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

Antonio Álvarez-López[1,2,*] Lorenzo Liverani[1] Enrique Zuazua[1,2,3]

antonio.alvarezl@uam.es lorenzo.liverani@fau.de enrique.zuazua@fau.de

Abstract

We study supervised regression with neural ODEs (NODEs) from a control-theoretic perspective to derive explicit population-risk bounds. We focus on a widely used class of non-autonomous models with constant parameters and explicit time dependence, which we call semi-autonomous NODEs (SA-NODEs). We constructively prove that SA-NODEs are capable of *exact* interpolation of admissible finite datasets, and even satisfy a stronger property that we call *simultaneous cell controllability* (SCC): their flows can map prescribed disjoint cells into arbitrarily small target balls. This property is the mechanism that upgrades interpolation into quantitative generalization, by allowing SA-NODEs to emulate piecewise-constant nonparametric estimators. Consequently, our risk bounds recover the rates of histogram and nearest-neighbor estimators, provided the network width satisfies a conservative scaling with the sample size. Numerical experiments show that trained SA-NODEs achieve competitive—often lower—test errors than these baselines. Finally, we show that the explicit time dependence is essential. Although two-layer autonomous NODEs can interpolate geometrically nondegenerate datasets, structural obstructions prevent them from achieving SCC. These limitations, further confirmed numerically, support the view that SA-NODEs provide a minimal effective architecture for learning.

Keywords. Neural ODEs; controllability; interpolation; nonparametric regression; generalization bounds.

MSC. 68T07, 93B05, 62G08, 34H05, 68Q32.

1 Introduction

1.1 Expressivity and generalization

Supervised learning is one of the central paradigms in machine learning, with successful applications across a wide range of problems in science and engineering [20]. The standard framework relies on a dataset of size $N \geqslant 1$, consisting of pairwise distinct inputs $\mathbf{x}_i \in \mathcal{X}$ and corresponding labels $\mathbf{y}_i \in \mathcal{Y}$,

$$\mathcal{D}_N = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \qquad \text{with } \mathbf{x}_i \neq \mathbf{x}_j \text{ for } i \neq j. \tag{1}$$

The goal is to approximate the unknown relationship between inputs and labels by fitting a predictor to the sample $\mathcal{D}_N$, typically via a parametric family $\{F_\theta\}_\theta$, where θ is the learnable parameter. The choice of $\mathcal{Y}$ determines the task. In this work, we focus on regression in a deterministic, noiseless setting. Accordingly, we assume throughout that $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^d$ for $d \geqslant 2$, and that there exists a ground-truth map $y : \mathcal{X} \to \mathcal{Y}$ such that $y(\mathbf{x}_i) = \mathbf{y}_i$ for all i.

Of course, fitting $\mathcal{D}_N$ alone does not constitute learning: the true goal is to approximate y accurately across $\mathcal{X}$. To assess performance beyond the training data, we fix a test probability measure $\mu \in \mathcal{P}(\mathcal{X})$, often viewed as the (unknown) data-generating distribution, and define the population (true) risk by

$$\mathcal{R}(\theta) \coloneqq \int \ell\left(F_\theta(\mathbf{x}), y(\mathbf{x})\right) \, \mathrm{d}\mu(\mathbf{x}), \tag{2}$$

where $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geqslant 0}$ is a chosen loss function. Since μ is unknown, the population risk $\mathcal{R}(\theta)$ cannot be computed, let alone minimized, directly from $\mathcal{D}_N$. One therefore introduces the empirical risk

$$\mathcal{R}_N(\theta) \coloneqq \frac{1}{N}\sum_{i=1}^N \ell\left(F_\theta(\mathbf{x}_i), \mathbf{y}_i\right) = \int \ell\left(F_\theta(\mathbf{x}), y(\mathbf{x})\right) \, \mathrm{d}\mu_N(\mathbf{x}), \qquad \mu_N \coloneqq \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i} \in \mathcal{P}(\mathcal{X}). \tag{3}$$

[1]Chair for Dynamics, Control, Machine Learning, and Numerics (Alexander von Humboldt Professorship), Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany.

[2]Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[3]Chair of Computational Mathematics, Universidad de Deusto, Av. de las Universidades 24, 48007 Bilbao, Basque Country, Spain.

*Corresponding author.

Training then amounts to selecting parameters θ that minimize the empirical risk. This approach introduces a fundamental tension, which is revealed by the following error decomposition:

$$\mathcal{R}(\theta) \leqslant \mathcal{R}_N(\theta) + |\mathcal{R}(\theta) - \mathcal{R}_N(\theta)|\,. \tag{4}$$

The first term measures the fit on the dataset; driving it to zero (exact interpolation) requires enough model *expressivity*. The second term is the *generalization gap*, measuring the discrepancy between the expected risk under μ and its empirical proxy μ_N. Classical approaches usually control this gap by constraining the parameter space. This strategy is not designed for the interpolation regime, where models are inherently overparameterized. Our purpose is to exhibit a different mechanism, based on controllability of macroscopic regions, by which exact interpolation and quantitative generalization can coexist.

Controlled flows and neural ODEs

We study expressivity and generalization within a class of predictors derived from dynamical systems. More precisely, we cast supervised learning under the guise of *controlled flows*, modeled by the ODE

$$\begin{cases} \dot{\mathbf{x}}(t) = v_{\theta(t)}(\mathbf{x}(t)), & t \in (0,T], \\ \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^d, \end{cases} \tag{5}$$

with $T > 0$ fixed, and the vector field $v_{\theta(\cdot)}(\cdot) : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ that we assume to be measurable in t and uniformly Lipschitz in $\mathbf{x}$. Under these conditions, (5) induces a unique flow

$$\Phi_t^\theta : \mathbb{R}^d \to \mathbb{R}^d, \qquad t \in [0,T],$$

and the predictor is $F_\theta = \Phi_T^\theta$. This formulation offers a control-theoretic interpretation [14, 28]. The parameters $\theta(\cdot)$ act as *controls* steering the dynamics, and empirical risk minimization becomes, in essence, a problem of *simultaneous controllability*: one seeks some $\theta(\cdot)$ such that Φ_T^θ drives each initial datum $\mathbf{x}_i$ to its prescribed target $\mathbf{y}_i$. This perspective places the rich machinery of control theory at the service of machine learning and provides a unified framework to study both expressivity and generalization.

From a machine learning perspective, it is natural to parameterize v_θ as a neural network. With this choice, (5) is termed a neural ODE, with many variants differing in parameterizations, readouts or discretizations [9, 14, 18, 23]. In the simple case of a single-layer ReLU network of width $p \in \mathbb{N}$, we obtain:

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w}_i(t)\big(\mathbf{a}_i(t)\cdot\mathbf{x}(t) + b_i(t)\big)_+, \qquad t \in [0,T]. \tag{NODE}$$

Here, $\theta = (\mathbf{w}_i, \mathbf{a}_i, b_i)_{i=1}^p : [0,T] \to (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$ are the controls, and $z_+ := \max\{z, 0\}$ is the ReLU.

While (NODE) is highly flexible, it is heavily parameterized, with its effective parameter count growing under finer temporal discretizations. This motivates the search for simpler architectures that preserve expressive power. We focus on one such model: the *semi-autonomous neural ODE*,

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w}_i\big(\mathbf{a}_i\cdot\mathbf{x}(t) + b_i t + c_i\big)_+, \qquad t \in [0,T], \tag{SANODE}$$

derived from (NODE) by restricting $\mathbf{w}_i$ and $\mathbf{a}_i$ to be time-independent, and $b_i(t) = b_i t + c_i$. Equivalently, this corresponds to parameterizing v_θ via a time-invariant network applied to the extended state $(\mathbf{x}, t)$. This design makes (SANODE) an appealing middle ground [15, 27]: its relative simplicity eases implementation with respect to (NODE) while retaining high expressivity [21].

1.2 Main results

Our goal is to establish expressivity and generalization guarantees for neural ODEs. We demonstrate that (SANODE) achieves both, avoiding structural limitations of purely autonomous models.

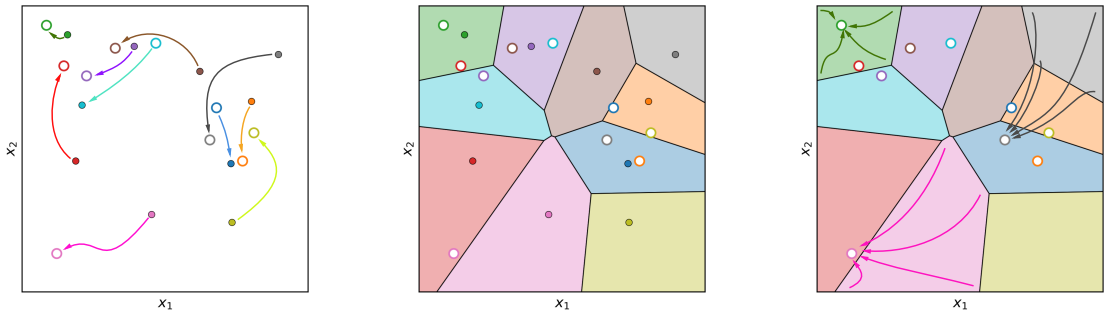

Figure 1: Overview of the proposed approach. **Left:** Interpolation of the training data from inputs (solid circles) to targets (hollow circles) via learned trajectories (arrows). **Center:** Voronoi partition of the input space, where each cell corresponds to a training sample. **Right:** The flow compresses entire cells toward their respective targets, illustrating the simultaneous cell controllability from Definition 2.

1.2.1 Expressivity.

The study of controllability for neural ODEs has garnered significant attention [1–3, 10, 28, 29]. A central concept in this literature is the *universal interpolation property* (UIP, [10]): given finitely many pairwise distinct inputs $\mathbf{x}_i$ and targets $\mathbf{y}_i$, does there exist a control $\theta(\cdot)$ such that $\Phi_T^\theta(\mathbf{x}_i) \approx \mathbf{y}_i$ for all i? While *approximate* interpolation (up to arbitrary precision) holds for various models, the exact UIP ($\Phi_T^\theta(\mathbf{x}_i) = \mathbf{y}_i$ for all i) is much more rigid. For time-dependent controls as in (NODE) or its linear-in-control version (see [1]), it is typically achieved by introducing temporal discontinuities [3, 28]. However, this machinery breaks down for (semi-)autonomous systems, leaving their exact UIP as a challenging open question.

Beyond exact interpolation of points, we introduce the stronger notion of *simultaneous cell controllability* to control entire regions of the input space. As illustrated in Figure 1, this requires the flow to map disjoint convex cells (e.g., a Voronoi partition) toward their respective targets.

With these notions in place, our expressivity results for (SANODE) establish the following:

- We prove the exact UIP for (SANODE) in Theorem 2.3. Our approach is constructive and yields explicit controls rather than merely proving their existence.
- We formalize simultaneous cell controllability and prove that (SANODE) exhibits this property in Theorem 2.5. Via Barron constant estimation, we derive explicit bounds on the required network width.

1.2.2 Generalization.

Under the setting of (5) and the squared Euclidean loss, the population and empirical risks (2)–(3) become:

$$\mathcal{R}(\theta) = \int \left\|\Phi_T^\theta(\mathbf{x}) - y(\mathbf{x})\right\|^2 \mathrm{d}\mu(\mathbf{x}), \qquad \mathcal{R}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left\|\Phi_T^\theta(\mathbf{x}_i) - \mathbf{y}_i\right\|^2. \tag{6}$$

The central challenge remains bounding the decomposition (4). Existing theoretical approaches typically bound the generalization gap uniformly over a restricted parameter space Θ. For instance, [22] proves that for the linear-in-control version of (NODE) under bounded Lipschitz controls, with probability $1-\delta$:

$$\mathcal{R}(\theta) \leqslant \mathcal{R}_N(\theta) + \sup_{\theta\in\Theta} |\mathcal{R}(\theta) - \mathcal{R}_N(\theta)| \lesssim \mathcal{R}_N(\theta) + \sqrt{\frac{(m+1)\log(R_\Theta m N)}{N}} + \frac{m\sqrt{R_\Theta}}{N^{1/4}} + \sqrt{\frac{\log(1/\delta)}{N}}$$

where m is the number of scalar control functions and $R_\Theta \geq 0$ encapsulates the bounds on Θ. If the controls are time-independent, this bound improves by removing the $O(N^{-1/4})$ term.

While informative, such bounds highlight an inherent tension: keeping the complexity (m and R_Θ) fixed as $N \to \infty$ shrinks the generalization gap but may preclude exact interpolation, leaving the empirical risk $\mathcal{R}_N(\theta)$ large. Conversely, driving $\mathcal{R}_N(\theta)$ to zero requires increasing complexity when N grows large, which deteriorates the bound. This motivates a natural **question**:

> *Can we design data-dependent controls θ_N, while appropriately scaling the network width, to guarantee that the population risk $\mathcal{R}(\theta_N)$ vanishes at a quantifiable rate as $N \to \infty$?*

We provide an affirmative answer by connecting (SANODE) to nonparametric statistics. Rather than assuming a fixed parametric form, nonparametric methods—such as kernel estimators [24, 33], histograms or k-nearest-neighbors [16, 31]—construct local approximations directly from the data. For Hölder spaces, these procedures yield explicit, minimax-optimal convergence rates [30]. This establishes a natural benchmark: any learning method claiming to generalize should ideally match these rates. We show that (SANODE) achieves this benchmark by using simultaneous cell controllability to emulate these estimators.

INFORMAL STATEMENT (Propositions 4.5 and 4.6). Given a dataset $\mathcal{D}_N$, a compactly supported probability measure μ, and an α-Hölder continuous target y, there exists a control θ_N such that

$$\mathcal{R}(\theta_N) \lesssim \mathsf{Err}_{\mathrm{np}} + \mathsf{Err}_{\mathrm{node}},$$

where $\mathsf{Err}_{\mathrm{np}}$ is the statistical error of a piecewise-constant nonparametric estimator built from $\mathcal{D}_N$, and $\mathsf{Err}_{\mathrm{node}}$ measures how accurately the flow of (SANODE) realizes this estimator. Provided the network width p grows sufficiently fast with N ($p \geq p_N$), the realization error is absorbed. Taking the expectation over the random draw of $\mathcal{D}_N$ (if the inputs $\mathbf{x}_i$ are sampled i.i.d. from μ), we recover the classical rates:

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

Our width requirement p_N—which depends on N and the geometry of the data partition—acts as an achievability guarantee rather than a tight bound. It proves that the architecture inherently supports generalization: given sufficient width, there exists a control whose true risk vanishes at an explicit rate. This circumvents the inherent tension of uniform bounds: exact interpolation and optimal statistical consistency can mathematically coexist.

Related work. Recent literature has explored generalization across various continuous-time models. Building on the uniform bounds of [22] for the linear-in-control version of (NODE), [32] extended this capacity-constrained approach to broader nonlinear neural ODEs. In parallel, other works have studied complementary mechanisms: [17] showed that incorporating feedback enhances robustness, while [8] derived generalization bounds for neural controlled differential equations driven by irregular time series.

1.2.3 The limits of autonomous models

Having established the capabilities of (SANODE), it is natural to ask whether explicit time dependence is truly necessary. We may remove it entirely, leading to the *autonomous neural ODE*:

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w}_i \left(\mathbf{a}_i \cdot \mathbf{x}(t) + b_i\right)_+, \qquad t \in [0, T]. \tag{ANODE}$$

To compensate for the resulting structural constraints, we also consider increasing the compositional depth, yielding the *two-layer autonomous neural ODE* (2ANODE):

$$\dot{\mathbf{x}}(t) = \sum_{j=1}^{p_2} \mathbf{w}_j \left(\sum_{i=1}^{p_1} u_{ji} \left(\mathbf{a}_{ij} \cdot \mathbf{x}(t) + b_{ij}\right)_+ + c_j \right)_+, \qquad t \in [0, T], \tag{2ANODE}$$

with $\mathbf{w}_j, \mathbf{a}_{ij} \in \mathbb{R}^d$ and $u_{ji}, b_{ij}, c_j \in \mathbb{R}$ for $p_1, p_2 \in \mathbb{N}$. Although not our primary focus, analyzing these autonomous models helps contextualize the advantages of explicit time dependence.

While we prove in Theorem 2.9 that (2ANODE) can achieve the exact UIP (provided the data satisfy the geometric condition (13)), we also discuss obstructions to simultaneous cell controllability in autonomous models like (ANODE) and (2ANODE). This reveals a gap between point interpolation and cell routing, a distinction that Section 5.2 illustrates experimentally. Thus, among the models we consider, (SANODE) provides the simplest setting in which both point interpolation and cell routing mechanisms coexist.

1.3 Organization

Section 2 introduces the core controllability notions and provides constructive proofs of the exact UIP for (SANODE) and (2ANODE). Section 3 reviews the nonparametric estimators that serve as our benchmarks. Section 4 then states and discusses our main generalization theorems. In Section 5, we numerically evaluate the capacity of neural ODEs to match or outperform these benchmarks on several regression tasks. Finally, Section 6 offers concluding remarks and future directions. Most proofs are deferred to Section 7.

Notation

For $n \in \mathbb{N}$, we use $[n] = \{1, \dots, n\}$. Vectors are bold (e.g. $\mathbf{u}, \mathbf{v}$) and we denote the Euclidean inner product of $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ by $\mathbf{u} \cdot \mathbf{v}$. The open ball in $\mathbb{R}^d$ of radius $r > 0$ centered at $\mathbf{x}$ is $B(\mathbf{x}, r)$, and the hypercube is $I_R := [-R, R]^d$ for $R > 0$. Let $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ be the space of absolutely continuous probability measures on $\mathbb{R}^d$. We set

$$\log_+(r) := \max\{\log(r), 0\} \quad \text{for } r > 0, \qquad \text{and} \qquad \log_+(0) = 0. \tag{7}$$

Concerning inequalities, we write $a \lesssim b$ (and $a \gtrsim b$) to indicate that $a \leqslant Cb$ (and $a \geqslant Cb$) for some constant $C > 0$ whose independence from the parameters of interest (such as N, ε, or p) will be clear from the context. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$.

2 Controllability of neural ODEs

2.1 Simultaneous point controllability

We begin with our central notion of expressivity. Consider the general model (5) with flow map Φ_t^θ. Since $\Phi_T^\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism, and hence injective, exact interpolation is impossible if distinct inputs $\mathbf{x}_i \neq \mathbf{x}_j$ are assigned the same target. To avoid this, we restrict our attention to datasets satisfying

$$\mathbf{x}_i \neq \mathbf{x}_j \quad \text{and} \quad \mathbf{y}_i \neq \mathbf{y}_j \qquad \text{for all } i \neq j. \tag{8}$$

Any dataset satisfying (8) will be called *admissible*.

Definition 1. *We say that the control system associated with* (5) *possesses the* approximate universal interpolation property (UIP) *if for every $\varepsilon > 0$, every $N \in \mathbb{N}$, and every dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ admissible in the sense of* (8)*, there exist $T > 0$ and an admissible control θ such that the flow map Φ_T^θ generated by* (5) *satisfies*

$$\|\Phi_T^\theta(\mathbf{x}_i) - \mathbf{y}_i\| \leq \varepsilon, \qquad \textit{for all } i \in [N]. \tag{9}$$

If (9) *holds with $\varepsilon = 0$, we say that the system possesses the* (exact) UIP.

When the vector field is a neural network, the admissible controls in Definition 1 are understood to range over all widths. Thus, the UIP is a property of the whole family, not of a fixed-width model.

Remark 2.1 (Time rescaling)**.** *In the definition above, the time horizon T is allowed to be chosen freely. However, this is equivalent to a fixed-horizon formulation whenever the family of vector fields under consideration is closed under positive scalar multiplication and time reparameterization. Indeed, for any $T_* > 0$, the rescaled curve $\mathbf{y}(s) := \mathbf{x}(sT/T_*)$ solves an analogous system on $[0, T_*]$ driven by the vector field $v_{\theta(sT/T_*)}T/T_*$, which remains in the same class. All models considered in this work satisfy this property, provided there are no uniform bounds imposed on the parameter space.*

Before detailing our contributions, we briefly review known results for the general model (NODE). The UIP was proved in **[28]** using $p = 1$ neurons. We state its generalization to arbitrary $p \geq 1$ from **[3]**.

Theorem 2.2 (Simultaneous point controllability for (NODE) **[3]**, Theorem 1])**.** *For every $d \geq 2$, $T > 0$ and $p \in \mathbb{N}$, the system* (NODE) *possesses the exact UIP. Furthermore, for any admissible dataset of size N (in the sense of* (8)*), the control can be chosen piecewise constant in time with $2\lceil N/p \rceil - 1$ jumps.*

The constructive proof of this theorem exploits the freedom afforded by time-dependent parameters. The key device is the orthogonality constraint $\mathbf{w}_i \cdot \mathbf{a}_i = 0$: geometrically, this forces the flow generated by each neuron to act as a shear transformation parallel to its own activation hyperplane. This decoupling makes the dynamics tractable and allows the control strategy to split naturally into two phases separated by a single switch. In the first interval, N parallel hyperplanes steer $d-1$ coordinates of the data simultaneously; after the switch, the hyperplanes reorient to handle the remaining coordinate.

By contrast, for (SANODE) the vectors $\mathbf{w}_i$ and $\mathbf{a}_i$ are time-independent, so the reorientation mechanism described above is no longer available. The compensating source of flexibility is the explicit time dependence in the argument $\mathbf{a}_i \cdot \mathbf{x} + b_i t + c_i$, which makes the activation hyperplanes translate at a constant velocity given by b_i. Our first new result—whose proof is deferred to Section 7.1.1—shows that this single degree of freedom is sufficient to recover exact controllability without requiring any temporal switching.

Theorem 2.3 (Simultaneous point controllability for (SANODE))**.** *For every $d \geq 2$ and $T > 0$, the system* (SANODE) *possesses the exact UIP. Moreover, for any admissible dataset of size N (in the sense of* (8)*), exact interpolation can be achieved with width $p = 2N$.*

2.2 Simultaneous cell controllability

The UIP captures the capacity to interpolate any admissible finite dataset. To control the population risk, however, one needs the flow to behave well across entire regions of the input space, rather than only at finitely many points. This leads to the following stronger controllability notion.

Definition 2. *We say that the control system associated with* (5) *possesses the property of* simultaneous cell controllability *(SCC) if, for every finite family of pairwise disjoint compact convex sets $\{A_k\}_{k=1}^K \subset \mathbb{R}^d$, every collection of target points $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$, and every $\eta > 0$, there exist $T > 0$ and an admissible control θ such that the flow map Φ_T^θ generated by* (5) *satisfies*

$$\Phi_T^\theta(A_k) \subset B(\mathbf{r}_k, \eta) \qquad \text{for all } k \in [K]. \tag{10}$$

Remark 2.4. *Although framed here in the context of flows, simultaneous cell controllability is an architecture-independent notion. It applies to any parametric map F_θ, from neural networks to classical regression models.*

The property of SCC is stronger than the approximate UIP, as setting $A_k = \{\mathbf{x}_k\}$ in (10) directly recovers condition (9). The fundamental shift is that the flow must now simultaneously route whole macroscopic regions rather than isolated points. Despite this added complexity, the following theorem establishes that (SANODE) exhibits this property, and provides a lower bound on the required network width. This is derived in Section 7.1.3 by constructing a reference vector field and applying quantitative Barron-type estimates (with the geometric constants detailed in Remark 2.6).

Theorem 2.5 (Simultaneous cell controllability for (SANODE))**.** *Let $d, K \geq 2$ and $R, T > 0$. Let $\{A_k\}_{k=1}^K$ be a family of pairwise disjoint, compact, and convex subsets of $I_R := [-R, R]^d$, and let $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$ be pairwise distinct target vectors. Then there exist*

$$\mathfrak{C} = \mathfrak{C}\left(d, T, R, (\mathbf{r}_k)_k, (A_k)_k\right) > 0, \qquad \mathfrak{M} = \mathfrak{M}\left(d, T, R, (\mathbf{r}_k)_k\right) > 0$$

and $\eta_0 \in (0, 1]$ such that for every $\eta \in (0, \eta_0]$ and every integer

$$p \geq \mathfrak{C}\, \eta^{-2-\mathfrak{C}} \left(1 - \log \eta\right)^2, \tag{11}$$

there exists a control θ of width p for (SANODE) *such that*

1. *$\Phi_T^\theta(A_k) \subset B(\mathbf{r}_k, \eta)$ for all $k \in [K]$;*
2. *$\|\Phi_T^\theta\|_{L^\infty(I_R)} \leq \mathfrak{M}$.*

If the targets are not necessarily pairwise distinct, the same conclusions hold for any fixed $\eta > 0$; however, the constant $\mathfrak{C}$ is no longer uniform and will depend on η.

Remark 2.6 (Explicit geometric constants). *The constants $\mathfrak{C}$, $\mathfrak{M}$, and η_0 in Theorem 2.5 may be chosen explicitly. A possible choice, obtained by combining Lemma 7.6 with Lemma 7.7, is governed by the geometry of the cells and their targets. Assume that the target vectors $\mathbf{r}_1, \dots, \mathbf{r}_K$ are pairwise distinct, let*

$$D_* := \max_{k\in[K]} \operatorname{diam}(A_k), \qquad s_* := \min_{i\neq j} \operatorname{dist}(A_i, A_j),$$

and let $\gamma_k : [0,T] \to \mathbb{R}^d$ be smooth pairwise disjoint curves such that $\gamma_k(0) \in A_k$, $\gamma_k(T) = \mathbf{r}_k$, γ_k is constant on $[0, 2T/3]$, and $\|\gamma_k(t)\| < 1 + \max_{k\in[K]} \|\mathbf{r}_k\| + \sqrt{d}\, R$ for all $t \in [0,T]$—their existence is guaranteed for $d \geq 2$, see Lemma 7.5. Define

$$m := \min_{t\in[0,T]} \min_{i\neq j} \|\gamma_i(t) - \gamma_j(t)\|, \qquad G := 1 + \max_{k\in[K]} \max_{1\leq j\leq d+5} \left\| \frac{\mathrm{d}^j \gamma_k}{\mathrm{d}t^j} \right\|_{L^\infty(0,T)},$$

and let

$$L := C_{d,T} \left[\frac{G}{m} + \frac{1 + D_*/s_*}{T} \left(1 + \log_+(4D_*)\right) \right],$$

and

$$B := C_{d,T}\, K \left[T G^{d+5}(m^d + m^{-4}) + \frac{(D_* + s_*)^{d+1}(1 + s_*^{-(d+4)})}{T} \left(1 + \log_+(4D_*)\right) \right].$$

Exploiting these quantities—see the proofs of Lemmas 7.6 and 7.7 for a derivation—one may take

$$\eta_0 := \min\left\{1, \frac{s_*}{2}, \frac{m}{2}\right\}, \quad \mathfrak{M} := 2 + \max\left\{T, \max_{k\in[K]} \|\mathbf{r}_k\| + \frac{3}{2}\sqrt{d}\, R\right\}, \tag{12}$$
$$\mathfrak{C} := \max\left\{3,\ 2TL,\ 4C_d^2\, \mathfrak{M}^4\, T^2\, B^2 e^{2TL}\right\}.$$

Thus, while $\mathfrak{M}$ depends solely on the domain size and target locations, $\mathfrak{C}$ captures the geometric bottleneck of the partition: the term $1/s_*$ in L forces an exponential penalty $\exp(c/s_*)$ in the factor e^{2TL} as $s_* \to 0$.

Remark 2.7 (Width requirement and sharpness). *The lower bound (11) provides a sufficient condition rather than a sharp complexity estimate. As noted in Remark 2.6, the geometric constant $\mathfrak{C}$ incurs an exponential penalty as the minimum cell separation s_* shrinks. Consequently, for the refined partitions used later to establish statistical consistency, this geometric bottleneck can force a super-polynomial threshold p_N as the dataset size N grows. This reflects the curse of dimensionality inherent in isolating many small regions.*

*We emphasize that this scaling acts purely as an achievability guarantee. The specific bound arises from our chosen proof strategy—specifically, applying the uniform approximation results of [**19**, Theorem 2]. Employing alternative tools (e.g., [**5**, Proposition 6]) could alter the trade-off between the convergence rate in p and the magnitude of $\mathfrak{C}$. Determining the sharp intrinsic width required for simultaneous cell controllability remains an open quantitative question.*

2.3 Autonomous models

Having established the UIP and SCC for (SANODE), it is natural to ask whether explicit time dependence is truly necessary. To answer this, we remove it entirely and analyze the autonomous model (ANODE) and its two-layer variant (2ANODE).

While [**3**] establishes the approximate UIP for (ANODE)—including explicit error decay rates as the width grows—bridging the gap from approximate to exact interpolation remains unresolved:

Open Problem. *Does the exact UIP hold for the single-layer autonomous system (ANODE)?*

The difficulty is structural: unlike time-dependent models, which can concatenate flow stages by switching controls, autonomous single-layer fields are highly rigid. Specifically, they cannot be spatially localized. Because they are finite superpositions of ridge functions—whose Fourier transforms are supported on one-dimensional sets—such nontrivial vector fields can never belong to $L^2(\mathbb{R}^d)$. Though elementary, this fact seems absent from standard references, so we provide a proof.

Lemma 2.8. *Let $d \geq 2$ and let $\sigma \in \mathscr{C}^0(\mathbb{R})$ have at most polynomial growth. For any $p \geq 1$, define*

$$g(\mathbf{x}) = \sum_{j=1}^{p} w_j\, \sigma(\mathbf{a}_j \cdot \mathbf{x} + b_j), \qquad (\mathbf{x} \in \mathbb{R}^d), \qquad \text{with } \mathbf{a}_j \in \mathbb{R}^d, \quad w_j, b_j \in \mathbb{R}.$$

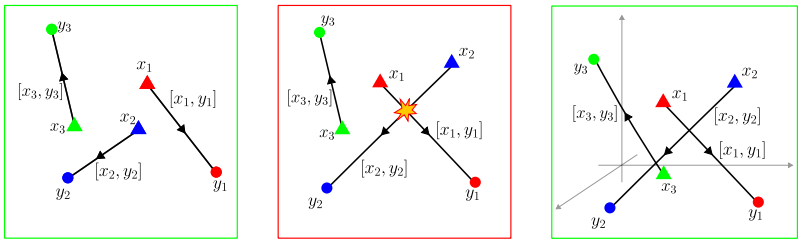

Figure 2: Illustration of the nondegeneracy condition (13). **Left & Right:** Admissible datasets for $d = 2$ and $d = 3$. **Center:** A degenerate data configuration. As discussed in Remark 2.10, condition (13) is generic for $d \geqslant 3$.

If $g \not\equiv 0$, then $g \notin L^2(\mathbb{R}^d)$. In particular, g cannot have compact support unless $g \equiv 0$.

Proof. For each j, the Fourier transform of the ridge function $\mathbf{x} \mapsto \sigma(\mathbf{a}_j \cdot \mathbf{x} + b_j)$ is supported, in the sense of tempered distributions, on the line $L_j = \{\lambda \mathbf{a}_j : \lambda \in \mathbb{R}\} \subset \mathbb{R}^d$. By linearity, $\widehat{g}$ is then supported in the finite union $\bigcup_{j=1}^{p} L_j$, which has Lebesgue measure zero in $\mathbb{R}^d$ since $d \geq 2$ by hypothesis.

Suppose $g \in L^2(\mathbb{R}^d)$. Then, by Plancherel, $\widehat{g} \in L^2(\mathbb{R}^d)$ as well. But $\widehat{g} = 0$ a.e. because an L^2–function supported on a null set must vanish almost everywhere, and therefore $g \equiv 0$. □

While Lemma 2.8 does not strictly rule out exact interpolation, it explains why "pointwise" steering strategies are exceptionally hard to implement with (ANODE) alone. A minimal way to restore localization without reintroducing time dependence is to add one compositional layer. By composing ReLU units, the two-layer field (2ANODE) can generate vector fields supported on convex polytopes (see Lemma 7.3 below), enabling the construction of localized controls.

For (2ANODE), we require the stronger assumption that input-output segments are mutually disjoint:

$$[\mathbf{x}_i, \mathbf{y}_i] \cap [\mathbf{x}_j, \mathbf{y}_j] = \varnothing \qquad \text{for all } i \neq j, \tag{13}$$

where $[\mathbf{a}, \mathbf{b}]$ denotes the line segment between $\mathbf{a}$ and $\mathbf{b}$. Note that this naturally implies $\mathbf{x}_i \neq \mathbf{y}_j$ for $i \neq j$.

Theorem 2.9 (Simultaneous point controllability for (2ANODE)). *Let $N \geq 1$, $d \geq 2$, and $T > 0$ be fixed. Consider any dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ admissible in the sense of (8), and additionally satisfying (13). Then there exists a (constant) control θ such that the flow map Φ_T^θ generated by (2ANODE), with $p_1 = 2d$ and $p_2 = N$, satisfies*

$$\Phi_T^\theta(\mathbf{x}_i) = \mathbf{y}_i, \qquad \textit{for all } i \in [N].$$

The proof is deferred to Section 7.1.2. Although (2ANODE) achieves exact UIP without any form of time dependence, its mechanism is less efficient than that of (SANODE): it requires $p_1 = 2d$ and $p_2 = N$, hence $p_1 \cdot p_2 = 2dN$ first-layer units and $p_2 = N$ second-layer units. By contrast, (SANODE) uses $p = 2N$ units.

Remark 2.10 (Dimensional constraints). *Condition (13) depends strongly on the dimension d. Because our current construction relies on vector fields localized strictly around straight segments $[\mathbf{x}_i, \mathbf{y}_i]$, these paths must not cross. In $d \geq 3$, linear segments generically do not intersect, so this condition is easily satisfied. In $d = 2$, however, straight-line collisions are frequent, making the assumption highly restrictive (see Figure 2).*

Nevertheless, the UIP can theoretically be established for all $d \geq 2$ without requiring (13). Instead of moving along straight lines, particles could be routed through intermediate waypoints to avoid collisions, using localized two-layer "connectors" to steer the flow around obstacles. While mathematically feasible, the technical length of such a construction places it beyond the scope of this paper, and we defer it to future work.

At this point, the distinction between point interpolation and cell routing becomes essential. While autonomous systems can map finite point clouds arbitrarily close to their targets by letting isolated trajectories bypass each other in $d \geq 2$ ([3, Theorem 7]), routing macroscopic cells A_k introduces structural obstructions. Since autonomous trajectories cannot cross, models like (ANODE) or (2ANODE) face a topological barrier. For example, if the target of A_1 lies behind A_2, the flow would need to push A_1 directly through the space initially occupied by A_2, which violates uniqueness. The explicit time dependence in (SANODE) circumvents this by allowing trajectories to cross safely in space-time.

Furthermore, Definition 2 strictly requires $\Phi_T^\theta(A_k) \subset B(\mathbf{r}_k, \eta)$. A weaker L^2-a.e. control might bypass the obstruction by sacrificing small-measure subsets of the cells: failing to route these small regions to

their targets would free up physical space to create "corridors" for other cells to pass through, potentially accommodating autonomous systems.

In summary, while autonomous systems can interpolate points—especially with added depth as in (2ANODE)—their lack of a time coordinate to schedule motion makes them structurally ill-suited for cell routing. Thus, among the architectures considered here, (SANODE) provides the simplest setting where exact interpolation and generalization coexist.

3 Nonparametric estimation

In this section, we recall the construction and fundamental properties of piecewise-constant nonparametric estimators. Our analysis focuses on two canonical approaches:

1. **Histogram estimators**, which partition the input domain using a fixed uniform grid.
2. **Nearest-neighbor estimators**, which rely on Voronoi tessellations [4, 7].

While the results in this section are standard, proofs are collected in Section 7.2, as the precise statements we need do not appear verbatim in the literature. We refer the interested reader to [16, 31] and references therein for a deeper dive into the theory of nonparametric estimators.

3.1 Histogram estimator

We first analyze the histogram approach. Fix $R > 0$, a resolution $h \in (0, 2R]$ and let

$$\mathcal{K} := \{0, \ \dots, \ \lceil 2R/h \rceil - 1\}^d \tag{14}$$

be the corresponding multi-index set. For each $k \in \mathcal{K}$, we define the cube

$$Q_k := I_R \cap \prod_{j=1}^{d} [-R + k_j h, \ -R + (k_j + 1)h) \,, \tag{15}$$

ensuring that the collection $\{Q_k\}_{k\in\mathcal{K}}$ forms a partition of I_R up to a μ-null boundary set. The total number of cells is given by

$$K_h := |\mathcal{K}| = \left\lceil \frac{2R}{h} \right\rceil^d \lesssim h^{-d}, \tag{16}$$

and the diameter of each cell satisfies $\operatorname{diam}(Q_k) \leqslant \sqrt{d}\, h$. We approximate $y(\cdot)$ by averaging its values within each cell. This leads to the definition of both a population-level and an empirical estimator:

- The population average $y_h \colon I_R \to \mathbb{R}^d$ is the best piecewise-constant approximation of y in $L^2(\mu)$:

$$y_h(\mathbf{x}) := \sum_{k\in\mathcal{K}} \mathbf{s}_k \, \mathbf{1}_{Q_k}(\mathbf{x}), \tag{17}$$

  where the coefficients $\mathbf{s}_k \in \mathbb{R}^d$ are defined by

$$\mathbf{s}_k := \begin{cases} \dfrac{1}{\mu(Q_k)} \displaystyle\int_{Q_k} y(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x}), & \text{if } \mu(Q_k) > 0, \\ 0, & \text{if } \mu(Q_k) = 0. \end{cases} \tag{18}$$

  Note that $y_h(\cdot)$ is in fact the best piecewise-constant approximation of $y(\cdot)$ in $L^2(\mu)$, as each $\mathbf{s}_k$ minimizes the local $L^2(\mu)$-error over Q_k.

- The empirical average $y_{N,h} : I_R \to \mathbb{R}^d$, instead, is computed directly from the dataset:

$$y_{N,h}(\mathbf{x}) := \sum_{k\in\mathcal{K}} \mathbf{S}_k \, \mathbf{1}_{Q_k}(\mathbf{x}). \tag{19}$$

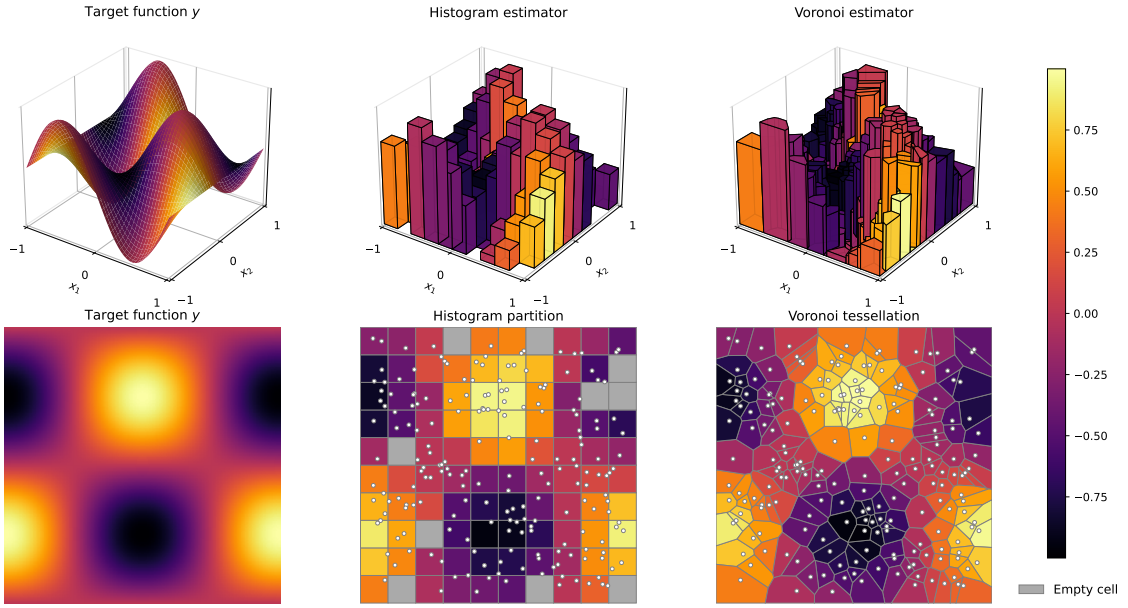

Figure 3: Histogram and Voronoi approximations of a function $y : \mathbb{R}^2 \to \mathbb{R}^2$, represented in both 3D (height and color) and 2D (color only). **Left:** The target function. **Center:** The histogram estimator, which assigns each cell the average value of its enclosed training points; gray cells indicate regions with no training data. **Right:** The Voronoi estimator, which assigns each cell the value of its nearest training point.

Letting $N_k \coloneqq \sum_{i=1}^{N} \mathbf{1}_{Q_k}(\mathbf{x}_i)$ denote the number of samples falling into Q_k, the coefficients $\mathbf{S}_k \in \mathbb{R}^d$ are now defined as

$$\mathbf{S}_k \coloneqq \begin{cases} \dfrac{1}{N_k} \displaystyle\sum_{i:\mathbf{x}_i \in Q_k} y(\mathbf{x}_i), & \text{if } N_k \geqslant 1, \\ 0, & \text{if } N_k = 0. \end{cases} \tag{20}$$

A schematic representation of $y_{N,h}(\cdot)$ when $d = 2$ can be seen in Figure 3 (central column). We now establish the convergence rate for the empirical histogram estimator.

Proposition 3.1. *Let $\mathcal{D}_N = \{(\mathbf{x}_i, y(\mathbf{x}_i))\}_{i=1}^N$ with $\mathbf{x}_1, \dots, \mathbf{x}_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{ac}(I_R)$, and $y \in \mathscr{C}^0(I_R; \mathbb{R}^d)$. Then, for any $h \in (0, 2R]$, the estimator $y_{N,h}(\cdot)$ defined in (19) satisfies*

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_{N,h}\|^2_{L^2(\mu)}\right] \lesssim \omega_y(\sqrt{d}\,h)^2 + \frac{\|y\|^2_{L^\infty(I_R)}}{Nh^d}, \tag{21}$$

where $\omega_y(t) \coloneqq \sup\{\|y(\mathbf{x}) - y(\mathbf{x}')\| : \mathbf{x}, \mathbf{x}' \in I_R,\ \|\mathbf{x} - \mathbf{x}'\| \leq t\}$ is the modulus of continuity of $y(\cdot)$.

The bound (21) captures the standard bias-variance trade-off in nonparametric estimation [16]. The bias term $\omega_y(\sqrt{d}\,h)^2$ measures the spatial approximation error and vanishes as $h \to 0$. Conversely, the statistical variance scales as $1/(Nh^d)$; as the grid becomes finer, fewer samples fall into each cell. Driving the total error to zero therefore requires the dataset size N to grow strictly faster than the number of cells.

The continuity assumption on y can be relaxed at the cost of losing an explicit algebraic decay rate for the bias. If $y \in L^\infty(I_R; \mathbb{R}^d)$, the bound (21) holds with the bias replaced by $\|y - y_h\|^2_{L^2(\mu)}$, which still vanishes as $h \to 0$ (e.g., via standard martingale convergence on nested grids [13, Theorem 4.4.6]). For $y \in L^2(\mu; \mathbb{R}^d)$, the bias similarly vanishes, though extending the $(Nh^d)^{-1}$ variance bound generally requires further assumptions. We do not pursue this level of generality here.

Conversely, by imposing stronger regularity on $y(\cdot)$ one can quantify the modulus of continuity. We focus on the Hölder case, characterized by

$$\omega_y(t) \leqslant L_y t^\alpha \qquad \text{for some } \alpha \in (0, 1] \text{ and } L_y > 0.$$

Optimizing the grid resolution h yields the following minimal statistical rate.

Corollary 3.2. *Suppose $y \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$ for some $\alpha \in (0,1]$. Then, the estimator y_h from (17) satisfies*

$$\|y - y_h\|^2_{L^2(\mu)} \lesssim h^{2\alpha}. \tag{22}$$

Moreover, $h = (2R) \wedge (N^{-\frac{1}{2\alpha+d}}) =: h_N$ yields the minimal convergence rate in (21),

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_{N,h_N}\|^2_{L^2(\mu)}\right] \lesssim N^{-\frac{2\alpha}{2\alpha+d}}. \tag{23}$$

Remark 3.3 (Sample complexity)**.** *The approximation by Riemann sums on a fixed grid of size h yields (22). To ensure a squared error below $\varepsilon > 0$, the number of cells must scale as $K_h \gtrsim \varepsilon^{-\frac{d}{2\alpha}}$. Consequently, the sample complexity to guarantee an expected risk below ε in (23) scales as:*

$$N \gtrsim \varepsilon^{-\frac{2\alpha+d}{2\alpha}} = \varepsilon^{-1} \cdot \varepsilon^{-\frac{d}{2\alpha}}. \tag{24}$$

We observe that the sample complexity naturally decomposes into the product of the Monte Carlo rate ε^{-1} (associated with estimating a mean with fixed variance) and a geometric factor $\varepsilon^{-d/(2\alpha)}$. This reflects the curse of dimensionality: whereas numerical integration computes a single global average, function recovery requires estimating K_h distinct local averages simultaneously to resolve the spatial structure of $y(\cdot)$ in $\mathbb{R}^d$.

Remark 3.4 (Minimax optimality)**.** *In terms of the number of cells $K_h \asymp h^{-d}$, the bias estimate (22) yields*

$$\|y - y_h\|_{L^2(\mu)} \lesssim K_h^{-\alpha/d}.$$

This matches the optimal rate among all linear *approximation spaces of dimension K_h, as captured by the Kolmogorov n-width*

$$d_n(\mathcal{F}) := \inf_{\dim(V)=n} \sup_{f\in\mathcal{F}} \inf_{g\in V} \|f - g\|_{L^2(\mu)}.$$

For the isotropic $\mathscr{C}^{0,\alpha}$-Hölder ball $\mathcal{F}$, one has $d_n(\mathcal{F}) \asymp n^{-\alpha/d}$ [25], an order attained by uniform piecewise-constant grids such as y_h. Moreover, this exponent cannot be improved even by adaptive *(nonlinear) methods with continuous parameter selection: the corresponding manifold n-widths exhibit the exact same decay rate on Besov (and hence Hölder) balls [12, Eq. (9.4)]. Adaptivity can only yield strictly faster rates on smaller, spatially inhomogeneous target classes [11].*

The rate $N^{-\frac{2\alpha}{2\alpha+d}}$ matches the minimax lower bound for nonparametric regression under additive label noise [30]. When $\mathbf{y}_i = y(\mathbf{x}_i) +$ noise, individual data points are unreliable, making local spatial averaging the optimal strategy to cancel out stochastic fluctuations.

In contrast, our setting is strictly *noiseless* because $\mathbf{y}_i = y(\mathbf{x}_i)$. Here, spatial averaging is overly conservative and wastes the precision of perfect data. Enforcing *strict interpolation* bypasses the statistical variance penalty, yielding faster rates. For instance, assuming $y \in \mathscr{C}^{k,\alpha}(I_R;\mathbb{R}^d)$ with $k \geq 0$ integer and $\alpha \in (0,1]$, spline interpolants m_N achieve [6]

$$\|y - m_N\|^2_{L^2(\mu)} \lesssim \|y - m_N\|^2_{L^\infty(I_R)} \lesssim \left(\frac{\log N}{N}\right)^{\frac{2(k+\alpha)}{d}}. \tag{25}$$

To exploit the lack of noise, we turn to the simplest interpolation scheme: the nearest-neighbor estimator.

3.2 Nearest-neighbor estimator

We now partition I_R by assigning each point to its nearest neighbor within $\{\mathbf{x}_i\}_{i=1}^N$. Unlike the histogram approach, where the grid is fixed and independent of the data, this partition naturally adapts to the local density of $\mathcal{D}_N$: dense regions produce smaller cells, while sparse regions yield larger ones. This adaptivity is exactly what yields faster approximation rates.

Formally, let $\{V_i\}_{i=1}^N$ denote the covering of I_R by the closed Voronoi cells generated by $\{\mathbf{x}_i\}_{i=1}^N$:

$$V_i := \{\mathbf{x} \in I_R : \|\mathbf{x} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_j\| \text{ for all } j \neq i\}. \tag{26}$$

For any $\mathbf{x} \in I_R$, we define its nearest neighbor within the dataset by the function:

$$x_{\mathrm{NN}}(\mathbf{x}) := \operatorname*{arg\,min}_{\mathbf{x}_i \in \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}} \|\mathbf{x} - \mathbf{x}_i\|. \tag{27}$$

where any ties on the cell boundaries are broken arbitrarily. Using this mapping, we define the nearest-neighbor interpolant of y, denoted by y_N^V, as

$$y_N^V(\mathbf{x}) \coloneqq y\left(x_{\mathrm{NN}}(\mathbf{x})\right) = \sum_{i=1}^{N} y(\mathbf{x}_i)\, \mathbf{1}_{V_i}(\mathbf{x}). \tag{28}$$

A schematic representation of y_N^V for $d = 2$ can be seen in the rightmost column of Figure 3. The accuracy of y_N^V is governed purely by the covering radius of $\mathcal{D}_N$, defined by

$$R_N \coloneqq \sup_{\mathbf{x}\in I_R} \min_{i\in[N]} \|\mathbf{x} - \mathbf{x}_i\| = \sup_{\mathbf{x}\in I_R} \|\mathbf{x} - x_{\mathrm{NN}}(\mathbf{x})\|. \tag{29}$$

The smaller R_N, the closer every point in I_R is to some input $\mathbf{x}_i$. For i.i.d. samples drawn from a density bounded below on I_R, the expected scaling of R_N is characterized by [**26**, Theorem 2.1], which yields

$$\mathbb{E}_{\mathcal{D}_N}[R_N^q] \lesssim \left(\frac{\log N}{N}\right)^{q/d} \quad \text{for any } q > 0. \tag{30}$$

The following proposition uses this rate to quantify the L^2-error for Hölder targets:

Proposition 3.5. *Let $N \geq 2$ and $\mathcal{D}_N = \{(\mathbf{x}_i, y(\mathbf{x}_i))\}_{i=1}^N$ with $\mathbf{x}_1, \dots, \mathbf{x}_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{\mathrm{ac}}(I_R)$. Assume that μ admits a density ρ with $\inf_{\mathbf{x}\in I_R} \rho(\mathbf{x}) > 0$, and let $y \in \mathscr{C}^{0,\alpha}(I_R; \mathbb{R}^d)$ for some $\alpha \in (0, 1]$. Then the nearest-neighbor interpolant y_N^V defined in (28) satisfies*

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_N^V\|_{L^2(\mu)}^2\right] \lesssim \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{d}}. \tag{31}$$

Note that the rate in (31) is strictly faster than the corresponding rate of $O\left(N^{-\frac{2\alpha}{2\alpha+d}}\right)$ derived in (23).

4 Generalization via simultaneous cell controllability

The previous two sections provide the ingredients needed for our subsequent analysis. On the one hand, Section 2 shows that (SANODE) can map prescribed convex cells into small target balls. On the other hand, Section 3 provides nonparametric estimators with explicit statistical risks. We now combine both ingredients to derive generalization bounds.

4.1 General bounds

To derive quantitative rates, we assume that $\mathbf{x}_1, \dots, \mathbf{x}_N$ are drawn i.i.d. from μ.[1] This induces the expected population risk over all realizations of the training sample:

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \coloneqq \int_{I_R} \cdots \int_{I_R} \mathcal{R}\left(\theta_N(\mathcal{D}_N)\right) \mathrm{d}\mu(\mathbf{x}_1) \cdots \mathrm{d}\mu(\mathbf{x}_N). \tag{32}$$

Our approach proceeds in two steps. First, we construct from $\mathcal{D}_N$ a piecewise-constant proxy $y_N(\cdot)$ that estimates $y(\cdot)$. Second, we design a control θ_N such that $\Phi_T^{\theta_N}$ approximates y_N on most of the domain.

Let $\{P_k\}_{k=1}^K$ be a partition of I_R into convex cells, let $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$ be target values computed from $\mathcal{D}_N$ (e.g., via cell averages), and define the piecewise-constant estimator

$$y_N(\mathbf{x}) \coloneqq \sum_{k=1}^{K} \mathbf{r}_k\, \mathbf{1}_{P_k}(\mathbf{x}).$$

Because Φ_T^θ is a homeomorphism, it cannot uniformly approximate the jump discontinuities of y_N. We therefore restrict the accuracy requirements to regions safely away from the cell boundaries. For a margin $\delta > 0$, we define the trimmed cores P_k^δ and the boundary layer Ω_δ (illustrated in Figure 4):

$$P_k^\delta \coloneqq \{\mathbf{x} \in P_k :\ \mathrm{dist}(\mathbf{x}, \partial P_k) \geqslant \delta\}, \qquad \Omega_\delta \coloneqq I_R \setminus \bigcup_{k=1}^{K} P_k^\delta. \tag{33}$$

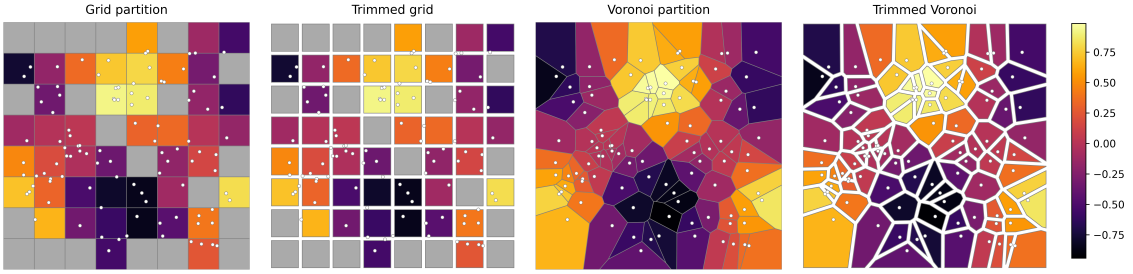

Figure 4: Comparison of grid and Voronoi partitions on the same dataset of $N = 100$ training points. **Left to right:** grid partition, trimmed grid partition, Voronoi partition, and trimmed Voronoi partition. In the trimmed cases, each cell is eroded inward by a margin $\delta > 0$, meaning points within distance δ of the boundary are removed to produce disjoint compact cores. The white regions indicate these removed boundary layers.

Since $y_N \equiv \mathbf{r}_k$ on each core P_k^δ, any control θ satisfying condition (10) up to tolerance η on these cores yields a natural decomposition of the error. This leads to the following model-agnostic bound.

Theorem 4.1 (Template bound)**.** *Assume $d \geqslant 2$ and that the flow Φ_T^θ of (5) is cell-wise controllable in the sense of Definition 2. Let $\mathcal{D}_N = \{(\mathbf{x}_i, y(\mathbf{x}_i))\}_{i=1}^N$ with $\mathbf{x}_1, \dots, \mathbf{x}_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{ac}(I_R)$, let $y \in \mathscr{C}^0(I_R; \mathbb{R}^d)$, and let y_N be a piecewise-constant estimator on $\{P_k\}_{k=1}^K$. Then, for any margin $\delta > 0$ and tolerance $\eta > 0$, there exists a data-dependent control $\theta_N = \theta_N(\mathcal{D}_N)$ such that*

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \mathbb{E}_{\mathcal{D}_N}\left[\|y - y_N\|_{L^2(\mu)}^2\right] + \mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] + \eta^2, \tag{34}$$

provided $\Phi_T^{\theta_N}$ remains uniformly bounded on I_R.

Proof. By the triangle inequality, for every control θ we have

$$\left\|\Phi_T^\theta - y\right\|_{L^2(\mu)}^2 \leqslant 2\left\|y - y_N\right\|_{L^2(\mu)}^2 + 2\left\|y_N - \Phi_T^\theta\right\|_{L^2(\mu)}^2.$$

To bound the realization error, we split the domain into the trimmed cores and the boundary layer:

$$\begin{aligned}\left\|y_N - \Phi_T^\theta\right\|_{L^2(\mu)}^2 &\leqslant \sum_{k=1}^K \mu(P_k^\delta) \sup_{\mathbf{x}\in P_k^\delta} \left\|\Phi_T^\theta(\mathbf{x}) - \mathbf{r}_k\right\|^2 + \int_{\Omega_\delta} \left\|y_N(\mathbf{x}) - \Phi_T^\theta(\mathbf{x})\right\|^2 \mathrm{d}\mu(\mathbf{x}) \\ &\leqslant \eta^2 + 2\mu(\Omega_\delta)\left(\|y_N\|_{L^\infty(I_R)}^2 + \|\Phi_T^\theta\|_{L^\infty(I_R)}^2\right),\end{aligned}$$

where the first term exploits the condition of Definition 2. By the uniform boundedness of both the estimator and the flow, taking the expectation over $\mathcal{D}_N$ yields the result. □

The template bound (34) cleanly isolates the three sources of error. The statistical error depends entirely on the chosen proxy y_N, the control error is dictated by the network capacity ($\eta \to 0$ as width grows), and the geometric error depends on the volume of the boundary layer Ω_δ. We treat the two proxy partitions of interest separately.

- **Uniform grid.** The boundary layer Ω_δ is simply a union of δ-strips along the fixed cubic cell faces.
- **Voronoi.** The boundary layer Ω_δ is the δ-neighborhood of the Voronoi skeleton $\Sigma_N := \bigcup_{i=1}^N \partial V_i$. Because the cells V_i are intersections of half-spaces, they are convex polytopes, ensuring that the trimmed cores V_i^δ are pairwise disjoint compact convex sets. This allows Definition 2 to be applied directly, even though Σ_N depends on the random spatial configuration of the sample points.

The next two lemmas establish quantitative bounds on the expected measure of Ω_δ for each type.

Lemma 4.2. *For $\mu \in \mathcal{P}_{ac}(I_R)$ with density bounded above, any partition $\{Q_k\}$ as in (15) with $h \in (0, 2R]$, and any margin $\delta \in (0, h/2)$, the boundary layer Ω_δ satisfies*

$$\mu(\Omega_\delta) \lesssim \frac{\delta}{h}. \tag{35}$$

[1]The deterministic core of the argument below, however, is purely geometric and applies to any fixed dataset.

Lemma 4.3. *For any integers $d \geq 2$ and $N \geq 1$, any $\mu \in \mathcal{P}_{ac}(I_R)$ with density bounded above and below by positive constants, and any sufficiently small $\delta > 0$, the boundary layer Ω_δ generated by $\{\mathbf{x}_i\}_{i=1}^N$ satisfies*

$$\mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] \lesssim \delta N^{1/d}. \tag{36}$$

Substituting these geometric bounds and the statistical risks from Section 3 into the decomposition (34) yields explicit rates for any system possessing the SCC property.

Corollary 4.4. *Under the hypotheses of Theorem 4.1, suppose that $y \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$ for some $\alpha \in (0,1]$. Furthermore, for the Voronoi estimator, assume that μ admits a strictly positive density on I_R. Then,*

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \begin{cases} h^{2\alpha} + \dfrac{1}{Nh^d} + \dfrac{\delta}{h} + \eta^2 & \text{(Histogram)}, \\ \left(\dfrac{\log N}{N}\right)^{2\alpha/d} + \delta N^{1/d} + \eta^2 & \text{(Voronoi)}. \end{cases} \tag{37}$$

4.2 Nonparametric rates with semi-autonomous neural ODEs

We now specialize these bounds to (SANODE). By Theorem 2.5, this model satisfies the SCC and admits realizing flows that remain uniformly bounded on I_R. To match the baseline rates, it suffices to balance the statistical, geometric, and control errors in (34) by appropriately scaling the margin δ, the tolerance η, and the width p_N. We first state the result for the histogram partition.

Proposition 4.5 (Histogram rate). *Let $d \geq 2$, let $y \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$ for some $\alpha \in (0,1]$, and let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be drawn i.i.d. from $\mu \in \mathcal{P}_{\mathrm{ac}}(I_R)$ with bounded density. Define*

$$h_N := N^{-\frac{1}{2\alpha+d}}, \qquad \delta_N := h_N^{1+2\alpha}, \tag{38}$$

and let p_N satisfy condition (11) for the histogram partition with cells $\{Q_k\}$ of side h_N and margin δ_N. Then, for all sufficiently large N, there exists a data-dependent control θ_N of width p_N for (SANODE) such that

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim N^{-\frac{2\alpha}{2\alpha+d}}.$$

Next, we establish the analogous result for the faster Voronoi partition.

Proposition 4.6 (Voronoi rate). *Let $d \geq 2$, let $y \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$ for some $\alpha \in (0,1]$, and let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be drawn i.i.d. from $\mu \in \mathcal{P}_{\mathrm{ac}}(I_R)$ with density bounded above and below on I_R. Define*

$$\delta_N := (\log N)^{\frac{2\alpha}{d}}\, N^{-\frac{2\alpha+1}{d}}, \tag{39}$$

and let p_N satisfy condition (11) for the Voronoi partition generated by $\{\mathbf{x}_i\}_{i=1}^N$ with margin δ_N. Then, for all sufficiently large N, there exists a data-dependent control θ_N of width p_N for (SANODE) such that

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{d}}.$$

The proofs of both propositions are provided in Section 7.3. In both cases, the width p_N ensures the existence of the required flow for each N. This required width scaling is directly tied to the geometry of the chosen partition—specifically, the minimum cell separation and maximum cell diameter. As partitions refine, these worst-case geometric constraints drive the super-polynomial growth discussed in Remark 2.6.

5 Numerical experiments

We complement the theoretical analysis of Section 4 with numerical experiments on (SANODE). The first experiment investigates whether the trained model can match nonparametric estimator baselines at a network width well below the theoretical prescription. This tests the sharpness of the sufficient width scaling. The second demonstrates that time-dependence of the vector field is necessary for topologically

non-trivial tasks, providing empirical support for the autonomous/non-autonomous separation discussed in Section 2.3. The code to reproduce these experiments is available in the GitHub repository ExGen.

Throughout, training points are sampled uniformly from $I_R = [-2, 2]^d$, and the population risk is estimated on a fixed test set using the squared ℓ^2-loss. All models are trained with the Adam optimizer at a fixed learning rate of 10^{-4}. Training ends after a maximum number of epochs (specified per experiment), or earlier if the loss falls below 10^{-6} or fails to improve by more than 10^{-8} over 5×10^3 consecutive steps.

5.1 Width scaling

The width prescriptions derived in Section 4 scale explosively with the dimension d, much like the storage costs of the nonparametric estimators they are calibrated against (see also Remark 2.7). In practice, however, neural ODEs are inherently nonlinear function approximators and may exploit this structure to achieve the same risk with far fewer parameters. We investigate to what extent this is the case by fixing the dataset size N and sweeping the network width p to determine the minimal width at which (SANODE) matches classical nonparametric baselines.

We consider two target functions $f, g \colon \mathbb{R}^d \to \mathbb{R}^d$:

- **Smooth target** ($\alpha = 1$): $f(\mathbf{x})^{(2k-1)} = \sin(x^{(2k-1)}) \cos(x^{(2k)})$ and $f(\mathbf{x})^{(2k)} = \cos(x^{(2k-1)}) \sin(x^{(2k)})$ for $k = 1, \dots, \lfloor d/2 \rfloor$, with $f(\mathbf{x})^{(d)} = x^{(d)}$ if d is odd.
- **Hölder-$1/2$ target** ($\alpha = 1/2$): $g(\mathbf{x})^{(i)} = \operatorname{sgn}(x^{(i)}) \sqrt{|x^{(i)}|}$ for $i \in [d]$.

We study two regimes: a low-dimensional setting with $d = 3$ and $N = 500$ training points, and a high-dimensional one where $d = 8$ with $N = 5000$ training points. The larger dataset for $d = 8$ compensates for the increased difficulty of the estimation problem in high dimension. The width is varied over a logarithmic grid $p \in \{5, 11, 23, 51, 109, 237, 512, 1024\}$. Training runs for at most 3×10^4 gradient steps, and the population risk is estimated on a fixed test set of 2×10^3 points. For reference, the histogram and Voronoi estimator errors are computed at the same (N, d) and displayed as horizontal baselines. The histogram estimator uses the optimal bandwidth $h_N = N^{-1/(2\alpha+d)}$, prescribed by Corollary 3.2.

In order to compare the complexity of (SANODE) with that of the nonparametric estimators, we report for each model the number of stored scalar values. For (SANODE) this coincides with the number of trainable parameters. For the histogram estimator this is at most Nd (at most N cells can be occupied, each storing a label in $\mathbb{R}^d$), and for the Voronoi estimator it is $2Nd$ (N input–label pairs, each in $\mathbb{R}^d$).

Figure 5 reports the test risk as a function of p for both targets and both dimensions. The quantitative results are collected in Table 1.

Case $d = 3$. For the smooth target, the (SANODE) already surpasses the histogram estimator at the smallest tested width ($p = 5$, 43 parameters) and converges to Voronoi-level error around $p = 512$ (4099 parameters), compared to the 3000 values stored by the Voronoi estimator. The crossover thus occurs at a parameter count comparable to the baseline, well below the theoretical prescription. The behavior on the Hölder-$1/2$ target is markedly different. Already at $p = 23$ (187 parameters), the (SANODE) achieves a test MSE of 1.2×10^{-4}, more than two orders of magnitude below the Voronoi baseline (2.8×10^{-2}), and the error plateaus around $p = 50$. The histogram estimator, with 500 stored values, achieves a test MSE far worse than the (SANODE) even at $p = 5$. This gap suggests that the (SANODE) exploits the component-wise, sign-symmetric structure of g, which the histogram partition cannot capture, and that the conservative worst-case width prescription is particularly loose for structured targets.

Case $d = 8$. The (SANODE) seems to do even better as the dimension increases, surpassing the histogram baseline (2.42×10^{-1}) around $p = 51$ (926 parameters) and approaching Voronoi-level error (8.95×10^{-2}) around $p = 237$ (4274 parameters), with almost an order of magnitude fewer parameters.

The picture for the Hölder-$1/2$ target is similar to the one in the low-dimensional setting. Despite the much higher dimension, the (SANODE) beats the histogram estimator (9.98×10^{-1}) already at $p = 5$ and drops below the Voronoi baseline (1.40×10^{-1}) between $p = 23$ and $p = 51$, reaching a plateau of around 10^{-5} from $p = 109$ onward, four orders of magnitude below the Voronoi baseline. This further confirms that the (SANODE) exploits the component-wise structure of g in a way that is largely insensitive to dimension, while the classical estimators degrade rapidly.

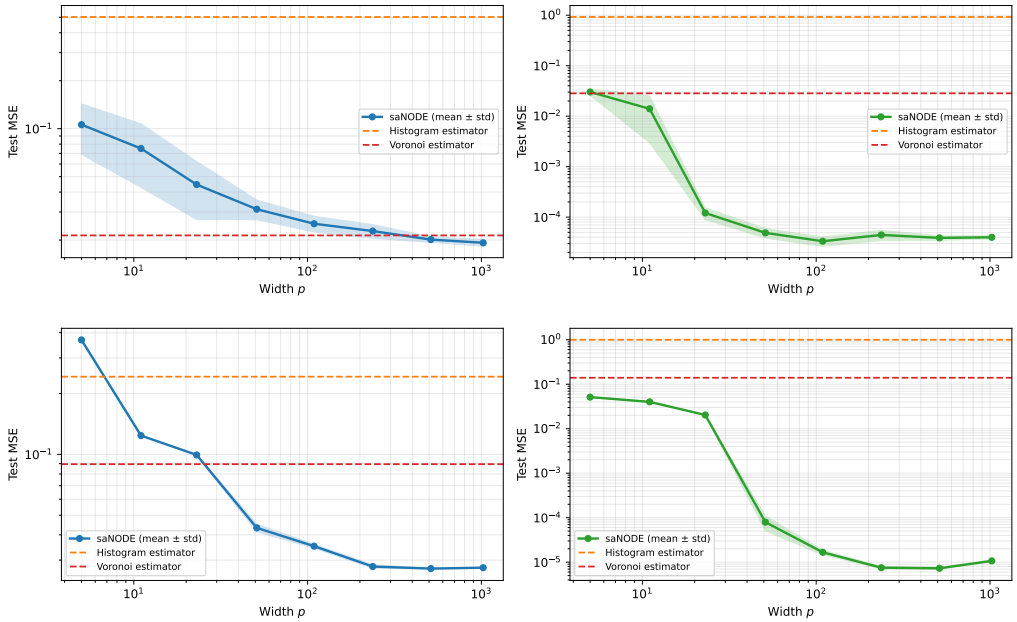

Figure 5: Test risk of (SANODE) as a function of network width p for the smooth target f (**left**) and the Hölder-$1/2$ target g (**right**), at $d = 3$, $N = 500$ (**top**) and $d = 8$, $N = 5000$ (**bottom**). Solid curves show the mean over 3 independent seeds and shaded bands indicate ± 1 standard deviation. Horizontal dashed lines mark the histogram and Voronoi estimator errors.

Table 1: Test MSE of (SANODE) versus p, averaged over 3 seeds. Complexity denotes the number of scalar degrees of freedom: trainable parameters for (SANODE), and Nd (histogram) or $2Nd$ (Voronoi) stored scalars for the baselines.

$d = 3$, $N = 500$

Method	p	Compl.	Smooth	Hölder-$1/2$
Histogram	–	1500	5.02×10^{-1}	9.22×10^{-1}
Voronoi	–	3000	2.14×10^{-2}	2.83×10^{-2}
(SANODE)	5	43	1.06×10^{-1}	3.02×10^{-2}
	11	91	7.51×10^{-2}	1.40×10^{-2}
	23	187	4.46×10^{-2}	1.21×10^{-4}
	51	411	3.12×10^{-2}	4.88×10^{-5}
	109	875	2.53×10^{-2}	3.33×10^{-5}
	237	1899	2.28×10^{-2}	4.44×10^{-5}
	512	4099	2.01×10^{-2}	3.88×10^{-5}
	1024	8195	1.92×10^{-2}	3.99×10^{-5}

$d = 8$, $N = 5000$

Method	p	Compl.	Smooth	Hölder-$1/2$
Histogram	–	40000	2.42×10^{-1}	9.98×10^{-1}
Voronoi	–	80000	8.95×10^{-2}	1.40×10^{-1}
(SANODE)	5	98	3.68×10^{-1}	5.13×10^{-2}
	11	206	1.24×10^{-1}	4.01×10^{-2}
	23	422	9.94×10^{-2}	2.04×10^{-2}
	51	926	4.33×10^{-2}	7.94×10^{-5}
	109	1970	3.52×10^{-2}	1.67×10^{-5}
	237	4274	2.79×10^{-2}	7.51×10^{-6}
	512	9224	2.72×10^{-2}	7.29×10^{-6}
	1024	18440	2.75×10^{-2}	1.07×10^{-5}

5.2 Necessity of time-dependence

As discussed in Section 2.3, because macroscopic trajectories cannot cross, autonomous flows face structural obstructions to cell routing whenever the target configuration is topologically incompatible with a continuous deformation of the input. We now provide direct empirical evidence for this limitation.

We consider a checkerboard sorting problem: given a $K \times K$ partition of $[-1, 1]^2$, the target assigns each point to $(+S, +S)$ if its cell indices (i, j) have an even sum, and to $(-S, -S)$ otherwise, setting $S = 0.7$. For $K \geqslant 2$, the alternating initial regions are heavily interleaved, forcing paths to cross to reach their respective targets. Time-dependent models like (SANODE) circumvent this spatial bottleneck by decoupling the motion in the extended space-time domain.

We compare (SANODE) against its autonomous counterpart (ANODE). To ensure a fair comparison, the width of (ANODE) is chosen so that the total parameter count matches the (SANODE):

$$\begin{cases} \text{(SANODE) params} = (2d+2)p + d \\ \text{(ANODE) params} = (2d+1)p + d \end{cases} \quad \Longrightarrow \quad p_{\text{aNODE}} = \left\lceil \frac{2d+2}{2d+1}\, p_{\text{saNODE}} \right\rceil .$$

Both models are trained on $N = 1000$ points with identical optimizer settings for at most 5×10^4 gradient steps, and evaluated on a separate test set of 10^3 points. We sweep the resolution $K \in \{2, 3, 4\}$.

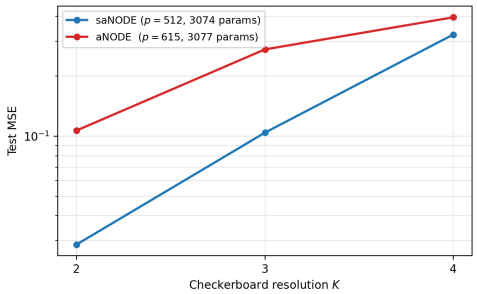

Figure 6: Test risk of (SANODE) and parameter-matched (ANODE) on the checkerboard sorting task with respect to K.

Figure 6 reports the test risk as a function of the grid resolution K. As expected, we observe a growing gap in test error between the two models. This behavior is consistent with the theoretical obstruction: the autonomous flow cannot cleanly resolve the increasingly fine interleaving of the cells.

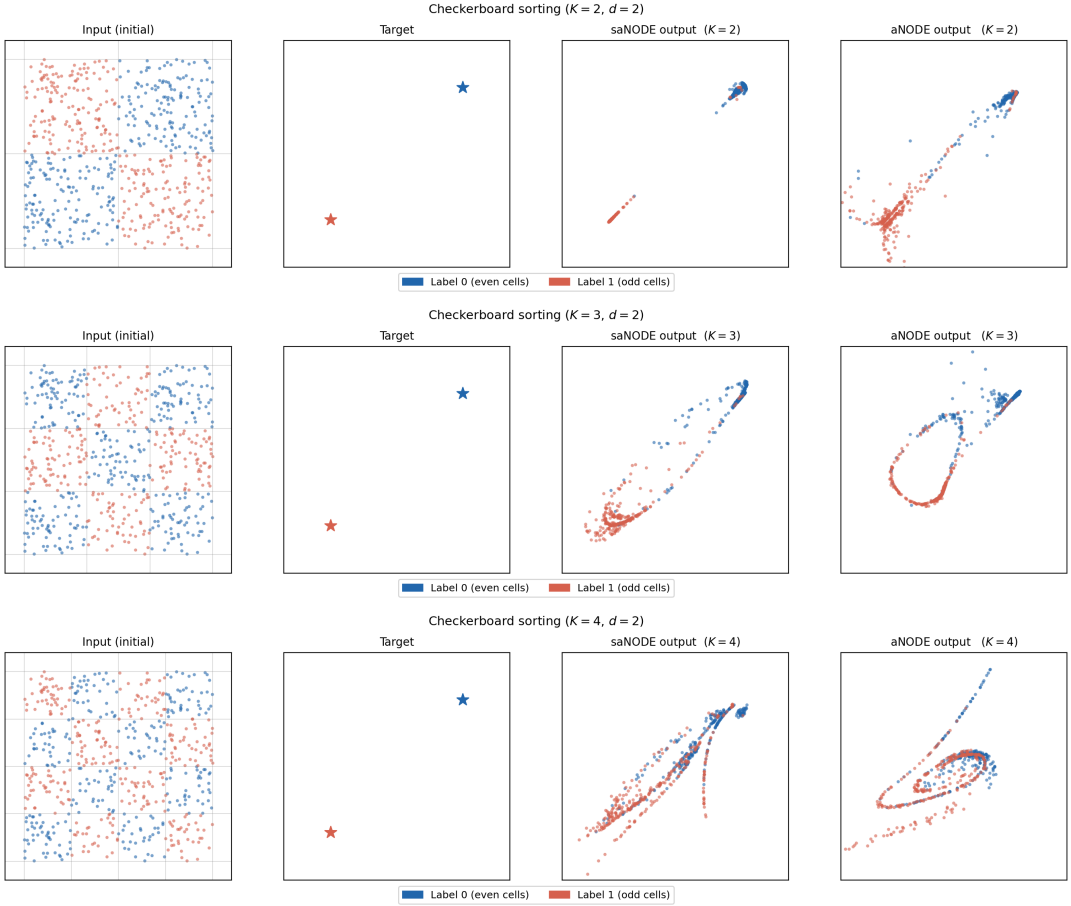

Figure 7: Checkerboard sorting for $d = 2$ at $K = 2$ (**top**), $K = 3$ (**middle**), and $K = 4$ (**bottom**). Each row shows, from left to right: input points, target assignment, (SANODE) output, and (ANODE) output. Points are colored by cell parity (even cells in blue, odd cells in red). The (SANODE) is able to separate the two classes at all resolutions, losing some precision only when $K = 4$. By contrast, the (ANODE) output becomes increasingly entangled as K grows.

Figure 7 provides a direct visualization of the learned mappings. The (SANODE) successfully routes the two classes into distinct clusters (showing only a minor loss of precision at $K = 4$). By contrast, the (ANODE) output remains entangled, with points from opposite classes overlapping. As the resolution in-

creases, this degradation becomes progressively more severe, confirming that the topological obstruction fundamentally limits autonomous models in region-routing tasks.

6 Conclusions and perspectives

We have developed a controllability-based framework to study generalization in neural ODEs. A central insight of our work is that point interpolation, while necessary for expressivity, cannot fully explain generalization. To bridge this gap, we introduced *simultaneous cell controllability*—the ability of the flow to compress entire input regions toward prescribed targets. We proved that the semi-autonomous model (SANODE) satisfies this property, enabling it to approximate piecewise-constant nonparametric estimators.

This connection yields explicit population-risk bounds. By constructing a single, data-dependent flow that realizes a statistically meaningful proxy, we demonstrate that exact interpolation and quantitative generalization can coexist. Our results thus establish that overparameterized neural ODEs can emulate nonparametric procedures, providing an achievability guarantee rather than an efficient scaling law. This approach departs from standard uniform generalization bounds, which evaluate the worst-case error over an entire hypothesis class and force a rigid trade-off between expressivity and global complexity.

Finally, we find that explicit time dependence is essential for our generalization mechanism. Without it, topological obstructions break simultaneous cell controllability, preventing autonomous architectures from routing macroscopic regions even when point interpolation is preserved.

We conclude by outlining natural directions opened for future work.

Saturation of regularity and higher-order flows. The bounds established here rely on piecewise-constant (order-0) approximations, whose statistical bias saturates at Lipschitz continuity ($\alpha = 1$). We deliberately favor this baseline because it naturally extends the exact UIP. Nevertheless, if the ground truth is smoother ($y \in \mathscr{C}^{m,\alpha}$), replacing the order-0 proxy with a degree-m cell-wise polynomial would yield faster statistical rates. Transferring this higher-order rate to Φ_T^θ, however, would require a complex analogue of Definition 2 capable of realizing local higher-degree polynomials up to tolerance η. Formulating and proving such higher-order controllability remains an interesting open problem.

Quantitative sharpness. While our proofs establish the vanishing of population risk, the required network width grows super-polynomially with N due to worst-case geometric bottlenecks in the partition. Closing the gap between these theoretical thresholds and the much smaller widths observed in practice requires sharper quantitative controllability estimates.

From static maps to trajectories. Finally, our controllability viewpoint suggests a natural extension from static regression to dynamical-system learning. While this work uses neural ODEs to approximate fixed terminal maps, applications often involve continuous sequences. Extending simultaneous cell controllability to steer entire trajectory tubes—rather than just terminal cells—would provide a framework for trajectory-level generalization bounds in sequence modeling.

7 Proofs

7.1 Proofs of Section 2

7.1.1 Proof of Theorem 2.3

We show that (SANODE) is simultaneously controllable, and we give an explicit construction. In particular, to steer N points to prescribed targets we use at most $p = 2N$ neurons. The proof requires several intermediate results.

A first key observation, that we will use throughout this section, is that each triple of parameters $(\mathbf{a}_j, b_j, c_j)$ defines a moving hyperplane

$$\mathsf{H}_j(t) \coloneqq \{x \in \mathbb{R}^d : \mathbf{a}_j \cdot \mathbf{x} + b_j t + c_j = 0\}, \tag{40}$$

which translates with constant normal velocity $-b_j \mathbf{a}_j / \|\mathbf{a}_j\|^2$.

The following preliminary lemma provides exact controllability of $d-1$ coordinates while keeping one coordinate fixed.

Lemma 7.1. *Let $N \geq 1$, $d \geq 2$, and let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ be any admissible finite dataset in the sense of (8) and such that $x_i^{(1)} \neq x_j^{(1)}$ for all $i \neq j$. For any $T > 0$ and any $R > \max_i x_i^{(1)}$, there exists*

$$\theta_R = (\mathbf{w}_i, \mathbf{a}_i, b_i, c_i)_{i=1}^N \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^N$$

such that the flow map $\Phi_T^{\theta_R}$ of (SANODE) satisfies, for each $i \in [N]$,

$$\left[\Phi_T^{\theta_R}(\mathbf{x}_i)\right]^{(1)} = x_i^{(1)} \qquad \text{and} \qquad \left[\Phi_T^{\theta_R}(\mathbf{x}_i)\right]^{(k)} = y_i^{(k)} \text{ for } k = 2, \dots, d$$

and, for each $i \in [N]$, we have $(\mathbf{a}_i \cdot \mathbf{x} + b_i t + c_i)_+ = 0$ for all $t \geq T$ and any $\mathbf{x} \in \mathbb{R}^d$ such that $x^{(1)} \leq R$.

Furthermore, for every $R_0 > \max_i x_i^{(1)}$, the family of controls $\{\theta_R\}_{R \geq R_0}$ can be chosen so that

$$\sup_{R \geq R_0} \max_{i \in [N]} \max_{t \in [0,T]} \max_{k=2,\dots,d} \left| \left[\Phi_t^{\theta_R}(\mathbf{x}_i)\right]^{(k)} \right| < \infty. \tag{41}$$

Proof. Up to relabeling the points, we may assume

$$x_1^{(1)} < x_2^{(1)} < \cdots < x_N^{(1)}. \tag{42}$$

We set $\mathbf{a}_i = \mathbf{e}_1$, and $w_i^{(1)} = 0$, for all $i \in [N]$. Then (SANODE) becomes

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^N \mathbf{w}_i \, (x^{(1)}(t) + b_i t + c_i)_+, \qquad t \in [0, T],$$

so, in components,

$$\dot{x}^{(1)}(t) = 0, \qquad \dot{x}^{(k)}(t) = \sum_{i=1}^N w_i^{(k)} \, (x^{(1)}(t) + b_i t + c_i)_+, \qquad k = 2, \dots, d. \tag{43}$$

In particular, $\left[\Phi_t^{\theta_R}(\mathbf{x}_i)\right]^{(1)} = x_i^{(1)}$ for all $t \in [0, T]$, so the first coordinate is frozen and (42) is preserved.

Step 1. Set

$$b_1 = \frac{x_1^{(1)} - 1 - R}{T}, \quad c_1 = 1 - x_1^{(1)}, \qquad \text{and} \qquad b_i = \frac{x_{i-1}^{(1)} - R}{T}, \quad c_i = -x_{i-1}^{(1)} \quad (i = 2, \dots, N).$$

Then each moving hyperplane $\mathsf{H}_i(t) := \{x \in \mathbb{R}^d \,:\, x^{(1)} + b_i t + c_i = 0\}$ satisfies:

(i) $\mathsf{H}_i(t)$ is orthogonal to $\mathbf{e}_1$ for all t.

(ii) For $i = 1$, $\mathsf{H}_1(0) : x^{(1)} = x_1^{(1)} - 1$ and $\mathsf{H}_1(T) : x^{(1)} = R$.

(iii) For $i \geq 2$, $\mathsf{H}_i(0) : x^{(1)} = x_{i-1}^{(1)}$ and $\mathsf{H}_i(T) : x^{(1)} = R$.

By the choice of (b_i, c_i) we have $b_i T + c_i = -R$ for all $i \in [N]$. Furthermore, the hyperplanes do not intersect on $[0, T)$.

Step 2. Fix $i \in [N]$. Since $x^{(1)}(t) \equiv x_i^{(1)}$, for each $k \geq 2$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \left[\Phi_t^{\theta_R}(\mathbf{x}_i)\right]^{(k)} = \sum_{j=1}^N w_j^{(k)} \, (x_i^{(1)} + b_j t + c_j)_+.$$

With the above choice of (b_j, c_j) and (42), if $j > i$ (hence $j \geq 2$), the argument $t \mapsto x_i^{(1)} + b_j t + c_j$ is equal to $x_i^{(1)} - x_{j-1}^{(1)} \leq 0$ at $t = 0$, and equal to $x_i^{(1)} - R < 0$ at $t = T$. By linearity,

$$x_i^{(1)} + b_j t + c_j \leq 0 \qquad \text{for all } t \in [0, T],$$

so neuron j is never active along the trajectory from $\mathbf{x}_i$. Therefore,

$$\left[\Phi_T^{\theta_R}(\mathbf{x}_i)\right]^{(k)} = x_i^{(k)} + \sum_{j=1}^{i} w_j^{(k)} \int_0^T (x_i^{(1)} + b_j t + c_j)_+ \,\mathrm{d}t =: x_i^{(k)} + \sum_{j=1}^{i} w_j^{(k)} M_{i,j}.$$

Conversely, if $j = 1$, the argument at $t = 0$ is $x_i^{(1)} + c_1 = x_i^{(1)} + 1 - x_1^{(1)} > 0$, while for $2 \le j \le i$ it is

$$x_i^{(1)} + c_j = x_i^{(1)} - x_{j-1}^{(1)} \ge x_i^{(1)} - x_{i-1}^{(1)} > 0.$$

Thus, for all $1 \le j \le i$, the integrand is strictly positive at $t = 0$ and continuous, so $M_{i,j} > 0$.

Step 3. Imposing $\left[\Phi_T^{\theta_R}(\mathbf{x}_i)\right]^{(k)} = y_i^{(k)}$ for $k = 2, \dots, d$ yields, for each fixed k,

$$y_i^{(k)} - x_i^{(k)} = \sum_{j=1}^{i} w_j^{(k)} M_{i,j}, \qquad i \in [N],$$

which is a lower-triangular linear system in $(w_1^{(k)}, \dots, w_N^{(k)})$ with strictly positive diagonal. Hence it has a unique solution for each $k = 2, \dots, d$.

Let $t \ge T$ and $\mathbf{x} \in \mathbb{R}^d$ with $x^{(1)} \le R$. Since $R > \max_i x_i^{(1)}$, Step 1 yields $b_i < 0$ and $b_i T + c_i = -R$ for all $i \in [N]$. Thus, for $t \ge T$ it holds that $b_i t + c_i \le -R$, which implies $(\mathbf{a}_i \cdot \mathbf{x} + b_i t + c_i)_+ = 0$ because

$$\mathbf{a}_i \cdot \mathbf{x} + b_i t + c_i \le x^{(1)} - R \le 0.$$

Finally, fix $R_0 > \max_i x_i^{(1)}$. The uniform boundedness follows from the explicit triangular construction. Indeed, the entries $M_{i,j}$ are of order R^{-1} as $R \to \infty$, while the corresponding weights grow at most linearly in R. Hence

$$w_j^{(k)} \int_0^t (x_i^{(1)} + b_j s + c_j)_+ \,\mathrm{d}s$$

remain uniformly bounded for $R \ge R_0$, uniformly in i, j, k and $t \in [0, T]$. Since the same quantities depend continuously on R and no $M_{i,i}$ vanishes for $R \ge R_0$, the bound extends to all $R \ge R_0$. □

Proof of Theorem 2.3. Since the dataset is finite and admissible, we can apply an arbitrarily small generic rotation to ensure that

$$x_i^{(1)} \ne x_j^{(1)} \quad \text{and} \quad y_i^{(2)} \ne y_j^{(2)} \quad \text{for } i \ne j.$$

Up to relabeling the pairs, we assume

$$y_1^{(2)} < y_2^{(2)} < \cdots < y_N^{(2)}. \tag{44}$$

We build $\theta \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^{2N}$ by concatenating two sets of N neurons which are active during the two phases given by the time intervals $[0, T/2]$ and $[T/2, T]$.

Step 1. Fix $R_0 > \max_i x_i^{(1)}$. By the last assertion of Lemma 7.1, applied on the interval $[0, T/2]$, there exists $C_0 > 0$ such that, for every $R_1 \ge R_0$,

$$\max_{i \in [N]} \max_{t \in [0, T/2]} \left| \left[\Phi_t^{\theta_{R_1}}(\mathbf{x}_i)\right]^{(2)} \right| \le C_0.$$

Choose

$$R_2 < (-C_0) \wedge \min_{i \in [N]} y_i^{(2)}. \tag{45}$$

Step 2. Now, for $\ell = N+1, \dots, 2N$, set $\mathbf{a}_\ell = -\mathbf{e}_2$ and $\mathbf{w}_\ell^{(k)} = 0$ for all $k = 2, \dots, d$. Choose (b_ℓ, c_ℓ) so that the hyperplanes $\mathsf{H}_\ell(t) \coloneqq \{x \in \mathbb{R}^d : -x^{(2)} + b_\ell t + c_\ell = 0\}$ all satisfy $\mathsf{H}_\ell(T/2) : x^{(2)} = R_2$. At $t = T$, they must reach the following prescribed levels: for $\ell = N+1, \dots, 2N-1$, impose

$$b_\ell \frac{T}{2} + c_\ell = R_2 \qquad \text{and} \qquad b_\ell T + c_\ell = y_{\ell+1-N}^{(2)},$$

and for $\ell = 2N$, impose

$$b_{2N}\,\frac{T}{2} + c_{2N} = R_2 \qquad \text{and} \qquad b_{2N}\,T + c_{2N} = y_N^{(2)} + 1.$$

Now consider the auxiliary dynamics on $[T/2, T]$ generated only by the second block, starting from

$$\mathbf{z}_i = (x_i^{(1)}, y_i^{(2)}, \dots, y_i^{(d)}).$$

Let $\Psi_i(t)$ denote the corresponding trajectory. Since $w_\ell^{(2)} = 0$ for all $\ell \geq N+1$ we have $[\Psi_i(t)]^{(2)} \equiv y_i^{(2)}$ on this interval, and the first coordinate satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\,[\Psi_i(t)]^{(1)} = \sum_{\ell=N+1}^{2N} w_\ell^{(1)}\,(-y_i^{(2)} + b_\ell t + c_\ell)_+.$$

Using (44) and the fact that $b_\ell t + c_\ell \leq b_\ell T + c_\ell$ for $t \in [T/2, T]$, we deduce that for $\ell = N+1, \dots, N+i-1$,

$$b_\ell t + c_\ell \leq b_\ell T + c_\ell = y_{\ell+1-N}^{(2)} \leq y_i^{(2)},$$

meaning $(-y_i^{(2)} + b_\ell t + c_\ell)_+ \equiv 0$ on $[T/2, T]$. Thus, only the neurons with $\ell \geq N+i$ contribute, yielding

$$[\Psi_i(T)]^{(1)} = x_i^{(1)} + \sum_{j=i}^{N} w_{N+j}^{(1)} \int_{T/2}^{T} (-y_i^{(2)} + b_{N+j} t + c_{N+j})_+ \,\mathrm{d}t =: x_i^{(1)} + \sum_{j=i}^{N} w_{N+j}^{(1)} \widetilde{M}_{i,j}.$$

By construction, $\widetilde{M}_{i,j} > 0$ for all $1 \leq i \leq j \leq N$. Imposing $[\Psi_i(T)]^{(1)} = y_i^{(1)}$ yields the upper-triangular linear system

$$y_i^{(1)} - x_i^{(1)} = \sum_{j=i}^{N} w_{N+j}^{(1)}\,\widetilde{M}_{i,j}, \qquad i \in [N],$$

which has a unique solution since the diagonal entries $\widetilde{M}_{i,i}$ are strictly positive. Let

$$B \coloneqq 1 + \max_{i\in[N]} \max_{t\in[T/2,T]} [\Psi_i(t)]^{(1)}.$$

Step 3. Choose $R_1 > \max\{R_0, B\}$ and apply Lemma 7.1 on $[0, T/2]$ with target level R_1 to construct the first block. Since $R_1 \geq R_0$, the bound defining C_0 applies. Hence, along the first-stage trajectories,

$$\left[\Phi_t^{\theta_{R_1}}(\mathbf{x}_i)\right]^{(2)} \geq -C_0 \qquad \text{for } t \in [0, T/2].$$

Moreover, since $b_\ell T/2 + c_\ell = R_2$, we have $b_\ell t + c_\ell \leq R_2$ for $t \in [0, T/2]$. Therefore, by $R_2 < -C_0$,

$$-\left[\Phi_t^{\theta_{R_1}}(\mathbf{x}_i)\right]^{(2)} + b_\ell t + c_\ell \leq C_0 + R_2 < 0,$$

so the second block is inactive on $[0, T/2]$. Since $R_1 > B$, the first block is inactive on $[T/2, T]$ along the auxiliary trajectories Ψ_i. By uniqueness, the full trajectories coincide with the concatenation of the first-stage trajectories and the auxiliary second-stage trajectories. Therefore

$$\Phi_T^\theta(\mathbf{x}_i) = \mathbf{y}_i, \qquad i \in [N].$$

□

Remark 7.2. *Theorem 2.3 extends to any neural ODE of the form*

$$\dot{\mathbf{x}}(t) = \sum_{j=1}^{p} \mathbf{w}_j\,(\mathbf{a}_j \cdot \mathbf{x}(t) + f_j(t) + c_j)_+, \qquad t \in [0, T], \tag{46}$$

provided each f_j is strictly monotone. Geometrically, the hyperplane velocities then vary over time rather than remain constant. We emphasize, however, that the linear choice $f_j(t) = b_j t$ is the simplest.

For instance, one may take $f_j(t) = d_j \tanh(b_j t)$, where $b_j > 0$ rescales time and $d_j \in \mathbb{R}$ bounds the hyperplane speed. In this case, the evolution splits into two phases:

***1. Control:** For small $b_j\,t$ one has $\tanh(b_j\,t) < 1$ and the exact-control objective is attained in this phase.*

***2. Stationary:** As $\tanh(b_i\,t) \to 1$, the system smoothly transitions to an autonomous regime.*

7.1.2 Proof of Theorem 2.9

The proof of Theorem 2.9 requires the following preliminary lemma.

Lemma 7.3. *Fix $d \geq 2$. Let $f_1 : \mathbb{R}^d \to \mathbb{R}^d$ be of the form (ANODE), let $\mathbf{u}_1, \dots, \mathbf{u}_d \in \mathbb{R}^d$ be linearly independent, and let $\alpha_i < \beta_i$ for each $i \in [d]$. For any $\delta_i > 0$ such that $\alpha_i + \delta_i < \beta_i - \delta_i$, define the sets*

$$K \coloneqq \{\mathbf{x} \in \mathbb{R}^d \,:\, \alpha_i \leq \mathbf{u}_i \cdot \mathbf{x} \leq \beta_i \,\forall i \in [d]\} \quad \text{and} \quad \tilde{K} \coloneqq \{\mathbf{x} \in \mathbb{R}^d \,:\, \alpha_i + \delta_i \leq \mathbf{u}_i \cdot \mathbf{x} \leq \beta_i - \delta_i \,\forall i \in [d]\}.$$

Then, there exists $f_2 : \mathbb{R}^d \to \mathbb{R}^d$ of the form (2ANODE) such that $f_2 \equiv f_1$ on $\tilde{K}$ and $f_2 \equiv 0$ on $\mathbb{R}^d \setminus K$.

Proof of Lemma 7.3. Let $\delta_{\min} \coloneqq \min_{i \in [d]} \delta_i > 0$. We define the barrier function $B : \mathbb{R}^d \to \mathbb{R}$ by

$$B(\mathbf{x}) \coloneqq \sum_{i=1}^{d} \left((\alpha_i + \delta_i - \mathbf{u}_i \cdot \mathbf{x})_+ + (\mathbf{u}_i \cdot \mathbf{x} - (\beta_i - \delta_i))_+ \right).$$

By definition, if $\mathbf{x} \in \tilde{K}$, all arguments inside the ReLUs are non-positive, yielding $B(\mathbf{x}) = 0$. Conversely, if $\mathbf{x} \notin K$, then for some index j, either $\mathbf{u}_j \cdot \mathbf{x} < \alpha_j$ or $\mathbf{u}_j \cdot \mathbf{x} > \beta_j$. In either case, the corresponding ReLU term is strictly greater than δ_j, ensuring that $B(\mathbf{x}) > \delta_{\min} > 0$ on $\mathbb{R}^d \setminus K$.

Let $f_1(\mathbf{x}) = \sum_{m=1}^{p_2} \mathbf{v}_m L_m(\mathbf{x})$, where $L_m(\mathbf{x}) \coloneqq (\mathbf{c}_m \cdot \mathbf{x} + d_m)_+$ and $\mathbf{v}_m \in \mathbb{R}^d$. Since $\mathbf{u}_1, \dots, \mathbf{u}_d$ span $\mathbb{R}^d$, then $B(\mathbf{x})$ grows linearly as $\|\mathbf{x}\| \to \infty$. Because each L_m has at most linear growth, the ratio L_m / B is bounded on $\mathbb{R}^d \setminus K$, allowing us to choose $\kappa > 0$ such that

$$L_m(\mathbf{x}) \leq \kappa B(\mathbf{x}) \quad \text{for all } \mathbf{x} \notin K \text{ and } m \in [p_2].$$

Finally, define

$$f_2(\mathbf{x}) \coloneqq \sum_{m=1}^{p_2} \mathbf{v}_m \, \left(L_m(\mathbf{x}) - \kappa B(\mathbf{x}) \right)_+ .$$

If $\mathbf{x} \in \tilde{K}$, then $B(\mathbf{x}) = 0$, which implies $f_2(\mathbf{x}) = \sum_m \mathbf{v}_m L_m(\mathbf{x}) = f_1(\mathbf{x})$. If $\mathbf{x} \notin K$, the bound $L_m(\mathbf{x}) \leq \kappa B(\mathbf{x})$ guarantees that every ReLU is inactive, yielding $f_2(\mathbf{x}) = 0$. Because $B(\mathbf{x})$ is a linear combination of ReLUs of affine functions, f_2 is representable in the two-layer form (2ANODE). □

Remark 7.4. *This result extends immediately to any pair of strictly nested convex polytopes: one replaces the defining inequalities of K in the construction of B with the affine constraints defining its facets.*

We are now ready to prove Theorem 2.9.

Proof of Theorem 2.9. Fix $T > 0$ and an admissible dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ satisfying (13). Without loss of generality, we may assume $\mathbf{x}_i \neq \mathbf{y}_i$ for all i. Let $\mathbf{u}_i \coloneqq (\mathbf{y}_i - \mathbf{x}_i)/T$ for each $i \in [N]$. The curve $\gamma_i(t) \coloneqq \mathbf{x}_i + t\mathbf{u}_i$ for $t \in [0, T]$ traces exactly the segment $S_i \coloneqq [\mathbf{x}_i, \mathbf{y}_i]$. Since all S_i are compact and pairwise disjoint, the minimum distance between them is positive:

$$\delta \coloneqq \min_{i \neq j} \operatorname{dist}(S_i, S_j) > 0.$$

We construct disjoint neighborhoods around each S_i as follows. For each i, fix an orthonormal basis $(\mathbf{e}_{i,1}, \dots, \mathbf{e}_{i,d})$ with $\mathbf{e}_{i,1} = \frac{\mathbf{y}_i - \mathbf{x}_i}{\|\mathbf{y}_i - \mathbf{x}_i\|}$. We define for each $\rho > 0$,

$$K_{i,\rho} \coloneqq \left\{ \mathbf{x} \in \mathbb{R}^d : -\rho \leq (\mathbf{x} - \mathbf{x}_i) \cdot \mathbf{e}_{i,1} \leq \|\mathbf{y}_i - \mathbf{x}_i\| + \rho \quad \text{and} \quad |(\mathbf{x} - \mathbf{x}_i) \cdot \mathbf{e}_{i,k}| \leq \rho \quad \text{for all } k \geq 2 \right\}.$$

Set $r \coloneqq \delta / (4\sqrt{d})$. By definition, every $\mathbf{x} \in S_i$ satisfies

$$0 \leq (\mathbf{x} - \mathbf{x}_i) \cdot \mathbf{e}_{i,1} \leq \|\mathbf{y}_i - \mathbf{x}_i\|, \qquad (\mathbf{x} - \mathbf{x}_i) \cdot \mathbf{e}_{i,k} = 0 \quad \text{for all } k \geq 2.$$

Hence $S_i \subset K_{i,r/2} \subset \operatorname{int}(K_{i,r})$. Moreover, the maximum distance from any $\mathbf{x} \in K_{i,r}$ to S_i is $\leqslant \sqrt{r^2 + (d-1)r^2} = r\sqrt{d} = \delta/4$, so $K_{i,r}$ lies entirely within the $\delta/4$-neighborhood of S_i.

By the triangle inequality, for any $i \neq j$,

$$\operatorname{dist}(K_{i,r}, K_{j,r}) \geq \operatorname{dist}(S_i, S_j) - 2(\delta/4) \geq \delta - \delta/2 = \delta/2 > 0.$$

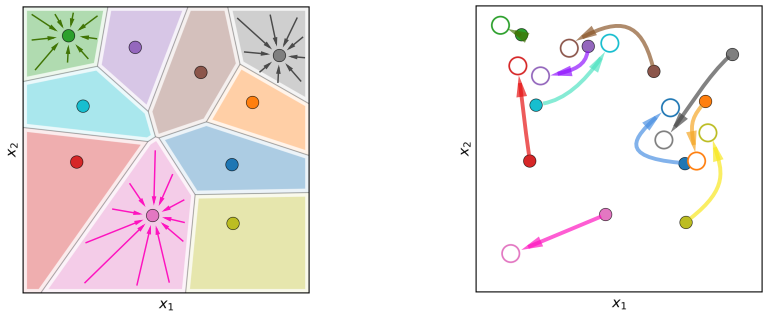

Figure 8: Two-step construction of the steering vector field. **Left:** Step 1 shows the initial pairwise disjoint convex sets A_k. The flow exponentially contracts these sets into small balls of radius δ centered around $\mathbf{x}_k$. **Right:** Step 2 depicts the rigid transport of these contracted balls along smooth, collision-free paths $\gamma_k(t)$ towards the target balls.

Hence, $K_{1,r}, \dots, K_{N,r}$ are pairwise disjoint.

We now construct the vector field. For each $i \in [N]$, we apply Lemma 7.3 to the constant field $g_i(\mathbf{x}) \equiv \mathbf{u}_i$—which is trivially of the form (ANODE)—setting the inner region to $K_{i,r/2}$ and the outer region to $K_{i,r}$. This yields a localized field v_i of the form (2ANODE) such that

$$v_i(\mathbf{x}) = \mathbf{u}_i \quad \text{for all } \mathbf{x} \in K_{i,r/2}, \qquad v_i(\mathbf{x}) = 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \setminus K_{i,r}.$$

Define the global vector field $v_\theta(\mathbf{x}) := \sum_{i=1}^N v_i(\mathbf{x})$. Since the neighborhoods $K_{1,r}, \dots, K_{N,r}$ are pairwise disjoint, for any fixed $i \in [N]$ we have

$$v_\theta(\mathbf{x}) = v_i(\mathbf{x}) + \sum_{j \neq i} \mathbf{0} = \mathbf{u}_i \qquad \text{for all } \mathbf{x} \in K_{i,r/2}.$$

Recall that the trace of the trajectory is $\gamma_i([0,T]) = S_i \subset K_{i,r/2}$. Therefore, the ODE dynamics exactly follow the constant velocity: $\dot\gamma_i(t) = v_\theta(\gamma_i(t)) = \mathbf{u}_i$. Integrating this yields $\Phi_T^\theta(\mathbf{x}_i) = \gamma_i(T) = \mathbf{y}_i$. Since this holds for all $i \in [N]$, we deduce that Φ_T^θ perfectly interpolates the dataset.

Finally, we determine the architecture. By the explicit construction in Lemma 7.3, since $g_i(\mathbf{x}) \equiv \mathbf{u}_i$ relies on a single constant term, the localized field takes the form

$$v_i(\mathbf{x}) = \mathbf{u}_i \left(1 - \kappa_i B_i(\mathbf{x})\right)_+,$$

where $B_i(\mathbf{x})$ is the barrier function. Because $K_{i,r}$ is defined by bounding the d projections onto the basis $\mathbf{e}_{i,k}$, the barrier B_i is a sum of exactly $2d$ ReLU terms (one for each upper and lower bound). Consequently, the global field v_θ exactly matches the two-layer architecture (2ANODE) with $p_1 = 2d$ and $p_2 = N$. □

7.1.3 Proof of Theorem 2.5

We first state some auxiliary lemmas. The first establishes that we can construct a smooth, compactly supported, time-varying vector field that steers each A_k to its target ball. We then give an explicit estimate of its Barron norm as a function of η and of the geometry of the sets and connecting trajectories. To this end, for a compactly supported vector field $v \in \mathscr{C}_c^\infty(\mathbb{R} \times \mathbb{R}^d; \mathbb{R}^d)$, we define the Barron norm as

$$\|v\|_{\mathcal{B}_2^{d+1}} := \sum_{\ell=1}^d \int_{\mathbb{R}\times\mathbb{R}^d} \|(\tau,\omega)\|_1^2 \, |\widehat{v^{(\ell)}}(\tau,\omega)| \, \mathrm{d}\tau \, \mathrm{d}\omega, \tag{47}$$

where $\hat v$ is the Fourier transform of v. This definition and the subsequent lemma build upon the framework of [19, Theorem 2], which we adapt to the setting of time-dependent vector fields.

Lemma 7.5 (Steering vector field). *Let $d \geq 2$, $K \geq 2$ and $T > 0$. Let $\{A_k\}_{k=1}^K$ be a family of pairwise disjoint compact convex subsets of $\mathbb{R}^d$, and let $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$ be pairwise distinct. Then, for every $\eta > 0$, there exists a vector field $v \in \mathscr{C}_c^\infty(\mathbb{R} \times \mathbb{R}^d; \mathbb{R}^d)$ whose flow map $\Phi_T : \mathbb{R}^d \to \mathbb{R}^d$ from time 0 to time T satisfies*

$$\Phi_T(A_k) \subset B(\mathbf{r}_k, \eta) \qquad \text{for all } k \in [K]. \tag{48}$$

Proof of Lemma 7.5. We first assume that the target points $\mathbf{r}_1, \dots, \mathbf{r}_K$ are pairwise distinct. Set

$$D_* := \max_{k\in[K]} \operatorname{diam}(A_k), \qquad s_* := \min_{i\neq j} \operatorname{dist}(A_i, A_j) > 0. \tag{49}$$

By definition of s_*, the open neighborhoods

$$V_k := \left\{ \mathbf{x} \in \mathbb{R}^d : \operatorname{dist}(\mathbf{x}, A_k) < \frac{s_*}{8} \right\} = A_k + \left(\frac{s_*}{8}\right) B(0,1) \tag{50}$$

are pairwise disjoint. Since $d \geq 2$, we may choose arbitrary points $\mathbf{x}_k \in A_k$ and connect each $\mathbf{x}_k$ to its target $\mathbf{r}_k$ via pairwise disjoint polygonal arcs. This is a standard topological consequence of working in dimensions $d \geq 2$, where finitely many paths can avoid intersections by introducing small perturbations. By smoothing the corners of these arcs and parameterizing them appropriately over time, we obtain smooth, pairwise disjoint curves $\gamma_k : [0,T] \to \mathbb{R}^d$ such that

$$\gamma_k(0) = \mathbf{x}_k, \quad \gamma_k \text{ is constant on } [0, 2T/3], \quad \gamma_k(T) = \mathbf{r}_k \tag{51}$$

and, by ensuring the polygonal arcs are contained within a reasonably tight bounding ball before smoothing, we also satisfy

$$\max_{t\in[0,T]} \|\gamma_k(t)\| < 1 + \max_{k\in[K]} \|\mathbf{r}_k\| + \max_{k\in[K]} \max_{\mathbf{x}\in A_k} \|\mathbf{x}\|. \tag{52}$$

Our proof is constructive. In particular, the final vector field will be the combination of a *compressing* vector field, collapsing each A_k to a small ball, and a *transporting* vector field, driving these balls to the targets $\mathbf{r}_k$. Figure 8 summarizes these steps visually. Accordingly, we divide this proof into three parts.

Compression. We fix a standard nonnegative radial mollifier $\psi \in \mathscr{C}_c^\infty(\mathbb{R}^d)$, with support in $B(0,1)$ and $\int \psi = 1$, and set

$$\psi_{s_*/32}(\mathbf{x}) := \left(\frac{32}{s_*}\right)^d \psi\left(\frac{32}{s_*}\mathbf{x}\right). \tag{53}$$

For each $k \in [K]$, define the smooth cutoff

$$\chi_k := \mathbf{1}_{A_k + (s_*/32)B(0,1)} * \psi_{s_*/32} \tag{54}$$

so that $\chi_k \in \mathscr{C}_c^\infty(\mathbb{R}^d; [0,1])$ with $\chi_k \equiv 1$ on A_k and $\operatorname{supp}(\chi_k) \subset A_k + (s_*/16)B(0,1) \subset V_k$. Since the supports of χ_k are pairwise disjoint, the vector field

$$u(\mathbf{x}) := \sum_{k=1}^K \chi_k(\mathbf{x})(\mathbf{x}_k - \mathbf{x}) \tag{55}$$

is well-defined and satisfies $u \in \mathscr{C}_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$. On each A_k, (55) simplifies exactly to $u(\mathbf{x}) = \mathbf{x}_k - \mathbf{x}$.

Transport. Let m be defined as

$$m := \min_{t\in[0,T]} \min_{i\neq j} \|\gamma_i(t) - \gamma_j(t)\|,$$

and denote

$$\delta := \min\left\{\eta, \frac{s_*}{8}, \frac{m}{8}\right\}, \qquad \lambda := \frac{3}{T} \log_+\left(\frac{D_*}{\delta}\right) \geq 0. \tag{56}$$

Furthermore, let $\xi \in \mathscr{C}_c^\infty(\mathbb{R}^d; [0,1])$ be a smooth radial cutoff with $\xi \equiv 1$ on $B(0,1)$ and $\operatorname{supp}\xi \subset B(0,2)$, and define

$$w(t,\mathbf{x}) := \sum_{k=1}^K \dot\gamma_k(t)\, \xi\left(\frac{8(\mathbf{x} - \gamma_k(t))}{m}\right), \qquad (t,\mathbf{x}) \in [0,T] \times \mathbb{R}^d. \tag{57}$$

For each $t \in [0,T]$, the spatial support of the k-th summand is contained in $B(\gamma_k(t), m/4)$, and these balls are pairwise disjoint. Moreover, $w(t,\mathbf{x}) = \dot\gamma_k(t)$ whenever $\|\mathbf{x} - \gamma_k(t)\| \leq \frac{m}{8}$.

Concatenation. To concatenate compression and transport, choose a nondecreasing $\varphi \in \mathscr{C}^\infty([0,T])$ with $\varphi(t) = 0$ on $[0,T/3]$ and $\varphi(t) = 1$ on $[2T/3,T]$, and define the time-dependent vector field

$$v_{[0,T]}(t,\mathbf{x}) := (1-\varphi(t))\,\lambda u(\mathbf{x}) + \varphi(t)\,w(t,\mathbf{x}), \qquad (t,\mathbf{x}) \in [0,T]\times\mathbb{R}^d. \tag{58}$$

Then $v_{[0,T]} \in \mathscr{C}_c^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$, and thus (58) generates a unique flow Φ_t on $[0,T]$.

We verify that Φ_t satisfies the condition (48). Since γ_k is constant on $[0,2T/3]$, we have $w \equiv 0$ on $[0,2T/3]\times\mathbb{R}^d$, and therefore (58) reduces there to $v_{[0,T]}(t,\mathbf{x}) = \lambda(1-\varphi(t))u(\mathbf{x})$. Besides, solving $\dot{\mathbf{x}}(t) = \lambda(1-\varphi(t))u(\mathbf{x})$, we see that the trajectory of any initial point $\mathbf{x}(0) \in A_k$ satisfies

$$\Phi_t(\mathbf{x}(0)) = \mathbf{x}_k + \exp\left(-\lambda\int_0^t (1-\varphi(s))\,\mathrm{d}s\right)(\mathbf{x}(0)-\mathbf{x}_k) \in A_k \qquad \text{for } t\in[0,2T/3]. \tag{59}$$

Since $\varphi(s) = 0$ on $[0,T/3]$ and $\varphi \le 1$ everywhere, the integral satisfies $\int_0^{2T/3}(1-\varphi(s))\,\mathrm{d}s \ge T/3$. Recalling our choice of λ in (56), this guarantees that at $t = 2T/3$,

$$\|\Phi_{2T/3}(\mathbf{x}(0)) - \mathbf{x}_k\| \le e^{-\lambda(T/3)}\|\mathbf{x}(0)-\mathbf{x}_k\| \le e^{-\lambda(T/3)}D_* \le \delta.$$

Consequently, we obtain

$$\Phi_{2T/3}(A_k) \subset B(\mathbf{x}_k,\delta). \tag{60}$$

On $[2T/3,T]$, we have $\varphi \equiv 1$, so $v_{[0,T]} = w$. Let $z_0 \in B(\mathbf{x}_k,\delta)$, and write $z_0 = \mathbf{x}_k + \mathbf{y} = \gamma_k(2T/3) + \mathbf{y}$ with $\|\mathbf{y}\| \le \delta \le m/8$. The path $z(t) := \gamma_k(t) + \mathbf{y}$ stays in the region where $w(t,\cdot) = \dot\gamma_k(t)$, hence solves the ODE; by uniqueness, the flow starting from z_0 equals $z(t)$, and in particular reaches $\mathbf{r}_k + \mathbf{y}$ at $t = T$. Combined with (60), this gives

$$\Phi_T(A_k) \subset B(\mathbf{r}_k,\delta) \subset B(\mathbf{r}_k,\eta).$$

To conclude, by a standard smooth extension-and-cutoff argument in the time variable, there exists a global field

$$v \in \mathscr{C}_c^\infty(\mathbb{R}\times\mathbb{R}^d;\mathbb{R}^d) \qquad \text{such that } v = v_{[0,T]} \qquad \text{on } [0,T]\times\mathbb{R}^d, \tag{61}$$

Since $v = v_{[0,T]}$ on $[0,T]\times\mathbb{R}^d$, the flow of v on $[0,T]$ coincides with Φ_t as above. In particular, (48) holds. □

Lemma 7.6 (Quantitative estimates on the steering field). *Let $v = v_\eta$ be the vector field constructed in the proof of Lemma 7.5 with tolerance η when $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$ are pairwise distinct (see (61)) and let Φ_t be its flow on $[0,T]$. Define the minimum path separation m and the maximum path derivative G associated with the trajectories γ_k as*

$$m := \min_{t\in[0,T]}\min_{i\ne j}\|\gamma_i(t)-\gamma_j(t)\|, \qquad G := 1 + \max_{1\le j\le d+5}\max_{k\in[K]}\left\|\frac{\mathrm{d}^j\gamma_k}{\mathrm{d}t^j}\right\|_{L^\infty(0,T)}, \tag{62}$$

and let δ be as in (56). Then there exists a constant $C_{d,T} > 0$, independent of η and $(A_k,\gamma_k)_k$, such that

$$\sup_{t\in[0,T]}\mathrm{Lip}_{\mathbf{x}}(v(t,\cdot)) \le C_{d,T}\left[\frac{G}{m} + \frac{1+D_*/s_*}{T}\log_+\left(\frac{D_*}{\delta}\right)\right]. \tag{63}$$

Moreover, the Barron norm of v, as defined in (47), satisfies

$$\|v\|_{\mathcal{B}_2^{d+1}} \le C_{d,T}\,K\left[TG^{d+5}(m^d+m^{-4}) + \frac{(D_*+s_*)^{d+1}(1+s_*^{-(d+4)})}{T}\log_+\left(\frac{D_*}{\delta}\right)\right]. \tag{64}$$

Finally, for $R > 0$ large enough so that $[-R,R]^d \supset \bigcup_{k\in[K]} A_k$, we also have:

$$\max_{(t,\mathbf{x})\in[0,T]\times[-R,R]^d}\|\Phi_t(\mathbf{x})\| \le 1 + \max_{k\in[K]}\|\mathbf{r}_k\| + \frac{3}{2}\sqrt{d}\,R. \tag{65}$$

Proof. Throughout the proof, C_d and $C_{d,T}$ denote positive constants that may change from line to line and depend only on d, T, and on the fixed smooth profiles ψ, ξ, and φ.

Lipschitz estimate (63). Using the decomposition (58) and the bound $0 \le \varphi \le 1$, we obtain

$$\sup_{t\in[0,T]} \mathrm{Lip}_{\mathbf{x}}(v(t,\cdot)) \le \lambda \|Du\|_{L^\infty(\mathbb{R}^d)} + \sup_{t\in[0,T]} \|D_{\mathbf{x}} w(t,\cdot)\|_{L^\infty(\mathbb{R}^d)}. \tag{66}$$

Recall the definition of u in (55). Writing $\chi_k = \mathbf{1}_{E_k} * \psi_{s_*/32}$ with $E_k := A_k + (s_*/32)B(0,1)$, by the classical Young's convolution inequality, we have

$$\|\nabla \chi_k\|_{L^\infty(\mathbb{R}^d)} = \|\mathbf{1}_{E_k} * \nabla \psi_{s_*/32}\|_{L^\infty(\mathbb{R}^d)} \le \|\mathbf{1}_{E_k}\|_{L^\infty} \|\nabla \psi_{s_*/32}\|_{L^1} \le C_d\, s_*^{-1}. \tag{67}$$

Since the functions χ_k have pairwise disjoint supports, and $\|\mathbf{x} - \mathbf{x}_k\| \le D_* + s_*/16$ for $\mathbf{x} \in \mathrm{supp}(\chi_k)$, we obtain

$$\|Du\|_{L^\infty(\mathbb{R}^d)} \le C_d \left(1 + \frac{D_*}{s_*}\right). \tag{68}$$

For the second term, differentiating (57) gives:

$$\partial_{x_j} w^{(\ell)}(t,\mathbf{x}) = \sum_{k=1}^{K} \dot\gamma_k^{(\ell)}(t)\, \frac{8}{m}\, \partial_j \xi\left(\frac{8(\mathbf{x} - \gamma_k(t))}{m}\right), \qquad \text{for all } j,\ell \in [d].$$

For any fixed t, the support of the k-th summand is contained in $B(\gamma_k(t), m/4)$. These balls are pairwise disjoint for all k, meaning at most one summand is nonzero at any given $(t,\mathbf{x})$. Since $|\dot\gamma_k(t)| \le G$ and $\|\nabla\xi\|_{L^\infty} < \infty$, we obtain

$$\sup_{t\in[0,T]} \|D_{\mathbf{x}} w(t,\cdot)\|_{L^\infty(\mathbb{R}^d)} \le C_d\, \frac{G}{m}. \tag{69}$$

Substituting (68), (69), and (56) into (66) yields the desired bound (63).

Barron estimate. Fix $n = d + 4$ and $g \in \mathscr{C}_c^\infty(\mathbb{R}^{d+1})$. Since $\frac{\|(\tau,\omega)\|_1^2}{(1+\|(\tau,\omega)\|_1)^n}$ is integrable on $\mathbb{R}^{d+1}$,

$$\int_{\mathbb{R}^{d+1}} \|(\tau,\omega)\|_1^2\, |\widehat{g}(\tau,\omega)|\, \mathrm{d}\tau\, \mathrm{d}\omega \le C_d \sup_{(\tau,\omega)\in\mathbb{R}^{d+1}} (1 + \|(\tau,\omega)\|_1)^n\, |\widehat{g}(\tau,\omega)|.$$

Expanding the weight as a sum of monomials and using the identity $(\tau,\omega)^\alpha \widehat{g} = c_\alpha \widehat{\partial^\alpha g}$ together with $\|\widehat{h}\|_{L^\infty} \le \|h\|_{L^1}$, we obtain

$$\int_{\mathbb{R}^{d+1}} \|(\tau,\omega)\|_1^2\, |\widehat{g}(\tau,\omega)|\, \mathrm{d}\tau\, \mathrm{d}\omega \le C_d \sum_{|\alpha|\le n} \|\partial^\alpha g\|_{L^1(\mathbb{R}^{d+1})}. \tag{70}$$

Moreover, for every component $\ell \in [d]$, the extension operator and the cutoff yield

$$\sum_{a+|\beta|\le n} \|\partial_t^a \partial_{\mathbf{x}}^\beta v^{(\ell)}\|_{L^1(\mathbb{R}\times\mathbb{R}^d)} \le C_{d,T} \sum_{a+|\beta|\le n} \|\partial_t^a \partial_{\mathbf{x}}^\beta v_{[0,T]}^{(\ell)}\|_{L^1((0,T)\times\mathbb{R}^d)}. \tag{71}$$

Applying (70) componentwise to v and using (71), we obtain

$$\|v\|_{\mathcal{B}_2^{d+1}} \le C_{d,T} \sum_{\ell=1}^{d} \sum_{a+|\beta|\le n} \|\partial_t^a \partial_{\mathbf{x}}^\beta v_{[0,T]}^{(\ell)}\|_{L^1((0,T)\times\mathbb{R}^d)}. \tag{72}$$

Thus, it suffices to bound the L^1 derivatives of $v_{[0,T]} = v^C + v^T$, where

$$v^C(t,\mathbf{x}) := (1 - \varphi(t))\, \lambda u(\mathbf{x}), \qquad v^T(t,\mathbf{x}) := \varphi(t)\, w(t,\mathbf{x}).$$

Since φ is a fixed smooth profile, its derivatives up to order n are bounded in $L^\infty(0,T)$ and in $L^1(0,T)$ by a constant $C_{d,T} > 0$; we will use this repeatedly below without further mention.

Compression. Recalling the definition of χ_k at (54) and of E_k above, Young's convolution inequality, combined with the isodiametric bound $|E_k| \le C_d(D_* + s_*)^d$ and the scaling $\|\partial^\gamma \psi_{s_*/32}\|_{L^1} \le C_d s_*^{-|\gamma|}$ yields, for every $|\gamma| \le n$,

$$\|\partial^\gamma \chi_k\|_{L^1(\mathbb{R}^d)} \le C_d\, (D_* + s_*)^d\, s_*^{-|\gamma|}. \tag{73}$$

Applying the product rule to $u^{(\ell)}(\mathbf{x}) = \sum_k \chi_k(\mathbf{x})(x_k^{(\ell)} - x^{(\ell)})$, and, using $\|\mathbf{x}_k - \mathbf{x}\| \le D_* + s_*$ for $\mathbf{x} \in \operatorname{supp}(\chi_k)$, (73) gives

$$\sum_{|\beta|\le n} \|\partial^\beta u^{(\ell)}\|_{L^1(\mathbb{R}^d)} \le C_d\, K\, (D_* + s_*)^{d+1} \left(1 + s_*^{-n}\right). \tag{74}$$

Multiplying by the temporal factor $\lambda(1-\varphi)$ and using the uniform bound on the L^1-derivatives of φ,

$$\sum_{a+|\beta|\le n} \|\partial_t^a \partial_x^\beta v^{C,(\ell)}\|_{L^1((0,T)\times\mathbb{R}^d)} \le C_{d,T}\, \lambda K\, (D_* + s_*)^{d+1} \left(1 + s_*^{-n}\right). \tag{75}$$

Transport. Differentiating $w^{(\ell)}(t,\mathbf{x}) = \sum_k \dot\gamma_k^{(\ell)}(t)\, \xi(8(\mathbf{x} - \gamma_k(t))/m)$ via the Leibniz and chain rules, every spatial derivative extracts a factor m^{-1}, and every time derivative either hits ξ (extracting another m^{-1}) or lands on the prefactor $\dot\gamma_k^{(\ell)}$. Furthermore, we note that

$$\operatorname{supp} \xi(8(\,\cdot\, - \gamma_k(t))/m) \subset B(\gamma_k(t), m/4),$$

and that the latter are pairwise disjoint at each t. Thus, at most one summand contributes at any $(t,\mathbf{x})$, so for all $a + |\beta| \le n$,

$$\left|\partial_t^a \partial_x^\beta w^{(\ell)}(t,\mathbf{x})\right| \le C_d\, G^{n+1} \sum_{r=0}^{a} m^{-(|\beta|+r)} \sum_{k=1}^{K} \mathbf{1}_{B(\gamma_k(t),m/4)}(\mathbf{x}). \tag{76}$$

Integrating over $(0,T)\times\mathbb{R}^d$, where each ball has volume $\lesssim m^d$, yields

$$\|\partial_t^a \partial_x^\beta w^{(\ell)}\|_{L^1((0,T)\times\mathbb{R}^d)} \le C_{d,T}\, K\, G^{n+1} \sum_{r=0}^{a} m^{d-|\beta|-r}. \tag{77}$$

Since $|\beta| + r \le n = d+4$, each exponent $d - |\beta| - r$ lies in $[-4, d]$, hence $m^{d-|\beta|-r} \le m^d + m^{-4}$. Summing over $a + |\beta| \le n$ and absorbing the φ-factor from $v^T = \varphi\, w$, we obtain

$$\sum_{a+|\beta|\le n} \|\partial_t^a \partial_x^\beta v^{T,(\ell)}\|_{L^1((0,T)\times\mathbb{R}^d)} \le C_{d,T}\, K\, G^{n+1}(m^d + m^{-4}). \tag{78}$$

Combining (75), (78), and (72) with $n = d+4$, and setting $\lambda \le (3/T)\log_+(D_*/\delta)$ from (56), proves (64).

Uniform bound. Let $\mathbf{x}_0 \in [-R,R]^d$. For $t \in [0, 2T/3]$, we have $v(t,\mathbf{x}) = \lambda(1-\varphi(t))u(\mathbf{x})$, since $w \equiv 0$ on $[0,2T/3]\times\mathbb{R}^d$. The flow either fixes $\mathbf{x}_0$ or moves it along a straight segment within the convex set $[-R,R]^d$. Therefore, $\|\Phi_t(\mathbf{x}_0)\| \le \sqrt{d}\, R$ for all $t \in [0, 2T/3]$.

For $t \in [2T/3, T]$ we have $v = w$, which vanishes outside $\bigcup_{k=1}^K B(\gamma_k(t), m/4)$. Therefore, whenever $\Phi_t(\mathbf{x}_0)$ is not stationary, it lies in $B(\gamma_k(t), m/4)$ for some k, yielding

$$\|\Phi_t(\mathbf{x}_0)\| \le \|\gamma_k(t)\| + \frac{m}{4}.$$

By (52) we have $\|\gamma_k(t)\| < 1 + \max_{j\in[K]} \|\mathbf{r}_j\| + \sqrt{d}\, R$. Additionally, since $\gamma_k(0) \in A_k \subset [-R,R]^d$, we trivially have $m \le 2\sqrt{d}\, R$. Combining these estimates gives

$$\|\Phi_t(\mathbf{x}_0)\| \le 1 + \max_{j\in[K]} \|\mathbf{r}_j\| + \frac{3}{2}\sqrt{d}\, R.$$

Since this bound dominates the compression phase and holds whenever the trajectory moves, we conclude

$$\max_{(t,\mathbf{x})\in[0,T]\times[-R,R]^d} \|\Phi_t(\mathbf{x})\| \le 1 + \max_{k\in[K]} \|\mathbf{r}_k\| + \frac{3}{2}\sqrt{d}\, R,$$

which proves (65). □

The next lemma establishes an $O(p^{-1/2})$ approximation bound for the flow generated by (SANODE). The result relies on [19, Theorem 2], which was subsequently adapted for dynamical systems in [21]. Here, we extend [21, Theorem 2.3] by making the dependence on the Barron norm of the reference vector field explicit.

Lemma 7.7. *Fix $d \geq 2$ and $T > 0$, let $v \in \mathscr{C}_c^\infty(\mathbb{R} \times \mathbb{R}^d; \mathbb{R}^d)$ with $\|v\|_{\mathcal{B}_2^{d+1}} < \infty$, and let $\Omega \subset \mathbb{R}^d$ be compact. Let $(\Psi_t)_{t\in[0,T]}$ be the flow induced on $[0,T]$ by $\dot{\mathbf{x}} = v(t,\mathbf{x})$, and assume that for some $R_\star \geq 1$,*

$$[0,T] \times \bigcup_{t\in[0,T]} \Psi_t(\Omega) \subset [-(R_\star - 1), R_\star - 1]^{d+1}. \tag{79}$$

Then there exists a constant $C_d > 0$, depending only on d, such that for every integer $p \geq 3$ satisfying

$$p > C_d^2\, R_\star^4\, T^2\, \|v\|_{\mathcal{B}_2^{d+1}}^2 \exp\left(2T \sup_{t\in[0,T]} \mathrm{Lip}_{\mathbf{x}}(v(t,\cdot))\right), \tag{80}$$

there exists a control θ of width p for the field v_θ of the form (SANODE) such that the flow Φ_t^θ satisfies

$$\sup_{\mathbf{x}\in\Omega} \sup_{t\in[0,T]} \|\Psi_t(\mathbf{x}) - \Phi_t^\theta(\mathbf{x})\| \leq \frac{C_d\, R_\star^2\, T\, \|v\|_{\mathcal{B}_2^{d+1}}}{\sqrt{p}} \exp\left(T \sup_{t\in[0,T]} \mathrm{Lip}_{\mathbf{x}}(v(t,\cdot))\right). \tag{81}$$

Proof. Denote $L := \sup_{t\in[0,T]} \mathrm{Lip}_{\mathbf{x}}(v(t,\cdot))$. By [19, Theorem 2], there exists a control θ of width p and a constant $C_d > 0$ such that[2]

$$\sup_{(t,\mathbf{x})\in[-R_\star,R_\star]^{d+1}} \|v(t,\mathbf{x}) - v_\theta(t,\mathbf{x})\| \leq \frac{C_d\, R_\star^2\, \|v\|_{\mathcal{B}_2^{d+1}}}{\sqrt{p}} =: \varepsilon.$$

For any $\mathbf{x} \in \Omega$, let $\tau \leq T$ be the maximal time such that $\Phi_t^\theta(\mathbf{x}) \in [-R_\star, R_\star]^d$ for all $t \in [0,\tau]$. On $[0,\tau]$, adding and subtracting $v(t, \Phi_t^\theta(\mathbf{x}))$ gives

$$\|\Psi_t(\mathbf{x}) - \Phi_t^\theta(\mathbf{x})\| \leq \int_0^t L\, \|\Psi_s(\mathbf{x}) - \Phi_s^\theta(\mathbf{x})\|\, \mathrm{d}s + \varepsilon t,$$

so Grönwall's lemma yields

$$\sup_{t\in[0,\tau]} \|\Psi_t(\mathbf{x}) - \Phi_t^\theta(\mathbf{x})\| \leq \varepsilon T e^{LT} < 1,$$

where the last inequality follows from (80). Hence $\Phi_t^\theta(\mathbf{x})$ stays within distance 1 of $\Psi_t(\mathbf{x})$, which by (79) lies in $[-(R_\star - 1), R_\star - 1]^d$. Thus $\Phi_t^\theta(\mathbf{x}) \in (-R_\star, R_\star)^d$ on $[0,\tau]$, forcing $\tau = T$ by continuity. Taking the supremum over $\mathbf{x} \in \Omega$ gives (81). □

We are now in a position to prove the theorem.

Proof of Theorem 2.5. Assume first that the targets $\mathbf{r}_k$ are pairwise distinct. Since $d \geq 2$, we can choose smooth, pairwise disjoint $\gamma_k : [0,T] \to \mathbb{R}^d$ satisfying (51)–(52). We use the construction in the proof of Lemma 7.5 with these fixed paths and tolerance $\eta/4$. Let v be the resulting field and $(\Psi_t)_{t\in[0,T]}$ its flow. Let s_* be defined by (49) and (m, G) by (62), set $\eta_0 := \min\{1, s_*/2, m/2\}$, and fix $\eta \in (0, \eta_0]$. Therefore,

$$\Psi_T(A_k) \subset B(\mathbf{r}_k, \eta/4) \qquad \text{for all } k \in [K]. \tag{82}$$

For $\mathfrak{M} > 0$ given by (12), it follows that:

$$[0,T] \times \bigcup_{t\in[0,T]} \Psi_t(I_R) \subset [-(\mathfrak{M} - 1), \mathfrak{M} - 1]^{d+1}. \tag{83}$$

[2] [19, Theorem 2] is scalar-valued, but applying it componentwise and distributing the hidden units among the d components only changes the dimensional constant C_d.

Moreover, if $\eta > 0$ is small enough that $2\eta < s_* \wedge m$, then $\delta = \min\left\{\frac{\eta}{4}, \frac{s_*}{8}, \frac{m}{8}\right\} = \frac{\eta}{4}$. Lemma 7.6 then yields:

$$\sup_{t\in[0,T]} \mathrm{Lip}_{\mathbf{x}}(v(t,\cdot)) \le L\,(1-\log\eta)\,, \qquad \|v\|_{\mathcal{B}_2^{d+1}} \le B\,(1-\log\eta)\,,$$

Increasing $\mathfrak{C}$, if necessary, we also ensure that the lower bound (11) implies the size condition (80) for every $\eta \in (0, \eta_0]$, for constants $L, B \ge 0$ that depend only on $d, T, R, (\mathbf{r}_k)_k, (A_k)_k$. Applying Lemma 7.7 with $\Omega = I_R$ and $R_\star = \mathfrak{M}$, we obtain that for any sufficiently large $p \ge 3$:

$$\begin{aligned}\sup_{\mathbf{x}\in I_R}\sup_{t\in[0,T]} \|\Psi_t(\mathbf{x}) - \Phi_t^\theta(\mathbf{x})\| &\le \frac{C_d\mathfrak{M}^2 TB(1-\log\eta)}{\sqrt{p}}\exp\left(LT(1-\log\eta)\right)\\ &\le \frac{C_d\mathfrak{M}^2 TBe^{LT}(1-\log\eta)}{\sqrt{p}}\eta^{-LT}.\end{aligned}$$

Now, choose $\mathfrak{C}$ sufficiently large so that $\mathfrak{C} > 2LT$ and $\sqrt{\mathfrak{C}} \ge 2C_d\mathfrak{M}^2 TBe^{LT}$. If p satisfies the bound in (11), we deduce:

$$\sup_{\mathbf{x}\in I_R}\sup_{t\in[0,T]} \|\Psi_t(\mathbf{x}) - \Phi_t^\theta(\mathbf{x})\| \le \frac{C_d\mathfrak{M}^2 TBe^{LT}}{\sqrt{\mathfrak{C}}}\eta^{1+\mathfrak{C}/2-LT} \le \frac{\eta}{2}.$$

For any $\mathbf{x} \in A_k$, combining this bound with (82) allows us to verify (i) via the triangle inequality:

$$\|\Phi_T^\theta(\mathbf{x}) - \mathbf{r}_k\| \le \|\Phi_T^\theta(\mathbf{x}) - \Psi_T(\mathbf{x})\| + \|\Psi_T(\mathbf{x}) - \mathbf{r}_k\| \le \frac{\eta}{2} + \frac{\eta}{4} < \eta.$$

Furthermore, using (83), we can easily verify (ii):

$$\|\Phi_T^\theta\|_{L^\infty(I_R)} \le \|\Psi_T\|_{L^\infty(I_R)} + \|\Phi_T^\theta - \Psi_T\|_{L^\infty(I_R)} \le (\mathfrak{M}-1) + \frac{\eta}{2} < \mathfrak{M},$$

which holds because $\eta \le 1$.

Finally, consider the case where the targets $\mathbf{r}_k$ are not necessarily pairwise distinct. We can slightly perturb them to choose auxiliary targets $\tilde{\mathbf{r}}_k$ that are pairwise distinct and satisfy:

$$\|\tilde{\mathbf{r}}_k - \mathbf{r}_k\| < \eta/4 \qquad \text{for all } k.$$

Applying the previous argument to these auxiliary targets $\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_K$ with a tighter tolerance of $\eta/2$, we obtain, for sufficiently large p, a control θ such that $\Phi_T^\theta(A_k) \subset B(\tilde{\mathbf{r}}_k, \eta/2)$. By the triangle inequality, it immediately follows that $\Phi_T^\theta(A_k) \subset B(\mathbf{r}_k, \eta)$. Since the mutual separation of the auxiliary targets may be of order η, the path-separation parameter m, and hence the constants produced by the previous argument, need not remain uniform as $\eta \downarrow 0$. □

7.2 Proofs of Section 3

Proof of Proposition 3.1. By the squared-sum inequality, the expected error can be decomposed into a bias and a variance component:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_{N,h}\|^2_{L^2(\mu)}\right] \le 2\,\|y - y_h\|^2_{L^2(\mu)} + 2\,\mathbb{E}_{\mathcal{D}_N}\left[\|y_h - y_{N,h}\|^2_{L^2(\mu)}\right],$$

where y_h and $y_{N,h}$ are the population and empirical estimators defined in (17) and (19), respectively. We bound each term separately.

Bias. Fix a cell Q_k with $p_k := \mu(Q_k) > 0$. The population coefficient $\mathbf{s}_k$ is the conditional expectation of $y(\cdot)$ on Q_k. By Jensen's inequality, for any $\mathbf{x} \in Q_k$:

$$\|y(\mathbf{x}) - \mathbf{s}_k\| = \left\|\frac{1}{p_k}\int_{Q_k}(y(\mathbf{x}) - y(\mathbf{x}'))\,\mathrm{d}\mu(\mathbf{x}')\right\| \le \frac{1}{p_k}\int_{Q_k}\|y(\mathbf{x}) - y(\mathbf{x}')\|\,\mathrm{d}\mu(\mathbf{x}').$$

For any $\mathbf{x}, \mathbf{x}' \in Q_k$, the distance satisfies: $\|\mathbf{x}-\mathbf{x}'\| \le \mathrm{diam}(Q_k) \le \sqrt{d}\,h$. Consequently, $\|y(\mathbf{x})-y(\mathbf{x}')\| \le \omega_y(\sqrt{d}\,h)$, which implies $\|y(\mathbf{x})-\mathbf{s}_k\| \le \omega_y(\sqrt{d}\,h)$. Since cells with $p_k = 0$ do not contribute to the $L^2(\mu)$ norm, integrating over the partition yields:

$$\|y - y_h\|^2_{L^2(\mu)} = \sum_{k\in\mathcal{K}}\int_{Q_k}\|y(\mathbf{x}) - \mathbf{s}_k\|^2\,\mathrm{d}\mu(\mathbf{x}) \le \omega_y(\sqrt{d}\,h)^2\sum_{k\in\mathcal{K}}\mu(Q_k) = \omega_y(\sqrt{d}\,h)^2, \tag{84}$$

where $\mathcal{K}$ is the index set defined in (14).

Variance. Since $y_h|_{Q_k} = \mathbf{s}_k$ and $y_{N,h}|_{Q_k} = \mathbf{S}_k$, the variance term becomes:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y_h - y_{N,h}\|^2_{L^2(\mu)}\right] = \sum_{k\in\mathcal{K}} p_k\, \mathbb{E}_{\mathcal{D}_N}\left[\|\mathbf{s}_k - \mathbf{S}_k\|^2\right]. \tag{85}$$

Fix an index k and let $N_k := \sum_{i=1}^N \mathbf{1}_{Q_k}(\mathbf{x}_i)$ denote the number of samples falling into Q_k. Because the inputs are drawn i.i.d. from μ, N_k follows a binomial distribution, $N_k \sim \mathrm{Binom}(N, p_k)$.

Conditionally on $N_k = n \geq 1$, the coefficient $\mathbf{S}_k$ is the empirical average of n i.i.d. random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ distributed as $y(\mathbf{X})$ for $\mathbf{X} \sim \mu(\cdot \mid Q_k)$. By definition, $\mathbb{E}[\mathbf{Y}_1] = \mathbf{s}_k$. The variance of this empirical average is exactly:

$$\mathbb{E}\left[\|\mathbf{S}_k - \mathbf{s}_k\|^2 \mid N_k = n\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n (\mathbf{Y}_j - \mathbf{s}_k)\right\|^2\right] = \frac{1}{n}\,\mathbb{E}\left[\|\mathbf{Y}_1 - \mathbf{s}_k\|^2\right] \leq \frac{\|y\|^2_{L^\infty(I_R)}}{n}.$$

If $N_k = 0$, we defined $\mathbf{S}_k = 0$, which gives $\|\mathbf{S}_k - \mathbf{s}_k\|^2 = \|\mathbf{s}_k\|^2 \leq \|y\|^2_{L^\infty(I_R)}$. Taking the unconditional expectation over the binomial variable N_k, we obtain:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|\mathbf{S}_k - \mathbf{s}_k\|^2\right] \leq \|y\|^2_{L^\infty(I_R)}\left(\mathbb{E}\left[\frac{\mathbf{1}_{\{N_k>0\}}}{N_k}\right] + \mathbb{P}(N_k = 0)\right).$$

We bound the two terms in the parenthesis. First, $\mathbb{P}(N_k = 0) = (1-p_k)^N \leq e^{-Np_k}$. Second, utilizing the elementary inequality $\frac{1}{n} \leq \frac{2}{n+1}$ for $n \geq 1$, we evaluate the expectation via an integral identity for the binomial generating function:

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{N_k>0\}}}{N_k}\right] \leq 2\,\mathbb{E}\left[\frac{1}{N_k+1}\right] = 2\int_0^1 \mathbb{E}[t^{N_k}]\,\mathrm{d}t = 2\int_0^1 (1-p_k+p_kt)^N\,\mathrm{d}t \leq \frac{2}{p_k(N+1)}.$$

Substituting these estimates into (85) yields:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y_h - y_{N,h}\|^2_{L^2(\mu)}\right] \leq \|y\|^2_{L^\infty(I_R)} \sum_{k\in\mathcal{K}}\left(\frac{2}{N+1} + p_k e^{-Np_k}\right).$$

Recall from (16) that the total number of cells is $K_h \lesssim h^{-d}$. Moreover, using $\sup_{p\geq 0} pe^{-Np} = (eN)^{-1}$, we can bound the sum uniformly:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y_h - y_{N,h}\|^2_{L^2(\mu)}\right] \leq \|y\|^2_{L^\infty(I_R)}\left(\frac{2K_h}{N+1} + \frac{K_h}{eN}\right) \lesssim \frac{\|y\|^2_{L^\infty(I_R)}K_h}{N} \lesssim \frac{\|y\|^2_{L^\infty(I_R)}}{Nh^d}.$$

Combining the bias and variance upper bounds directly gives (21). □

Proof of Corollary 3.2. Hölder regularity gives $\omega_y(t) \leqslant L_y t^\alpha$ where $L_y > 0$ is the Hölder constant of $y(\cdot)$ on I_R. Hence (84) gives

$$\|y - y_h\|^2_{L^2(\mu)} \leq \omega_y(\sqrt{d}\,h)^2 = L_y^2(\sqrt{d}\,h)^{2\alpha} = L_y^2 d^\alpha h^{2\alpha} \lesssim h^{2\alpha}.$$

Substituting this into (21),

$$\mathbb{E}_{\mathcal{D}_N}\|y - y_{N,h}\|^2_{L^2(\mu)} \lesssim h^{2\alpha} + (Nh^d)^{-1}.$$

The choice $h_N = (2R) \wedge N^{-1/(2\alpha+d)}$ balances both terms and gives

$$\mathbb{E}_{\mathcal{D}_N}\|y - y_{N,h_N}\|^2_{L^2(\mu)} \lesssim N^{-2\alpha/(2\alpha+d)}.$$ □

Proof of Proposition 3.5. If $y(\cdot)$ is α-Hölder with constant $L_y > 0$, then

$$\|y(\mathbf{x}) - y_N^V(\mathbf{x})\| = \|y(\mathbf{x}) - y(x_{\mathrm{NN}}(\mathbf{x}))\| \leq L_y\|\mathbf{x} - x_{\mathrm{NN}}(\mathbf{x})\|^\alpha \leq L_y R_N^\alpha, \tag{86}$$

where R_N is the covering radius defined in (29). Therefore,

$$\|y - y_N^V\|^2_{L^2(\mu)} \leq \|y - y_N^V\|^2_{L^\infty(I_R)} \leq L_y^2 R_N^{2\alpha}.$$

Since ρ is bounded below on the cube I_R, there exists $c > 0$ such that $\mu(B(\mathbf{x}, r) \cap I_R) \geq cr^d$ for all $\mathbf{x} \in I_R$ and all sufficiently small $r > 0$. Combining this estimate with (30) directly proves (31). □

7.3 Proofs of Section 4

Proof of Lemma 4.2. For a single cell Q_k of side length h, the volume of its trimmed boundary rim is $h^d - (h-2\delta)^d \lesssim h^{d-1}\delta$. Summing over the $K_h \asymp h^{-d}$ cells comprising the partition, the total volume is bounded by $|\Omega_\delta| \lesssim (h^{-d})(h^{d-1}\delta) = \delta/h$. Because the probability density of μ is bounded from above by a constant, the measure directly satisfies $\mu(\Omega_\delta) \lesssim |\Omega_\delta| \lesssim \delta/h$. □

To prove Lemma 4.3, we first isolate a preliminary estimate controlling the probability that the gap between the first two nearest-neighbor distances is small.

Lemma 7.8. *Assume $N \geq 2$, fix $\mathbf{x} \in I_R$, and let $D_1(\mathbf{x}) \leq D_2(\mathbf{x})$ be the distances from $\mathbf{x}$ to its nearest and second-nearest points within $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. For $r > 0$, denote $F_{\mathbf{x}}(r) := \mu(B(\mathbf{x}, r))$. Then, for every $\delta > 0$,*

$$\mathbb{P}_{\mathcal{D}_N}\left(D_2(\mathbf{x}) - D_1(\mathbf{x}) \leq 2\delta\right) \leq N(N-1)\int_0^\infty \left[F_{\mathbf{x}}(r+2\delta) - F_{\mathbf{x}}(r)\right]\left(1 - F_{\mathbf{x}}(r)\right)^{N-2} \mathrm{d}F_{\mathbf{x}}(r).$$

Proof. For each $i \in [N]$, set $Z_i := \|\mathbf{x} - \mathbf{x}_i\|$. Writing the order statistics as $Z_{[1]} \leq Z_{[2]} \leq \cdots \leq Z_{[N]}$, we have $D_1(\mathbf{x}) = Z_{[1]}$ and $D_2(\mathbf{x}) = Z_{[2]}$. By symmetry,

$$\mathbb{P}\left(D_2(\mathbf{x}) - D_1(\mathbf{x}) \leq 2\delta\right) \leq N\,\mathbb{P}\left(Z_1 = Z_{[1]},\ Z_{[2]} \leq Z_1 + 2\delta\right).$$

Conditioning with respect to Z_1, we see that on the event $\{Z_1 = r,\ Z_1 = Z_{[1]},\ Z_{[2]} \leq r + 2\delta\}$, at least one of the remaining $N-1$ distances must belong to $[r, r+2\delta]$, while all the others must be at least r. For a fixed index $j \in \{2, \dots, N\}$,

$$\mathbb{P}\left(r \leq Z_j \leq r + 2\delta\right) = F_{\mathbf{x}}(r+2\delta) - F_{\mathbf{x}}(r),$$

and for each $k \notin \{1, j\}$, $\mathbb{P}(Z_k \geq r) = 1 - F_{\mathbf{x}}(r)$. By independence, these probabilities multiply, and a union bound over the $N-1$ possible choices of j yields

$$\mathbb{P}\left(Z_1 = Z_{[1]},\ Z_{[2]} \leq r + 2\delta \mid Z_1 = r\right) \leq (N-1)\left[F_{\mathbf{x}}(r+2\delta) - F_{\mathbf{x}}(r)\right]\left(1 - F_{\mathbf{x}}(r)\right)^{N-2}.$$

Integrating with respect to the law of Z_1, whose distribution function is $F_{\mathbf{x}}$, proves the claim. □

Proof of Lemma 4.3. We first treat the case $N = 1$. Then $V_1 = I_R$, so $\Omega_\delta = \{\mathbf{x} \in I_R : \mathrm{dist}(\mathbf{x}, \partial I_R) < \delta\}$, and since μ has density bounded above, $\mathbb{E}_{\mathcal{D}_N}[\mu(\Omega_\delta)] = \mu(\Omega_\delta) \lesssim \delta \leq \delta N^{1/d}$.

We may therefore assume $N \geq 2$. Let $\mathbf{x} \sim \mu$ be an independent test point; by Fubini's theorem,

$$\mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] = \mathbb{P}\left(\mathbf{x} \in \Omega_\delta(\mathcal{D}_N)\right),$$

where the probability is taken jointly over $\mathcal{D}_N$ and $\mathbf{x}$. The bound is trivial whenever $\delta N^{1/d} \geq 1$ since probabilities are at most 1, so we restrict to

$$\delta N^{1/d} \leq 1. \tag{87}$$

Let $D_1(\mathbf{x})$ and $D_2(\mathbf{x})$ denote the distances from $\mathbf{x}$ to its nearest and second-nearest sample points among $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, respectively. We claim that

$$\{\mathbf{x} \in \Omega_\delta,\ \mathrm{dist}(\mathbf{x}, \partial I_R) > \delta\} \subset \{D_2(\mathbf{x}) - D_1(\mathbf{x}) \leq 2\delta\}. \tag{88}$$

Suppose that $\mathbf{x} \in \Omega_\delta$ and $\mathrm{dist}(\mathbf{x}, \partial I_R) > \delta$, and let V_i be the Voronoi cell containing $\mathbf{x}$. Since $\mathbf{x} \notin V_i^\delta$, there exists $\mathbf{z} \in \partial V_i$ with $\|\mathbf{x} - \mathbf{z}\| \leq \delta$. The constraint $\mathrm{dist}(\mathbf{x}, \partial I_R) > \delta$ excludes $\mathbf{z} \in \partial I_R$, so $\mathbf{z}$ lies on an internal Voronoi boundary and is equidistant from at least two sample points $\mathbf{x}_i, \mathbf{x}_j$. The triangle inequality gives

$$\|\mathbf{x} - \mathbf{x}_j\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_i\| + 2\|\mathbf{x} - \mathbf{z}\|.$$

Taking $\mathbf{x}_i$ as a nearest neighbor and $\mathbf{x}_j$ as a second-nearest neighbor of $\mathbf{x}$, we obtain $D_2(\mathbf{x}) - D_1(\mathbf{x}) \leq 2\|\mathbf{x} - \mathbf{z}\| \leq 2\delta$, which proves (88). Consequently,

$$\mathbb{P}\left(\mathbf{x} \in \Omega_\delta(\mathcal{D}_N)\right) \leq \mathbb{P}\left(\mathrm{dist}(\mathbf{x}, \partial I_R) \leq \delta\right) + \mathbb{P}\left(D_2(\mathbf{x}) - D_1(\mathbf{x}) \leq 2\delta\right). \tag{89}$$

Since μ has density bounded above and the Lebesgue measure of the δ-strip near ∂I_R is $O(\delta)$,

$$\mathbb{P}\left(\operatorname{dist}(\mathbf{x}, \partial I_R) \le \delta\right) \lesssim \delta. \tag{90}$$

For the second term, fix $\mathbf{x} \in I_R$ and define $F_{\mathbf{x}}(r) := \mu(B(\mathbf{x}, r))$. The density bound on μ controls the μ-mass of a spherical annulus by its Euclidean volume:

$$F_{\mathbf{x}}(r + 2\delta) - F_{\mathbf{x}}(r) \lesssim \delta\,(r + 2\delta)^{d-1} \lesssim \delta\, r^{d-1} + \delta^d. \tag{91}$$

Substituting (91) into Lemma 7.8 and using the identity $(N-1)\int_0^\infty (1 - F_{\mathbf{x}})^{N-2}\,\mathrm{d}F_{\mathbf{x}} = 1$ (immediate from the change of variable $s = F_{\mathbf{x}}(r)$),

$$\mathbb{P}\left(D_2(\mathbf{x}) - D_1(\mathbf{x}) \le 2\delta \mid \mathbf{x}\right) \lesssim N\delta\, I_{\mathbf{x}} + N\delta^d, \qquad I_{\mathbf{x}} := (N-1)\int_0^\infty r^{d-1}\left(1 - F_{\mathbf{x}}(r)\right)^{N-2}\,\mathrm{d}F_{\mathbf{x}}(r). \tag{92}$$

It remains to bound $I_{\mathbf{x}}$ uniformly in $\mathbf{x} \in I_R$. Stieltjes integration by parts, whose boundary terms vanish (at $r = 0$ since $d \ge 2$, at $r = \infty$ since $\operatorname{supp}\mu \subset I_R$), gives

$$I_{\mathbf{x}} = (d-1)\int_0^\infty r^{d-2}(1 - F_{\mathbf{x}}(r))^{N-1}\,\mathrm{d}r.$$

Let $\rho_{\min} > 0$ be a uniform lower bound for the density of μ. Since I_R is a cube, there exist constants $c > 0$ and $r_0 > 0$ such that $|B(\mathbf{x}, r) \cap I_R| \ge c\, r^d$ for all $\mathbf{x} \in I_R$ and $0 < r \le r_0$, hence $F_{\mathbf{x}}(r) \ge c\, r^d$ on the same range. Using $1 - z \le e^{-z}$, we get $(1 - F_{\mathbf{x}}(r))^{N-1} \le e^{-c(N-1)r^d}$ for $r \le r_0$, while for $r \ge r_0$, $1 - F_{\mathbf{x}}(r) \le 1 - F_{\mathbf{x}}(r_0)$ is uniformly bounded away from 1. Setting $D_R := \operatorname{diam}(I_R)$ and splitting at r_0,

$$I_{\mathbf{x}} \lesssim \int_0^{r_0} r^{d-2} e^{-c(N-1)r^d}\,\mathrm{d}r + \int_{r_0}^{D_R} r^{d-2}(1 - F_{\mathbf{x}}(r_0))^{N-1}\,\mathrm{d}r.$$

The change of variables $u = (N-1)r^d$ in the first integral yields a bound $\lesssim N^{-(d-1)/d}$, and the second is exponentially small in N and absorbed into the same bound. Hence

$$I_{\mathbf{x}} \lesssim N^{-(d-1)/d} \qquad \text{uniformly in } \mathbf{x} \in I_R. \tag{93}$$

Substituting (93) into (92) yields

$$\mathbb{P}\left(D_2(\mathbf{x}) - D_1(\mathbf{x}) \le 2\delta \mid \mathbf{x}\right) \lesssim \delta N^{1/d} + N\delta^d \lesssim \delta N^{1/d},$$

where the last step uses $N\delta^d = (\delta N^{1/d})^d \le \delta N^{1/d}$ from (87) and $d \ge 2$. Taking expectations and combining with (89) and (90), we conclude

$$\mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] \lesssim \delta + \delta N^{1/d} \lesssim \delta N^{1/d},$$

since $N^{1/d} \ge 1$. This proves (36). □

Proof of Corollary 4.4. For the histogram partition, apply Theorem 4.1 with $y_N = y_{N,h}$. Then Proposition 3.1 and Lemma 4.2 give

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_{N,h}\|^2_{L^2(\mu)}\right] \lesssim h^{2\alpha} + \frac{1}{Nh^d}, \qquad \mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] \lesssim \frac{\delta}{h}.$$

Substituting these bounds into (34) gives the first estimate.
For the Voronoi partition, apply Theorem 4.1 with $y_N = y_N^V$. The expectation bound in Proposition 3.5 and the boundary-layer estimate in Lemma 4.3 give

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_N^V\|^2_{L^2(\mu)}\right] \lesssim \left(\frac{\log N}{N}\right)^{2\alpha/d}, \qquad \mathbb{E}_{\mathcal{D}_N}[\mu(\Omega_\delta)] \lesssim \delta N^{1/d}.$$

Substituting these bounds into (34) gives the Voronoi estimate. □

Proofs of Propositions 4.5 and 4.6

Proof of Proposition 4.5. Define $\eta_N^2 := N^{-\frac{2\alpha}{2\alpha+d}}$, and $h_N := N^{-\frac{1}{2\alpha+d}}, \delta_N := h_N^{1+2\alpha}$, and apply the histogram bound in Corollary 4.4 with $h = h_N$ and $\delta = \delta_N$. By Theorem 2.5, there exists $p_N^* \in \mathbb{N}$ given by (11), such that for every $p_N \geq p_N^*$ one can choose a control for (SANODE) of width p_N with $\eta^2 \asymp N^{-\frac{2\alpha}{2\alpha+d}}$, whose flow is uniformly bounded on I_R. Moreover, Proposition 3.1 and Corollary 3.2 give

$$\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_{N,h_N}\|^2_{L^2(\mu)}\right] \lesssim h_N^{2\alpha} + \frac{1}{Nh_N^d} \asymp N^{-\frac{2\alpha}{2\alpha+d}},$$

and Lemma 4.2 yields $\mathbb{E}_{\mathcal{D}_N}[\mu(\Omega_{\delta_N})] \lesssim \delta_N h_N^{-1} = h_N^{2\alpha} = N^{-\frac{2\alpha}{2\alpha+d}}$. If $N > 2^{1+d/(2\alpha)}$ then $\delta_N < h_N/2$ and substituting these bounds into (37) yields the claim. □

Proof of Proposition 4.6. Let p_N satisfy condition (11) for the Voronoi partition generated by $\{\mathbf{x}_i\}_{i=1}^N$, with margin δ_N and tolerance η_N defined by

$$\eta_N^2 := \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{d}}, \qquad \delta_N := (\log N)^{\frac{2\alpha}{d}} N^{-\frac{2\alpha+1}{d}}.$$

Substituting the bounds (31) from Proposition 3.5 and (36) from Lemma 4.3 into (37) yields the conclusion. □

References

[1] A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. *Applied Mathematics & Optimization*, 92(2):33, 2025. doi: 10.1007/s00245-025-10299-7.

[2] 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. doi: 10.4208/jml.241114.

[3] A. Álvarez-López, A. H. Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural ODEs. *Neural Networks*, 180:106640, 2024. doi: 10.1016/j.neunet.2024.106640.

[4] F. Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991. doi: 10.1145/116873.116880.

[5] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[6] B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, and H. Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017. doi: 10.1016/j.jmva.2017.05.010.

[7] G. Biau and L. Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015. ISBN 9783319253862. doi: 10.1007/978-3-319-25388-6.

[8] L. Bleistein and A. Guilloux. On the generalization and approximation capacities of neural controlled differential equations. In *International Conference on Learning Representations*, 2024.

[9] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pages 6572–6583, 2018.

[10] J. Cheng, Q. Li, T. Lin, and Z. Shen. Interpolation, approximation, and controllability of deep neural networks. *SIAM Journal on Control and Optimization*, 63(1):625–649, 2025. doi: 10.1137/23M1599744.

[11] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Applied and Computational Harmonic Analysis*, 11(2):192–226, 2001. doi: 10.1006/acha.2001.0336.

[12] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. doi: 10.1017/S0962492900002816.

[13] R. Durrett. *Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019. ISBN 9781108591034. doi: 10.1017/9781108591034.

[**14**] W. E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017. doi: 10.1007/s40304-017-0103-z.

[**15**] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

[**16**] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer, 2002. doi: 10.1007/b97848.

[**17**] J. Jia, Z. Yang, M. Wang, K. Guo, J. Yang, X. Yu, and L. Guo. Feedback favors the generalization of Neural ODEs. In *International Conference on Learning Representations*, 2025.

[**18**] P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707, 2020.

[**19**] J. Klusowski and A. Barron. Approximation by combinations of relu and squared relu ridge functions with l1 and l0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018. doi: 10.1109/TIT.2018.2874447.

[**20**] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 2015.

[**21**] Z. Li, K. Liu, L. Liverani, and E. Zuazua. Universal approximation of dynamical systems by semiautonomous neural odes and applications. *SIAM Journal on Numerical Analysis*, 64:193–223, 2026. doi: 10.1137/24M1679690.

[**22**] P. Marion. Generalization Bounds for Neural Ordinary Differential Equations and Deep Residual Networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 48918–48938, 2023.

[**23**] S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama. Dissecting neural ODEs. In *Advances in Neural Information Processing Systems*, volume 33, pages 3952–3963, 2020.

[**24**] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

[**25**] A. Pinkus. *n-Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer, Berlin, Heidelberg, 1985. ISBN 978-3-540-13638-5. doi: 10.1007/978-3-642-69894-1.

[**26**] A. Reznikov and E. B. Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2016. doi: 10.1093/imrn/rnv342.

[**27**] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[**28**] D. Ruiz-Balet and E. Zuazua. Neural ODE Control for Classification, Approximation, and Transport. *SIAM Review*, 65(3):735–773, 2023. doi: 10.1137/21M1411433.

[**29**] E. Sontag and H. Sussmann. Complete controllability of continuous-time recurrent neural networks. *Systems & Control Letters*, 30(4):177–183, 1997. doi: https://doi.org/10.1016/S0167-6911(97)00002-9.

[**30**] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[**31**] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 978-0-387-79051-0.

[**32**] M. Verma and M. Kumar. Analysis of generalization capacities of Neural Ordinary Differential Equations. *Transactions on Machine Learning Research*, 2025.

[**33**] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.