohl et al. ICLR (2019).

Continuous Normalizing Flows

Many SOTA models share this continuity-equation formulation, differing in their training objectives.

a) Likelihood-trained.[1] Maximize the endpoint (log-)likelihood at the sample, evaluated by solving

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \rho_t^\theta \left(\Phi_t^\theta(x)\right) = -\operatorname{div} v_{\theta(t)}\left(\Phi_t^\theta(x)\right).$$

b) Flow matching.[2] Prescribe particle paths X_t and learn the induced field u_t by regression:

$$\min_\theta \mathbb{E}_{t,\, X_t \sim \rho_t} \left[\left\| v_{\theta(t)}(X_t) - u_t(X_t) \right\|^2 \right].$$

[1] R. T. Q. Chen et al. NeurIPS (2018); W. Grathwohl et al. ICLR (2019).
[2] Y. Lipman et al. ICLR (2023); X. Liu et al. ICLR (2023); M. S. Albergo & E. Vanden-Eijnden, ICLR (2023).

Continuous Normalizing Flows

Many SOTA models share this continuity-equation formulation, differing in their training objectives.

a) Likelihood-trained.[1] Maximize the endpoint (log-)likelihood at the sample, evaluated by solving

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \rho_t^\theta \left(\Phi_t^\theta(x)\right) = -\operatorname{div} v_{\theta(t)} \left(\Phi_t^\theta(x)\right).$$

b) Flow matching.[2] Prescribe particle paths X_t and learn the induced field u_t by regression:

$$\min_\theta \mathbb{E}_{t,\, X_t \sim \rho_t} \left[\left\| v_{\theta(t)}(X_t) - u_t(X_t) \right\|^2 \right].$$

c) Score-based diffusion models.[3] A forward SDE transforms data into noise. Learn the score $\nabla \log \rho_t$ to run the reverse-time SDE. Its densities satisfy a Fokker-Planck equation, but the same marginals are generated by the **probability flow ODE** (hence, by a continuity equation).

[1] R. T. Q. Chen et al. NeurIPS (2018); W. Grathwohl et al. ICLR (2019).
[2] Y. Lipman et al. ICLR (2023); X. Liu et al. ICLR (2023); M. S. Albergo & E. Vanden-Eijnden, ICLR (2023).
[3] J. Sohl-Dickstein et al. ICML (2015); J. Ho et al. NeurIPS (2020); Y. Song et al. ICLR (2021).

CHAPTER 4



Controllability for Measure Transport

A. ÁLVAREZ-LÓPEZ, J. CHENG, L. ZHANG, AND Q. LI

«Flow-Based Approximation and Control of Measures»

From neural ODEs to general controlled flows

We now drop the neural ODE ansatz and fix a Lipschitz control family $\mathcal{V} = \{v_\theta : \theta \in \Theta\}$.

$$\text{Let} \quad \mathcal{A}_{\mathcal{V}} = \bigcup_{T>0} \{\Phi_T^\theta : \mathbb{R}^d \to \mathbb{R}^d : \theta \in L^\infty(0,T;\Theta)\}, \qquad \mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}}) = \{\Phi_{\#}\mu_{\mathrm{B}} : \Phi \in \mathcal{A}_{\mathcal{V}}\}.$$

Main question.

Conditions on $(\mu_{\mathrm{B}}, \mathcal{V})$ for density (UAP) of $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ in $\mathcal{P}(\mathbb{R}^d)$?
Relation with simultaneous point controllability?[1]

The answer splits by metric. We focus on Wasserstein and total variation.

- Duprez, Morancey & Rossi. SICON (2019).
  Controllability of continuity equations with localized Lipschitz fields.
- Elamvazhuthi et al. IEEE Control Syst. Lett. (2022).
  Trajectory approximation and measure controllability with neural ODEs.
- Ruiz-Balet & Zuazua. SIAM Rev. (2023).
  $\mathrm{W_p}$-UAP with one-neuron ReLU neural ODEs.
- Ishikawa et al. JMLR (2023).
  Sobolev universality and *UAP* with invertible networks.
- Ruiz-Balet & Zuazua. J. Math. Pures Appl. (2024).
  TV-UAP for the same ansatz.
- Marzouk et al. JMLR (2024).
  Statistical guarantees for distribution learning via neural ODEs.
- Scagliotti & Farinelli. Nonlinear Anal. (2025).
  Approximation of W_2-optimal maps by linear-control neural ODEs.

[1] Our work extends the functional framework of Q. Li, T. Lin & Z. Shen. JEMS (2023) and J. Cheng et al. SICON (2025) to the setting of measures.

Wasserstein: from point controllability to universality

Universal Interpolation Property (UIP): $\mathcal{A}_{\mathcal{V}}$ *can map any* N *distinct points onto any* N *distinct targets. We refer to it as the* π*-UIP if it is only achieved up to some permutation of the targets.*

Theorem (A. Á-L, J. Cheng, L. Zhang, Q. Li)

- *On empirical measures, the* W_p*-UAP (*$1 \leq p < \infty$*) is equivalent to approximate* π*-UIP.*
- *On continuous measures, for a nondecreasing Lipschitz budget* $\Lambda : [1, \infty) \to [0, \infty)$*, the* Λ-W_p*-UAP (*$1 \leq p < \infty$*) holds if and only if the* Λ*-UIP holds (both with* $\mathrm{Lip}(\Phi) \leq \Lambda$*).*

Why the budget? Divergence-free flows have UIP but not the W_p-UAP. $N \to \infty$ forces a blow-up:

$$0 < \mathrm{gap} \leq \mathrm{Lip}(\Phi_N) \underbrace{W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N)}_{\to 0} + \underbrace{W_p\big((\Phi_N)_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N\big) + W_p(\mu_\star^N, \mu_\star)}_{\to 0} \implies \mathrm{Lip}(\Phi_N) \to \infty.$$

[2] $W_p^p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, \mathrm{d}\gamma(x, y)$, with $\Gamma(\mu, \nu)$ the set of couplings of μ, ν.

Wasserstein: from point controllability to universality

Universal Interpolation Property (UIP): $\mathcal{A}_{\mathcal{V}}$ *can map any* N *distinct points onto any* N *distinct targets. We refer to it as the* π*-UIP if it is only achieved up to some permutation of the targets.*

Theorem (A. Á-L, J. Cheng, L. Zhang, Q. Li)

- *On empirical measures, the* W_p*-UAP (*$1 \leq p < \infty$*) is equivalent to approximate* π*-UIP.*
- *On continuous measures, for a nondecreasing Lipschitz budget* $\Lambda : [1, \infty) \to [0, \infty)$*, the* Λ-W_p*-UAP (*$1 \leq p < \infty$*) holds if and only if the* Λ*-UIP holds (both with* $\mathrm{Lip}(\Phi) \leq \Lambda$*).*

Why the budget? Divergence-free flows have UIP but not the W_p-UAP. $N \to \infty$ forces a blow-up:

$$0 < \mathrm{gap} \leq \mathrm{Lip}(\Phi_N) \underbrace{W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N)}_{\to 0} + \underbrace{W_p\big((\Phi_N)_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N\big) + W_p(\mu_\star^N, \mu_\star)}_{\to 0} \implies \mathrm{Lip}(\Phi_N) \to \infty.$$

Alternative path. Via L^p-approximation of transport maps, affine-invariant $\mathcal{V}$ with one nonlinearity (e.g. neural ODEs) is W_p-universal on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with compact support.

[2] $W_p^p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, \mathrm{d}\gamma(x, y)$, with $\Gamma(\mu, \nu)$ the set of couplings of μ, ν.

Total variation: mixing coordinates

- Total variation [3] is rigid: $|\Phi_{\#}\mu - \Phi_{\#}\nu|_{\mathrm{TV}} = |\mu - \nu|_{\mathrm{TV}}$ for any flow Φ, so the UIP is vacuous.
- Flows fixing the origin have TV-universality[4] but fail the UIP.

Theorem (A. Á-L, J. Cheng, L. Zhang, Q. Li)

Let $d \geq 2$. Suppose $\mathcal{A}_{\mathcal{V}}$ is closed under inversion and approximates, on every compact,

(i) *marginals: coordinate-wise nondecreasing maps $\phi(x^{(k)})$ in $W^{1,1}$;*

(ii) *shears: $x^{(k)} \mapsto x^{(k)} + \psi(x^{(\ell)})$ in L^1, with ψ nondecreasing.*

Then the system has the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

[3] $|\mu - \nu|_{\mathrm{TV}} := \sup_A |\mu(A) - \nu(A)| = \frac{1}{2}\|\rho_\mu - \rho_\nu\|_{L^1}$ for $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with densities ρ_μ, ρ_ν.

[4] B. Dacorogna and J. Moser. Ann. Inst. H. Poincaré (1990).

Total variation: mixing coordinates

- Total variation is rigid: $|\Phi_{\#}\mu - \Phi_{\#}\nu|_{\mathrm{TV}} = |\mu - \nu|_{\mathrm{TV}}$ for any flow Φ, so the UIP is vacuous.
- Flows fixing the origin have TV-universality but fail the UIP.

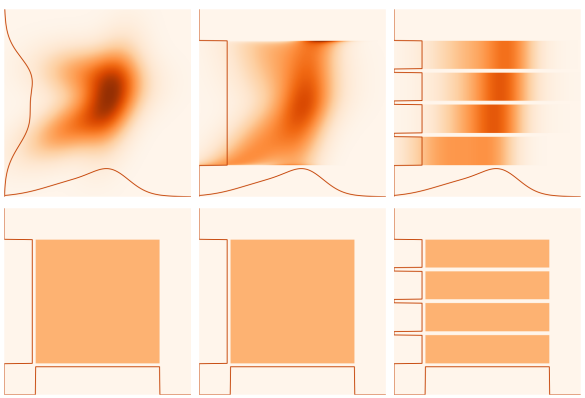
Theorem (A. Á-L, J. Cheng, L. Zhang, Q. Li)

Let $d \geq 2$. Suppose $\mathcal{A}_{\mathcal{V}}$ is closed under inversion and approximates, on every compact,

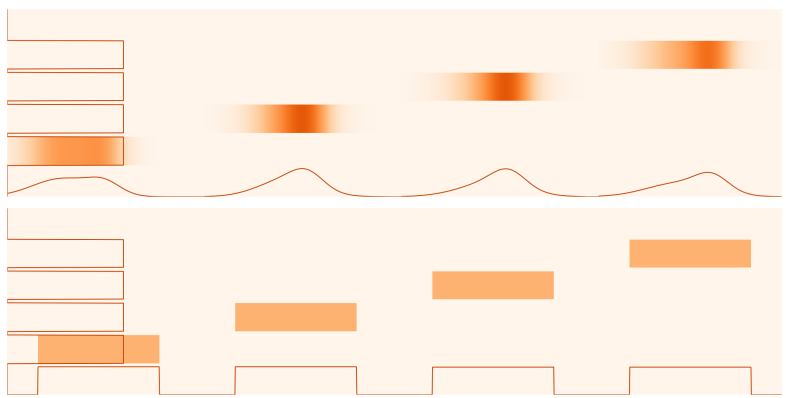
(i) *marginals: coordinate-wise nondecreasing maps $\phi(x^{(k)})$ in $W^{1,1}$;*

(ii) *shears: $x^{(k)} \mapsto x^{(k)} + \psi(x^{(\ell)})$ in L^1, with ψ nondecreasing.*

Then the system has the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.



Marginal matching and discretization into layers of equal mass.



Shearing: separating supports for independent 1D rearrangement.

This strategy is inspired by D. Ruiz-Balet and E. Zuazua. J. Math. Pures Appl. (2024).

Total variation: mixing coordinates

- Total variation is rigid: $|\Phi_{\#}\mu - \Phi_{\#}\nu|_{\mathrm{TV}} = |\mu - \nu|_{\mathrm{TV}}$ for any flow Φ, so the UIP is vacuous.
- Flows fixing the origin have TV-universality but fail the UIP.

Theorem (A. Á-L, J. Cheng, L. Zhang, Q. Li)

Let $d \geq 2$. Suppose $\mathcal{A}_{\mathcal{V}}$ is closed under inversion and approximates, on every compact,

(i) *marginals: coordinate-wise nondecreasing maps $\phi(x^{(k)})$ in $W^{1,1}$;*

(ii) *shears: $x^{(k)} \mapsto x^{(k)} + \psi(x^{(\ell)})$ in L^1, with ψ nondecreasing.*

Then the system has the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

Corollary.

Affine-invariant $\mathcal{V}$ with one nonpolynomial $\mathscr{C}^1$/ReLU component has the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

Affine invariance: $v \in \mathcal{V} \Longrightarrow (x \mapsto Wv(Ax + b)) \in \mathcal{V}$ for all $W, A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$.

CHAPTER 5



Continuous Normalizing Flows

A. ÁLVAREZ-LÓPEZ, B. GESHKOVSKI, AND D. RUIZ-BALET
«Constructive approximate transport maps with normalizing flows»
Applied Mathematics & Optimization, 92(2):33, 2025



Problem

The **Kullback-Leibler divergence** between two probability measures μ, ν with densities ρ_μ, ρ_ν is

$$\mathrm{KL}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\mu}{\rho_\nu}\right)\rho_\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

The **optimization problem** can be reformulated (at the population level) as $\min_\theta \ \mathrm{KL}(\mu_\star\|\mu_T^\theta)$.

Why?

$$\max_\theta \rho_T^\theta(x_1,\dots,x_N) \overset{\text{i.i.d.}}{=} \max_\theta \prod_{i=1}^N \rho_T^\theta(x_i) \iff \max_\theta \frac{1}{N}\sum_{i=1}^N \log \rho_T^\theta(x_i)$$

$$\downarrow \ (LLN)$$

$$\min_\theta \mathrm{KL}(\mu_\star\|\mu_T^\theta) = \min_\theta \Big(-\mathbb{E}_{\mu_\star}[\log \rho_T^\theta] + \underbrace{\mathbb{E}_{\mu_\star}[\log \rho_\star]}_{\text{indep. of } \theta} \Big) = \max_\theta \mathbb{E}_{\mu_\star}[\log \rho_T^\theta]$$

Problem

The **Kullback-Leibler divergence** between two probability measures μ, ν with densities ρ_μ, ρ_ν is

$$\mathrm{KL}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\mu}{\rho_\nu}\right)\rho_\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$ and $T > 0$, find a control $\theta = (\theta(t))_{t\in[0,T]}$ such that $\mathrm{KL}(\mu_\star\|\mu_T^\theta) < \varepsilon$, subject to:

$$\begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\left(\rho_t^\theta(x)\, v_{\theta(t)}(x)\right) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x), \end{cases}$$

where we take $v_{\theta(t)}(x) = w(t)(a(t)^\top x + b(t))_+$.

Problem

The **Kullback-Leibler divergence** between two probability measures μ, ν with densities ρ_μ, ρ_ν is

$$\mathrm{KL}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\mu}{\rho_\nu}\right)\rho_\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$ and $T > 0$, find a control $\theta = (\theta(t))_{t\in[0,T]}$ such that $\mathrm{KL}(\mu_\star\|\mu_T^\theta) < \varepsilon$, subject to:

$$\begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\left(\rho_t^\theta(x)\, v_{\theta(t)}(x)\right) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x), \end{cases}$$

where we take $v_{\theta(t)}(x) = w(t)(a(t)^\top x + b(t))_+$.

- **Difficulty 1: KL is not a distance.** Non-symmetric and no triangle inequality.

Problem

The **Kullback-Leibler divergence** between two probability measures μ, ν with densities ρ_μ, ρ_ν is

$$\mathrm{KL}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\mu}{\rho_\nu}\right)\rho_\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$ and $T > 0$, find a control $\theta = (\theta(t))_{t\in[0,T]}$ such that $\mathrm{KL}(\mu_\star\|\mu_T^\theta) < \varepsilon$, subject to:

$$\begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\left(\rho_t^\theta(x)\, v_{\theta(t)}(x)\right) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x), \end{cases}$$

where we take $v_{\theta(t)}(x) = w(t)(a(t)^\top x + b(t))_+$.

- **Difficulty 1: KL is not a distance.** Non-symmetric and no triangle inequality.
- **Difficulty 2: Blow-up** whenever:
  - **a)** $\mu_\star \not\ll \mu_T^\theta$ (in particular, if $\exists A \subset \mathbb{R}^d$ with $\mu_\star(A) > 0$ and $\mu_T^\theta(A) = 0$).
  - **b)** $\int_{\mathbb{R}^d} \rho_\star \left[\log\left(\rho_\star/\rho_T^\theta\right)\right]_+ = +\infty$.

Problem

The **Kullback-Leibler divergence** between two probability measures μ, ν with densities ρ_μ, ρ_ν is

$$\mathrm{KL}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\mu}{\rho_\nu}\right)\rho_\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$ and $T > 0$, find a control $\theta = (\theta(t))_{t\in[0,T]}$ such that $\mathrm{KL}(\mu_\star\|\mu_T^\theta) < \varepsilon$, subject to:

$$\begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\left(\rho_t^\theta(x)\, v_{\theta(t)}(x)\right) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x), \end{cases}$$

where we take $v_{\theta(t)}(x) = w(t)(a(t)^\top x + b(t))_+$.

$\Longrightarrow$ *We need control over the support of μ_T^θ and over the relative tail decay.*

Main result

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$, and let $\rho_\star$ be a probability density with $\rho_\star \log \rho_\star \in L^1(\mathbb{R}^d)$ and

$$\exists M, \sigma_\bullet > 0 : \qquad \rho_\star(x) \leq \rho_\bullet(x) := \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \qquad (\|x\| \geq M).$$

Then for all $T > 0$ and $\varepsilon > 0$ there exists a piecewise-constant control θ such that

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int_{\mathbb{R}^d} \rho_\star \log \frac{\rho_\star}{\rho_T^\theta} < \varepsilon.$$

Main result

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$, and let $\rho_\star$ be a probability density with $\rho_\star \log \rho_\star \in L^1(\mathbb{R}^d)$ and

$$\exists M, \sigma_\bullet > 0 : \qquad \rho_\star(x) \le \rho_\bullet(x) := \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \qquad (\|x\| \ge M).$$

Then for all $T > 0$ and $\varepsilon > 0$ there exists a piecewise-constant control θ such that

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int_{\mathbb{R}^d} \rho_\star \log \frac{\rho_\star}{\rho_T^\theta} < \varepsilon.$$

Number of switches (L): We can choose

$$L(\varepsilon) \;\le\; (d+10) \underbrace{\left\lceil \frac{2R_\varepsilon}{h_\varepsilon} \right\rceil^d}_{\text{Number of cells}} + 2d$$

where $[-R_\varepsilon, R_\varepsilon]^d$ contains $\ge 1-\varepsilon$ mass of both μ_{B} and $\mu_\star$, and h_ε is the mesh size of a uniform grid on which the Riemann approximation has TV-error $\le \varepsilon$.

Proof sketch: (i) + (ii) + (iii)

(i) TV-approximation[1] on $[0, T/2]$:

$$\left|\mu^{\theta}_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^{\theta}_{T/2} - \rho_\star\|_{L^1} < \varepsilon \qquad \text{(constructive } \theta\text{)}.$$

[1] D. Ruiz-Balet and E. Zuazua. J. Math. Pures Appl. (2024).

Proof sketch: (i) + (ii) + (iii)

(i) TV-approximation[1] on $[0, T/2]$:

$$\left|\mu^\theta_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^\theta_{T/2} - \rho_\star\|_{L^1} < \varepsilon \qquad \text{(constructive } \theta\text{)}.$$

(ii) Classical Pinsker inequality: $|\mu_\star - \mu^\theta_T|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} \cdot \mathrm{KL}(\mu_\star \,\|\, \mu^\theta_T)}$.

[1] D. Ruiz-Balet and E. Zuazua. J. Math. Pures Appl. (2024).

Proof sketch: (i) + (ii) + (iii)

(i) TV-approximation[1] on $[0, T/2]$:

$$\left|\mu^\theta_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^\theta_{T/2} - \rho_\star\|_{L^1} < \varepsilon \qquad \text{(constructive } \theta\text{)}.$$

(ii) Classical Pinsker inequality: $|\mu_\star - \mu^\theta_T|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\cdot \mathrm{KL}(\mu_\star \,\|\, \mu^\theta_T)}$.

Reverse Pinsker inequality[2] **lifts TV $\to$ KL:**

$$\mathrm{KL}(\mu_\star\|\mu^\theta_T) \le \frac{\log\left\|\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T}\right\|_{L^\infty}}{1 - \left\|\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T}\right\|^{-1}_{L^\infty}} \cdot |\mu_\star - \mu^\theta_T|_{\mathrm{TV}} = \frac{1}{2}\cdot \frac{\log\left\|\frac{\rho_\star}{\rho^\theta_T}\right\|_{L^\infty}}{1 - \left\|\frac{\rho_\star}{\rho^\theta_T}\right\|^{-1}_{L^\infty}} \cdot \|\rho_\star - \rho^\theta_T\|_{L^1}$$

[1] D. Ruiz-Balet and E. Zuazua. J. Math. Pures Appl. (2024).

[2] I. Sason. *On Reverse Pinsker Inequalities* (2015).

Proof sketch: (i) + (ii) + (iii)

(i) TV-approximation[1] on $[0, T/2]$:

$$\left|\mu^\theta_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^\theta_{T/2} - \rho_\star\|_{L^1} < \varepsilon \qquad \text{(constructive } \theta\text{)}.$$

(ii) Classical Pinsker inequality: $|\mu_\star - \mu^\theta_T|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\cdot \mathrm{KL}(\mu_\star \,\|\, \mu^\theta_T)}$.

Reverse Pinsker inequality[2] **lifts TV $\to$ KL:**

$$\mathrm{KL}(\mu_\star\|\mu^\theta_T) \le \frac{\log\left\|\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T}\right\|_{L^\infty}}{1 - \left\|\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T}\right\|^{-1}_{L^\infty}} \cdot |\mu_\star - \mu^\theta_T|_{\mathrm{TV}} = \frac{1}{2}\cdot \underbrace{\frac{\log\left\|\frac{\rho_\star}{\rho^\theta_T}\right\|_{L^\infty}}{1 - \left\|\frac{\rho_\star}{\rho^\theta_T}\right\|^{-1}_{L^\infty}}}_{\text{Finite thanks to (iii)}} \cdot \|\rho_\star - \rho^\theta_T\|_{L^1}$$

(iii) Positivity & Tail control on $[T/2, T]$:

$$\rho^\theta_T > 0 \quad \text{and} \quad \lim_{\|x\|\to\infty} \frac{\rho_\star(x)}{\rho^\theta_T(x)} < \infty.$$

[1] D. Ruiz-Balet and E. Zuazua. J. Math. Pures Appl. (2024).

[2] I. Sason. *On Reverse Pinsker Inequalities* (2015).

Toolbox

The solution of the continuity equation

$$\partial_t \rho_t^\theta(x) + \operatorname{div}\left(\rho_t^\theta(x)\, w(a^\top x + b)_+\right) = 0, \qquad \rho_0 = \rho_{\mathrm{B}}$$

has a **closed-form** (by the method of characteristics):

$$\rho_t^\theta(x) = \begin{cases} \rho_{\mathrm{B}}(x), & x \in H^- := \{a^\top x + b \le 0\}, \\ e^{-t\, w^\top a}\, \rho_{\mathrm{B}}\left(\mathrm{A}_t(x)\right), & x \in H^+ := \{a^\top x + b > 0\}, \end{cases}$$

where the transformation $\mathrm{A}_t(x)$ is **affine** in x:

- **General case** ($w^\top a \neq 0$):

$$\mathrm{A}_t(x) = e^{-t\, wa^\top} x - \frac{b}{w^\top a}\left(1 - e^{-t w^\top a}\right) w.$$

- **Degenerate case** ($w^\top a = 0$):

$$\mathrm{A}_t(x) = (I - t\, wa^\top)x - t\, b\, w.$$

Further extensions

1. Reverse KL

Assume that $\rho_\star > 0$, $\rho_\star \log \rho_\star \in L^1$ and

$$\exists \sigma_\bullet > 0: \quad \rho_\star(x) \gtrsim e^{-\frac{\|x\|^2}{2\sigma_\bullet^2}} \quad \text{for large } \|x\|.$$

Then, for any $\varepsilon > 0$, we build θ for which

$$\mathrm{KL}(\mu_T^\theta \,\|\, \mu_\star) < \varepsilon.$$

2. Beyond KL

$\mathrm{KL}(\mu_\star \,\|\, \mu_T^\theta) < \varepsilon$ yields approximation in:

- any Csiszár f-divergence with $\sup \frac{f(t)}{t \log t} < \infty$.
- any Rényi divergence D_λ with $\lambda \in (0, 1]$.

This applies to TV and Hellinger distances.

For $\mu \ll \nu$: $D_f(\mu\|\nu) = \int f(\mathrm{d}\mu/\mathrm{d}\nu)\,\mathrm{d}\nu$, and $D_\lambda(\mu\|\nu) = (\lambda - 1)^{-1} \log \int (\mathrm{d}\mu/\mathrm{d}\nu)^\lambda \,\mathrm{d}\nu$, with f convex, $f(1) = 0$.

3. Beyond Gaussians

Let $\rho_{\mathrm{B}} = e^{-U_{\mathrm{B}}}$ and $\rho_\star = e^{-U_\star}$ with full supp.

- Linear growth: $U_\star(x) \gtrsim \|x\|$ for large $\|x\|$.
- Compatibility: For any $A \in \mathbb{R}^{d\times d}$ with $\sigma(A) \subset (0, \infty)$ there is $\lambda_A > 0$ s.t.

$$\frac{\mathrm{conv}\, U_\star(\lambda_A x)}{U_{\mathrm{B}}(Ax)} \to +\infty \quad \text{as } \|x\| \to \infty.$$

4. Beyond Lipschitz

To interpolate different tail decays in $d = 1$ such as $e^{-|x|^p} \to e^{-|x|^q}$ for any $p, q > 0$, use:

$$\sigma(x) = x \log x \cdot \mathbf{1}_{\{x>1\}} \quad \text{(locally Lipschitz).}$$

Characteristics grow exponentially but *do not blow up in finite time*.

PART III



Applications

CHAPTER 6



Mean-field Landscape of Transformers

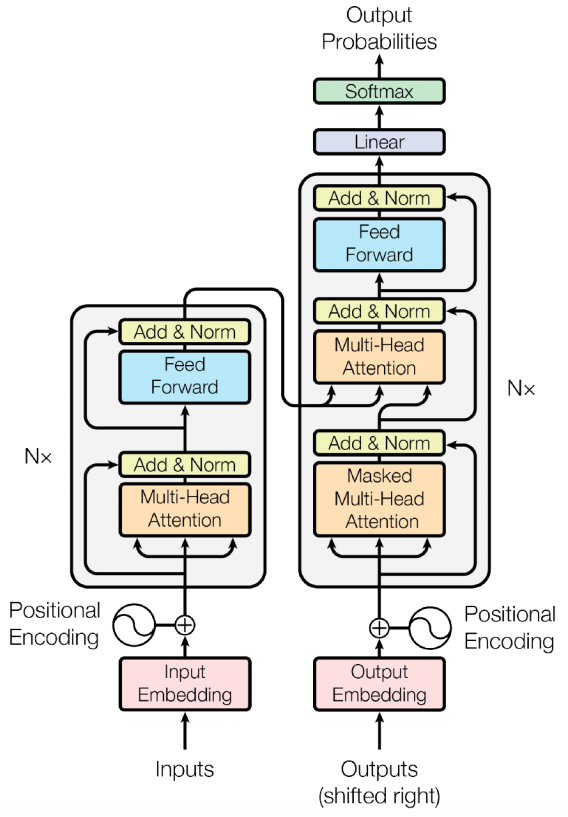
A. ÁLVAREZ-LÓPEZ, B. GESHKOVSKI, AND D. RUIZ-BALET

«Perceptrons and localization of attention's mean-field landscape»

International Conference on Machine Learning, 2026

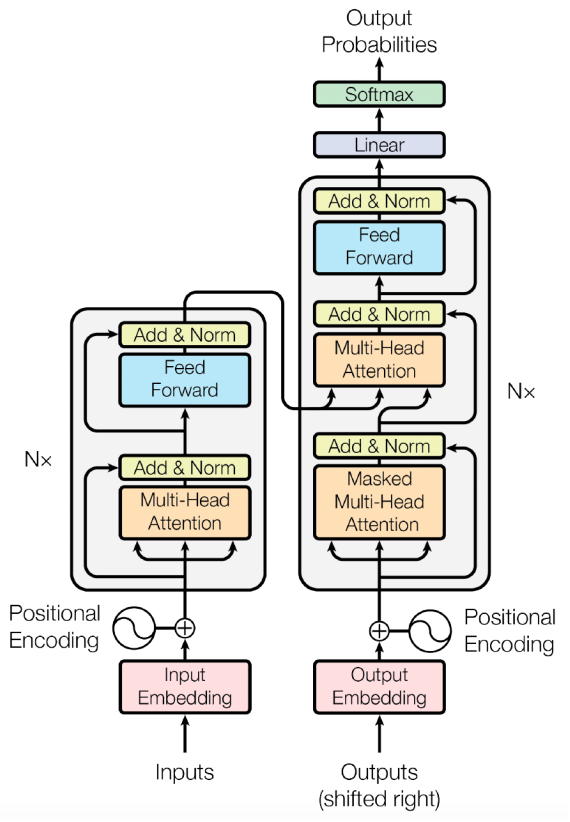


Motivation



A. Vaswani et al. *Attention Is All You Need* (2017)

Motivation



A. Vaswani et al. *Attention Is All You Need* (2017)

Core mechanism:

Maps an input sequence of n tokens $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ to an output sequence $(y_1, \ldots, y_n) \in (\mathbb{R}^d)^n$ via stacked blocks of[5]

(i) Attention

(ii) Feed-forward networks

(iii) Layer normalization

Today, they are the backbone of modern AI:

Language: GPT, Claude, Gemini, LLaMA...

Vision: ViT, DiT, SAM

Science: AlphaFold, ESM

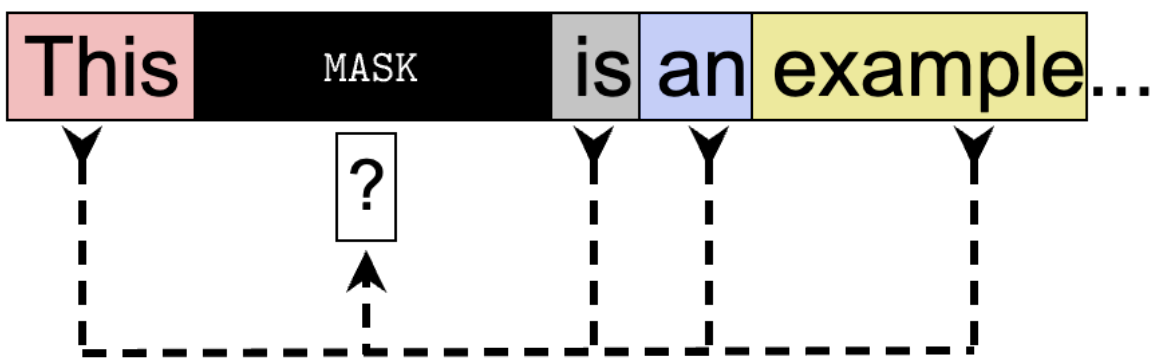
. . . yet the internal dynamics are still poorly understood.

[6] T. Lin et al. *A Survey of Transformers.* AI Open (2022). Comprehensive overview of variants.

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[7] **encoder**[8] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell,\, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell,\, x_k^\ell\rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell,\, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell,\, x_k^\ell\rangle}} x_j^\ell \right\|}}_{=: y_i^\ell \quad \text{(attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma\left(A^\ell y_i^\ell + b^\ell\right)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma\left(A^\ell y_i^\ell + b^\ell\right)\right\|}}_{=: x_i^{\ell+1} \quad \text{(perceptron)}}$$



[7] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[8] BERT model J. Devlin et al., NAACL (2019) (see [PH22, Algorithm 9] as well) and vision Transformers A. Dosovitskiy et al., ICLR (2021).

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[7] **encoder**[8] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell,\, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell,\, x_k^\ell\rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell,\, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell,\, x_k^\ell\rangle}} x_j^\ell \right\|}}_{=:\, y_i^\ell \quad \text{(attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma\left(A^\ell y_i^\ell + b^\ell\right)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma\left(A^\ell y_i^\ell + b^\ell\right)\right\|}}_{=:\, x_i^{\ell+1} \quad \text{(perceptron)}}$$

$\xrightarrow[\text{as } \Delta t \to 0]{\text{first-order}}$ **Interacting particle ODE on the sphere**[9]

$$\dot{x}_i(t) = \mathbf{P}^\perp_{x_i(t)} \left(V(t) \sum_{j=1}^n \frac{e^{\langle B(t)x_i(t),\, x_j(t)}}{\sum_{k=1}^n e^{\langle B(t)x_i(t),\, x_k(t)}} x_j(t) \;+\; W(t)\, \sigma(A(t)x_i(t) + b(t)) \right)$$

where $\mathbf{P}^\perp_x = I - xx^\top$ is the orthogonal projection onto $\mathrm{T}_x\mathbb{S}^{d-1}$.

layers $\leftrightarrow$ time | tokens $\leftrightarrow$ particles | PostLayerNorm $\leftrightarrow$ sphere

[7] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[8] BERT model J. Devlin et al., NAACL (2019) (see [PH22, Algorithm 9] as well) and vision Transformers A. Dosovitskiy et al., ICLR (2021).

[9] Y. Lu et al., arXiv:1906.02762 (2019); S. Dutta et al., NeurIPS (2021). First to interpret stacked Transformer layers as a numerical ODE solver (Euler's method + Lie-Trotter splitting) and as an interacting particle system.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(V(t) \sum_{j=1}^{n} \frac{e^{\langle B(t)x_i(t),\, x_j(t)\rangle}}{\sum_k e^{\langle B(t)x_i(t),\, x_k(t)\rangle}} x_j(t) \;+\; W(t)\,\sigma(A(t)x_i(t) + b(t)) \right)$$

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \sum_{j=1}^{n} \frac{e^{\beta\, x_i(t)^\top x_j(t)}}{\sum_k e^{\beta\, x_i(t)^\top x_k(t)}}\, x_j(t) \;+\; W\, \sigma(A x_i(t)) \right)$$

1. Simplifications

Attention: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).

Perceptron: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

Attention: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).

Perceptron: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d \times d})^2$ and zero bias $b = 0$.

2. Gradient structure

Unnormalized attention[10]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$

Conservative perceptron: $W\sigma(Ax) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \operatorname{diag}(\omega_i)$).

[10] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

Attention: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).

Perceptron: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

Unnormalized attention[10]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$

Conservative perceptron: $W\sigma(Ax) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \operatorname{diag}(\omega_i)$).

Example: $\mathrm{V}_\theta(x) = \sum_i \omega_i (a_i^\top x)_+^2$ for the ReLU perceptron.

[10] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

Attention: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).

Perceptron: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

Unnormalized attention[10]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$

Conservative perceptron: $W\sigma(Ax) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \operatorname{diag}(\omega_i)$).

Note: Our main results will extend to B symmetric invertible, $b \neq 0$, and softmax attention.

[10] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\partial_t \mu_t \;+\; \mathrm{div}\left(\mu_t\, \mathbf{P}_x^{\perp}\left[\pm \int_{\mathbb{S}^{d-1}} e^{\beta\, x^\top y}\, y \,\mathrm{d}\mu_t(y) \;+\; \frac{1}{2}\nabla \mathrm{V}_\theta(x)\right]\right) \;=\; 0$$

1. Simplifications

Attention: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).

Perceptron: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

Unnormalized attention[10]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \;\longmapsto\; \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$

Conservative perceptron: $W\sigma(Ax) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \mathrm{diag}(\omega_i)$).

Note: Our main results will extend to B symmetric invertible, $b \neq 0$, and softmax attention.

3. Mean-field limit

$\mu_t^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)} \xrightarrow[n\to\infty]{W_2} \mu_t$ unif. on $[0, T]$, and the limit solves a continuity equation.

[10] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

A primer on Wasserstein gradient flows[11]

For an energy functional $\mathsf{E} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R}$:

First variation: $\frac{\delta \mathsf{E}}{\delta \mu} : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by $\lim_{\varepsilon \to 0} \frac{\mathsf{E}[\mu + \varepsilon \xi] - \mathsf{E}[\mu]}{\varepsilon} = \int \frac{\delta \mathsf{E}}{\delta \mu}[\mu](x) \, \mathrm{d}\xi(x).$

Wasserstein gradient: $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu] \in \mathsf{T}(\mathbb{S}^{d-1}).$

Let our energy and its first variation be defined canonically:

$$\mathsf{E}_{\beta,\theta}[\mu] = \frac{1}{2\beta} \iint e^{\beta\, x^\top y} \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \; + \; \frac{1}{2} \int \mathsf{V}_\theta \, \mathrm{d}\mu, \qquad \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x) = \frac{1}{\beta} \int e^{\beta\, x^\top y} \, \mathrm{d}\mu(y) + \frac{1}{2} \mathsf{V}_\theta(x).$$

The mean-field PDE is a **Wasserstein gradient flow**

($+$ ascent / $-$ descent determined by $V(t) = \pm I$, with sign perturbations absorbed in V_θ)

$$\partial_t \mu_t \; \pm \; \mathrm{div} \left(\mu_t \, \underbrace{\mathbf{P}_x^\perp \left[\int_{\mathbb{S}^{d-1}} e^{\beta\, x^\top y} \, y \, \mathrm{d}\mu_t(y) \; + \; \frac{1}{2} \nabla \mathsf{V}_\theta(x) \right]}_{= \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu_t](x)} \right) \; = \; 0$$

[11] R. Jordan, D. Kinderlehrer, F. Otto. SIMA (1996). L. Ambrosio, N. Gigli, G. Savaré. Lectures in Mathematics (2008).

State of the art

For **pure attention**, the asymptotic picture is well understood:

Ascent (attractive dynamics): Metastability emerges first, and later tokens collapse into a single Dirac mass (consensus/clustering/synchronization).

Descent (repulsive dynamics): Particles equidistribute across the sphere (the uniform surface measure is the unique global minimizer).

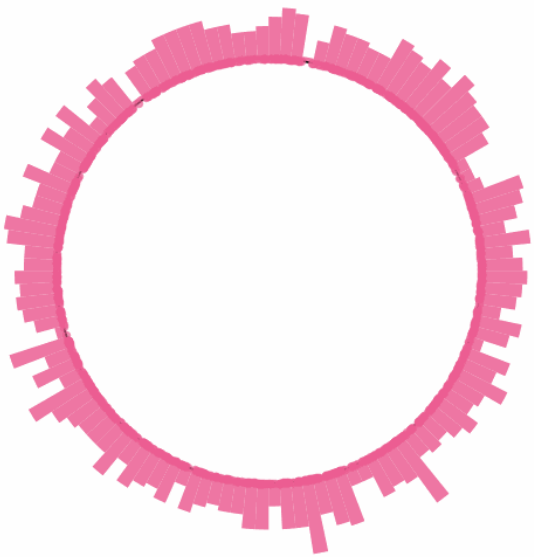
- Sander et al. PMLR (2022).
  Wasserstein gradient flows with doubly stochastic attention.
- Geshkovski et al. NeurIPS (2023).
  First clustering study for attention viewed as interacting particles.
- Castin, Ablin, Peyré. ICML/PMLR (2024).
  Sharp local-Lipschitz bounds for masked/unmasked attention, including the mean-field regime.
- Geshkovski et al. arXiv (2024).
  Metastability: long-lived multi-cluster configurations.
- Karagodin et al. NeurIPS (2024).
  Masked attention clusters (depending only on $\mathrm{spec}(V)$!) despite losing the mean-field gradient-flow structure.
- Yu et al. JMLR (2024).
  Interpretation via sparse rate reduction/compression.
- Abella et al. ICML/PMLR (2025).
  The essential trend towards consensus in attention.
- Alcalde, Fantuzzi, Zuazua. SIMODS (2025).
  Clustering in hardmax transformers.
- Bruno et al. ICLR (2025).
  Formalization of metastable clustered states.
- Bruno et al. NeurIPS (2025).
  Multi-scale analysis: structure formation vs coarsening.
- Burger et al. Phil. Trans. R. Soc. A (2025).
  Weighted gradient flow for general softmax attention.
- Castin, Ablin, Carrillo, Peyré. FoCM (2026).
  Mean-field limit for multi-head, L^2, Sinkhorn, sigmoid and masked attention, plus invariant Gaussian dynamics.
- Chen et al. arXiv (2025).
  Quantitative rates for mean-field clustering.
- Cowsik et al. Phys. Rev. E (2025).
  Signal propagation geometry and depth scaling.
- Geshkovski et al. Bull. AMS (2025).
  Framework: attention as interacting particles and gradient flows.
- Polyanskiy et al., arXiv (2025).
  Synchronization on $\mathbb{S}^1$ for all $\beta \geq -0.16$, but for no $\beta < -2/3$.
- Alcalde, Bungert, Riedl, Roith. arXiv (2026).
  W_2 concentration bounds for mean-field transformers as $1/\beta \to 0+$.
- Alcalde, Fantuzzi, Zuazua. Math. Found. Mach. Learn. (2026).
  Hardmax attention + perceptron for exact sequence interpolation.
- Alcalde, Geshkovski, Ruiz-Balet. ICML (2026).
  Hardmax attention as Frank--Wolfe on the token convex hull.
- Criscitiello et al. SIADS (2026).
  Synchronization criteria on spheres.

State of the art

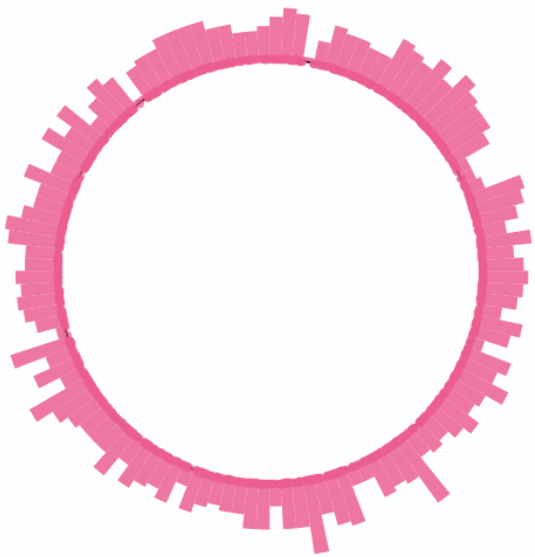
For **pure attention**, the asymptotic picture is well understood:

Ascent (attractive dynamics): Metastability emerges first, and later tokens collapse into a single Dirac mass (consensus/clustering/synchronization).

Descent (repulsive dynamics): Particles equidistribute across the sphere (the uniform surface measure is the unique global minimizer).



Pure attention in $\mathbb{S}^1$ + gradient ascent.



Pure attention in $\mathbb{S}^1$ + gradient descent.

Our problem

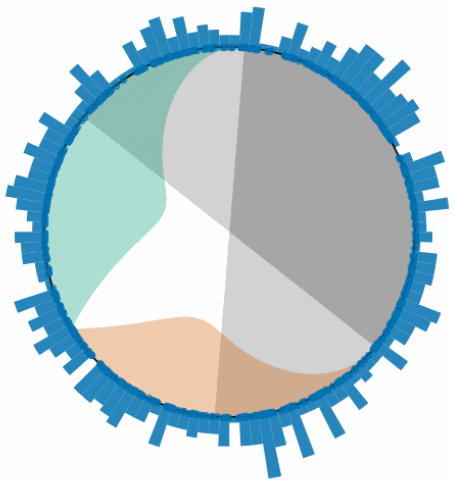
Main question.

How does adding the perceptron modify the equilibria of pure attention?

Our problem

Main question.

How does adding the perceptron modify the equilibria of pure attention?

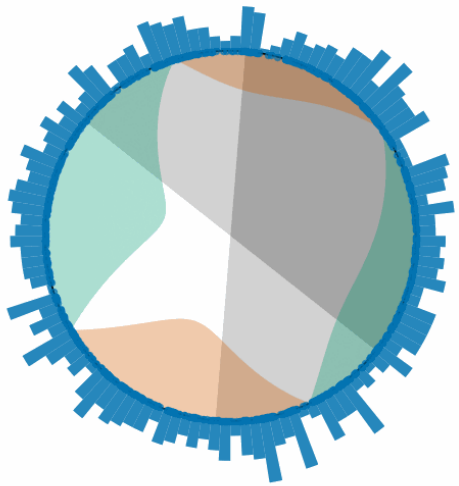


Radial profile: V_θ

$V_\theta(x) > 0$

$V_\theta(x) < 0$

$(Ax)_i \leq 0$



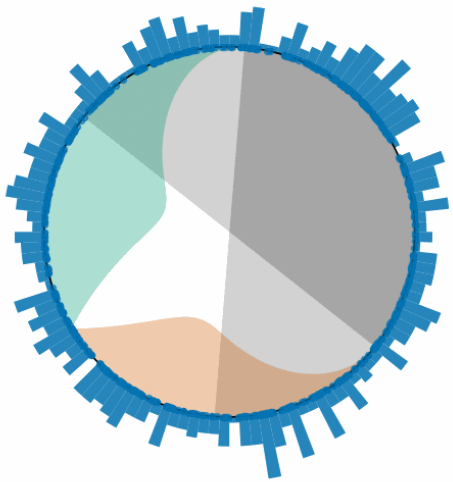
ReLU perceptron in $\mathbb{S}^1$
+ gradient ascent.

GELU perceptron in $\mathbb{S}^1$
+ gradient ascent.

Our problem

Main question.

How does adding the perceptron modify the equilibria of pure attention?

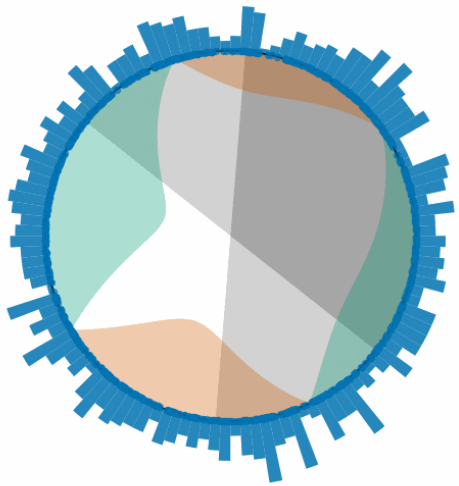


Radial profile: V_θ

$\mathrm{V}_\theta(x) > 0$

$\mathrm{V}_\theta(x) < 0$

$(Ax)_i \leq 0$



ReLU perceptron in $\mathbb{S}^1$
+ gradient descent.

GELU perceptron in $\mathbb{S}^1$
+ gradient descent.

Equilibrium points

Stationary measures:

$$\operatorname{div}\left(\mu \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu]\right) = 0 \qquad \text{(weakly)}$$

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad (\mu\text{-a.e.})$$

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Second-order positive-definite (SOPD) critical points:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \;\geq\; 0 \qquad \forall\, \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}).$$

$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathsf{E}(\mu(t))$, with $\mu(t)$ the W_2-geodesic along $\xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla_{\mathbb{S}} \phi \colon \phi \in C^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$

Strict SOPD: $\operatorname{Hess}_\mu \mathsf{E}(\xi,\xi) \geq \kappa \|\xi\|^2_{L^2(\mu)}$ for some $\kappa > 0$ independent of ξ.

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Second-order positive-definite (SOPD) critical points:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \;\geq\; 0 \qquad \forall\, \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}).$$

$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathsf{E}(\mu(t))$, with $\mu(t)$ the W_2-geodesic along $\xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla_{\mathbb{S}} \phi \colon \phi \in C^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$

Strict SOPD: $\operatorname{Hess}_\mu \mathsf{E}(\xi,\xi) \geq \kappa \|\xi\|^2_{L^2(\mu)}$ for some $\kappa > 0$ independent of ξ.

Why do we need both notions?

Stationarity yields enough rigidity when V_θ is piecewise polynomial (e.g., for $\sigma = \mathrm{ReLU}$).

Second-order analysis is vital when V_θ is real-analytic (e.g., for $\sigma = \mathrm{GELU}$), where stationarity is too soft and only stable configurations carry actual information.

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Proof sketch.

1. The vector field $\nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu]$ is real-analytic on each cell $I \subset \mathbb{S}^1$ where $\operatorname{sign}(Ax)$ is constant.
2. By the identity theorem, if $\operatorname{supp}(\mu)$ were infinite in any I, the analytic extension of the stationary field associated with I would vanish globally.
3. Integrating this extended field forces $f^\mu(x) := \int e^{\beta x \cdot y} \, d\mu(y)$ to be a global quadratic polynomial.
4. By the lemma, μ has a polynomial density, implying full support ($\implies$ $\operatorname{supp}(\mu) = \mathbb{S}^1$).
5. Full support enforces stationarity everywhere. Comparing the global quadratic identities obtained from adjacent activation cells forces V_θ to be globally real-analytic $\implies$ *contradiction*.

Identity theorem (S. G. Krantz and H. R. Parks. A Primer of Real Analytic Functions (2002))

Let M be a connected real-analytic manifold and let $f : M \to \mathbb{R}$ be real-analytic. If the zero set of f contains a nonempty open subset, then $f \equiv 0$ on M. If $\dim M = 1$, it suffices that the zero set has an accumulation point in M.
B. S. Mityagin. Math. Notes (2020): We can relax "open subset" to "a subset of positive measure on M".

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Proof sketch.

1. The vector field $\nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu]$ is real-analytic on each cell $I \subset \mathbb{S}^1$ where $\operatorname{sign}(Ax)$ is constant.
2. By the identity theorem, if $\operatorname{supp}(\mu)$ were infinite in any I, the analytic extension of the stationary field associated with I would vanish globally.
3. Integrating this extended field forces $f^\mu(x) := \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y)$ to be a global quadratic polynomial.
4. By the lemma, μ has a polynomial density, implying full support ($\implies$ $\operatorname{supp}(\mu) = \mathbb{S}^1$).
5. Full support enforces stationarity everywhere. Comparing the global quadratic identities obtained from adjacent activation cells forces V_θ to be globally real-analytic $\implies$ *contradiction*.

Lemma

For $\beta > 0$, the map $\mu \mapsto \left(f^\mu : x \mapsto \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y) \right)$ is injective on finite signed measures on $\mathbb{S}^{d-1}$, and if f^μ is a polynomial of degree $\leq k$, then μ has a density that is a polynomial of degree $\leq k$.

Tool: By the Funk--Hecke formula, spherical harmonics $Y_{j\ell}$ are eigenfunctions of the kernel: $\int e^{\beta x \cdot y} Y_{j\ell}(y) \mathrm{d}\sigma(y) = \lambda_j(\beta) Y_{j\ell}(x)$ with $\lambda_j(\beta) > 0$.

Atomicity on the circle (GELU)

Under real-analyticity, stationarity is not enough; strict stability collapses the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, and σ be real-analytic. Every strict SOPD critical point $\mu \in \mathcal{P}(\mathbb{S}^1)$ of $\mathsf{E}_{\beta,\theta}$ is purely atomic with finitely many atoms.

Atomicity on the circle (GELU)

Under real-analyticity, stationarity is not enough; strict stability collapses the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, and σ be real-analytic. Every strict SOPD critical point $\mu \in \mathcal{P}(\mathbb{S}^1)$ of $\mathsf{E}_{\beta,\theta}$ is purely atomic with finitely many atoms.

Proof sketch.

1. Again by the identity theorem, $\mathrm{supp}(\mu)$ being infinite would give $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0$ on $\mathbb{S}^1$, collapsing the Hessian to a bilinear form:

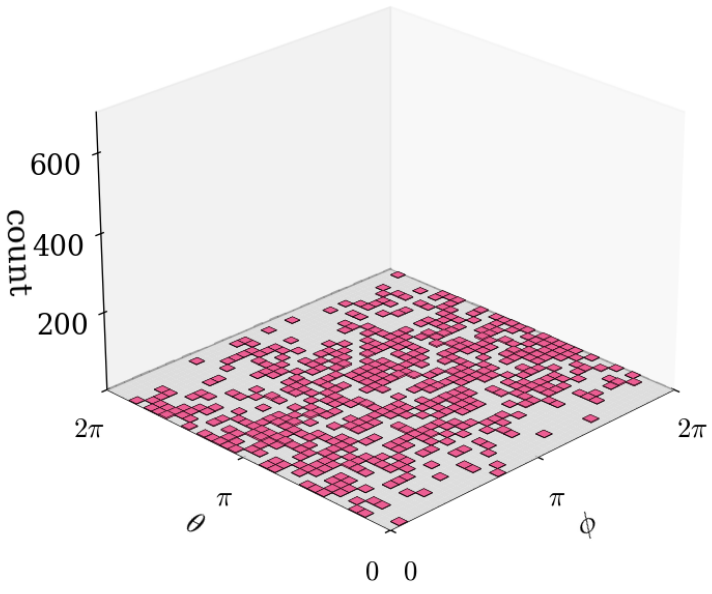
$$\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = -\frac{1}{\beta} \iint \mathsf{K}''_\beta(\theta-\phi)\, \xi(\theta)\xi(\phi) \mathrm{d}\mu(\theta)\mathrm{d}\mu(\phi)$$

2. Around any accumulation point, we build unit-norm, zero-mean tangent vectors ξ_δ of support size δ.

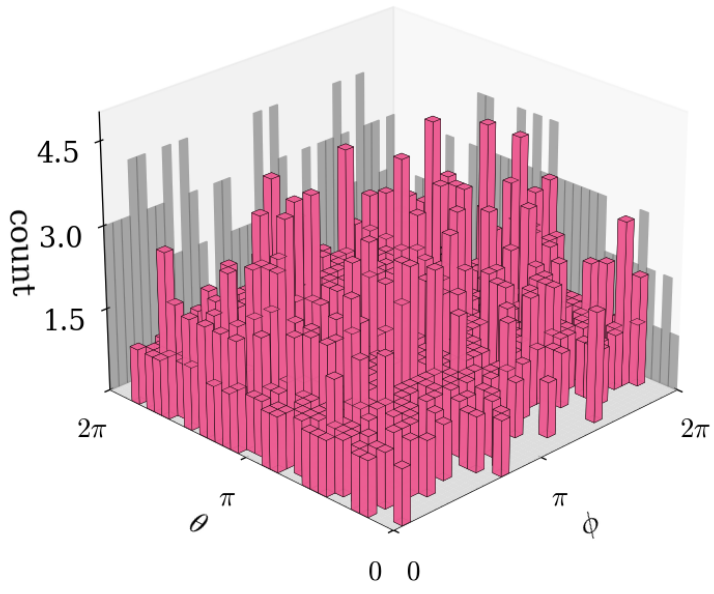
3. Since $\int \xi_\delta \mathrm{d}\mu = 0$, we can center the kernel ($\mathsf{K}''_\beta \to \mathsf{K}''_\beta - \mathsf{K}''_\beta(0)$). The energy cost then vanishes as the support shrinks, contradicting the strict SOPD coercivity:

$$\left|\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi_\delta,\xi_\delta)\right| \lesssim \frac{1}{\beta} \sup_{|h|\le\delta} \left|\mathsf{K}''_\beta(h) - \mathsf{K}''_\beta(0)\right| \xrightarrow{\delta\to 0} 0$$

What happens if $d \geq 3$?

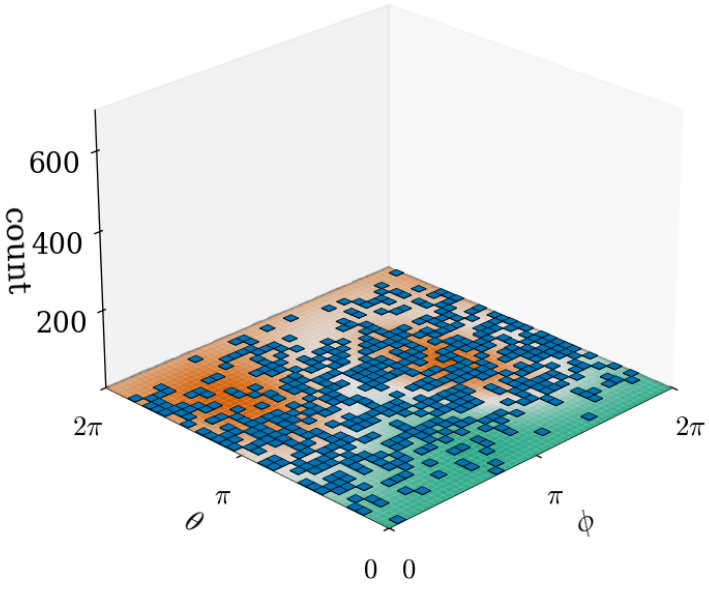


Pure attention in $\mathbb{S}^2$
+ gradient ascent.



Pure attention in $\mathbb{S}^2$
+ gradient descent.

What happens if $d \geq 3$?

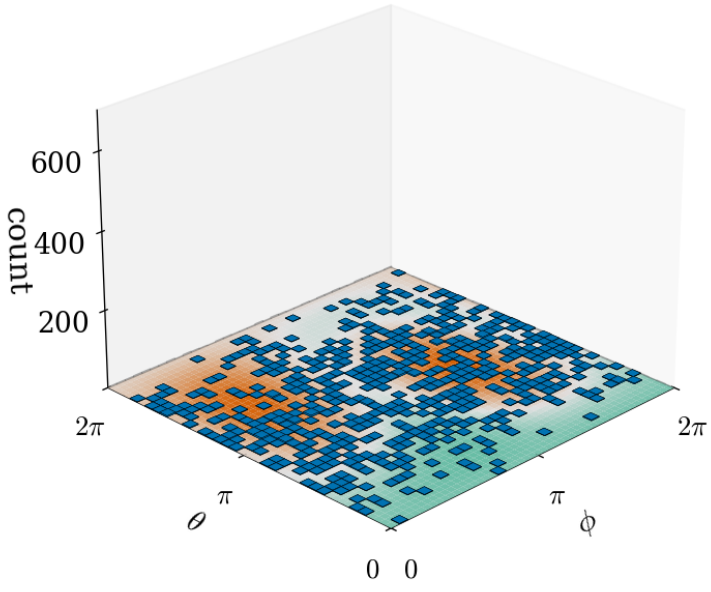


Surface profile: V_θ

$\mathrm{V}_\theta(x) > 0$

$\mathrm{V}_\theta(x) < 0$

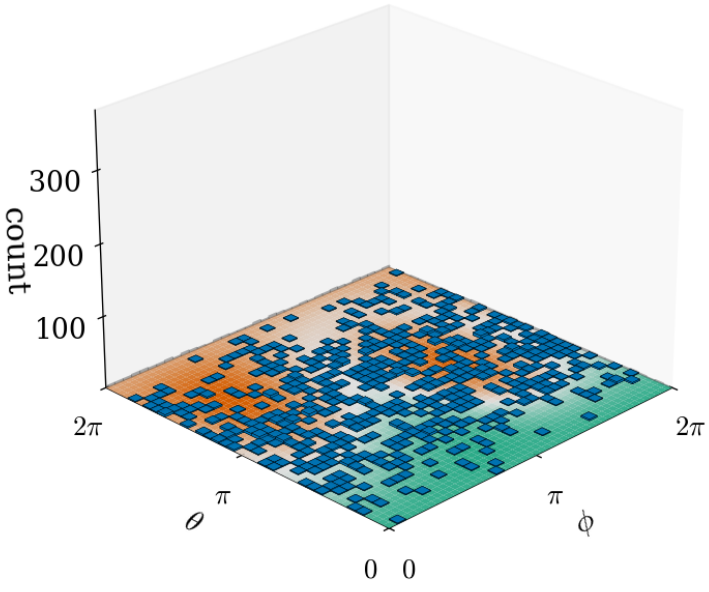
$(Ax)_i \leq 0$



ReLU perceptron in $\mathbb{S}^2$
+ gradient ascent.

GELU perceptron in $\mathbb{S}^2$
+ gradient ascent.

What happens if $d \geq 3$?

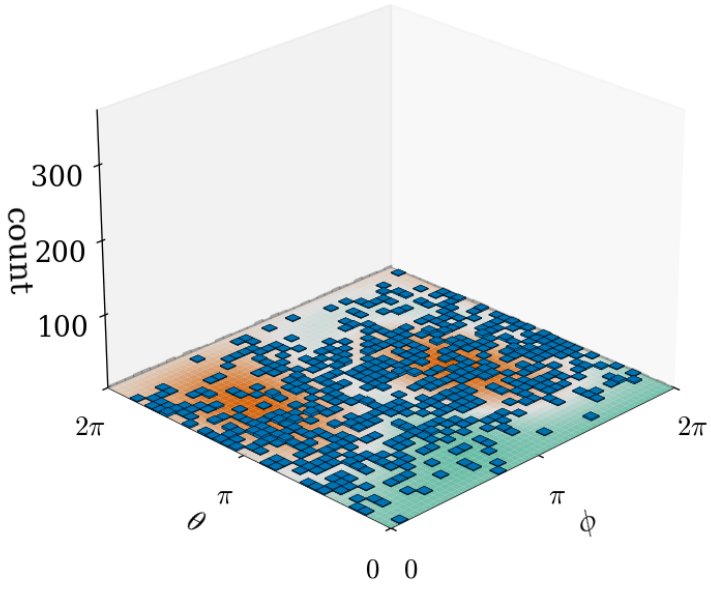


Surface profile: V_θ

$\mathsf{V}_\theta(x) > 0$

$\mathsf{V}_\theta(x) < 0$

$(Ax)_i \leq 0$



ReLU perceptron in $\mathbb{S}^2$
+ gradient descent.

GELU perceptron in $\mathbb{S}^2$
+ gradient descent.

What happens if $d \geq 3$?

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d \geq 2$. Suppose that μ is either stationary for σ the ReLU and V_θ is not real-analytic, or that μ is a strict SOPD critical point for σ the GELU. Then:

(a) *Singular support:* $\mathrm{supp}(\mu)$ *has zero surface measure.*

(b) *Generic atomicity: For a dense set of (β, θ), μ is purely atomic with finitely many atoms for GELU, and with countably many atoms inside $\{x : (Ax)_+ \neq 0\}$ for ReLU.*

What happens if $d \geq 3$?

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d \geq 2$. Suppose that μ is either stationary for σ the ReLU and V_θ is not real-analytic, or that μ is a strict SOPD critical point for σ the GELU. Then:

- ***(a)*** *Singular support:* $\mathrm{supp}(\mu)$ *has zero surface measure.*
- ***(b)*** *Generic atomicity: For a dense set of (β, θ), μ is purely atomic with finitely many atoms for GELU, and with countably many atoms inside $\{x : (Ax)_+ \neq 0\}$ for ReLU.*

Proof sketch. **(a)** Analogous to $d = 2$: a positive-measure support forces a global analytic identity.

(b) Set $g_{\beta,\theta}(x) := \nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu](x)$. To apply parametric transversality, we perturb any a_j along $h \in T_x\mathbb{S}^{d-1}$:

$$D_{a_j} g_{\beta,\theta}(x)[h] \;=\; \omega_j \sigma(a_j^\top x)\, h.$$

This is surjective whenever there is an active index ($\omega_j \sigma(a_j^\top x) \neq 0$). Thus, for generic parameters, the zeros are non-degenerate (hence isolated by the inverse function theorem).

For GELU, the zero set is finite by compactness. For ReLU, the argument is applied on every active stratum, yielding at most countably many zeros.

Quantitative anti-concentration

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Quantitative anti-concentration

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Proof sketch. Test against ξ supported on $\mathcal{S}$ with zero mean ($\sum_{i \in \mathcal{S}} m_i \xi_i = 0$), and split the Wasserstein Hessian into three components:

$$\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi, \xi) = \underbrace{\mathcal{T}_{\mathrm{in}}}_{\text{intra-cluster}} + \underbrace{\mathcal{T}_{\mathrm{out}}}_{\text{inter-cluster}} + \underbrace{\mathcal{T}_{\theta}}_{\text{perceptron}} \geq 0.$$

Intra-cluster: The kernel $e^{\beta \cos\theta}$ is strictly concave for $|\theta| \lesssim \beta^{-1/2}$. Thus, $\mathcal{T}_{\mathrm{in}} < -e^{\beta - 1/8} \sum_{i \in \mathcal{S}} m_i$.

Inter-cluster: $\mathcal{T}_{\mathrm{out}} > 0$ but bounded by $e^{\beta - 3/2} \sum_{k \notin \mathcal{S}} m_k$.

Perceptron: $\mathcal{T}_{\theta} > 0$ and bounded using the Lipschitz constant of σ.

Enforcing $\mathcal{T}_{\mathrm{in}} + \mathcal{T}_{\mathrm{out}} + \mathcal{T}_{\theta} \geq 0$ limits the mass $\sum_{i \in \mathcal{S}} m_i$.

Quantitative anti-concentration

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

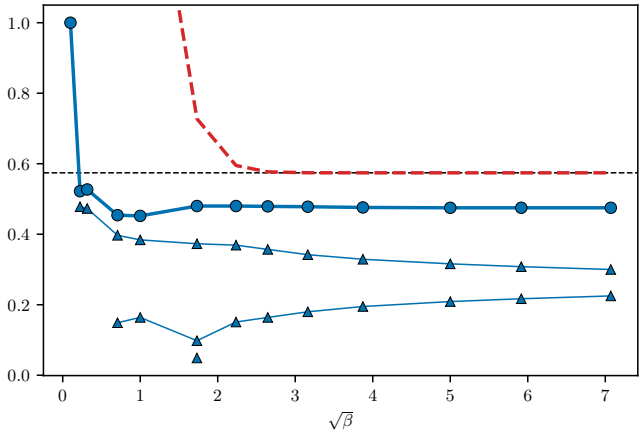
Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Verified empirically in $d = 2$



Gradient descent + GELU

$N = 1000$ particles.

- Largest cluster
- Other clusters

Red: full bound.

Black: mass $= 0.5742$.

Quantitative anti-concentration

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

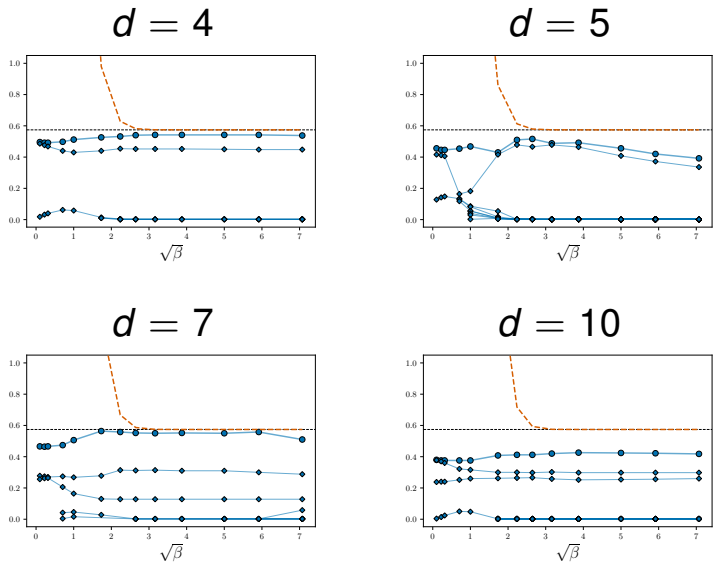
Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Even in higher dimensions! (empirically)



Gradient descent + GELU

$N = 1000$ particles.

Largest cluster

Other clusters

Red: full bound.

Black: mass $= 0.5742$.

CHAPTER 7



Distributional Models in Digital Health

A. ÁLVAREZ-LÓPEZ AND M. MATABUENA

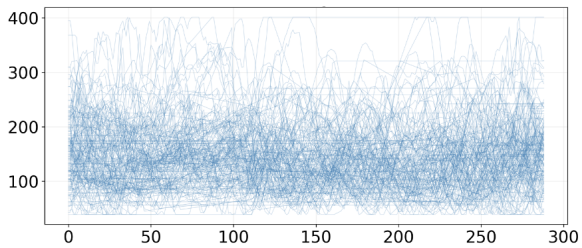
«Continuous-time learning of probability distributions via neural ODEs with applications in CGM data»

Preprint & Submitted, 2026

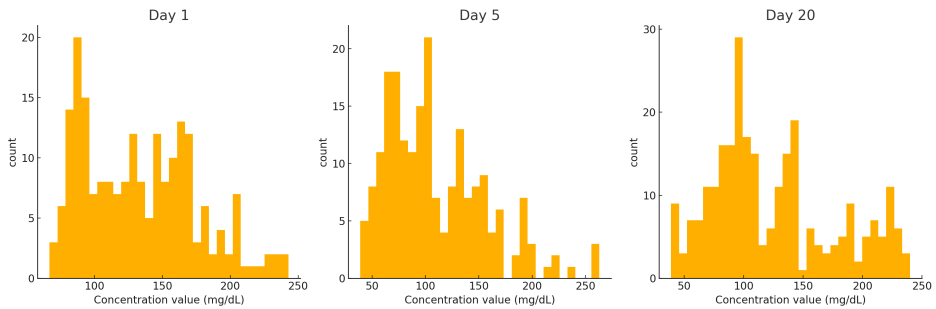


Clinical motivation

A continuous glucose monitor worn by a patient produces data streams of ∼288 glucose measurements per day. For each daily series, we build a glucodensity[12] from the proportion of time spent at each glucose level.



Raw glucose concentration over time.
Source: Wadwa et al. NEJM (2023).

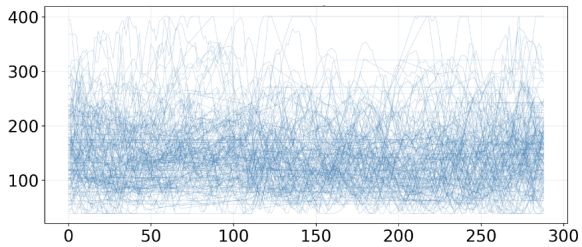


Estimated (empirical) glucodensity.

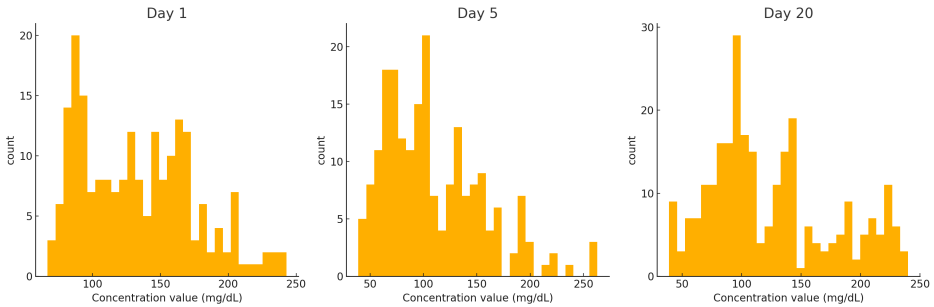
[12] M. Matabuena et al., Stat. Methods. Med. Res. (2021).

Clinical motivation

A continuous glucose monitor worn by a patient produces data streams of $\sim$288 glucose measurements per day. For each daily series, we build a glucodensity[12] from the proportion of time spent at each glucose level.



Raw glucose concentration over time.
Source: Wadwa et al. NEJM (2023).



Estimated (empirical) glucodensity.

Mathematical goal: *Recover an evolving probability law μ_t on $\mathbb{R}^d$ from discrete samples.*

Clinical demands: *Track glucose distribution, reflecting disease progression or treatment efficacy, ensuring that the model remains clinically interpretable.*

Difficulties: Irregular, noisy data streams. Multimodality (different metabolic states are mixed).

[12] M. Matabuena et al., Stat. Methods. Med. Res. (2021).

Proposed model

Model.

$$\rho_{\theta(t)}(x) = \sum_{k=1}^{K} \alpha_k(t)\, \mathcal{N}(x \mid \mu_k, \Sigma_k), \qquad (t, x) \in [0,1] \times \mathbb{R}^d,$$

$$\dot{z}(t) = \mathrm{NeuralODE}_\phi(z(t), t), \qquad \alpha(t) = \frac{z(t)_+}{\|z(t)_+\|_1}.$$

- **Interpretability**: Fixed Gaussian components (μ_k, Σ_k) act as reference glycaemic states.
- **Universality**[13]: The model can approximate in $L^1(\mathbb{R}^d)$ any continuous density path within

$$\left\{ \rho : \mathbb{R}^d \times [0, T] \to \mathbb{R}_{\geq 0} \;\middle|\; \int_{\mathbb{R}^d} \rho = 1 \right\}.$$

[13] Universal approx. of neural networks + Density of Gaussians (N. Wiener, Tauberian theorems, 1932).

Proposed model

Model.

$$\rho_{\theta(t)}(x) = \sum_{k=1}^{K} \alpha_k(t)\,\mathcal{N}(x \mid \mu_k, \Sigma_k), \qquad (t, x) \in [0,1] \times \mathbb{R}^d,$$

$$\dot{z}(t) = \mathrm{NeuralODE}_\phi(z(t), t), \qquad \alpha(t) = \frac{z(t)_+}{\|z(t)_+\|_1}.$$

Two-stage optimization procedure

- **Step 1 (Static):** Compute (μ_k, Σ_k) (by SGD) and $\alpha^{(i)}$ (least squares at each t_i). We alternate between minimizing the Maximum Mean Discrepancy on the global dataset and on each local dataset (respectively).

$$\mathrm{MMD}^2(\mu, \nu) = \iint k(x, x')\,\mathrm{d}\mu(x)\mathrm{d}\mu(x') + \iint k(y, y')\,\mathrm{d}\nu(y)\mathrm{d}\nu(y') - 2\iint k(x, y)\,\mathrm{d}\mu(x)\mathrm{d}\nu(y).$$

Kernel: $k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$ with σ selected using the median heuristic.

Advantages: Closed-form expressions for Gaussian mixtures vs empirical measures. Robust to outliers.

Proposed model

Model.

$$\rho_{\theta(t)}(x) = \sum_{k=1}^{K} \alpha_k(t)\, \mathcal{N}(x \mid \mu_k, \Sigma_k), \qquad (t, x) \in [0,1] \times \mathbb{R}^d,$$

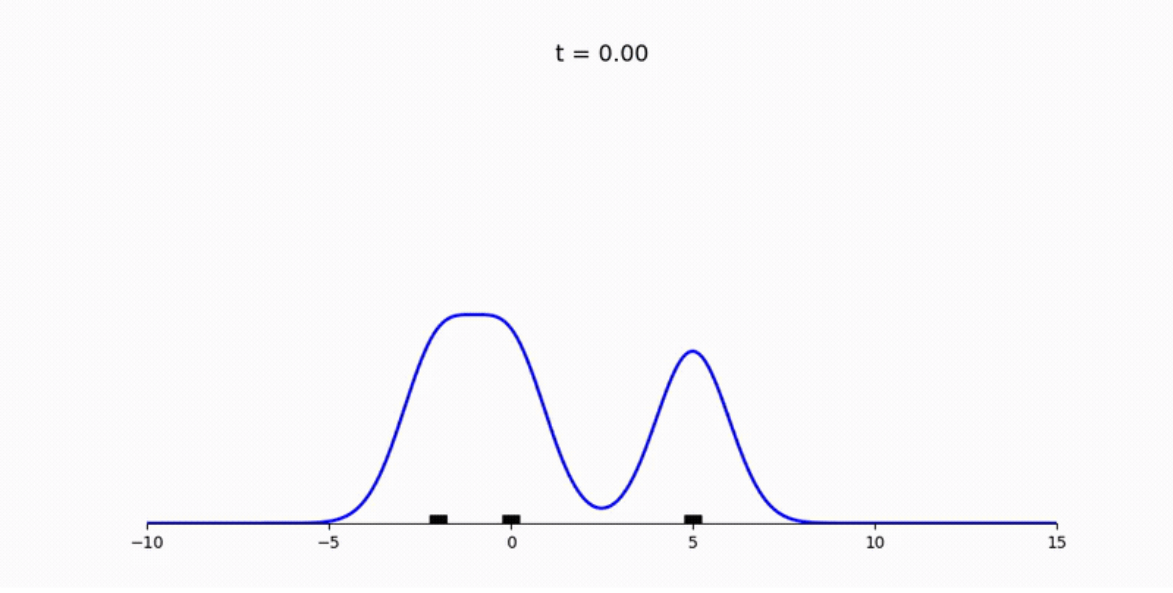
$$\dot{z}(t) = \mathrm{NeuralODE}_\phi(z(t), t), \qquad \alpha(t) = \frac{z(t)_+}{\|z(t)_+\|_1}.$$

Two-stage optimization procedure

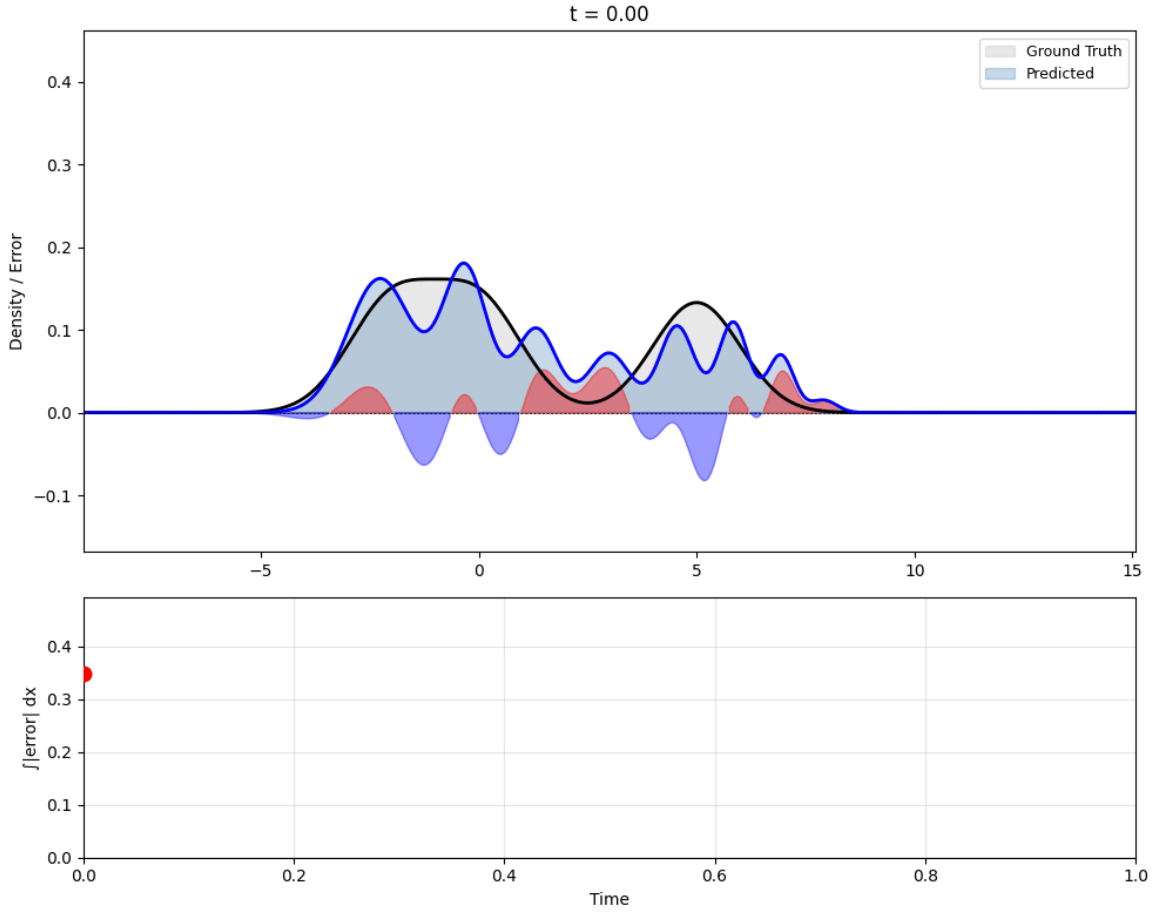
- **Step 2 (Dynamics):** Fit the neural ODE parameters ϕ to the estimated weights:

$$\min_{\phi} \sum_i \left\| \alpha(t_i; \phi) - \alpha^{(i)} \right\|^2 + \nu \|\phi\|^2, \qquad (\nu \geq 0 \text{ fixed}).$$

Validation (synthetic data)



Validation (synthetic data)



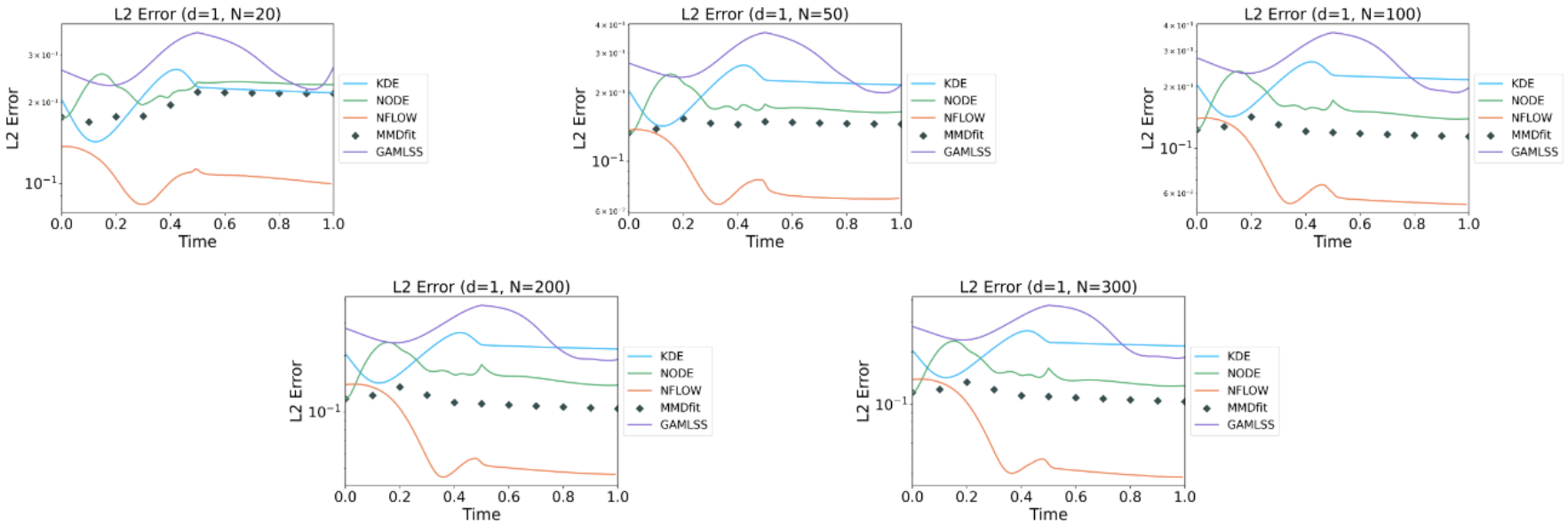
Step 1 - Parameters

Parameter	Value	Parameter	Value
Max iterations	20	*Rel. tol. param.*	10^{-9}
Rel. tol. err.	10^{-7}	*Abs. tol.*	10^{-6}
Components (K)	10	*Learning rate*	0.01
Grad. steps	10	*Ridge reg.*	0.1

Step 2 - Parameters

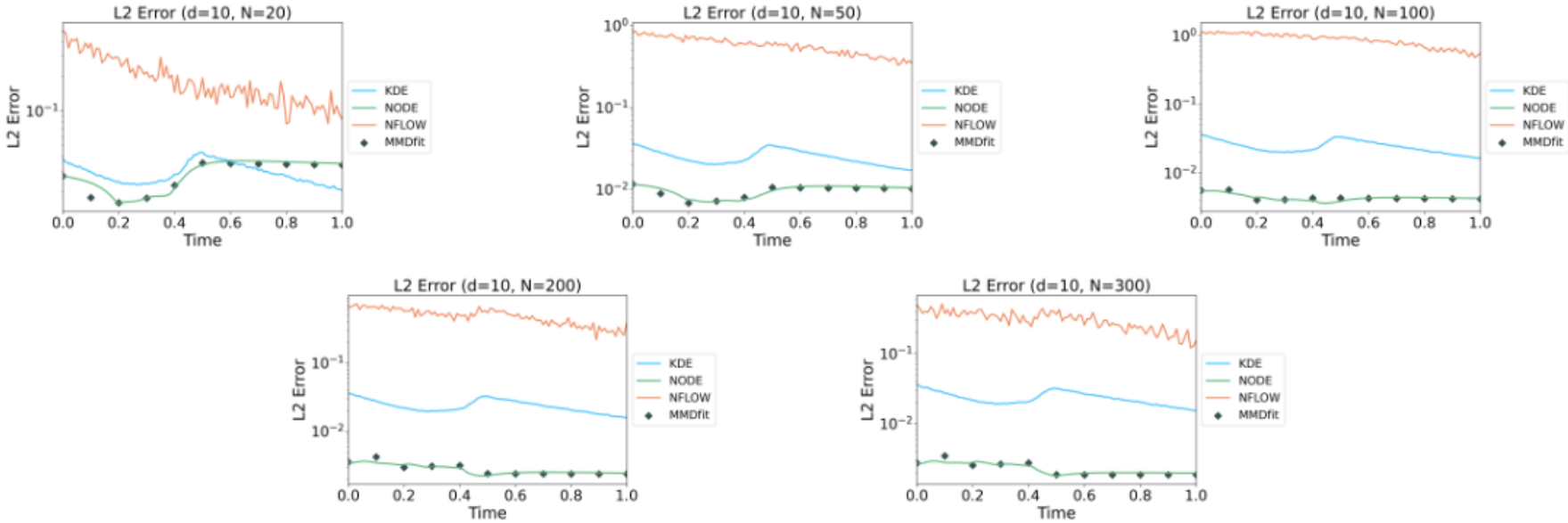
Parameter	Value	Parameter	Value
Seed	42	*Hidden dim*	200
Activation	ReLU	*Layers*	2
Optimizer	Adam	*Integration time*	1.0
Step size	0.01	*Integrator*	RK4
Learning rate	10^{-3}	ℓ^2 *reg.*	0.001
Max epochs	2000	*Monte Carlo samples*	10000
Rel. tol.	10^{-6}	*Abs. tol.*	10^{-6}

Comparison with alternative methods



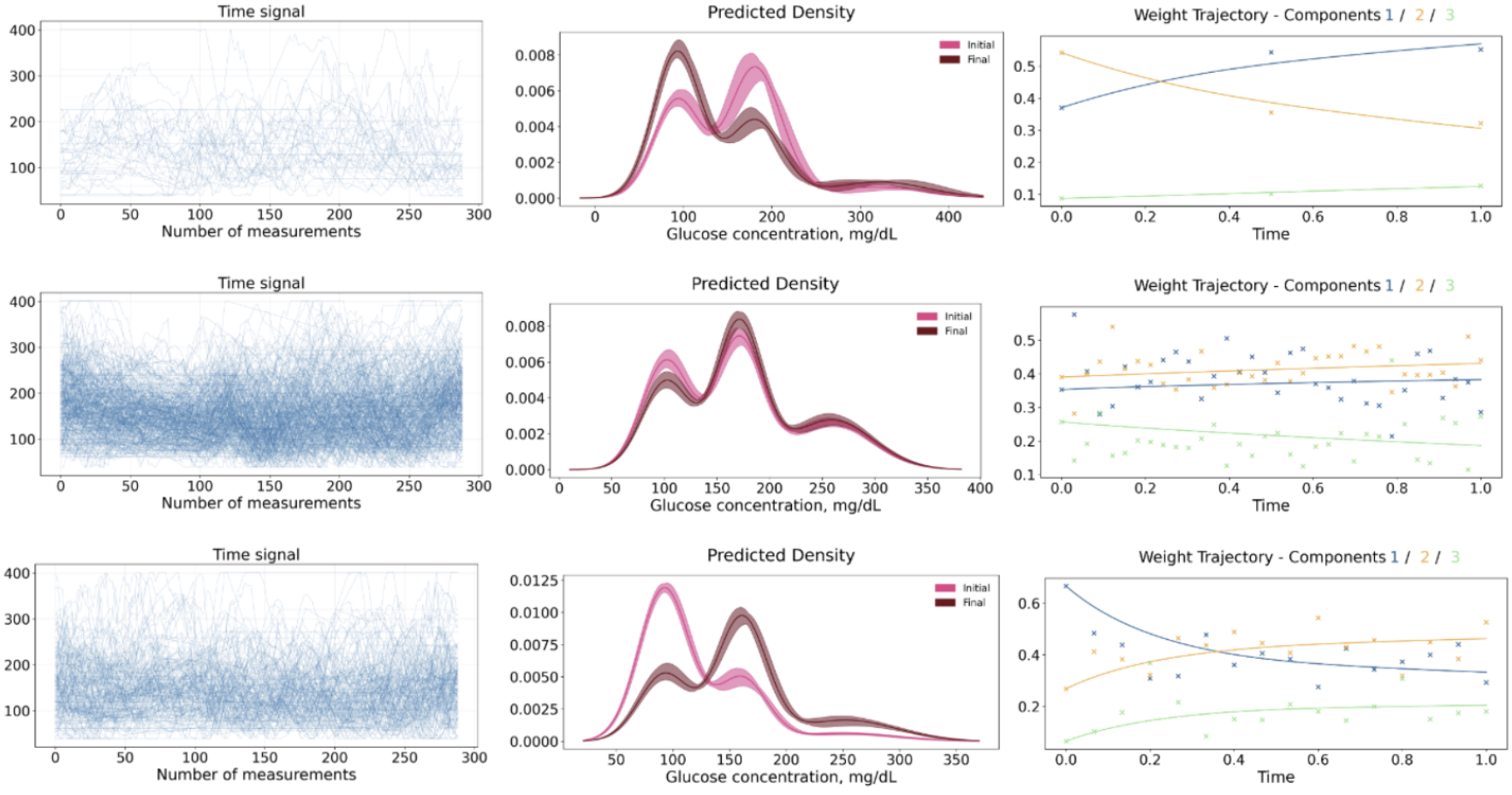
$d = 1$: Competitive performance.

Comparison with alternative methods



$d = 10$**:** Lowest L^2 error among the compared methods.

Clinical case study



Thank you.

Full Thesis

Full Thesis

A. Álvarez-López, R. Orive-Illera, and E. Zuazua

Cluster-based classification with neural ODEs via control

Journal of Machine Learning, 4(2):128--156, 2025

A. Álvarez-López, A. H. Slimane, and E. Zuazua

Interplay between depth and width for interpolation in neural ODEs

Neural Networks, 180:106640, 2024

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet

Constructive approximate transport maps with normalizing flows

Applied Mathematics & Optimization, 92(2):33, 2025

A. Álvarez-López, L. Liverani, and E. Zuazua

Constructive interpolation and generalization rates for neural ODEs: a control perspective

Preprint & Submitted, 2026

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet

Perceptrons and localization of attention's mean-field landscape

International Conference on Machine Learning, 2026

A. Álvarez-López and M. Matabuena

Continuous-time learning of probability distributions: a case study in a digital trial of young children with T1D

Preprint & Submitted, 2026