NEURAL ODE CONTROL FOR CLASSIFICATION, APPROXIMATION AND TRANSPORT

DOMÈNEC RUIZ-BALET[†] AND ENRIQUE ZUAZUA[⋆†]

ABSTRACT. We analyze Neural Ordinary Differential Equations (NODEs) from a control the-oretical perspective to address some of the main properties and paradigms of Deep Learning (DL), in particular, data classification and universal approximation. These objectives are tackled and achieved from the perspective of the simultaneous control of systems of NODEs. For instance, in the context of classification, each item to be classified corresponds to a different initial datum for the control problem of the NODE, to be classified, all of them by the same common control, to the location (a subdomain of the euclidean space) asso-ciated to each label. Our proofs are genuinely nonlinear and constructive, allowing us to estimate the complexity of the control strategies we develop. The nonlinear nature of the activation functions governing the dynamics of NODEs under consideration plays a key role in our proofs, since it allows deforming half of the phase space while the other half remains invariant, a property that classical models in mechanics do not fulfill. This very property allows to build elementary controls inducing specific dynamics and transformations whose concatenation, along with properly chosen hyperplanes, allows achieving our goals in finitely many steps. The nonlinearity of the dynamics is assumed to be Lipschitz. Therefore, our results apply also in the particular case of the ReLU activation function. We also present the counterparts in the context of the control of neural transport equations, establishing a link between optimal transport and deep neural networks.

Keywords. Neural ODEs, simultaneous control, universal approximation, data classifi-cation, Deep Learning, transport equations, Wasserstein distance, Optimal Transport.

1. INTRODUCTION

In this paper, we prove that the dynamical control properties of Neural Ordinary Differential Equations (NODEs) allow for the understanding of some of the main Deep Neural Networks' properties. In particular, we will focus on data classification and universal approximation.

Given a dataset $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$, x_i being the input and y_i the corresponding label, Supervised Learning (SL) aims to find a map $\phi : \mathbb{R}^d \to \mathbb{R}^d$ that associates x to y for later use on the allocation of a label to new unknown data. In classification, the labels take values in a finite set of $\mathbb{R}^d$ (typically associated with a subset of the natural numbers), while in regression, the labels correspond to continuous values of $\mathbb{R}^d$.

Neural networks (NN) are widely used in this context. A shallow neural network can be written as a linear combination of the form:

$$(1.1) \qquad y(x) = \sum_{i=1}^n W_i \boldsymbol{\sigma}(A_i x + b_i)$$

†Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.
 Chair of Computational Mathematics, Fundación Deusto Av. de las Universidades 24, 48007 Bilbao, Basque-Country, Spain.
⋆Chair in Applied Analysis, Alexander von Humboldt-Professorship, Department of Data Science Friedrich-Alexander-Universität, Erlangen-Nürnberg, 91058 Erlangen, Germany.

where $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^d$ is the so called activation function (more details will be given later on), and $W_i, A_i \in \mathbb{R}^{d \times d}$, $b_i \in \mathbb{R}^d$, $i \in \{1, ..., n\}$.

In the seminal paper by Cybenko [10], it was proved that, under certain assumptions on $\boldsymbol{\sigma}$, any continuous function in a compact set can be approximated by NN of the form (1.1) in the supremum norm as $n \to \infty$. The product dn is the width of the network. This density result is the so called Universal Approximation Theorem and it has been the object of intensive study and several improvements and generalizations (see [19, 22, 29]).

A deep neural network is obtained from the composition of shallow neural networks of the type (1.1) and in recent years they have played an important role in diverse applications, [21]. Each of this composition steps is refereed to as a *layer* and the depth of the network is the total number of compositions operated. Deep neural networks can also be interpreted as discrete dynamical systems and can be employed, in particular, to prove universal approximation theorems, [11]. The parameters W, A and b entering in the design of the networks are often found via optimization, using, for instance, a least-squares error type functionals [8], in the so-called *learning* process.

Residual neural networks (ResNets) (see [18]) refer to the variant in which an "*inertia*" term is added in each layer, leading to a discrete dynamical system of the form

$$(1.2) \qquad\qquad x_{l+1} = x_l + hW_l\boldsymbol{\sigma}(A_l x_l + b_l),$$

where h is a positive real number. ResNets can also be understood as an Euler discretization scheme of an ODE, the so called Neural ODEs (NODEs) ([6, 17, 42]):

$$(1.3) \qquad\qquad \dot{x}(t) = W(t)\boldsymbol{\sigma}(A(t)x(t) + b(t)).$$

This paper aims to develop a dynamical control theoretical analysis of NODEs (1.3). The time-dependent coefficients $A(t)$, $W(t)$ and $b(t)$ are not chosen using least squares and optimization approaches (as in [14] and the references therein). Rather, we adopt a dynamic controllability approach according to which the time-dependent parameters $A(t)$, $W(t)$ and $b(t)$ are viewed as controls and are built in a constructive way, being defined in a piecewise-constant switching manner, adapted to the distribution of data to be classified. This allows for algorithmic constructions that permit to measure the complexity of the controls needed to achieve the main goals, in particular, classification and universal approximation. Our approach does not require either the regularity of the nonlinearity that classical ODE control methods need, often based on linearization and Lie brackets. In this way our results apply also in the case of the Rectified Linear Unit (ReLU) activation function which is Lipschitz but fails to be C^1.

The properties and problems we study are addressed in an ordered and systematic manner, according to their complexity.

The first one we address is the classification problem. From a control perspective, it can be formulated as a simultaneous control problem. The data to be classified are viewed as the initial data of the Cauchy problem associated to (1.3). Classification requires to build controls $A(t)$, $W(t)$ and $b(t)$ that simultaneously drive all initial data to their corresponding final destination, allocated depending on their label. The final targets that we prefix are disjoint sets corresponding to a partition of the space $\mathbb{R}^d$. The choice of the partition is not unique, and the specific control problem to be addressed depends on it. Although our methods allow to handle arbitrary partitions, for the sake of simplicity, we consider the particular one in which the space is partitioned in parallel strips, which suffices to achieve the goal of classification. In Theorem 1, Section 3, we develop our new constructive proof of classification. As mentioned above, our construction allows to master the complexity of the required classifying controls, in terms of the distribution and structure of the dataset to be classified.

Theorem 1 assures that each initial data can be driven simultaneously to the strip corresponding to its label. But, in fact, our constructions allow driving the system to much more precise targets so that each trajectory approaches arbitrarily a prescribed target point in $\mathbb{R}^d$. This is proved in Section 4, Theorem 2.

Both results hold in an arbitrary time horizon $T > 0$. This can be easily achieved by scaling arguments, the general case $T > 0$ being reduced to $T = 1$. Obviously, the L^∞-norm of the

controls employed depends then monotonically and linearly on $1/T$, T being the length of time-horizon.

From a control perspective, these results can be viewed as particular instances of simultaneous or ensemble controllability properties. This topic has been considered in the control community for a variety of systems [24, 25, 39]. From that perspective, the results in the present paper are new since the ODE system under consideration is the same for all initial data and the number N of the controlled initial data is arbitrarily large, while the number of available controls is limited: $2d^2 + d$. This kind of simultaneous control result would be impossible for linear finite or infinite-dimensional systems, where the simultaneous control never occurs when different trajectories solve the same system. The specific nature of the nonlinearity of the NODEs and, as explained above, the very structure of the activation function is the key of our construction.

Our proofs are genuinely nonlinear and strongly rely on the structure of the activation function $\boldsymbol{\sigma}$ that models the nonlinearity of the NODE (1.3) under consideration. The key ingredient of our proofs is that the activation function allows splitting the phase-space into two half-spaces, so that one remains invariant under the action of the control, while the other one evolves in the direction we wish, approaching the prescribed final target. Such construction is not possible for the classical nonlinear ODE systems in Mechanics in which all trajectories evolve simultaneously, driven by the nonlinear flow. The proofs are constructive, allowing us to estimate the complexity of the needed piecewise-constant controls for achieving such results. Moreover, our techniques are of dynamic nature, and they are not based on optimization or linearization arguments, that would not allow to obtain the global results we prove.

Once the classification problem is solved and its upgraded simultaneous controllability version is proved in Theorems 1 and 2, our second main objective is to show that these techniques, when refined, allow also to prove the Universal Approximation Theorem (Section 5). More precisely, in Theorem 4 we prove that, given a simple function in $f : \Omega \to \mathbb{R}^d$ (a linear finite combination of characteristic functions of measurable sets), under the assumption that characteristic sets have finite perimeter, for any bounded set $\Omega \subset \mathbb{R}^d$ and $\epsilon > 0$ arbitrary being given, we can find controls $A(t), W(t)$ and $b(t)$ (depending in particular on ϵ) such that the associated flow ϕ_T of (1.3) satisfies:

$$\|f - \phi_T(\cdot; A, W, b)\|_{L^2(\Omega)} < \epsilon.$$

The proof, which is of constructive nature, yields a connection between the complexity of the function to be approximated and the cost of control in terms of the box-counting dimension (see [15, Chapter 2]). The control complexity is measured in terms of its L^∞-norm and the number of switches and it depends on, from one side, the level of approximation ϵ we desire, and from the other, on the complexity of the target function. In fact, any simple function can be approximated and its control cost approximated provided that the box-counting dimension of the boundaries of its characteristic sets is smaller than the dimension of the ambient space.

Our third main objective is to formulate these results in terms of control of transport equations, where the dynamical control perspective is made more evident. In Section 6 (see Theorems 5 and 6), we prove that the Neural Transport equation

$$(1.4) \qquad \partial_t \rho + \mathrm{div}_x \big[\left(W(t) \boldsymbol{\sigma}(A(t)x + b(t)) \right) \rho \big] = 0$$

is approximately controllable. Furthermore, we will see that one can achieve a simultaneous control result for a Neural Transport system of equations of the form (1.4) under certain conditions on the initial data and targets (roughly, having disjoint supports), which mimic the configuration of classification problems.

1.1. **Roadmap and related work.** This paper is organized as follows. In Section 2, we introduce the notation and formulate the main problems more precisely, presenting some of the fundamental tools that will systematically be used in our proofs. Section 3 is devoted to the classification problem and in Section 4 we reinterpret and further develop it in the simultaneous control context. Section 5 is devoted to proving the Universal Approximation Theorem, while Section 6 to analyze neural transport equations. Section 7 is devoted to present some open problems and further directions of research.

For the sake of simplicity of the exposition most of our results are presented in the particular case of the ReLU activation function, although the methods we develop apply for larger classes of activation functions that will be made clear below.

We conclude this introduction with a discussion of some of the related literature.

Control issues for NODEs have been addressed from different perspectives. In [35], for instance, the authors analyze the classical control problem of NODEs. Note however that this is done in the context in which one typically considers one single trajectory and not the simultaneous control problems that classification requires.

In the context of simultaneous control of Neural ODEs, in [9] a Lie bracket technique is developed for certain smooth activation functions that allow to achieve universal interpolation results, similar to those in Section 4. We also refer to [1] for further developments on the geometric Lie bracket interpretation of these concepts in the context of classification, also using smooth vector fields. Note however that our approach can be implement for Lipschitz continuous activation functions as the ReLU as in Figure 2.1.

In [14], the authors show a simultaneous control result for sufficiently regular activations under smallness conditions on the targets, and employing linearization arguments, to later derive Turnpike results for NODEs. Our techniques, of constructive nature, are of independent interest and pave the path to analyze Turnpike properties of NODEs in a much broader context.

Recently, in [23], universal approximation results were proved by means of NODEs, see also [37, 38]. In Section 5 we provide a simpler and more comprehensive proof of those results which allows, in particular, to handle the complexity of the controls in terms of the data to be classified or the target functions.

The results in Section 6, which provide a transport formulation of Deep Learning, making use of the classical link between ODEs and transport equations, can be seen as bilinear-type simultaneous control results for Neural transport equations and systems (Theorems 5 and 6). The bilinear control of transport equations has also been considered (see, for instance, [12, 13]) with controls localized in space motivated by collective behavior problems [28]. The nature of the controls we employ here and the dynamics that the systems under consideration, exploiting the very features of activation functions, are completely different. Furthermore, the use of Neural transport equations allows a probabilistic formulation of the classification problem.

Our results for Neural Transport equations can also be related to optimal transport, that can be formulated as a minimal kinetic energy controllability problem as in [3], [2] and the references therein.

1.2. **Notation.** Throughout this article we will use the following notation.

- We understand by $E \lesssim_q F$ the existence of a constant $C(q)$ depending on q such that $E \leq C(q)F$. Analogously, we denote by $E \sim_q F$ the existence of $C(q)$ such that $E = C(q)F$.
- Let $x \in \mathbb{R}^d$, we denote by $x^{(k)}$ the k-th component of the vector x.
- Let $\omega \subset \mathbb{R}^d$ we denote by $\omega^{(k)}$ the following set

$$\omega^{(k)} := \{x \in \mathbb{R} : \exists y \in \omega \quad \text{such that } y^{(k)} = x\}.$$

- Given $\omega \subset \mathbb{R}^d$, we denote by diam the quantity

$$\mathrm{diam}(\omega) := \sup_{x_1, x_2 \in \omega} |x_1 - x_2|$$

and by $\mathrm{diam}_{(k)}$ the quantity

$$\mathrm{diam}_{(k)}(\omega) := \sup_{x_1, x_2 \in \omega} |x_1^{(k)} - x_2^{(k)}|.$$

2. Mathematical setting

2.1. **Setting.** Consider a Lipschitz activation function $\sigma : \mathbb{R} \to \mathbb{R}$. For the clarity of the exposition, most of the presentation will be developed in the context of the particular activation function

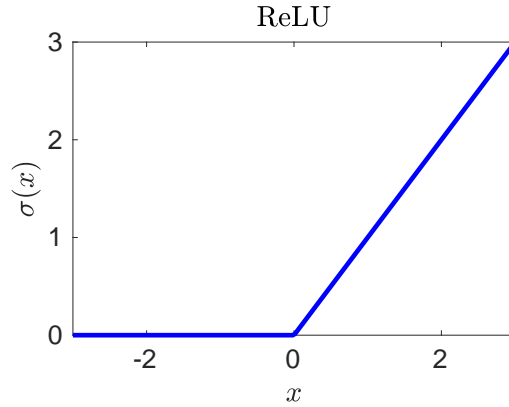$$(2.1) \qquad \qquad \sigma(x) = \max\{x, 0\},$$

FIGURE 2.1. The Rectified Linear Unit (ReLU) activation function (2.1)

the so called ReLU displayed in Figure 2.1. Later, in Remark 3.2, we shall discuss the extension to more general activation functions (roughly to functions which are Lipschitz, vanishing in $(-\infty, 0]$ and positive in $(0, +\infty)$.) We shall also employ the vector-valued version of the activation function, denoted by $\boldsymbol{\sigma}$, built applying σ to every component of a vector:

$$(2.2) \qquad \boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^d, \qquad \boldsymbol{\sigma}(x) = \begin{pmatrix} \sigma(x^{(1)}) \\ \sigma(x^{(2)}) \\ \vdots \\ \sigma(x^{(d)}) \end{pmatrix}.$$

Consider now the Neural ODE

$$(2.3) \qquad \begin{cases} \dot{x} = W(t)\boldsymbol{\sigma}(A(t)x + b(t)) \\ x(0) = x_0. \end{cases}$$

The coefficients A, W and b play the role of controls, $A, W \in L^\infty((0,T), \mathbb{R}^{d \times d})$, $b \in L^\infty((0,T), \mathbb{R}^d)$. With the notations above (2.3) is a system of d coupled differential equations.

The first question we address is the classification problem, which, in the context of the NODEs (2.3), can be reformulated as follows. We fix a time horizon $[0,T]$. Consider N distinct data $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ to be classified by their labels $\{y_i\}_{i=1}^N \subset \mathbb{R}^d$ using (2.3). In fact $\{x_i\}$, the data to be classified, will be taken to be the initial data at time $t = 0$ for the Cauchy problem (2.3), while $\{y_i\}$ are the corresponding labels that will allow us to define the targets at the final time $t = T$.

To reformulate this problem we consider a partition of $\mathbb{R}^d$ into M (the number of distinct labels) disjoint open sets with nonempty interior. We allocate one of these sets to each label. From a control theoretical perspective the classification problem can be reformulated as follows: to find controls W, A, b bringing each initial data to the corresponding set along the trajectories of (2.3).

Note that this problem can be considered as a simultaneous or ensemble control one, in the sense that the same controls are aimed to control all trajectories.

The Partition of the Euclidean Space. Our techniques can be applied for an arbitrary partition of the euclidean space. But the complexity of the control maps will depend on the structure, geometry and topology of the partition. To simplify the construction of the controls and our presentation we shall consider a particular partition of the space into parallel strips that we introduce now.

Let M be the number of different labels y_i, i.e.

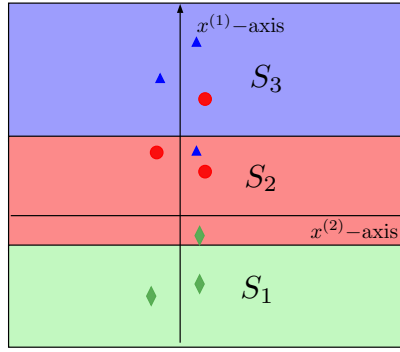$$M = |\{y_i : \quad i \in \{1, ..., N\}\}|.$$

FIGURE 2.2. Partition of the Euclidean plane $(d = 2)$ into three parallel strips $(M = 3)$. Nine points to be classified are denoted by three different colors and shapes corresponding to their labels (black triangles, red circles and green diamonds) and they are distributed in an unclassified manner. The goal of the approximation process, that we aim to conduct by controlling the flow of the Neural ODE, is to drive each subclass of three points to the corresponding label, i.e. to the strip corresponding to its color.

We split the space $\mathbb{R}^d$ into M disjoint sets $\mathcal{S} := \{S_m\}_{m=1}^M$, each set associated with one label. For simplicity of the presentation we set the partition into strips

$$(2.4) \qquad S_m := \{\alpha_{m-1} < x^{(1)} \le \alpha_m\} \qquad 1 \le m \le M$$

with $\alpha_0 = -\infty$, $\alpha_M = +\infty$ and $\alpha_m < \alpha_{m+1}$.

The Control Process. Given controls W, A, b we denote the solution of (2.3) by $\phi_T(x; A, W, b)$. Given initial data $\{x_i\}_{i=1}^N$ with labels $\{y_i\}_{i=1}^N$, the goal is to find controls W, A, b such that

$$\phi_T(x_i; A, W, b) \in \mathcal{S}_{m(i)} \qquad 1 \le i \le N$$

where $m(i) \in \{1, ..., M\}$ denotes the index of the label corresponding to the input x_i.

The Prediction. Once this map ϕ_T has been built out of the corresponding controls A, W, b, it can be used for prediction purposes, assigning to each new datum x the corresponding label y_m associated with the subset S_m of the partition satisfying $\phi_T(x; A, W, b) \in S_m$.

2.2. **The fundamental operations: Place, freeze, compress/expand and translate.** All our results will be proved by the same methodology. It consists of building piece-wise constant controls which, concatenated, will assure that the NODE fulfills the needed requirements. Each of these constant actions exploits the activation function's properties, the ReLU, in this presentation. In each step the available control parameters W, A, and b are chosen so that the ODE induces a transformation of the Euclidean space $\mathbb{R}^d$ which leaves invariant half of the space while the other evolves towards the desired direction. Obviously, the choice of the correct controls is essential to assure that the invariant and the moving half-spaces are the appropriate ones, together with the direction of motion.

For simplicity of the proofs we will also choose separating hyperplanes parallel to the Cartesian axes. The algorithm could be optimized by choosing oblique hyperplanes adapted to the datasets to be classified (in the context of the classical problem) or the function to be approximated (in the context of universal approximation). But, for simplicity of the presentation, we will only consider hyperplanes parallel to the Cartesian axes.

(1) Consider a hyperplane of the form:

$$\left\{ x^{(k)} - c = 0 \right\}.$$

This hyperplane can be represented as $Ax + b = 0$ with the choice of the controls A, b as below

$$A_{ij} = \pm \delta_{ik} \delta_{jk}, \quad b_i = \mp c \delta_{ik}.$$

The hyperplane divides the Euclidean space in two half-spaces. Here and in the sequel δ_{jk} stands for the Kronecker delta.

(2) When applying the vector-valued activation function $\boldsymbol{\sigma}$ to $Ax + b$, taking into account that the ReLU activation function σ vanishes for negative inputs, the choice of the sign of A and b is crucial to determine which half-space remains invariant;

$$\sigma(Ax + b)^{(i)} = \delta_{ik} \max\{\pm(x^{(k)} - c), 0\}.$$

See Figure 2.3.A.

(3) Choosing W to be a rotation matrix, when applied to $\boldsymbol{\sigma}(Ax + b)$, we can change the direction of the vector field in the direction we desire, see Figure 2.3.B. Note that, in particular, we can choose W so that the hyperplane is attractive or repulsive with respect to the moving half-space, but also we can generate a movement parallel to the hyperplane.

Throughout the proofs, we will only consider controls W such that the resulting vector fields point in some of the canonical euclidean directions. As it occurs for the choice of the hyperplanes, proofs could be implemented with a lower number of concatenations if W were chosen in oblique directions, depending on the dataset. But, for simplicity of the exposition, we shall only employ these elementary W's.

Summarizing, we can:

(1) **Place** and orient the dividing hyperplane arbitrarily.

(2) Choose the invariant and the active half-spaces. In particular we may choose the half-space that we aim to **freeze**.

(3) Choose the orientation of the vector-field $\boldsymbol{\sigma}(Ax + b)$ so that either
  - it induces an **expansion or contraction** of the active half-space.
  or
  - it induces a **translation** on the active half-space parallel to the dividing hyperplane.

As we shall see, the concatenation of these elementary transformations will suffice to achieve our classification and approximation goals.

3. Classification

This section is devoted to present Theorem 1 and its proof, which allows to classify an arbitrary finite dataset in strips, according to their label. The proof is based on combining the essential movements explained in subsection 2.2. For simplicity we state and prove the Theorem for the ReLU (2.1) activation function. However our arguments apply to more general activation functions σ, as we will make precise later in Remark 3.2.

Theorem 1. *Let $d \geq 2$ and $M \geq 2$ be natural numbers and let σ be as in (2.1). Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ be the dataset to be classified. Assume that $x_i \neq x_j$ if $i \neq j$. Then, for every $T > 0$, there exist control functions $A, W \in L^\infty\big((0,T); \mathbb{R}^{d \times d}\big)$ and $b \in L^\infty\big((0,T), \mathbb{R}^d\big)$ such that the flow associated to (2.3), when applied to all initial data $\{x_i\}_{i=1}^N$, classifies the trajectories by strips (2.4), according to their labels $\{y_i\}_{i=1}^N$, i.e.*

$$\phi_T(x_i; A, W, b) \in S_{m_i},$$

S_{m_i} being the subset corresponding to the label y_i, with $i \in \{1, ..., M\}$. Furthermore, controls are piecewise constant with a finite number of switches of the order of $\mathcal{O}(N)$. Therefore, they also lie in BV.

Remark 3.1 ($d = 1$). *Theorem 1 requires that $d \geq 2$. If $d = 1$ and $x_1 < x_2$, then for all choices of the controls W, A, b, the associated flow will fulfill that $\phi_T(x_1; A, W, b) < \phi_T(x_2; A, W, b)$ for all $T > 0$. This is due to the ODE formulation and the uniqueness of solutions for the corresponding Cauchy problem. Note that this restriction does not necessarily limit time-discrete NNs as, for instance, the one in (1.2), since it does not necessarily fulfill this monotonicity property when the step-size is large. Our proofs apply in an ODE setting and therefore are limited to $d \geq 2$. As a corollary, they apply as well for the time-discrete ResNets (1.2) provided the step-size h is small enough.*
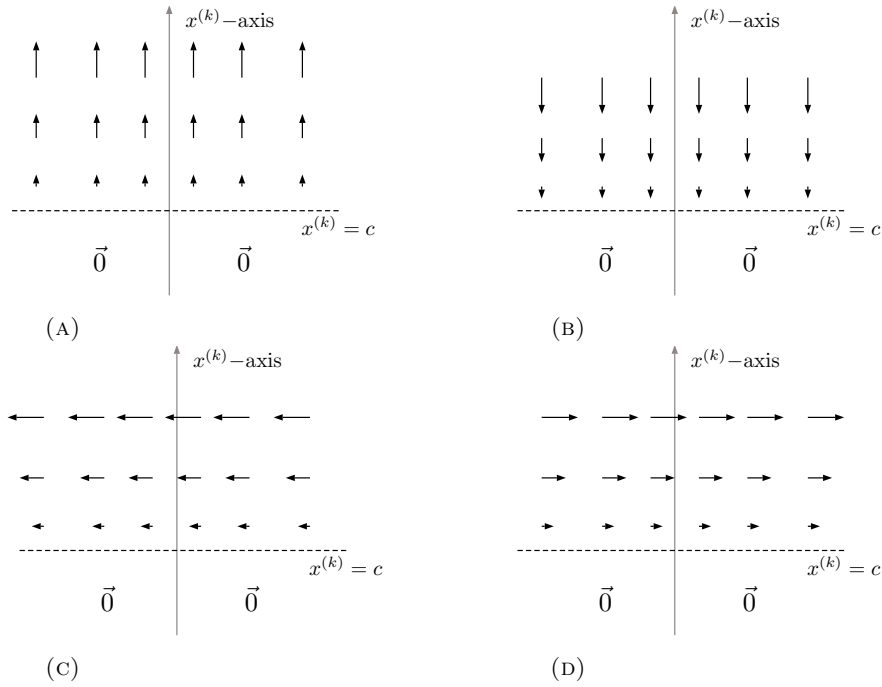
FIGURE 2.3. Representation of the elementary vector-fields achieved with $\boldsymbol{\sigma}(Ax+b)$ as in in Subsection 2.2. While the lower half-space remains invariant, the other one experiences a transformation of the form: (A) Expansion; (B) Contraction; (C) and (D) correspond to translations in opposite directions.

Proof. As we mentioned above, the actual value of the final time $T > 0$ is irrelevant since, by scaling, it can be set to be, in particular, $T = 1$.

Therefore we present a strategy that in a finite number of steps allows to achieve the classification in a finite time $T^* > 0$. The scheme of the proof is the following:

(1) **Preparation of the dataset**. The goal of this first step is to assure that the data to be classified fulfill the condition:

$$\forall i \neq j, \quad x_i^{(1)} \neq x_j^{(1)}. \tag{3.1}$$

Obviously, for that to hold, one needs to exploit the dynamics of the system by properly choosing the controls. This condition is relevant to assure that data can be efficiently separated by hyperplanes parallel to the Cartesian axes.

(2) **Classification** of the points so that the corresponding trajectories reach the allocated strips.

(3) **Time rescaling** allows to assure that the goal is achieved in the given final time $T > 0$.

We now present in more detail the process of each of these steps.

(1) **Preparation of the dataset.** We need the data to be classified to fulfill the condition (3.1). If the data $\{x_i\}_{i=1}^N$ do not fulfill this property, by a proper choice of the controls A, W, b in a time $\tau > 0$, we may assure that (3.1) is satisfied by $\phi_\tau(x_i; A, W, b)$, $i \in \{1, ..., N\}$.

The construction of the controls guaranteeing that this separation property is fulfilled can be made by induction.

- Suppose we have only two points, x_i and x_j, sharing the first component $x^{(1)}$. Since $x_i \neq x_j$, there is a coordinate k for which $x_i^{(k)} \neq x_j^{(k)}$. We then take the mean value $r = (x_i^{(k)} + x_j^{(k)})/2$ and consider the matrix A and the vector b defined by

$$A_{ij} = \delta_{jk}\delta_{ik} \quad i, j \in \{1, ..., d\}, \qquad b_j = r\delta_{jk} \quad j \in \{1, ..., d\}.$$
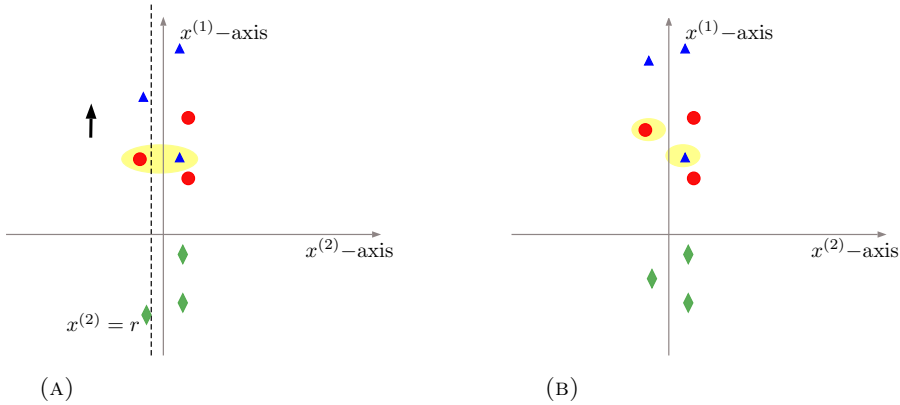
FIGURE 3.1. (A) Initial configuration in which two points within the yellow cloud share the same component $x^{(1)}$. The hyperplane parallel to the $x^{(1)}$-axis is chosen separating the two points. The vector field induces a vertical upwards translation of the left half-space. (B) Represents the final distribution of points. The yellow clouds indicate the points that initially shared the same $x^{(1)}$-component, a fact that is avoided after the application of the dynamic deformation.

The hyperplane $\{Ax + b = 0\}$ separates the two points x_i and x_j. We choose W such that the resulting vector field $\boldsymbol{\sigma}(Ax + b)$ can change the $x^{(1)}$-component of one of the elements. For doing so, we set

$$W = \begin{pmatrix} 1 & 1 & ... & 1 \\ 0 & 0 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & ... & 0 \end{pmatrix}$$

acting as a translation in the half-space. Solving the NODE in a sufficiently small time $\tau > 0$, we can assure that

(3.2)
$$\phi_\tau(x_j; A, W, b)^{(1)} \neq \phi_\tau(x_i; A, W, b)^{(1)}, \quad \forall i \neq j$$

as shown in Figure 3.1.qquad
Note that $\tau > 0$ needs to be taken small enough so to guarantee that when assuring that (3.2) holds, the other points do not lose the distinction of all their coordinates.

- This argument can be applied recursively on all pairs of data sharing the component $x^{(1)}$.

Once the separation property is guaranteed we can reset $t = 0$ to proceed to classification. This first step shows that, without loss of generality, we can assume that the dataset to be classified fulfills the separation property (3.1).

(2) Classification. By the previous step we have a set of points that, in particular, do not share the $x^{(1)}$-component.

Recall also that to each initial datum x_i, $i \in \{1, ..., N\}$, it corresponds the strip $S_{m(i)}$ as a target, associated to its label $m(i) \in \{1, ..., M\}$. Whenever a datum x_j lies in the set corresponding to its target, i.e. $x_j \in S_{m(j)}$, no action is needed, which corresponds to simply taking $W = 0$, i.e. the trivial dynamics. But for each $j \in \{1, ..., N\}$ such that $x_j \notin S_{m(j)}$ one needs to force the allocation of the points to the corresponding strip by a suitable choice of the controls.

We do it iteratively as follows. We start with any point $x_j \notin S_{m(j)}$. To simplify the notation, let us assume that this is the point x_1 corresponding to the index $j = 1$. Then:
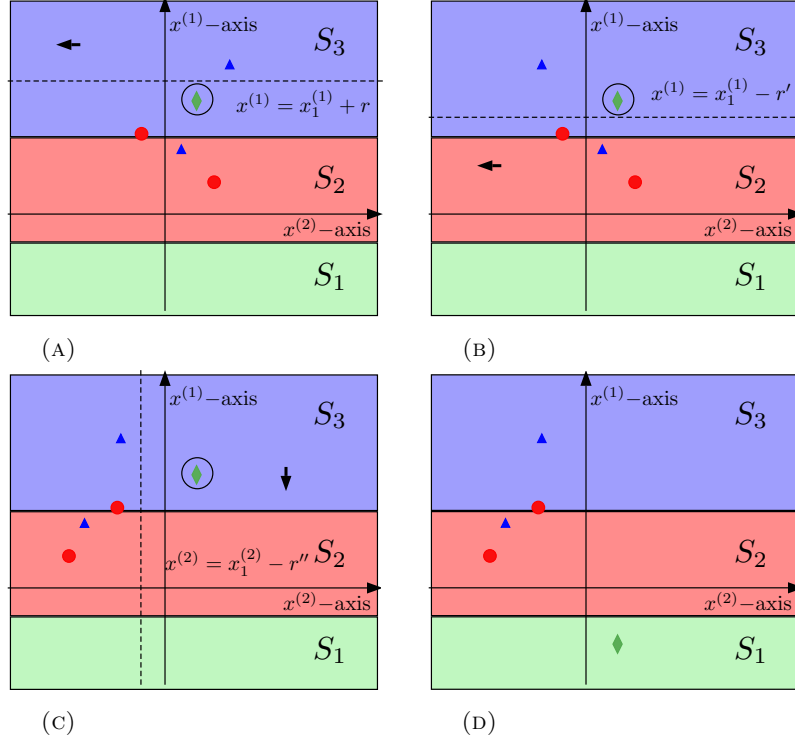
FIGURE 3.2. (A), (B) and (C) represent the three main steps (a), (b) and (c) and the concatenation leading to the final configuration of the dataset in (D). In this manner the green diamond that was originally located in the strip S_3 ends up lying in the one corresponding to its label (green color) S_1.

(a) we choose the hyperplane $\{Ax + b = 0\}$ with A and b of the form

$$(3.3) \qquad A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad b = - \begin{pmatrix} x_1^{(1)} + r \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where $r > 0$ is such that $x_1^{(1)} + r < x_i^{(1)}$ for every i such that $x_1^{(1)} < x_i^{(1)}$. We choose W so that the field is parallel to the hyperplane, for simplicity:

$$(3.4) \qquad W = \begin{pmatrix} 0 & 0 & \dots & 0 \\ -1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

as in Figure 3.2. We solve (2.3) up to a time T_1 so that

$$(3.5) \qquad \phi_{T_1}(x_i, A, W, b)^{(2)} < x_1^{(2)}, \quad \forall i \text{ s.t. } x_1^{(1)} < x_i^{(1)}.$$

(b) Then, we set r' to be such that $x_1^{(1)} - r' > x_i^{(1)}$ for any i such that $x_1^{(1)} > x_i^{(1)}$ and we choose our controls A' and b' as

$$A' = \begin{pmatrix} -1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad b' = \begin{pmatrix} x_1^{(1)} - r' \\ 0 \\ \vdots \\ 0. \end{pmatrix}$$
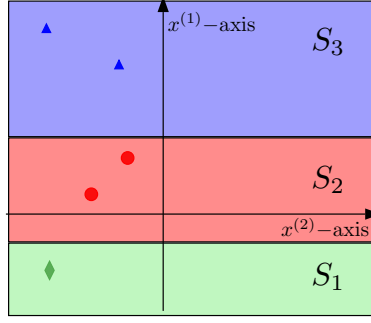
FIGURE 3.3. Final configuration after applying recursively the steps shown in Figure 3.2.

In this way $\boldsymbol{\sigma}(Ax + b)$ yields a null field ($= \mathbf{0}$) in all points that satisfy $x_i^{(1)} \geq x_1^{(1)} + r'$.

Now we choose $W' = W$ as before and we solve (2.3) up to a time T_2 so that:

$$\phi_{T_2}(\phi_{T_1}(x_i; A, W, b); A', b', W')^{(2)} < x_1^{(2)}.$$

(c) Then, we can choose a hyperplane represented by

$$A'' = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad b'' = \begin{pmatrix} 0 \\ -x_1^{(2)} + r'' \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

with r'' so that the image of x_1 under the previous transformations, the one to be properly classified, lies alone to one side of the hyperplane. We can then choose W'' so that the corresponding vector field pushes this point towards $S_{m(1)}$. Note that this procedure not only assures that the image of x_1 gets into $S_{m(1)}$ but also allows to determine exactly the location of the first coordinate of that point. We choose a target location $z_1^{(1)}$ in $S_{m(1)}$ such that there is no point x_i fulfilling $x_i^{(1)} = z_1^{(1)}$. Moreover, note that his process keeps the $x^{(1)}$ coordinate of all points $i \neq 1$ unaltered while allocating x_1 to $S_{m(1)}$ (see Figure 3.3 for the final result). Observe that the process above will not change the proper allocation of the points that already lie in the corresponding strip. Applying this construction recursively, in finitely many steps, we can guarantee the correct collocation all points that initially did not fulfill the classification criterion, i. e. such that $x_j \notin S_{m(j)}$.

(3) **Time rescaling.** The argument above assures that the complete classification can be achieved in some time $T^* > 0$. Classification can be assured in any time $T > 0$, by time-scaling, $s = (tT)/T^*$. Indeed, we then have

$$\frac{d}{ds}x = \frac{T^*}{T}W(s)\boldsymbol{\sigma}\left(A(s)x(s) + b(s)\right) \quad s \in (0, T)$$

and the multiplicative term T^*/T can be absorbed within the control W, keeping A and b unchanged (other than time rescaling).

$\square$

Remark 3.2 (On the activation function). *Although the proof was presented in the particular case of the ReLU, the properties of the activation function σ that we used in the proof are the following:*

(1) σ is Lipschitz continuous,
(2) σ vanishes in $(-\infty, 0]$,
(3) σ is positive in $(0, +\infty)$.

The above properties of σ suffice to assure that the main steps of our proof apply as represented in Figure 3.2. In fact, these conditions could also be further relaxed to consider activation functions that do not necessarily vanish in $(-\infty, 0]$ but simply decay as s tends to $-\infty$. But this would increase the cost of control. We shall not pursue this case here but the main steps of the proof would be essentially the same.

Remark 3.3 (Complexity of the controls I). *The controls are piecewise constant, and their complexity can be measured by the number and amplitude of the required switches. The number of required switches for the classification task is $4N$, where N is the number of points: N for the preparation and $3N$ for the classification.*

Our proof shows that the role of the control b is very much related to the distribution and spreading of the data along the Euclidean space. On the other hand, the controls W and A are of norm one in L^∞. Their complexity can be estimated in terms of the number of switchings, which is reflected on the BV-norm.

In our proof, the cost of the control is measured in terms of the needed time of control, which is the accumulation of the time spent in each iteration. The length of the n-th time interval T_n for which the control is constant is larger when points whose ordering needs to be swapped are very close to each other, and this leads to a higher cost. Note also that in our proof, we did not exploit the possible structuring of data into clusters. Indeed, when a given point to be classified is embedded into a cluster corresponding to the same label, one may take advantage of that fact to classify the whole cluster simultaneously and, in this manner, reduce the number of switchings.

By the rescaling step 3 of Theorem 1, we have seen that classification can be achieved in any time horizon, in particular in $T = 1$. A natural way of measuring the cost of control is to fix $T = 1$ and then to estimate the L^∞ and BV norm of the controls.

A complete analysis of the cost of classification in terms of the complexity of the data-set is a challenging problem that requires substantial further work. However, we will give an insight of this issue relating it with the control cost in Section 5, Remark 5.1.

4. SIMULTANEOUS CONTROL

The proof of the previous classification Theorem shows that, other than classifying the data to the corresponding strips, we can also, for instance, completely control the first coordinate $x^{(1)}$ of all points to be classified at the final time, with the same controls.

Based on that observation, in this section, we explore to which extent the classification can be enhanced to allocate to each datum to be classified a given point of destination. As we shall see, this can be done in an approximate manner, finding controls driving the ensemble of data to be classified as close as we wish from the given final locations chosen. This is an approximate simultaneous or ensemble controllability result. In fact, as we shall see, the exact simultaneous control can be achieved when initial data and targets are well separated.

Theorem 2 (Simultaneous Control). *Suppose $d \geq 2$. Fix $T > 0$ and let the activation function σ be the ReLU (2.1). Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ be distinct initial data and $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$ be distinct final data. Then there exist piecewise-constant control functions $A, W \in L^\infty\big((0,T); \mathbb{R}^{d \times d}\big)$ and $b \in L^\infty\big((0,T), \mathbb{R}^d\big)$ such that the associated flow to (2.3) with the controls A, W, b fulfills:*

$$\phi_T(x_i; A, W, b) = z_i \quad \forall i \in \{1, ..., N\}.$$

Remark 4.1. *Several remarks are in order:*
- *Note that it is impossible to bring two different points to the same target by means of a Lipschitz vector field. Therefore, the control to distinct targets is the sharpest result one can achieve.*
- *By continuity, this also implies the approximate simultaneous control when the initial data or targets are not completely distinct. Indeed, let us assume for instance that $z_i = z_j$. We can then slightly modify these targets to new points points $z'_j = z_j + \boldsymbol{\epsilon}$, with $|\boldsymbol{\epsilon}| \leq \epsilon$ arbitrarily small. The exact simultaneous control of the system to the targets $\{z'_i\}_{i=1}^N \subset \mathbb{R}^d$ is then feasible and assures the control to a distance ϵ of the original targets $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$.*

(A)   (B)

FIGURE 4.1. (A) Initial configuration for the simultaneous control before applying Theorem 1. The big colored circles represent the target locations for each point, denoted by small circles. (B) Configuration after controlling the $x^{(1)}$-component of each point.

In other words, even when the targets $\{z_i\}_{i=1}^N \subset \mathbb{R}^d$ are not distinct, given any $\epsilon > 0$, there exist control functions $A, W \in L^\infty\left((0,T);\mathbb{R}^{d\times d}\right)$ and $b \in L^\infty\left((0,T),\mathbb{R}^d\right)$ (depending on ϵ) such that the associated flow to (2.3) with the controls A, W, b fulfills:

$$\phi_T(x_i; A, W, b) = z_i + \boldsymbol{\epsilon} \quad \forall i \in \{1, ..., N\}, \quad \|\boldsymbol{\epsilon}\| = \epsilon.$$

Proof. We proceed in several steps.

(1) **Preparing the target.** The backward NODE

$$\begin{cases} \dot{x} = W'\boldsymbol{\sigma}(A'x + b') \\ x(T) = z \end{cases}$$

is also a solution of the forward NODE (2.3) by the controls $W(t) = -W'(T-t)$, $A(t) = A'(T-t)$ and $b(t) = b'(T-t)$. Therefore, we can apply the same argument done in Step 1 of Theorem 1 for preparing the targets. At the price of N extra switches, we can find controls W, A and b such that the targets fulfill

$$z_i^{(1)} \neq z_j^{(1)} \qquad \forall j \neq i.$$

(2) **Control the first component.** Applying the arguments in the proof of Theorem 1, we may control the first component $x^{(1)}$ of each trajectory. In this way we can deduce there existence of A, W and b such that:

$$\phi_T(x_i; A, W, b)^{(1)} = z_i^{(1)} \quad \forall i \in \{1, ..., N\}.$$

(3) **Controlling to the targets.** Now, we consider the hyperplanes $\{x^{(1)} = c\}$, for c negative enough. Let us take

$$A' = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & ... & 0 \end{pmatrix}, \quad b' = \begin{pmatrix} -c \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

We use W' to rotate the field in the direction to control the point $\phi_T(x_1; A, W, b)$ to its target z_1 with time $T' > 0$.

Then we set $c = z_1^{(1)} + \delta$ for δ small enough and we consider the hyperplane $x^{(1)} = c$ and apply the same argument, finding W'' that controls $\phi_{T'}(\phi_T(x_2; A, W, b); A', W', b')$ to z_2. Note that z_1 remains static along this deformation (see Figure 4.2).

Since $z_i^{(1)} \neq z_j^{(1)}$ for all i, j holds, we can proceed inductively. After a time-rescaling, as in Theorem 1, we conclude the proof of the simultaneous controllability.
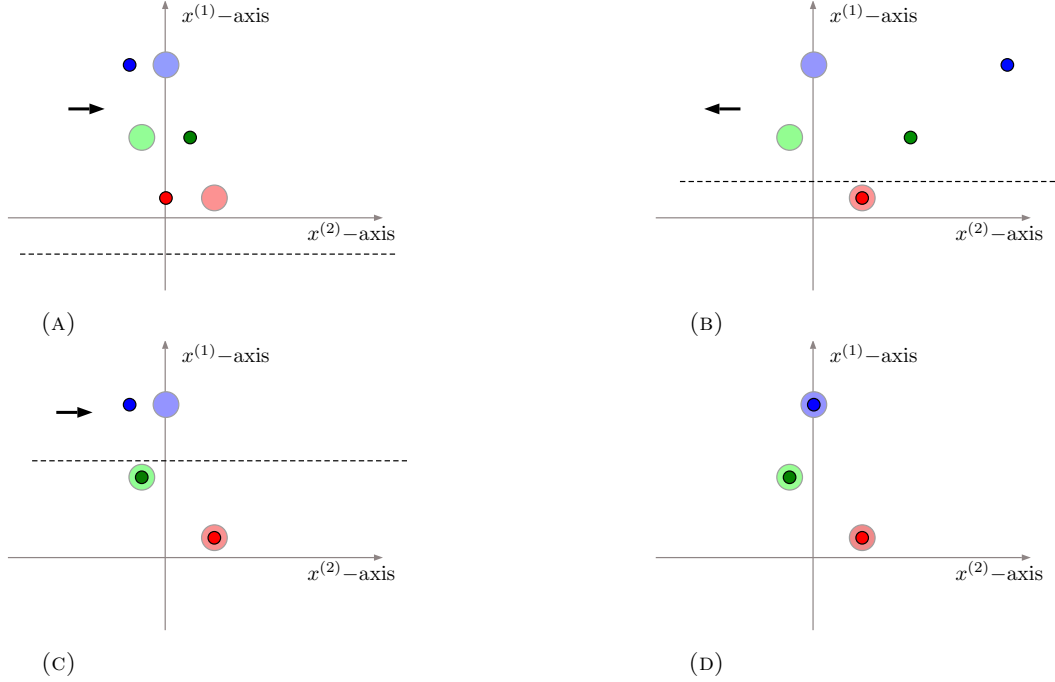
□

FIGURE 4.2. Iterative process leading to the simultaneous controllability. The colored big circles represent the targets, the small ones the data to e controlled, the dashed line is the chosen hyperplanes.

Remark 4.2 (On the activation function). *As in the context of Theorem 1, although Theorem 2 was stated for the ReLU (2.1) activation function, it can be extended to more general activation functions as in Remark 3.2.*

Remark 4.3. *The classification result in Theorem 1 is a direct consequence of the simultaneous controllability result in Theorem 2. In fact, as a corollary of Theorem 2, the classification can be achieved for an arbitrary partition $\mathcal{S}$ of the Euclidean space and any allocation of labels.*

Remark 4.4 (Complexity of the controls II). *Note that, with respect to the classification result and Remark 3.3, the number of switches in this simultaneous control result has increased by $2N$, remaining of the order of N. However, its norm will depend also on the chosen targets. The reason is that the gap for the choice of a hyperplane in step 3 of the proof of Theorem 2 can be small, depending on the target configuration. This will imply the smallness of the vector field chosen, and therefore the needed time will be larger.*

Note however, that this discussion is specific for the controls we have constructed. The analysis of optimal control strategies is a challenging open problem.

Remark 4.5 (Nonlinearities with drift). *Note that the simultaneous control result we have proved would also hold in an approximate manner for systems*

$$\dot{x} = f(x) + W(t)\boldsymbol{\sigma}(A(t)x + b(t))$$

with f bounded. Indeed, the control W can enhance the effect of the nonlinearity making f negligible. Whether an exact simultaneous control result in the spirit of Theorem 2 can be obtained in this setting is an interesting open problem.

Remark 4.6 (Sparsity). *The proof above is built-up in consecutive steps. In each step we employ only an hyperplane and a direction of motion for the flow. The proof does not require to activate at all time the whole ensemble of components of the controls A, W, b.*

To be more precise, let us consider a hyperplane with normal vector a and displaced by b from the origin. Then, the sign of scalar product $\langle a, x - b \rangle$ determines whether x is located at one side or the other of the hyperplane. Therefore, applying the scalar function σ to $\langle a, x - b \rangle$,
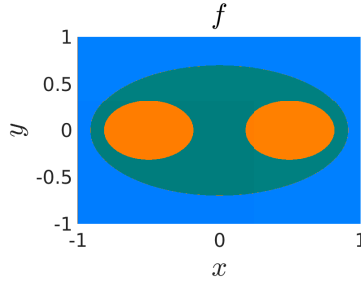
FIGURE 5.1. Simple function in the square $\Omega = [-1, 1] \times [-1, 1]$, $M = 3$, each color represents a set Ω_m.

one obtains either 0 if x is in one side of the hyperplane or a positive one if x is in the other one.

One can then set the direction of motion of the active half-space by choosing a vector $w \in \mathbb{R}^d$, generating a dynamics of the form:

$$\dot{x} = w\sigma\left(\langle a, x - b\rangle\right)$$

In the expression above only $3d$ controls are activated. But, taking into account that $\langle a, b\rangle$ is a scalar value, the total number of active controls is $2d + 1$.

As a conclusion of this discussion we see that our proofs hold with controls W, A and b with the following particular structure:

$$W = \begin{pmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_d(t) \end{pmatrix} (1, 1, \cdots, 1), \quad A = \mathrm{diag}\begin{pmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_d(t) \end{pmatrix}, \quad b(t) = \tilde{b}(t)\begin{pmatrix} a_1(t) \\ \vdots \\ a_d(t) \end{pmatrix}$$

where $w_i, a_i, \tilde{b} \in L^\infty((0, T), \mathbb{R})$, involving a total of $2d + 1$ controls.

The form of dynamics (2.3) is richer since it allows for more general matrices W, A. But the arguments above show that the actual dimension of the needed controls can be drastically diminished. This is a manifestation of the fact that the results own this paper can be achieved assuring further sparsity or polarization conditions on the controls.

5. UNIVERSAL APPROXIMATION

In this section we prove the Universal Approximation Theorem using the methods developed to treat simultaneous controllability.

We mainly consider simple functions of the form

$$(5.1) \qquad\qquad f = \sum_{m=1}^{M} \alpha_m \chi_{\Omega_m},$$

as in Figure 5.1. Our goal is to show how, controlling a NODE, such a function can be approximated in L^2_{loc}. Obviously, by density, the result extends immediately to any bounded measurable function. But, as we shall see, the complexity of the controls we will build depends on the structure of the simple function and, more precisely, on the box-counting dimension of the boundaries of the sets Ω_m.

In the context of NODEs the goal of approximating a function can be interpreted as a simultaneous control problem for an uncountable ensemble of trajectories. We do it by viewing the values f at each point x as the target or label that each x have to be associated with along the trajectories of the NODE. Consequently, the proofs of the main results of this section will employ the ideas of the previous ones, but with substantial added work.

For simplicity, we assume that all the boundaries of the sets Ω_m constituting the simple target function f, have a finite perimeter. Later on in Remark 5.1 we shall show that the same proof

works under weaker conditions, to the price of increasing the complexity of the needed controls. We will see how the box-counting dimension of the boundary of the characteristic sets arises in the control cost. The approximation can only be guaranteed over bounded sets Ω, as we shall see in Section 6 (Remark 6.8). Otherwise the use of unbounded controls would be required.

More precisely, the main result of this section is as follows:

Theorem 3 (Universal Approximation). *Let $d \geq 2$ and $T > 0$. Let $\Omega \subset \mathbb{R}^d$ be a bounded set, and σ be the ReLU activation function in (2.1). Then, for any $f \in L^2(\Omega; \mathbb{R}^d)$ and $\epsilon > 0$ there exist $A, W \in L^\infty((0,T); \mathbb{R}^d)$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that the flow generated by (2.3), $\phi_T(\cdot; A, W, b)$, satisfies:*

$$(5.2) \qquad \|\phi_T(\cdot; A, W, b) - f(\cdot)\|_{L^2(\Omega)} < \epsilon.$$

Theorem 3 is a direct consequence of the following theorem:

Theorem 4 (Approximation for simple functions). *Let $d \geq 2$ and $T > 0$. Let $\Omega \subset \mathbb{R}^d$ be a bounded set, and σ be the ReLU activation function in (2.1). Consider any f in the form of (5.1) with characteristic sets Ω_m of finite perimeter. Then, for any $\epsilon > 0$ there exist piecewise-constant controls $A, W \in L^\infty((0,T); \mathbb{R}^d)$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that the flow generated by (2.3), $\phi_T(\cdot; A, W, b)$, satisfies:*

$$\|\phi_T(\cdot; A, W, b) - f(\cdot)\|_{L^2(\Omega)} < \epsilon.$$

Setting $T = 1$, the norm the controls and their switches satisfy:

$$(5.3) \qquad \|W\|_{L^\infty((0,1); \mathbb{R}^{d \times d})} \lesssim_{\Omega,f} \epsilon^{-4d(d-1)} \qquad as \ \epsilon \to 0$$

$$(5.4) \qquad \|A\|_{L^\infty((0,1); \mathbb{R}^{d \times d})} = 1, \quad \|b\|_{L^\infty((0,1); \mathbb{R}^d)} \lesssim_{\Omega,f} \epsilon^{-2d(d-1)} \qquad as \ \epsilon \to 0$$

$$(5.5) \qquad The \ number \ of \ switches \ of \ the \ controls \ A, W, b \ is \ of \ the \ order \ of \ \epsilon^{-2d(d-1)}.$$

Proof. Let us first explain the main ideas of the proof.

In the NODE formulation of this result, each input $x \in \mathbb{R}^d$ has a target $f(x) \in \mathbb{R}^d$. The actual value of $f(x)$ depends on the characteristic set in which x lies. Ideally, we would like to find controls A, W and b such that for every $x \in \Omega$ we have that

$$\begin{cases} \dot{x} = W(t)\boldsymbol{\sigma}(A(t)x(t) + b(t)) \\ x(0) = x \\ \phi_T(x; A, W, b) = f(x). \end{cases}$$

But, due to the discontinuity of $f(\cdot)$ and the continuous dependence of the solutions of the ODE with respect to the initial data, this task is impossible. Thus, we have to relax the problem so to formulate our goal in an approximate manner.

First, we will define for any $h > 0$ small enough, a partition of Ω in a way that

$$\Omega = \Omega_c \sqcup \Omega_h, \qquad \text{meas}(\Omega_h) \sim h$$

Then, the control problem we seek to solve is the following. For all $\nu > 0$, find A, W and b (that depend on h) satisfying that the associated controlled flow $\phi_T(\cdot, A, W, b)$ satisfies:

$$\begin{cases} |\phi_T(x; A, W, b) - f(x)| < \nu & x \in \Omega_c \\ |\phi_T(x; A, W, b)| < K & x \in \Omega_h. \end{cases}$$

with K independent of h and ν.

This result can be achieved by means of a careful implementation and further development of the ideas of the proof of the simultaneous controllability result of Theorem 2. The proof is structured as follows.

In **Step 1** we construct the partition (Ω_h, Ω_c). Furthermore, we will split Ω_c in d-dimensional hyperrectangles, each of them having a single target associated in $\mathbb{R}^d$. **Step 2** concerns the compression of Ω_c and of its hyperrectangles plus an adaptation of Theorem 1. The goal is to arrive to a configuration in which one will be able to apply the simultaneous controllability, Theorem 2. Finally, in **Step 3**, we apply the simultaneous controllability, Theorem 2.

Step 1: Construction of Ω_h.

1.1 Covering of boundaries of the characteristic sets. Thanks to the compactness of the boundary $\Gamma = \bigcup_{m=1}^{M} \partial\Omega_m$ it can be covered with a finite number of hypercubes of side h. Let us denote this cover by Γ_h. The cover Γ_h can be chosen in a way that each edge of each hypercube is oriented in the direction of an element of the canonical basis.

Since the length of Γ is finite, and its dimension is $d-1$, the number, N_Γ, of needed hypercubes of size h in the cover Γ_h is:

$$N_\Gamma \lesssim_\Gamma h^{-(d-1)} \quad \text{as} \quad h \to 0.$$

Since each hypercube has volume h^d, the volume of the cover Γ_h is of the order of h. Note that, Γ_h (represented by in white in Figure 5.2a) contains the jumps of the simple function f to be approximated.

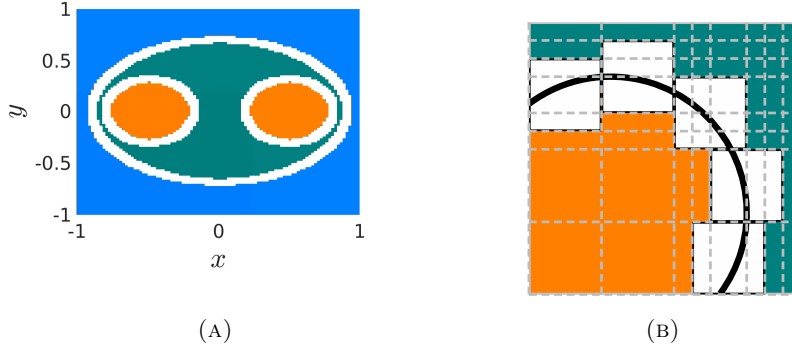(A)   (B)

FIGURE 5.2. (A) The resulting function after not considering the set Γ_h showed in white. (B) Qualitative representation d-dimensional hyperrectangles' generation. The black curve represents part of the boundary of a characteristic set. The white hypercubes are part of the cover of such boundary and the grey dotted lines represent the hyperplanes (5.6) generated out of the sides of the hypercubes of the cover. The colors orange and dark green represent the values of the target function in coherence with Figure 5.2a.

1.2 Determine the final d-dimensional hyperrectangles. From the cover Γ_h we want to find a partition of $\Omega \setminus \Gamma_h$ made out of d-dimensional hyperrectangles $\{\mathcal{H}_l\}_{l=1}^N$ satisfying the following:

$$\forall x, y \in \mathcal{H}_l, \quad f(x) = f(y),$$

i.e. each hyperrectangle has at most a single associated value α_m.

For finding such partition we proceed as follows. We consider all hyperplanes that define the hypercubes in Γ_h (see Figure 5.2b for a qualitative illustration). Altogether we have

$$(5.6) \qquad \left\{ x^{(k)} = c_{k,n} \right\} \qquad 1 \leq n \leq N_{\Gamma,k} \qquad 1 \leq k \leq d, \qquad c_{k,n} < c_{k,n+1}$$

where $N_{\Gamma,k}$ is the number of hyperplanes orthogonal to the k-th vector of the canonical basis and let $c_{k,N_{\Gamma,k}+1} = +\infty$. By construction one has that $N_{\Gamma,k} \leq 2N_\Gamma$ and therefore

$$(5.7) \qquad N_{\Gamma,k} \lesssim_\Gamma h^{-(d-1)} \quad \text{as} \quad h \to 0 \qquad \text{for every } 1 \leq k \leq d.$$

All these hyperplanes generate d-dimensional hyperrectangles $\{\mathcal{H}_l\}_{l=1}^N$ with the property that every $\mathcal{H}_l$ has a associated target in $\alpha_{m(l)} \in \mathbb{R}^d$. Then, the number of final hyperrectangles N can be estimated by:

$$N \leq \prod_{k=1}^{d} N_{\Gamma,k} \leq 2^d N_\Gamma^d.$$

Hence

$$(5.8) \qquad N \lesssim_\Gamma h^{-d(d-1)} \quad \text{as} \quad h \to 0.$$
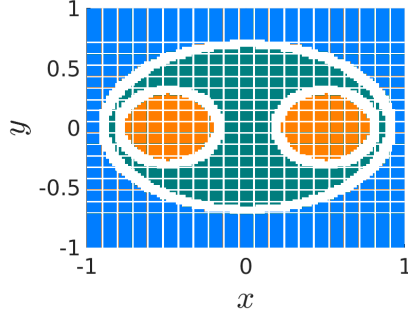
FIGURE 5.3. Qualitative representation of f with removed strips around the meshing hyperplanes. The white regions correspond to Ω_h.

We consider sufficiently thin d-dimensional strip around each hyperplane $\{x^{(k)} = c_{k,n}\}$ of side $\zeta = h^d$ given by

$$(5.9) \qquad \mathscr{S}_{n,k} := (\mathbb{R} \times \mathbb{R} \times \cdots \times [c_{n,k} - \zeta, c_{n,k} + \zeta] \times \cdots \times \mathbb{R}) \cap \Omega.$$

These strips, of measure of the order of h^d, belong to Ω_h together with Γ_h (see Figure 5.3.). The total number of hyperplanes is of the order of $h^{-(d-1)}$, (5.7), and therefore also the number of the thin strips. Hence, one can see that the measure of Ω_h is of the order of h. After removing the d-dimensional strips, we are left with at most N connected sets.

Then, making an abuse of notation, we denote the number by N the number of connected components of $\Omega \setminus \cup_{k,n} \mathscr{S}_{n,k}$ and its connected components by $\{\mathcal{H}_l\}_{l=1}^N$.

Now the L^2 norm can be split in two:

$$\|f - \phi\|_{L^2(\Omega)}^2 = \sum_{i=1}^N \int_{\mathcal{H}_i} |f - \phi|^2 dx + \int_{\Omega_h} |f - \phi|^2 dx$$
$$\lesssim_\Gamma N\nu^2 + h^{-(d-1)}h^d.$$

We see that it suffices to pick:

$$(5.10a) \qquad h \lesssim_\Gamma \epsilon^2$$

$$(5.10b) \qquad \nu \lesssim_\Gamma \epsilon N^{-\frac{1}{2}} \lesssim_\Gamma \epsilon h^{\frac{d(d-1)}{2}} \lesssim_\Gamma \epsilon^{1+d(d-1)}$$

to ensure that

$$\|f - \phi\|_{L^2(\Omega)} \lesssim_\Gamma \epsilon.$$

Step 2: Compression of Ω_c. In this Step we apply a compression Lemma that allows us to apply the simultaneous controllability of Theorem 2. Given a set $\omega \subset \mathbb{R}^d$, let us define

$$\text{diam}(\omega) := \max_{x_1,x_2 \in \omega} |x_1 - x_2|$$

$$\text{diam}_{(k)}(\omega) := \max_{x_1,x_2 \in \omega} |x_1^{(k)} - x_2^{(k)}|.$$

Furthermore, let us denote by $(\omega)^{(k)}$ the k-th component of the set ω, i.e.

$$\omega^{(k)} := \{x \in \mathbb{R} : \exists y \in \omega \quad \text{such that } y^{(k)} = x\}.$$

As done in Theorem 2, in step 1, since the backward NODE is a solution for the forward NODE for the appropiate controls, we can find an equivalent set of targets $\{\alpha_m\}_{m=1}^M$ such that $\alpha_m^{(1)} \neq \alpha_{m'}^{(1)}$ if $m \neq m'$. Without loss of generality, we will assume that the set of targets fulfills such contidition.

Let us summarize the properties we have built in the following

$(\mathbf{H})$
$\begin{cases}
\text{(a) Consider hyperplanes } \{x^{(k)} = c_{k,n}\} \text{ in the form of (5.6).} \\
\text{(b) Let } \zeta > 0 \text{ be the width of the strips } \mathscr{S}_{n,k} \text{ in (5.9).} \\
\text{(c) Consider the sets } \{\mathcal{H}_l\}_{l=1}^N \text{ defined by the connected components of } \Omega \setminus \cup_{n,k}\mathscr{S}_{n,k}. \\
\text{(d) Assume that each set } \mathcal{H}_l \text{ has a single associated target } \alpha_{m(l)}.
\end{cases}$

Lemma 1 (Fundamental Compression Lemma). *Let $\Omega \subset \mathbb{R}^d$ be bounded, and consider hyperplanes, strips and sets as in ($\mathbf{H}$). Assume that the set of targets $\{\alpha_m\}_{m=1}^M \subset \mathbb{R}^d$ fulfil $\alpha_m^{(1)} \neq \alpha_{m'}^{(1)}$ if $m \neq m'$. For any $T > 0$ and any $0 < \eta < {}^1/_N$ small enough, there exist piecewise-constant controls $A, W \in L^\infty((0,T); \mathbb{R}^{d \times d})$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that*

$$\phi_1(\mathcal{H}_l; A, W, b)^{(1)} \subset [\alpha_{m(l)}^{(1)} - C\eta, \alpha_{m(l)}^{(1)} + C\eta] \qquad \text{for every } 1 \leq l \leq N,$$
$$\mathrm{diam}_{(k)}\phi_1(\Omega; A, W, b) < \eta \qquad \text{for every } 2 \leq k \leq d,$$
$$\phi_1(\Omega; A, W, b) \subset \mathbb{B}(0, K)$$

with K and C independent of N, N_Γ and η. Furthermore, fixing $T = 1$, one has that:

$$\|A\|_{L^\infty((0,1);\mathbb{R}^{d \times d})} = 1, \qquad \|b\|_{L^\infty((0,1);\mathbb{R}^d)} \lesssim_\Omega N,$$
$$\|W\|_{L^\infty((0,1);\mathbb{R}^{d \times d})} \lesssim_{\Omega,\alpha,d} N \left(\frac{1}{\eta} + \frac{1}{\zeta} + \log\left(\frac{1}{\eta\zeta}\right)\right).$$

The number of discontinuities of the controls A, W, b is of the order of N.

Here, the subscript $\lesssim_\alpha$ denotes the dependence on the set $\{\alpha_m\}_{m=1}^M$.

For the sake of readability, the proof is postponed to Appendix A. The proof uses a concatenation of flows as the ones described in Subsection 2.2 to achieve the desired result.

Note that, Lemma 1, aside of compressing, for η small enough, also contains a generalization of Theorem 1 for the hyperrectangles.[1]

Step 3: Simultaneous Control. For η small enough, one can find a collection of M sets, $\{Q_m\}_{m=1}^M$ such that:

$$\mathrm{diam}(Q_m) \leq \max\{1, 2C\}\eta, \quad 1 \leq m \leq M$$

$$\phi_T(\mathcal{H}_l) \subset Q_m \qquad \text{if the target associated to } \mathcal{H}_l \text{ is } \alpha_m.$$

Therefore we can apply Theorem 2. Let us denote the flow from Theorem 2 by ψ_T. The controllability time of Theorem 2, T_M does not depend on h and by Grönwall inequality to obtain that

$$\nu := \max_{1 \leq m \leq M} \sup_{x \in \psi_{T_M}(Q_m)} |x - \alpha_m| \leq C_M \eta$$

where C_M is a constant that depends on M and on the set $\{\alpha_m\}_{m=1}^M$ but not on h. Hence $\nu \sim_{M,\alpha} \eta$. Furthermore, the extra number of switches will depend on M as in Theorem 2 but not on N.

We summarize the above discussion in a lemma that will be used in the next section.

Lemma 2. *Let $\Omega \subset \mathbb{R}^d$ be bounded, and consider hyperplanes, strips and sets as in ($\mathbf{H}$). Let f be a simple function as in (5.1). For all $\nu > 0$ small enough, there exist piecewise-constant controls $A, W \in L^\infty((0,T); \mathbb{R}^{d \times d})$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that the associated flow of (2.3) satisfies*

$$\begin{cases}
|\phi_T(x; A, W, b) - f(x)| < \nu & x \in \Omega_c \\
|\phi_T(x; A, W, b)| < K & x \in \Omega_h.
\end{cases}$$

---

[1] observe that, now, the sets are classified by strips, this is analogous to the result of Theorem 1.

Moreover, control norms are with K and C independent of N, N_Γ and η. Furthermore, fixing $T = 1$, one has that:

$$\|A\|_{L^\infty((0,1);\mathbb{R}^{d\times d})} = 1, \qquad \|b\|_{L^\infty((0,1);\mathbb{R}^d)} \lesssim_{\Omega,\alpha,d} N,$$

$$\|W\|_{L^\infty((0,1);\mathbb{R}^{d\times d})} \lesssim_{\Omega,\alpha,d,M} N\left(\frac{1}{\nu} + \frac{1}{\zeta} + \log\left(\frac{1}{\eta\zeta}\right)\right).$$

The number of discontinuities of the controls A, W, b is of the order of N.

Step 4: Estimates. Recalling (5.10a)(5.10b)(5.8) plus the requirement that $\eta < N^{-1}$ applied to (5.12) we obtain that:

$$\|A\|_{L^\infty((0,1);\mathbb{R}^{d\times d})} = 1, \qquad \|b\|_{L^\infty((0,1);\mathbb{R}^d)} \lesssim_{\Omega,d} \epsilon^{-2d(d-1)}, \qquad \|W\|_{L^\infty((0,1);\mathbb{R}^{d\times d})} \lesssim_{\Omega,f,d} \epsilon^{-4d(d-1)}.$$

The number of switches of the controls A, W, b is of the order of $\epsilon^{-2d(d-1)}$.

$\square$

Remark 5.1 (General simple functions). *In the previous theorem we have assumed that the perimeter of the characteristic sets is finite. However, the same proof works for certain sets with infinite perimeter.*

Let D be the upper box-counting dimension of Γ (see [15, Chapter 2], see also [4, Chapter 1]). As in the proof above, let $N_\Gamma(h)$ be the number of hypercubes of side h needed to cover the boundary Γ, the upper box-counting dimension is defined as

$$D := \limsup_{h\to 0} \frac{\log N_\Gamma(h)}{\log\left(\frac{1}{h}\right)}.$$

Hence, by definition one has that:

$$N_\Gamma \lesssim h^{-D} \text{ as } h \to 0.$$

Therefore, the estimates (5.3)(5.4) and (5.5) are replaced by

$$\|W\|_{L^\infty} \lesssim \epsilon^{-\frac{4Dd}{d-D}}, \qquad \|b\|_{L^\infty} \lesssim \epsilon^{-\frac{2dD}{d-D}} \qquad \text{as } \epsilon \to 0$$

The number of switches of A, W, b will be of the order of $\epsilon^{-\frac{2dD}{d-D}}$.

Note that the estimates blow up when $D = d$, in such case, we cannot provide an estimate. Boundaries with dimension $D = d$ exist, for instance, the Mandelbrot set [34]. Generally speaking, one can have a Jordan curve that has dimension equals to the ambient dimension [26].

Taking the box-counting dimension as a measure of complexity of the simple function we see that as the characteristic sets are more complex the cost of the control grows.

6. Control and simultaneous control of Neural transport equations

In this section we use the techniques of the proof of Theorem 4 to control a Neural transport equation (NTE) of the form

$$(6.1) \qquad \begin{cases} \partial_t\rho + \mathrm{div}_x\left[\left(W(t)\boldsymbol{\sigma}(A(t)x + b(t))\right)\rho\right] = 0 & (x,t) \in \mathbb{R}^d \times (0,T) \\ \rho(0) = \rho^0 \in C_c(\mathbb{R}^d, \mathbb{R}^+) \end{cases}$$

where $C_c(\mathbb{R}^d, \mathbb{R}^+)$ stands for the space of compactly supported continuous nonnegative functions. Our proof of the Universal Approximation Theorem can be interpreted in the context of the simultaneous control of these NTEs. This will also give rise to a transport formulation of the classification problem.

In this section the approximation will be measured in the sense of the Wasserstein-1 distance:

Definition 1. *Let $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$ be probability measures. The Wasserstein-1 distance $\mathcal{W}_1(\mu, \nu)$ is defined by as:*

$$\mathcal{W}_1(\mu, \nu) = \sup_{Lip(g)\leq 1} \left\{\int_{\mathbb{R}^d} g\,d\mu - \int_{\mathbb{R}^d} g\,d\nu\right\}$$

where $Lip(g) \leq 1$ stands for the class of Lipschitz functions with Lipschitz constant less or equal than 1.

Wasserstein distances play a central role in the theory of optimal transport (see [41, Chapter 5] for a general reference in this subject).

6.1. Control of the Neural Transport equation.

Since the vector field $W\boldsymbol{\sigma}(Ax+b)$ of the Neural Transport equation (6.1) is Lipschitz, it preserves the mass for each set $\mathcal{H}$ along the characteristic flow $\phi_T(\mathcal{H})$

$$\int_{\phi_T(\mathcal{H})} \rho(T)dx = \int_{\mathcal{H}} \rho^0 dx.$$

The control of the (6.1) can be done by controlling its characteristics plus a compression argument as done in Lemma 1 in Theorem 4.

We will consider target measures ρ^* that are finite combinations of Dirac masses:

$$(6.2) \qquad \rho^* = \sum_{m=1}^{M} \beta_m \delta_{\alpha_m}$$

with $\beta_m > 0$ and $\alpha_m \in \mathbb{R}^d$.

Theorem 5. *Let $T > 0$, $d \geq 2$, σ be as (2.1) and ρ^* as in (6.2) be a target positive measure satisfying:*

$$\int_{\mathbb{R}^d} d\rho^* = \int_{\mathbb{R}^d} \rho^0 dx = 1.$$

Then, for every $\epsilon > 0$, there exist piecewise-constant control functions $W, A \in L^\infty((0,T); \mathbb{R}^{d\times d})$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that the solution of (6.1) satisfies:

$$\mathcal{W}_1(\rho(T), \rho^*) < \epsilon$$

where $\mathcal{W}_1$ is the Wasserstein-1 distance. Furthermore, one has

$$\|W\|_{L^\infty((0,T),\mathbb{R}^{d\times d})} \lesssim_{\rho^0,\rho^*} \epsilon^{-1} \qquad \text{as } \epsilon \to 0.$$

Moreover, $\|b\|_{L^\infty((0,T),\mathbb{R}^d)}$, $\|A\|_{L^\infty((0,T),\mathbb{R}^{d\times d})}$ and the number of discontinuities of W, A and b are bounded independently of ϵ.

Proof. We proceed in several steps:

(1) Assigning targets to the initial datum. Consider $F : \mathbb{R} \to \mathbb{R}$ be defined as:

$$(6.3) \qquad F(s) = \int_{-\infty}^{s} \left(\int_{\mathbb{R}^{d-1}} \rho^0 dx^{(2)}...dx^{(d)} \right) dx^{(1)}.$$

Using the fact that $\rho^0 \in C_c(\mathbb{R}^d, \mathbb{R}^+)$, there exist two real numbers, name them c_0 and c_M, such that

$$\forall s \leq c_0, \quad F(s) = 0, \qquad \forall s \geq c_M, \qquad F(s) = 1.$$

Since $\rho^0 \in C_c(\mathbb{R}^d, \mathbb{R}^+)$, F is continuous and monotone increasing, i.e. we can find $M-1$ real numbers such that:

$$F(c_m) - F(c_{m-1}) = \beta_m \qquad 1 \leq m \leq M.$$

Hence, we will divide ρ^0 with $M-1$ hyperplanes of the form:

$$\left\{ x^{(1)} = c_m \right\}, \quad 1 \leq m \leq M-1$$

(see Figure 6.1). The intention is to find a flow such that, by means of the characteristics, can transport all the mass β_m contained between the hyperplanes $\{x^{(1)} = c_m\}$ and $\{x^{(1)} = c_{m-1}\}$ to the target α_m. As in the previous section, due to the continuity with respect of the initial data of the characteristics, this task will have to be carried out in an approximate manner.

(2) Compression and simultaneous control. Similarly as in **Step 1** of Theorem 4 , for every $\epsilon > 0$, we will consider a d-dimensional strip around each hyperplane $\{x^{(1)} = c_m\}$ for $m = 1, ..., M-1$ of the form
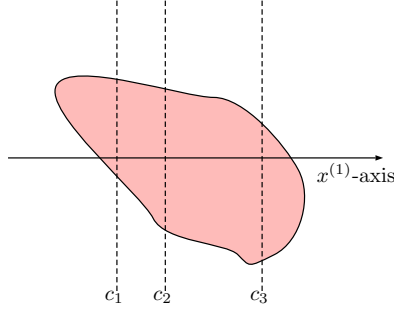
$$[c_m - \zeta, c_m + \zeta] \times \mathbb{R}^{d-1}$$

FIGURE 6.1. The support of the initial data ρ^0 and its division by hyperplanes.

for some $\zeta > 0$ in a way that:

$$F(c_m + \zeta) - F(c_m - \zeta) \le \frac{\epsilon}{M} \qquad 1 \le m \le M - 1.$$

In this way, we ensure that the mass lying in the strips is of the order of ϵ.

Using Lemma 2, we can ensure that the mass lying in the d-dimensional strips will lie in a bounded set independently of ζ and consequently independently of ϵ.

Let $\mathcal{H} \subset \mathbb{R}^d$ be a subset, and let $\rho(t)$ be the solution of (6.1) for certain A, W, b, let ϕ_T be the solution of (2.3) for the same controls. Due to the divergence structure of Equation (6.1) and that the field $W\boldsymbol{\sigma}(Ax + b)$ is Lipschitz, one has that the mass is preserved along the characteristics and

$$\int_{\phi_T(\mathcal{H})} \rho(T)dx = \int_{\mathcal{H}} \rho^0 dx.$$

Define $\mathcal{H}_m = \mathrm{supp}(\rho^0) \cap ([c_{m-1} + \zeta, c_m - \zeta] \times \mathbb{R}^{d-1})$. Now we may apply Lemma 2 to obtain the approximate controllability of all characteristics in every $\mathcal{H}_m$ while having the mass allocated in the strips has been transported in a bounded set $\mathbb{B}(0, K)$ with K independent of ζ and ν. Moreover, choose ν small enough so that

$$\mathbb{B}(\alpha_m, \nu) \cap \mathbb{B}(\alpha_{m'}, \nu) = \varnothing \qquad \text{if } m \ne m'.$$

Now, the objective is to quantify, in terms of the Wasserstein distance, the proximity of $\rho(T)$ to ρ^*. The conclusion will follow from a straightforward computation using the triangular inequality with a suitable ρ^ϵ. We will find a ρ^ϵ of the form

$$\rho^\epsilon = \sum_{m=1}^{M} \overline{\beta}_m \delta_{\alpha_m} + \sum_{m=1}^{M} \underline{\beta}_m \delta_{z_m}.$$

Let us proceed with the construction of ρ^ϵ.

(a) **Choice of the masses $\{(\overline{\boldsymbol{\beta}}_{\boldsymbol{m}}, \underline{\boldsymbol{\beta}}_{\boldsymbol{m}})\}_{\boldsymbol{m=1}}^{\boldsymbol{M}}$.** Consider ξ_m to be the minimizer of the following quantity

(6.4)
$$\xi_m := \min_{0 < \xi \le \nu} \left| \int_{\mathbb{B}(\alpha_m, \xi)} \rho(T)dx - \beta_m \right|, \qquad 1 \le m \le M.$$

Observe that the function $g_m : \mathbb{R}^+ \to [0, 1]$ defined as

$$g_m(\xi) = \int_{\mathbb{B}(\alpha_m, \xi)} \rho(T)dx$$

is an increasing function. Therefore, if the minimizer (6.4) is unique and if (6.4) satisfies $\xi_m < \nu$, then we have that $g_m(\xi_m) = \beta_m$. We, then, define

$$\overline{\beta}_m := g(\xi_m) = \int_{\mathbb{B}(\alpha_m, \xi_m)} \rho(T)dx.$$

By definition, $\overline{\beta}_m > 0$. Moreover, taking into consideration that the mass transported, through the application of Lemma 2, into $\mathbb{B}(\alpha_m, \nu)$ is at least $\beta_m - 2\epsilon/M$, one has that

$$\beta_m - \frac{2\epsilon}{M} \le \overline{\beta}_m \le \beta_m \qquad 1 \le m \le M$$

and

$$1 - 2\epsilon \leq \sum_{m=1}^{M} \overline{\beta}_m \leq 1.$$

Now define the remainders of mass $\underline{\beta}_m$ as follows

$$\underline{\beta}_m = \beta_m - \overline{\beta}_m, \qquad 1 \leq m \leq M$$

note that, by construction, one has

$$0 \leq \underline{\beta}_m \leq \frac{2\epsilon}{M}, \qquad 1 \leq m \leq M.$$

(b) **Choice of the locations $\{z_m\}_{m=1}^{M}$.** Let $\mathcal{B} = \bigcup_{m=1}^{M} \mathbb{B}(\alpha_m, \xi_m)$ and define $G : \mathbb{R} \to \mathbb{R}$ as

$$G(s) = \int_{-\infty}^{s} \left(\int_{\mathbb{R}^{d-1}} (1 - \chi_{\mathcal{B}}) \rho(T) dx^{(2)} ... dx^{(d)} \right) dx^{(1)}.$$

As in step 1 of this proof, there exist real numbers c_0' and c_M' such that:

$$\forall s \leq c_0', \quad G(s) = 0, \qquad \forall s \geq c_M', \qquad G(s) = 1$$

and as before, we find $M - 1$ real numbers such that

$$G(c_m') - G(c_{m-1}') = \underline{\beta}_m, \qquad \mathcal{T}_m = [c_{m-1}', c_m'] \times \mathbb{R}^{d-1}.$$

Take a collection of $\{z_m\}_{m=1}^{M}$ such that for every $m \in \{1, ..., M\}$, satisfies $z_m \in \mathcal{T}_m \cap \mathbb{B}(0, K)$.

Then, choosing $\nu = \epsilon$ the result follows by the triangular inequality.

$$\mathcal{W}_1(\rho(T), \rho^\epsilon) = \sup_{Lip(g) \leq 1} \left| \int_{\mathbb{R}^d} g(x) \rho(T) dx - \sum_{m=1}^{M} \overline{\beta}_m g(\alpha_m) - \sum_{m=1}^{M} \underline{\beta}_m g(z_m) \right|$$

$$\leq \sup_{Lip(g) \leq 1} \left| \sum_{m=1}^{M} \int_{\mathbb{B}(\alpha_m, \xi_m)} (g(x) - g(\alpha_m)) \rho(T) dx \right| + \sup_{Lip(g) \leq 1} \left| \sum_{m=1}^{M} \int_{\mathcal{T}_m} (g(x) - g(z_m)) \rho(T) dx \right|$$

$$\leq 2\nu + 2K\epsilon$$

$$\mathcal{W}_1(\rho^*, \rho^\epsilon) = \sup_{Lip(g) \leq 1} \left| \sum_{m=1}^{M} \overline{\beta}_m g(\alpha_m) + \sum_{m=1}^{M} \underline{\beta}_m g(\alpha_m) - \sum_{m=1}^{M} \overline{\beta}_m g(\alpha_m) - \sum_{m=1}^{M} \underline{\beta}_m g(z_m) \right|$$

$$= \sup_{Lip(g) \leq 1} \left| \sum_{m=1}^{M} \underline{\beta}_m g(\alpha_m) - \underline{\beta}_m g(z_m) \right| \leq 4K\epsilon.$$

(3) **Estimates.** The control cost, also depends on the initial density function ρ^0. Define the map $\epsilon_c : \mathbb{R}^+ \to \mathbb{R}^+$ as

$$\epsilon_c(\zeta) = \int_{c-\zeta}^{c+\zeta} \left(\int_{\mathbb{R}^{d-1}} \rho^0 dx^{(2)} ... dx^{(d)} \right) dx^{(1)}.$$

Note that $\epsilon_c(0) = 0$, differentiating with respect to ζ and evaluating at 0 we obtain that:

$$\frac{d}{d\zeta} \epsilon_c(0) = 2 \int_{\mathbb{R}^{d-1}} \rho^0(c, x^{(2)}, ..., x^{(d)}) dx^{(2)} ... dx^{(d)}.$$

Formally, the derivative of inverse map of ϵ_c, which we denote by $\zeta_c : \mathbb{R}^+ \to \mathbb{R}^+$, at zero fulfills:

$$\frac{d}{d\epsilon} \zeta_c(0) = \left(2 \int_{\mathbb{R}^{d-1}} \rho^0(c, x^{(2)}, ..., x^{(d)}) dx^{(2)} ... dx^{(d)} \right)^{-1}.$$

Therefore, it suffices to require that

(6.5) $$\zeta \leq \min_{m \in \{1, ..., M-1\}} \left\{ \left(2 \int_{\mathbb{R}^{d-1}} \rho^0(c_m, x^{(2)}, ..., x^{(d)}) dx^{(2)} ... dx^{(d)} \right)^{-1}, 1 \right\} \epsilon \qquad \text{as } \epsilon \to 0.$$

From Lemma 2 we obtain that:

$$\|W\|_{L^\infty} \lesssim_{\rho^*} \epsilon^{-1} \qquad \text{as } \epsilon \to 0.$$

Furthermore, $\|A\|_{L^\infty} = 1, \|b\|_{L^\infty} \leq C$ and the number of discontinuities of the controls is bounded independently of ϵ but dependent on the target ρ^*.

$\square$

Remark 6.1 (More general target configuarations)**.** *We stated the theorem by setting targets that are Dirac masses, but, it is well known that, in particular, one can approximate any compactly supported probability measure by a finite number of Dirac masses.*

Remark 6.2 (Influence of the initial condition on the control cost)**.** *The proof is valid for more general settings, namely for $\rho^0 \in L_c^1(\mathbb{R}^d; \mathbb{R}^+)$, where by $L_c^1(\mathbb{R}^d; \mathbb{R}^+)$ we understand the L^1 compactly supported nonnegative functions. However, the quantification of the control cost depends on the singularities of ρ^0.*

In (6.5) one can observe that as the density at a point c_m increases, ζ should be smaller. If $\rho^0 \in L_c^1(\mathbb{R}^d)$ the cost will depend on the "$\mathtt{strength}$" of the singularities of ρ^0. The precise quantification of the control cost depending on the regularity of F (or singularity of ρ^0) will not be treated in this article. For more details and a fine analysis of singularities we refer to [20, Chapter I and II].

Remark 6.3. *Due to the uniqueness of the solution of the underlying characteristic system, we cannot consider, in general, initial data that are Dirac deltas. More precisely, by the uniqueness of the characteristics, it is impossible to bring the initial datum $\rho^0 = \delta_0$ to any configuration of the type $\rho^* = \lambda\delta\alpha_1 + (1 - \lambda)\delta\alpha_2$ for any $\lambda \in (0, 1)$ and any $\alpha_1, \alpha_2 \in \mathbb{R}^d$.*

Remark 6.4. *In Theorems 4 and 5, to object of study was the input output map given by the flow ϕ_T associated to (2.3). In Theorem 5 we use the same principles for another objective, to control in approximate manner from a given density function to another.*

In Theorem 4 and Remark 5.1 we observed how the geometry of the supports plays a crucial role in the control cost. However, in Theorem 5, since we are controlling a single probability density, we do not observe such dependence. In contrast, in (6.5) we observe how high concentrations of mass can increase the cost of control.

Remark 6.5. *The restriction of $d \geq 2$ comes from the limitation pointed out in Remark 3.1. However, the transport equation*

$$\partial_t \rho + \mathrm{div}_x[V(x,t)\rho] = 0$$

with V as a control, can be controlled in the one-dimensional case approximately with respect to the Wasserstein-1 distance. It would be enough to design appropriate locations of attractors and repulsors depending on the mass distribution of ρ^0. This would allow concentrating the mass in a finite number of points. Later, one needs to design a dynamical system that brings each mass approximately on the approximation of the target. Neural Transport Equations can achieve this in dimension $d \geq 2$ just with the controls A, W and b.

Remark 6.6. *If the target Dirac masses have the same mass, one does not need the simultaneous controllability. Given M initial data $\{x_m\}_{m=1}^M$ and M targets $\{\alpha_m\}_{m=1}^M$, it is enough that for every $\epsilon > 0$ one can find controls A, W and b such that*

$$\forall j \in \{1, ..., M\} \quad \exists! m : \quad \phi_T(x_m; A, b, W) \in \mathbb{B}(\alpha_j, \epsilon).$$

Note that, one can also approximate any continuous compactly supported function by a finite number of Dirac masses whose support is disjoint and all of them have the same mass.

Remark 6.7 (Other Wasserstein metrics)**.** *The definition we gave of the Wasserstein-1 metric relies on the Kantorovich-Rubinstein Theorem [40, Theorem 1.14]. In the Euclidean space, a Wasserstein-p distance is defined as:*

$$\mathcal{W}_p(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \left\{ \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\gamma(x, y) \right\}$$

where $\Gamma(\mu_1, \mu_2)$ denotes the set of all measures γ that satisfy $\mu_1 = \int_{\mathbb{R}^d} \gamma(x, y)dx$ and $\mu_2 = \int_{\mathbb{R}^d} \gamma(x, y)dy$. Theorem 5 states the result in the Wasserstein-1 metric, however, since the supports of the final datum $\rho(T)$ and the target are compact, one has equivalence of all $\mathcal{W}_p$ distances [40, Chapter 7, Section 7.1.2].

Remark 6.8. *For the ReLU case, and for any Lipschitz nonlinearity, fixing controls A, W and b, we see that the characteristics, at most, have an exponential decay. This is the impediment for Theorem 3 to be a universal approximation theorem for functions in $L^2(\mathbb{R}^d, \mathbb{R}^d)$.*

6.2. **Simultaneous control of Neural transport equations and classification.**

The scalar transport equation does not allow distinguishing among labels. One should take a vectorial structure in order to have a transport formulation for classification.

Let us consider M classes, and M compactly supported probability densities ρ_m for $m = 1, ..., M$. Moreover, we assume, as in Section 2, that for every x there exists, at most, a unique label y. This implies that the supports of the probability measures ρ_m are disjoint:

$$\text{supp}(\rho_m) \cap \text{supp}(\rho_{m'}) = \varnothing, \qquad \text{if } m \neq m'.$$

Theorem 4 is based on approximating a simple function. Therefore, in Theorem 4 we have already treated the problem of sending uniform disjoint compactly supported distributions to precisely located Dirac deltas (in an approximate manner). Combining the ideas of Theorem 4 and 5 one has the following: Let $M \in \mathbb{N}$ be a natural number and $\Omega \subset \mathbb{R}^d$ for $d \geq 2$ and consider the system:

$$(6.6) \qquad \begin{cases} \partial_t \rho_m + \text{div}_x \left[(W(t)\boldsymbol{\sigma}(A(t)x + b(t))\rho_m) \right] = 0, & m \in \{1, ..., M\} \\ \rho_m(0) = \rho_m^0 \in C_c^0(\mathbb{R}^d), & m \in \{1, ..., M\}. \end{cases}$$

Theorem 6. *Let $T > 0$ and ρ_m^* be as in (6.2) the target measure for the the m-th equation. Assume that the initial conditions satisfy:*

$$\text{supp}(\rho_m^0) \cap \text{supp}(\rho_{m'}^0) = \varnothing \quad \text{if } m \neq m', \qquad \text{Per}\left(\partial \text{supp}(\rho_m^0)\right) < +\infty, \quad m \in \{1, ..., M\}$$

where by Per *we understand the perimeter. In addition, assume that the target functions satisfy:*

$$\text{supp}(\rho_m^*) \cap \text{supp}(\rho_{m'}^*) = \varnothing \qquad \text{if } m \neq m', \qquad \int \rho_m^* = \int \rho_m^0 dx = 1 \qquad m \in \{1, ..., M\}.$$

Then, for any $\epsilon > 0$, there exist controls $W, A \in L^\infty((0,T), \mathbb{R}^{d \times d})$ and $b \in L^\infty((0,T), \mathbb{R}^d)$ such that the solution of (6.6) *satisfies:*

$$(6.7) \qquad \qquad \mathcal{W}_1(\rho_m(T), \rho_m^*) < \epsilon \quad m \in \{1, ..., M\}$$

where $\mathcal{W}_1$ is the Wasserstein-1 distance.

The proof follows combining the arguments of Theorem 4 and 5. The only difference with respect to Theorem 5 is that all initial data have to be separated by several extra hyperrectangles as done in **Step 1** in Theorem 4. This step was not needed for controlling just one transport equation, but it had to be done in the universal approximation Theorem 4. This implies that, in this case, the control cost will depend both, on the geometry of the supports of each density function, and on the concentration of mass of such density functions.

The classification interpretation is the following. Let $\mathcal{S} = \{S_m\}_{m=1}^M$ be a partition of $\mathbb{R}^d$ and consider system (6.6). The classification problem consists of: for every $\epsilon > 0$, being able to find controls (that depend on ϵ) $W, A \in L^\infty((0,T), \mathbb{R}^{d \times d})$ and $b \in L^\infty((0,T), \mathbb{R}^d)$ such that the solution of (6.6) satisfies:

$$\int_{S_m} \rho_m(T)dx = 1 - \epsilon, \qquad \int_{\mathbb{R}^d} x\rho_m(T)dx \in S_m \qquad m \in \{1, ..., M\}$$

Roughly speaking, the classification problem, in terms of transport equations, is equivalent to send most of the mass of each probability measure to prefixed sets and to control its expectation to the prefixed set.

Remark 6.9. *Note that, since the supports are disjoint, if we consider $\rho = \sum_{m=1}^N \rho_m$, the controls A, W and b of Theorem 6 control also the scalar equation* (6.1).

Remark 6.10 (Robustness with respect to large number finite dimensional samples)**.** *The controls obtained from Theorem 6 are robust with respect to large number finite dimensional samples that follow the probabilities distributions of the initial data of system* (6.6)*. This implies*

that, given an error ϵ, the control cost of classification for finite dimensional samples as in Theorem 1 does not blow up when the cardinality of sample tends to infinity.

Remark 6.11. *Note that Remark 5.1 also applies in this case for having more irregular boundaries. However, if the boundaries of the supports of each distribution are at positive distance, one does not need to consider a fine meshing of the boundary.*

7. CONCLUSIONS AND PERSPECTIVES

We now present a number of conclusions and open problems.

(1) **The role of the activation function.** The key property that allows to prove all the results in this paper is that the activation functions are able to leave one half-space invariant while moving the other half. As mentioned in the Introduction, the simultaneous controllability is a rare property for dynamical systems. In particular, it can never occur in linear systems when all trajectories solve the same system. Our methods exploit this nonlinear feature of activation functions to obtain simultaneous control properties of NODEs that lead to the desired results in the Machine Learning context.

(2) **Algorithmic complexity.** All proofs developed in this paper are algorithmic.

The proofs in Theorem 1 and 2 require a number of steps of the order of $\mathcal{O}(N)$, where N is the number of points to be classified (see Remarks 3.3 and 4.4). In the classification Theorem 1 each point to be classified is handled in an iterative manner. Thus, our estimate on the number of needed iterations corresponds to the worst-case scenario. In practice, the optimal number of required switches will be lower, depending on the structure of the data to be classified (the initial data of the Neural ODE) and also on the partition chosen (or target in the case of Theorem 2). In particular, when the initial data are structured into clusters our proof can take advantage of that to classify the data in fewer iterations, keeping them grouped in packages during the evolution.

Note that, even in the context of Theorem 1 and 2, where the number of switches is of order of $\mathcal{O}(N)$, the control cost depends on the initial data and target destination. In particular, the cost of control increases when the distance between data corresponding to different labels decreases or their mixing increases.

The analysis is more intricate for the universal approximation theorem (Section 5) and for the control of Neural transport equations (Section 6). Even if the proofs are also algorithmic, the L^∞-norm of the control depends on the precision required when approximating the target as well as on the complexity of initial and final data, see Remark 5.1.

(3) **Optimal control strategies.** As we have explained above, by time-scaling, all results are valid for an arbitrary $T > 0$. For the analysis of the complexity of the control dynamics and the cost of control it is convenient to normalize the time to $T = 1$. The complexity of the controls can be then estimated in terms of the number of switches and their TV-norm.

The controls we build are not necessarily the optimal ones. In fact, as we have mentioned above, when the data to be classified present clustering phenomena, the complexity of the needed control diminishes.

Optimal controls could be defined setting the time horizon to be $T = 1$ and minimizing their L^∞-norm. This would lead to an optimality system or Pontryagin maximum principle ([30]) characterizing optimal controls that could be expected to present a bang-bang structure. The analysis of the complexity of optimal controls through such optimality system, which will strongly depend on the configuration and structure of the data to be classified, is an interesting and complex open problem. Of course such characterization does not lead to any explicit expression, it is of a purely implicit nature, making its posterior use rather complex ([36]). Note that this issue is even more complex in the context of the ReLU activation function, because of its lack of regularity.

Other norms and cost criteria can also be used to define optimal controls. For instance, other than penalizing solely the norm of the control, one could also penalize the controlled trajectories, enhancing the stability and the turnpike properties of the control processes, as shown in [14] or in [43] where the L^1 penalization of the state gives rise to sparse bang-bang controls as the ones shown in this article.

In summary, the development of computational methodologies to derive control strategies of minimal complexity is an interesting and challenging topic.

(4) Optimal activation functions. In the previous point, we were proposing to find optimal controls with prescribed initial and final data. One general open question would be the following. How to construct a vector field such that the norm of the control or the number of switches is minimal? or more precisely, what kind of vector fields can classify with minimal cost for any target function? This question is analogous to the optimal observation and location of sensors in linear systems, addressed for instance in [31, 32], but optimizing with respect to the nonlinearity.

This question is not typically addressed in the control literature since, normally, in mechanics, we are not allowed to choose the dynamics that we want to control.

(5) Varying dimension. The scaling of the control cost with respect to the precision ϵ and the number of switches in Theorem 4 depend on the dimension d. The number of switches and the L^∞ norm of the control suffer from the curse of dimensionality. However, the whole control process is made keeping d invariant.

Typically in ML, one can embed the system into a larger space, allowing more hyperplanes and benefiting from the extra directions. An open question would be: can we decrease the control cost by embedding the system into a larger dimension? How can one formulate a dimension varying Neural ODE? What is the impact on the control cost of this variability?

(6) NODEs vs ResNets. The continuous setting of NODEs allows for a better understanding and mastery of the exponential growth and decay of solutions with respect to time. These qualitative properties play an important role in the proofs of Theorems 4,5 and 6. On the other hand, the transport interpretation made in Section 6 allows for a more synthetic understanding of supervised learning. The time-continuous perspective can also be more adequate to analyze optimal control strategies.

Once the NODE setting is well understood, a posteriori, the time-discrete setting allows achieving similar results by means of fine discretizations of the Neural ODE, i.e., with sufficiently many deep layers.

On the other hand, as mentioned in Remark 3.1, when the time step is large enough, the discrete dynamics is richer in the sense that they are able to express features that the continuous NODEs cannot.

(7) On the topology approximation. Our proof of universal approximation guarantees approximation in L^2. But it does not allow to obtain density results in spaces of higher regularity since it is fundamentally based on the approximation of simple functions. The approximation in $W^{k,p}$, if possible, would require more sophisticated techniques. In the discrete setting, such results have been obtained ([5, 11, 16]).

(8) Other Neural ODEs. In our poofs we have considered controls $W(t), A(t), b(t)$. But as we have seen, in each step of the control iteration only a few of the available control components were activated. This suggests that similar results can be achieved with fewer controls. For instance,
  - $W = Id$. We could consider the model

$$\dot{x} = \boldsymbol{\sigma}(A(t)x + b(t))$$

    on which the vector field $\boldsymbol{\sigma}$ is always pointing towards the first quadrant, which makes it impossible to achieve purposes such as the Universal Approximation (Section 5) or the approximate control of NTEs (Section 6).
  - We could also consider the variant

$$\dot{x} = W(t)\boldsymbol{\sigma}(x) + b(t)$$

Our results could be extended for this type of Neural ODEs, using in an essential way the fact that a quadrant of the phase space remains invariant under the dynamics. However, this structure of the NODE introduces further rigidity in the dynamics and the proofs require more complex arguments.

Another perspective is to generalize the results for second-order Neural ODEs such as continuous versions of Momentum Residual Neural Networks [33].

(9) Optimal Transport. Theorems 5 and 6 naturally establish a path towards optimal transport theory.

Optimal transport can be formulated in a dynamic framework. Namely, given two positive measures ρ^0 and ρ^1 with the same mass, the goal is to find an optimal vector field V and a solution ρ of the continuity equation

$$(7.1) \qquad \begin{cases} \partial_t \rho + \mathrm{div}_x \left[V(x,t)\rho \right] = 0 \\ \rho(t=0) = \rho^0 \\ \rho(t=T) = \rho^1 \end{cases}$$

so that ρ transports ρ^0 into ρ^1 ([2, 3, 7, 27]). This can be interpreted as an optimal controllability problem, the vector field V being aimed to optimize the energy

$$\mathcal{W}_2(\rho^0, \rho^1) = T \inf \int_0^T \int_{\mathbb{R}^d} \|V(x,t)\|^2 d\rho(t) dt$$

subject to (7.1). Here $\mathcal{W}_2$ stands for the Wasserstein-2 distance. The control problem (7.1) is equivalent on finding geodesics in the measure space.

Theorems 5 and 6 guarantee that Neural transport equations (6.1) and (6.6) allow to transport approximately ρ^0 into ρ^1. Analyzing how close are the optimal flows of Neural transport equations from the actual optimal solution of the dynamic optimal transport is an interesting and challenging problem.

<div align="center">ACKNOWLEDGMENTS</div>

<div align="center">APPENDIX A. PROOF OF THE FUNDAMENTAL COMPRESSION LEMMA</div>

Lemma 1 (Fundamental Compression Lemma). *Let $\Omega \subset \mathbb{R}^d$ be bounded, and consider hyperplanes, strips and sets as in ($\mathbf{H}$). Assume that the set of targets $\{\alpha_m\}_{m=1}^M \subset \mathbb{R}^d$ fulfil $\alpha_m^{(1)} \neq \alpha_{m'}^{(1)}$ if $m \neq m'$. For any $T > 0$ and any $0 < \eta < 1/N$ small enough, there exist piecewise-constant controls $A, W \in L^\infty((0,T); \mathbb{R}^{d \times d})$ and $b \in L^\infty((0,T); \mathbb{R}^d)$ such that*

$$\phi_1(\mathcal{H}_l; A, W, b)^{(1)} \subset [\alpha_{m(l)}^{(1)} - C\eta, \alpha_{m(l)}^{(1)} + C\eta] \qquad \textit{for every } 1 \leq l \leq N,$$

$$\mathrm{diam}_{(k)}\phi_1(\Omega; A, W, b) < \eta \qquad \textit{for every } 2 \leq k \leq d,$$

$$\phi_1(\Omega; A, W, b) \subset \mathbb{B}(0, K)$$

with K and C independent of N, N_Γ and η. Furthermore, fixing $T = 1$, one has that:

$$\|A\|_{L^\infty((0,1);\mathbb{R}^{d \times d})} = 1, \qquad \|b\|_{L^\infty((0,1);\mathbb{R}^d)} \lesssim_\Omega N,$$

$$\|W\|_{L^\infty((0,1);\mathbb{R}^{d \times d})} \lesssim_{\Omega,\alpha,d} N\left(\frac{1}{\eta} + \frac{1}{\zeta} + \log\left(\frac{1}{\eta\zeta}\right)\right).$$

The number of discontinuities of the controls A, W, b is of the order of N.

Here, the subscript $\lesssim_\alpha$ denotes the dependence on the set $\{\alpha_m\}_{m=1}^M$.

Proof. In the whole proof we fix the L^∞ norm of W to be 1, and we study the time needed for each step to be completed. By rescaling, this will give us the L^∞ norm of W.

1. Compression of Ω_c.

In this step we compress Ω_c. Notice that in all this process we will be transforming also Ω, being able to bound the transformation of Ω_h whose diameter will remain very close to the diameter of Ω_c. We will compress the sets $\mathcal{H}_l$ so that:

- All hyperrectangles, after applying the controls, will have a diameter of the order of η as small as we want.
- The compressed sets will be separated by a fix constant independent of η.

This will allow us to apply the simultaneous controllability after a refinement of Theorem 1 made in the next step of the proof.

There will be two phases, firstly we compress the $x^{(k)}$-coordinates for $k = 1, ..., d - 1$. As we will see later on, with this compression the sets $\mathcal{H}_l$ will have been transformed and moved in a particular configuration that will allow the the compression in the $x^{(d)}$-coordinate. Moreover, while compressing the the $x^{(d)}$-coordinate we will create a separation between the compressed sets.

For each k from 1 to $d - 1$ apply successively the following steps:

1.1 Compression of the $x^{(k)}$-coordinates $1, ..., d - 1$.

Here below we describe a procedure that will be recursively applied using successively all hyperplanes $\{x^{(k)} = c_{k,n}\}$ starting from $n = 1$ until $n = N_{\Gamma,k}$.

1.1.1 First Compression.
We set the hyperplane $\{x^{(k)} = c_{k,n}\}$ and the controls

$$A_{ij} = -\delta_{ki}\delta_{kj}, \quad b_i = c_{k,n}\delta_{k,i}.$$

We choose W so that the field points towards the hyperplane, generating a contraction. We solve (2.3) for T large enough so that the hyperrectangles that have been affected by the nonzero vector field have a $\mathrm{diam}_{(k)}$ lower than δ (see Figure A.1), i.e.

$$\mathrm{diam}_{(k)}\phi_T(\mathcal{H}_l) < \delta \quad \text{if } \max_{x \in \mathcal{H}_l} x^{(k)} < c_{k,n}.$$

By $\phi_T(\mathcal{H}_l)$ we understand the flow applied to the set $\mathcal{H}_l$ until time T with the aforementioned controls.

We make an abuse of notation, and we redefine $\mathcal{H}_l := \phi_T(\mathcal{H}_l)$. This will be done after each step in the proof to avoid the long notation of applying sequentially different flows with different controls.

The time needed for making this compression is of the order of:

$$(A.3) \qquad\qquad\qquad T_{\text{step 1.1.1}} \sim_\Omega \log\left(\frac{1}{\delta}\right).$$
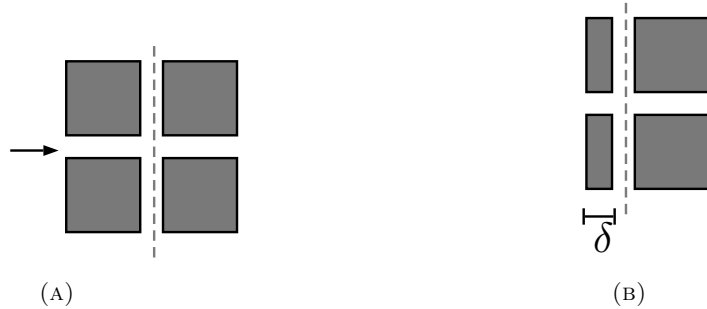

FIGURE A.1. Qualitative representation of Step 1.1.1 (A) Initial meshing of hypercubes where a vector field is applied to compress, (B) qualitative representation of the resulting transformation.

1.1.2 Separation.
We choose a parallel hyperplane to the one in the step above $\{x^{(k)} = c_{k,n} + (1/4)\zeta\}$ and the matrix

$$A_{ij} = -\delta_{ki}\delta_{kj}, \quad b_i = \left(c_{k,n} + \frac{1}{4}\zeta\right)\delta_{k,i}.$$

We choose W so that the field points to $-\infty$ in the $x^{(d)}$-axis. We solve (2.3) for T large enough so that the sets that have been affected by the nonzero vectorfield, have

lower values in the $x^{(d)}$-coordinate than the sets that have not been affected by the vectorfield (in the region in which the field has been **0**) (see Figure A.2), i.e.

$$\max_{x \in \phi_T(\mathcal{H}_i)} x^{(d)} < \min_{\xi \in \mathcal{H}_j} \xi^{(d)} \quad \text{if} \max_{x \in \mathcal{H}_i} x^{(k)} < c_{k,n}, \quad \max_{x \in \mathcal{H}_j} x^{(k)} > c_{k,n}.$$

Given that the side of the removed strip is of the order of ζ, the field will be bounded by below by a quantity of the order of ζ. The space needed for the translation is a constannt depending on Ω. Therefore, the time needed for achieving the separation is of the order of

(A.4)
$$T_{\text{step 1.1.2}} \sim_\Omega \frac{1}{\zeta}.$$

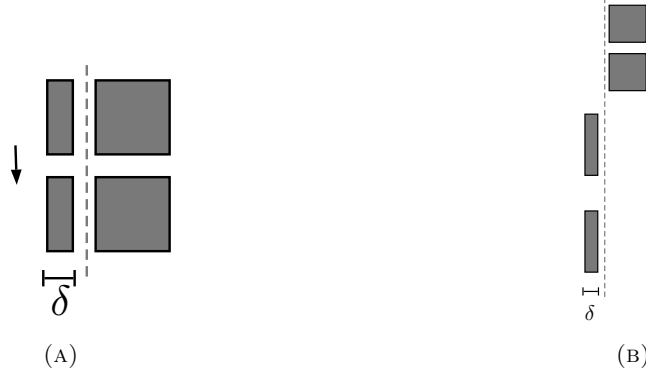(A)   (B)

FIGURE A.2. Qualitative representation of step 1.1.2 (A) vector field that is going to be applied to separate, (B) qualitative representation of the resulting transformation.

1.1.3 **Second Compression.** We apply a compression in the $x^{(k)}$ coordinate, by setting a hyperplane $\{x^{(k)} = c_{k,n+1}\}$. Setting

$$A = -\delta_{ki}\delta_{kj}, \quad b_i = c_{k,n+1}\delta_{k,i}$$

and W so that the field points towards the hyperplane, making it attractive (see Figure A.3). We solve (2.3) for T large enough so that:

$$\sup_{x \in \phi_T(\mathcal{H}_l)} |x^{(k)} - c_{k,n+1}| < \delta, \quad \text{if} \max_{x \in \mathcal{H}_l} x^{(k)} < c_{k,n+1}.$$

The needed time is:

(A.5)
$$T_{\text{step 1.1.3}} \sim_\Omega \log\left(\frac{1}{\delta}\right).$$

As said before, this procedure is applied sequentially for every $1 \leq k \leq d-1$ and every $1 \leq n \leq N_{\Gamma,k} - 1$ obtaining that the resulting sets satisfy

$$\sup_{x \in \phi_T(\mathcal{H}_l)} |x^{(k)} - c_{k,N_{\Gamma,k}}| < \delta, \quad \forall k \in \{1,...,d-1\}, \quad \forall l \in \{1,...,N\}.$$

The set Ω has transformed, and $\text{diam}_{(k)}(\phi_T(\Omega)) < \delta$ for $1 \leq k \leq d-1$.

Before continuing, we will "*move the sets away from the targets*". The reason of this particular action will be better understood in Step 2 of this proof.

Let $c = \text{diam}(\{0\} \cup \{\alpha_m\}_{m=1}^M) + 1$, consider the ball $\mathbb{B}(0,c)$ and pick the hyperplane

(A.6)
$$\left\{x^{(1)} = c\right\}.$$

Choose A and W so that the vector field, in its nonzero region in the phase space, follows the Cartesian direction of the first element of the canonical basis and makes the hyperplane (A.6) attractive. We solve (2.3) for T long enough so that:

$$\sup_{x \in \phi_T(\mathcal{H}_l)} |x^{(1)} - c| < \delta, \quad \forall l \in \{1,...,N\}.$$

FIGURE A.3. Second compression phase of step 1.1.3 (A) a vector field is generated towards an attractive hyperplane, (B) the qualitative result

This transformation does not affect the smallness of the k coordinates from $k = 1, ..., d - 1$.

The overall required time of this step is the following

$$T_{\text{step 1.1}} \sim_{\Omega, d, \alpha} N_\Gamma \left(\log \left(\frac{1}{\delta} \right) + \frac{1}{\zeta} \right)$$

where by $\sim_\alpha$ we denote the dependence on the targets.

In the separation step, Step 1.1.2, the set deforms since the field is not homogeneous along the set. We must ensure that a hyperplane with normal vector equal to an element of the canonical basis can fit between the transformed sets. This is the main reason of the first compression step, Step 1.1.1. This step reduces the $\text{diam}_{(k)}$ of the set making that the field in the next step, Step 1.1.2 is almost constant on the hyperrectagle. Let us quantify the present discussion.

Before Step 1.1.2 the distance between two neighbor rectangles is ζ. After applying Step 1.1.2 $(d - 1)N_\Gamma$ times the separation is:

$$\zeta - (d - 1)N_\Gamma T_{\text{step 1.1.2}} \zeta \delta$$

so it will be enough if

$$\delta \leq \mathcal{C} N_\Gamma^{-1} \zeta$$

for a certain constant $\mathcal{C}$ to guarantee that after the whole process the distance between two neighbor rectangles is of the order of ζ.

Notice that A has been chosen to be unitary, also b is bounded since we have chosen only among the original hyperplanes defined in **Step 1**. Furthermore, the number of switches we employed are explicit:

(A.7)   The number of switches in the A control is $d - 1$.

(A.8)   The number of switches in the b control and W control are of the order of N_Γ.

1.2 **Compression of the d-coordinate.**

Now, the task is to compress the $x^{(d)}$- coordinate and separate enough each set while keeping a small diameter. After the transformations made above, the sets can be separated by $N - 1$ hyperplanes of the form

$$\left\{ x^{(d)} = C_{d,l} \right\} \qquad 1 \leq l \leq N$$

For doing so, for l from 1 until N we proceed as follows:

1.2.1 **Compression.** Take $\{x^{(d)} = C_{d,l}\}$ and $A_{ij} = \delta_{di}\delta_{dj}$, $b_i = -C_{d,l}\delta_{di}$ and we choose W so that the hyperplane $\{x^{(d)} = C_{d,l}\}$ is attractive. We choose $T > 0$ so that

$$\sup_{x \in \phi_T(\mathcal{H}_l)} |x^{(d)} - C_{l,d}| < \delta \qquad \forall l \text{ such that } \inf_{x \in \mathcal{H}_l} x^{(d)} > C_{d,l}$$

see Figure A.4.

FIGURE A.4. Compression the $x^{(d)}$ coordinate, step 1.2.1 (A) a vector field is generated towards an attractive hyperplane. (B) The resulting effect.
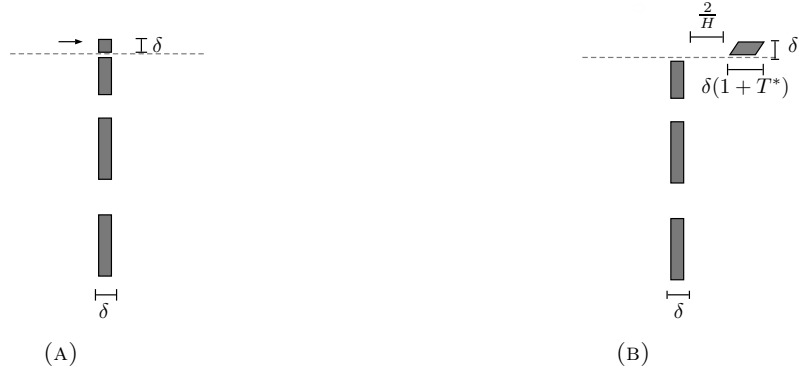

FIGURE A.5. Displacement in the $x^{(1)}$ direction, Step 1.2.2 (A) A vector field is generated in order to push the set in the $x^{(1)}$-coordinate (B) Qualitative representation of the resulting effect.

The diameter of $\phi_T(\mathcal{H}_l)$ is lower than $\operatorname{diam}(\phi_T(\mathcal{H}_l)) < \sqrt{d}\delta$.

1.2.2 **Separation.** We fix a parallel hyperplane to the one before. We set it slightly below at a distance of the order of the removed band, ζ, $\{x^{(d)} = C_{d,l} - (1/4)\zeta\}$, and we choose $A_{ij} = \delta_{di}\delta_{dj}$, $b_i = (-C_{d,l} + (1/4)\zeta)\delta_{di}$ and W such that the field points to the $x^{(1)}$-coordinate. We solve (2.3) for T^* large enough so that

$$\inf_{x \in \phi(\mathcal{H}_l)} \left| c_{1,N_\Gamma} - x^{(1)} \right| = \frac{2}{N}.$$

This would require $T^* = 8/(N\zeta)$.

When separating the sets one should ensure that at the end of the process we stay in a bounded set independently of N.

After applying this movement along the $x^{(1)}$-coordinate, the diameter of the set $\mathcal{H}_l$ has grown being:

$$\operatorname{diam}(\phi(\mathcal{H}_l))^2 \leq \left((d-1) + (1+T^*)^2 \right) \delta^2$$

$$\left((d-1) + \left(1 + \frac{8}{N\zeta} \right)^2 \right) \delta^2.$$

The time needed to perform this steps, the number of switches and the norm of b are estimated by the following:

$$(A.9) \qquad T_{\text{step 1.2}} \sim_\Omega N \left(\log \left(\frac{1}{\delta} \right) + \frac{1}{\zeta} \right), \quad \|b\|_\infty \sim N$$
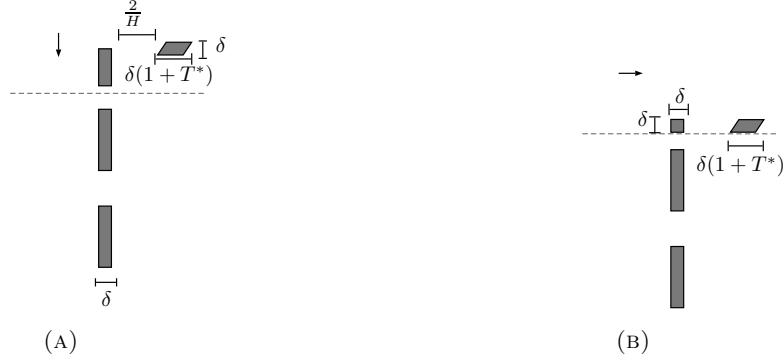
FIGURE A.6. Compression phase for the $x^{(d)}$ direction, (A) a vector field is generated towards an attractive hyperplane. (B) Qualitative representation of the resulting effect.

(A.10)
$$\text{The number of switches of the } A \text{ control has been } 1.$$
$$\text{The number of switches of the } b \text{ and } W \text{ control is of the order of } N.$$

We apply iteratively this process (see Figure A.6) for all hyperplanes $\{x^{(d)} = C_{d,l}\}$ (obtaining at the end Figure A.7). Since the velocity field has different strength at different parts of the set, the distance between two $\mathcal{H}_l$ can diminish.

In order to guarantee that after the whole process we have enough distance between the sets, we have to require δ to be small enough. When we apply the horizontal translation, the difference in the vector field inside each set is δ. Hence, we know that the distance in their $x^{(1)}$-coordinate between the sets reduces as:

$$\min_{\substack{x \in \phi_t(\mathcal{H}_l) \\ z \in \phi_t(\mathcal{H}_{l+1})}} |x^{(1)} - z^{(1)}| \geq \frac{2}{N} - (\delta t).$$

If we impose that, after T^*N units of time, the distance in the $x^{(1)}$-coordinate is bigger than $1/N$, this implies that the requirement is the following:

(A.11)
$$\delta < \frac{\zeta}{8N}$$

The requirement (A.11) will be not the only one. Now the task is to guarantee that $\text{diam}(\phi_T(\Omega_h))$ is uniformly bounded for h small enough and that the maximum diameter of all the sets $\phi_T(\mathcal{H}_l)$ is as small as we desire.

- The requirement for the diameter of each set follows from
$$\text{diam}(\phi(\mathcal{H}_l))^2 \leq \left((d-1) + (1 + NT^*)^2\right)\delta^2$$

(A.12)
$$\text{diam}(\phi(\mathcal{H}_l))^2 \leq \left((d-1) + \left(1 + \frac{8}{\zeta}\right)^2\right)\delta^2 \leq \eta^2$$

- For the diameter of $\phi_T(\Omega_h)$ one has that
$$\text{diam}(\phi_T(\Omega_h))^2 \leq \delta^2(d-1) + N^2\left(\frac{2}{N} + \max_l \text{diam}(\mathcal{H}_l)\right)^2$$
$$\leq \delta^2(d-1) + 4 + N^2\eta^2 + 4N\eta.$$

Choosing any η such that

(A.13)
$$\eta < \frac{1}{N}$$

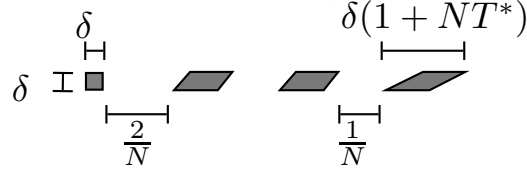we would guarantee that $\phi_T(\Omega)$ remains bounded.

FIGURE A.7. Qualitative representation of the final configuration of the compression process. The sets $\phi_T(\mathcal{H}_l)$ are uniformly separated by $1/N$ and their diameter is of the order of δ.

With all the process described, we have moved the sets downwards in the $x^{(d)}$-component, this translation depends on N. However, by putting an attractive hyperplane at the of the form $x^{(d)} = c$ at the desired place, we can move the set without having a dependence on N.

Now the sets are at distances bigger than $1/N$ between them and each one of them has a bound in its diameter (A.12). We can choose δ as small as we want and we can guarantee that:

$$\text{diam}(\phi(\mathcal{H}_l)) < \eta \quad \Longleftrightarrow \quad \delta^2 < \eta^2 \left((d-1) + \left(1 + \frac{8}{\zeta} \right)^2 \right)^{-1}.$$

Using the expressions (A.9) and (A), we find that the time horizon needed in Step 1 of this proof scales in the following manner:

$$(A.14) \qquad T_{\text{step 1}} \sim_{\Omega,d,\alpha} N \left(\log \left(\frac{1}{\delta} \right) + \frac{1}{\zeta} \right) + N_\Gamma \left(\log \left(\frac{1}{\delta} \right) + \frac{1}{\zeta} \right)$$

The norm of the control b depends on N as seen in the treatment of the $x^{(d)}$-coordinate (Step 1.2) (A.9)

$$(A.15) \qquad \qquad \|b\|_\infty \sim N$$

Considering (A.10)(A.7) the total number of switches on the control A is

(A.16) The number of switches of the A control has been d,

while for b and W we use (A.8) and (A.10)

(A.17) The number of switches of the b and W control is of the order of N

2. Ordering and Grouping.

Let ϕ_T be the flow associated to the compression phase, and let Ω and $\mathcal{H}_l$ be defined as in **Step 1** of this proof. We redefine Ω and $\mathcal{H}_l$ as $\Omega := \phi_T(\Omega)$ and $\mathcal{H}_l := \phi_T(\mathcal{H}_l)$. Without loss of generality, we make a change of coordinates, for simplifying notation, in a way that:

$$\Omega \subset \{x : \quad |x^{(1)}| \leq 2, |x^{(k)}| < \eta \quad k \geq 2\}.$$

Moreover, let $K = 2 \, \text{diam} \left(\Omega \cup \{\alpha_m\}_{m=1}^M \cup \{0\} \right)$ and consider the ball $\mathbb{B}(0, K)$. Naturally one has that:

$$\Omega \subset \mathbb{B}(0, K), \qquad \{\alpha_m\}_{m=1}^M \subset \mathbb{B}(0, K).$$

Note that by Step 1.1, all the sets are *away* from the targets and in particular, the following condition holds:

$$(A.18) \qquad \qquad \alpha_{m'}^{(1)} \notin \mathcal{H}_l^{(1)} \quad \text{if } \alpha_{m(l)} \neq \alpha_{m'}.$$

Remind that by $(\omega)^{(k)}$ we denote the k-th component of the set ω, i.e.

$$\omega^{(k)} := \{x \in \mathbb{R} : \exists y \in \omega \quad \text{such that } y^{(k)} = x\}.$$

Now we build a control that will control the first component of $\mathcal{H}_l$ to their corresponding target while ensuring that after this control process, (A.18) still holds. This will allow us to proceed recursively.

The procedure we are going to show is, in particular, an improvement of Theorem 1 for keeping track properly of Ω_h while applying our controls.
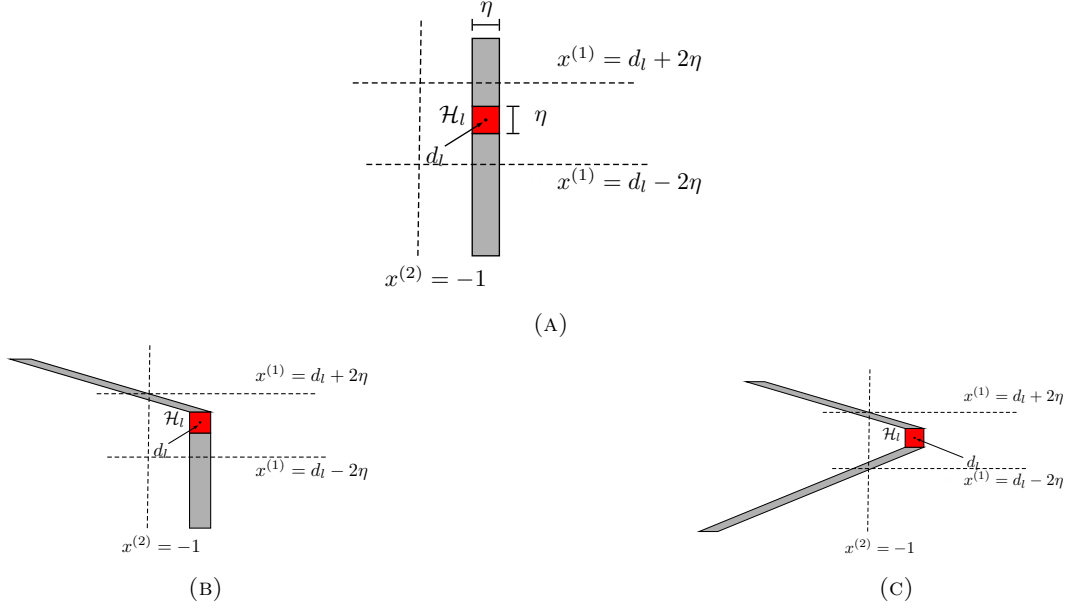
(A)

(B)

(C)

FIGURE A.8. (A) Representation of the bounds (A.19) and the hyperplane (A.20), Ω is contained in the hyperrectangle, the set $\mathcal{H}_l$ in red and $\Omega \setminus \mathcal{H}_l$ in grey. (B) The resulting set after applying the flow $\psi_{T_{Sep1}}^{Sep1}$ to Ω. (C) The resulting set after applying $\psi_{T_{Sep2}}^{Sep2}$ to $\psi_{T_{Sep1}}^{Sep1}(\Omega)$.

The time horizon will depend on the number of elements N. The mission is to make use of the explicit controlled flows to obtain sharper bounds and independent on the number of elements. Since the time horizon depends on N, the direct application of Grönwall's inequality is not of practical use for bounding Ω_h independently of N. This is the reason why we examine the dynamics in more detail.

This Lemma that we are proving, leaves the system in a configuration in which the simultaneous controllability of Theorem 2 can be applied. In the application of Theorem 2, after applying Theorem 1, we only set controls that endow flows that follow directions in the orthogonal subspace of $\mathrm{Span}((1, 0, ..., 0))$. Due to the Cartesian structure of our flows see that we only have to ensure that after the refinement of Theorem 1 the condition $\mathrm{diam}_{(k)}(\Omega) = \mathcal{O}(\eta)$ for any $k \geq 2$ still holds.

For every $l \in \{1, ..., N\}$ consider the center of $\mathcal{H}_l$,

$$d_l = \frac{1}{|\mathcal{H}_l|} \int_{\mathcal{H}_l} x \, dx$$

and we proceed sequentially.

(1) Observe that Ω is bounded in the second coordinate by:

$$x^{(2)} = \eta, \qquad x^{(2)} = -\eta.$$

Fix the following bounds:

(A.19a) $$x^{(1)} = d_l^{(1)} + 2\eta$$

(A.19b) $$x^{(1)} = d_l^{(1)} - 2\eta$$

and the hyperplane(A.20).

(A.20) $$\left\{ x^{(2)} = -1 \right\}$$

as Figure A.8.A shows.

(2) We select the hyperplane $\{x^{(1)} = d_l + \eta\}$. We apply the flow as in Theorem 1, pointing at the negative side in the $x^{(2)}$-component (as done in Theorem 1 in equation (3.4) and selecting the control A as in equation (3.3)). Then, the bound in $x^{(2)}$ will evolve as

(A.21) $$x^{(2)}(t) = -\max(x^{(1)} - d_l^{(1)} - \eta, 0)t + \eta.$$

(3) Let us denote by ψ_T^{Sep1} the associated flow with such controls. We want to find T_{Sep1} such that:

$$\text{diam}_{(1)}\left(\psi_{T_{Sep1}}^{Sep1}(\Omega) \cap \{x^{(2)} \geq -1\} \cap \{x_1 \geq d_l + \eta\}\right) = \eta$$

For doing so, we compute the intersection between (A.21) and (A.20), taking $x^{(1)}$ as in (A.19a) and we find a lower bound for the time in which the translation has to be applied (see Figure A.8.B):

(A.22) $$T_{Sep1} \sim 1 + \frac{1}{\eta}.$$

(4) We can do an analogous procedure for the hyperplane $x^{(1)} = d_l - \eta$, and we obtain also the same lower bound for the time (A.22). Let us denote by $\psi_{T_{Sep2}}^{Sep2}$ the resulting flow. After this point, we obtain a structure similar to Figure A.8.C. Let us define, by convenience,

$$\psi^{Sep} := \psi_{T_{Sep2}}^{Sep2} \circ \psi_{T_{Sep1}}^{Sep1}.$$

Now we have that:

$$\text{diam}_{(1)}(\psi^{Sep}(\Omega) \cap \{x_2 \geq -1\}) \leq 4\eta, \qquad (\psi^{Sep}(\Omega))^{(1)} = (\Omega)^{(1)}.$$

(5) Now we choose the hyperplane $\{x^{(2)} = -1\}$ and we apply a translation movement. Denote by $\psi_{T_{Mov}}^{Mov}$ the translation movement flow. We choose a vector field in the $x^{(1)}$ direction pointing to the first component of the target of d_l, $\alpha_{m(l)}^{(1)}$, during the required time T_{Mov} to reach the target. Since the distance between the hyperplane $\{x^{(1)} = -1\}$ and d_l is independent of η, T_{Mov} does not depend on the parameter η.

$$T_{Mov} = \frac{|d_l^{(1)} - \alpha_{m(l)}^{(1)}|}{|d_l^{(2)} + 1|}.$$

Then we have that, defining $a_1 := \min\{\alpha_{m(l)}^{(1)}, d_l^{(1)}\}$ and $a_2 := \max\{\alpha_{m(l)}^{(1)}, d_l^{(1)}\}$, there exists a constant C independent of η such that:

$$\left(\psi_{T_{Mov}}^{Mov}\left(\psi^{Sep}(\Omega) \cap \{x_2 \geq -1\}\right)\right)^{(1)} \subset [a_1 - C\eta, a_2 + C\eta] \subset \mathbb{B}(0, K)^{(1)}$$

for η small enough. Moreover

$$\psi_{T_{Mov}}^{Mov}\left(\psi^{Sep}(\Omega) \cap \{x_2 < -1\}\right) = \psi^{Sep}(\Omega) \cap \{x_2 < -1\}$$

$$\max_{x \in \psi_{T_{Mov}}^{Mov}(\psi^{Sep}(\mathcal{H}_l))} |x - \alpha_{m(l)}| \leq C_2\eta$$

where $C_2 = ((d-1) + C\eta)^{1/2}$.

For guaranteeing that after this transformation we still fulfill condition (A.18), we have to choose η small enough, depending only on the target configuration, i.e. η small so that

$$\alpha_{m'}^{(1)} \notin [\alpha_{m(l)}^{(1)} - C\eta, \alpha_{m(l)}^{(1)} - C\eta] \qquad \text{if } \alpha_{m'} \neq \alpha_{m(l)}.$$

(6) We have deformed substantially the set in the $x^{(2)}$-component. Now, in order to apply the argument for the next set $\mathcal{H}_l$ we need to compress around that coordinate. So we choose the hyperplane $\{x^{(2)} = 0\}$ and we apply a contraction, denote it by ψ_T^{Cont}, until time T_{Cont} for which the following is satisfied

$$\text{diam}_{(k)}\left(\psi_{T_{Cont}}^{Cont}\left(\psi_{T_{Mov}}^{Mov}\left(\psi^{Sep}(\Omega)\right)\right)\right) < \eta.$$

Furthermore, one has that:

$$\left(\psi_{T_{Cont}}^{Cont}\left(\psi_{T_{Mov}}^{Mov}\left(\psi^{Sep}(\Omega) \cap \{x_2 < -1\}\right)\right)\right)^{(1)} \subset (\Omega)^{(1)}.$$

(7) Redefining $\Omega := \psi_{T_{Cont}}^{Cont}\left(\psi_{T_{Mov}}^{Mov}\left(\psi^{Sep}(\Omega)\right)\right)$, one can apply the argument recursively for the next $\mathcal{H}_l$ until all $\mathcal{H}_l$'s have its first component approximately controlled.

Now there are M sets of diameter of the order of η. The control time of the whole process $T_{\text{step 2}} = N(T_{Sep1} + T_{Sep2} + T_{Mov} + T_{Cont})$ is of the order of

$$(A.23) \qquad T_{\text{step 2}} \sim_\alpha N\eta^{-1} + N \log\left(\frac{1}{\eta}\right) + N,$$

while the norms of the control b remained uniformly bounded.

(A.24)   The number of switches of the controls A, W, b has been of the order of N

Comparing the two steps, the final time horizon is of the order of

$$T \lesssim_{\Omega,\alpha,d} N\left(\frac{1}{\eta} + \frac{1}{\zeta} + \log\left(\frac{1}{\eta\zeta}\right)\right).$$

The norms of the control b is dependent on N from Step 1.2

$$\|b\|_{L^\infty} \lesssim_\Omega N.$$

Finally the number of switches is of the order of N

The number of switches of the controls A, W, b has been of the order of N.

$\square$

Remark A.1 (The Bottleneck). *The compression process done in the Lemma above is constituted by several stages, combining contractions and separations in different manner. However, there is a control stage which is giving the bound on the control cost which is the Ordering and Grouping step (Step 2 of the proof). In producing certain flows, such as separations, NODEs require to have higher controls than when producing contractions.*

References

[1] A. Agrachev and A. Sarychev. Control on the manifolds of mappings as a setting for deep learning. *arXiv preprint arXiv:2008.12702*, 2020.

[2] J.-D. Benamou and Y. Brenier. The optimal time-continuous mass transport problem and its augmented Lagrangian numerical resolution. 1998.

[3] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[4] C. Bishop and Y. Peres. *Fractals in probability and analysis*, volume 162. Cambridge University Press, 2017.

[5] M. Burger and A. Neubauer. Error bounds for approximation with neural networks. *Journal of Approximation Theory*, 112(2):235–250, 2001.

[6] R. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[7] Y. Chen, T. Georgiou, and A. Tannenbaum. Matrix optimal mass transport: a quantum mechanical approach. *IEEE Transactions on Automatic Control*, 63(8):2612–2619, 2017.

[8] L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[9] C. Cuchiero, M. Larsson, and J. Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020.

[10] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[11] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) relu networks. *arXiv preprint arXiv:1905.02199*, 2019.

[12] M. Duprez, M. Morancey, and F. Rossi. Approximate and exact controllability of the continuity equation with a localized vector field. *SIAM Journal on Control and Optimization*, 57(2):1284–1311, 2019.

[13] M. Duprez, M. Morancey, and F. Rossi. Minimal time for the continuity equation controlled by a localized perturbation of the velocity vector field. *Journal of Differential Equations*, 269(1):82–124, 2020.

[14] C. Esteve, B. Geshkovski, D. Pighin, and E. Zuazua. Large-time asymptotics in deep learning. *arXiv preprint arXiv:2008.02491*, 2020.

[15] K. Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[16] I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05):803–859, 2020.

[17] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] K. Hornik, M. Stinchcombe, H. White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[20] S. Jaffard and Y. Meyer. *Wavelet methods for pointwise regularity and local oscillations of functions*, volume 587. American Mathematical Soc., 1996.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[22] M. Leshno, V. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[23] Q. Li, T. Lin, and Z. Shen. Deep learning via dynamical systems: An approximation perspective. *arXiv preprint arXiv:1912.10382*, 2019.

[24] J.-L. Lions. Exact controllability, stabilization and perturbations for distributed systems. *SIAM review*, 30(1):1–68, 1988.

[25] J. Lohéac and E. Zuazua. From averaged to simultaneous controllability. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 25, pages 785–828, 2016.

[26] W. Osgood. A Jordan curve of positive area. *Transactions of the American Mathematical Society*, 4(1):107–112, 1903.

[27] G. Peyré, M. Cuturi, et al. Computational optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[28] B. Piccoli, F. Rossi, and E. Trélat. Control to Flocking of the Kinetic Cucker–Smale Model. *SIAM J. Math. Anal.*, 47(6):4685–4719, 2015.

[29] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8(1):143–195, 1999.

[30] L. Pontryagin. *Mathematical theory of optimal processes*. Routledge, 2018.

[31] Y. Privat, E. Trélat, and E. Zuazua. Optimal shape and location of sensors for parabolic equations with random initial data. *Arch. Ration. Mech. Anal.*, 216(3):921–981, 2015.

[32] Y. Privat, E. Trélat, and E. Zuazua. Optimal observability of the multi-dimensional wave and schrödinger equations in quantum ergodic domains. *Journal of the European Mathematical Society*, 18(5):1043–1111, 2016.

[33] M. Sander, P. Ablin, M. Blondel, and G. Peyré. Momentum residual neural networks. *arXiv preprint arXiv:2102.07870*, 2021.

[34] M. Shishikura. The Hausdorff dimension of the boundary of the Mandelbrot set and Julia sets. *Annals of Mathematics*, pages 225–267, 1998.

[35] E. Sontag and H. Sussmann. Complete controllability of continuous-time recurrent neural networks. *Systems & control letters*, 30(4):177–183, 1997.

[36] H. Sussmann. A bang-bang theorem with bounds on the number of switchings. *SIAM Journal on Control and Optimization*, 17(5):629–651, 1979.

[37] P. Tabuada and B. Gharesifard. Universal approximation power of deep neural networks via nonlinear control theory. *arXiv preprint arXiv:2007.06007*, 2020.

[38] T. Teshima, K. Tojo, M. Ikeda, I. Ishikawa, and K. Oono. Universal approximation property of neural ordinary differential equations. *arXiv preprint arXiv:2012.02414*, 2020.

[39] M. Tucsnak and G. Weiss. Simultaneous exact controllability and some applications. *SIAM Journal on Control and Optimization*, 38(5):1408–1427, 2000.

[40] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[41] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[42] E. Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

[43] C. E. Yagüe and B. Geshkovski. Sparse approximation in learning via neural odes. *arXiv preprint arXiv:2102.13566*, 2021.

Email address: domenec.ruiz@deusto.es, domenec.ruizi@uam.es

Email address: enrique.zuazua@fau.de