Optimal control for neural ODE in a long time horizon and applications to the classification and simultaneous controllability problems

Jon Asier Bárcena-Petisco

Optimal control for neural ODE in a long time horizon and applications to the classification and ensemble controllability problems

Jon Asier Bárcena-Petisco*

June 7, 2022

Abstract: We study the optimal control in a long time horizon of neural ordinary differential equations which are control-affine or whose activation function is homogeneous. When considering the classical regularized empirical risk minimization problem we show that, in long time and under structural assumption on the activation function, the final state of the optimal trajectories has zero training error if the data can be interpolated and if the error can be taken to zero with a cost proportional to the error. These hypotheses are fulfilled in the classification and ensemble controllability problems for some relevant activation and loss functions. Our proofs are mainly constructive combined with a proof by contradiction: We find that in long time horizon if the final error is not zero, we can construct a less expensive control which takes the error to zero. Moreover, we prove that the norm of the optimal control is constant. Finally, we show the sharpness of our hypotheses by giving an example for which the error of the final state of the optimal trajectory, even if it decays, is strictly positive for any time.

Key words: data classification, exact controllability, neural ODE, nonlinear systems, optimal control, ensemble controllability

AMS subject classification: 34H05, 49N10, 93B05

Abbreviated title: Optimal control for neural ODE

*Department of Mathematics, University of the Basque Country UPV/EHU, Barrio Sarriena s/n, 48940, Leioa, https://orcid.org/0000-0002-6583-866X Spain. E-mail: jonasier.barcena@ehu.eus.

1 Introduction

In this paper we study the optimal control of neural ordinary differential equations for a long time horizon. Neural ODE have been used in Machine Learning in the last five years, a trend started with [Wei17, HR17]. However, they date back to the 90s, when they were already used for the construction of controls (see the survey [Son92]) and when their controllability properties were first studied (see, for example, [Zbi93] and [SS97]). The control systems governed by neural ODE have considerably better controllability properties than linear control systems. In fact, as pointed out in [RBZ21], for a fixed $d \in \mathbb{N}$, if chosen the right neural ODE we can control an arbitrarily large amount of data in $\mathbb{R}^d$, whereas in linear systems we can at most control an amount of data equal to the dimension of the control.

Roughly, though we do it in an abstract setting, the problem under study is the following: Given a dataset $\mathbf{x} = (x^1, \dots, x^N) \in (\mathbb{R}^d)^N_*$, for

$$(\mathbb{R}^d)^N_* := \{(x^1, \dots, x^N) \in (\mathbb{R}^d)^N : x^i \neq x^j \quad \forall i, j \in \{1, \dots, N\} : i \neq j\},$$

we seek to take simultaneously the data set to some target points or regions in $\mathbb{R}^d$ in a given time $T > 0$. The distance to those targets is measured with an error function (also known as *loss function*). The control is the minimizer of the risk minimization functional, which provides a balance between a small cost for the control and a small value for the loss function at the final state of the optimal trajectory.

We study the controllability on control-affine neural networks, which are given by the following equations:

$$\begin{cases} \dot{y}(t) = w(t)\sigma(y(t)) + b(t), \\ y(0) = x, \end{cases} \tag{1.1}$$

for $x \in \mathbb{R}^d$ the initial value, and $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d$ a Lipschitz function, which is called the *activation function*. The functions (w, b) are the controls and they belong to $L^2(0, T; \mathcal{U})$, for $\mathcal{U}$ defined by:

$$\mathcal{U} := \mathcal{M}_{d\times d} \times \mathcal{M}_{d\times 1}.$$

If we want to emphasize the dependence of (1.1) to the initial value and the control, we write $y(\cdot; x, w, b)$. Similarly, we denote the sequence of solutions of (1.1) for some fixed control (w, b) and a data set $\mathbf{x}$ as:

$$y(\cdot; \mathbf{x}, w, b) := \left(y(\cdot; x^1, w, b), \dots, y(\cdot; x^N, w, b)\right).$$

Since σ is Lipschitz, (1.1) is well-posed by the Cauchy-Lipschitz Theorem.

In addition, we also study more compound neural networks, which are given by the equations:

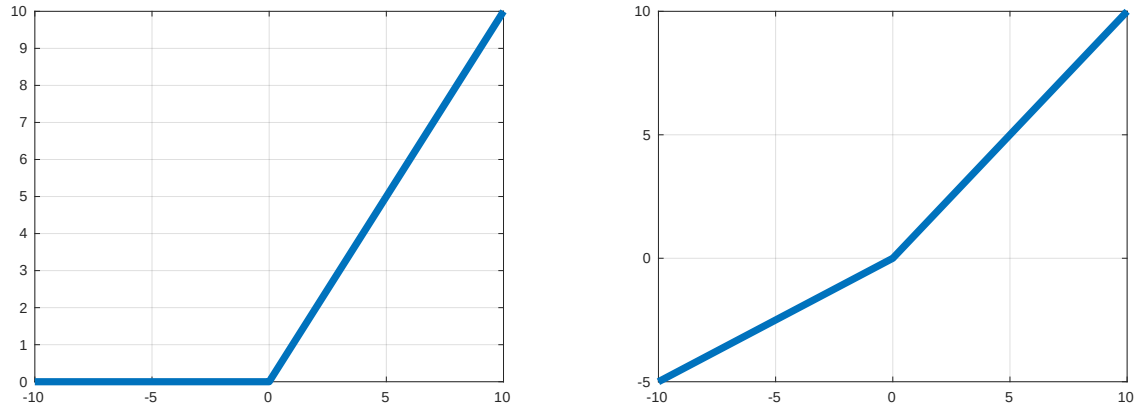
$$\begin{cases} \dot{y}(t) = r(t)\sigma\left(w(t)y(t) + b(t)\right), \\ y(0) = x. \end{cases} \tag{1.2}$$

Here x is the initial value and (r, w, b) is the control, which belongs to $L^2(0, T; \tilde{\mathcal{U}})$, for:

$$\tilde{\mathcal{U}} := \mathcal{M} \times \mathcal{M}_{d\times d} \times \mathcal{M}_{d\times 1},$$

for:

$$\mathcal{M} \subseteq \{M \in \mathcal{M}_{d\times d} : \mathcal{M}_{i,i} \in \{1, -1\}, \quad \forall i = 1, \dots, d, \quad \mathcal{M}_{i,j} = 0, \quad \forall i \neq j\}. \tag{1.3}$$



(a) Rectified linear units.

(b) Parametric linear units.

Figure 1: Some usual activation functions for $d = 1$.

In fact, the intensity of the flow is modelled by (w, b), and the direction of the flow, by r. We may take $\mathcal{M} = \{I\}$, which makes sense when σ admits negative values. However, we have considered the general setting to have relevant results also for the case in which σ is a positive function; that is, in which $\sigma \geq 0$. We assume that the activation function σ is Lipschitz and homogeneous in the sense that:

$$\sigma(\lambda x) = \lambda\sigma(x), \quad \forall\lambda > 0, \quad \forall x \in \mathbb{R}^d. \tag{1.4}$$

This includes important *activation functions* such as rectified linear units, which are given by:

$$\sigma(x) = (\max\{x_1, 0\}, \dots, \max\{x_d, 0\}),$$

see [NH10]; parametric rectified units, given by:

$$\sigma(x) = (\alpha x_1 1_{x_1<0} + x_1 1_{x_1>0}, \dots, \alpha x_d 1_{x_d<0} + x_d 1_{x_d>0}),$$

see [HZRS15]; and, of course, the identity, $\sigma(x) = x$ (see Figure 1 for the graphs of such activations functions in one dimension). As in the previous system, $y(\cdot; x, r, w, b)$ and $y(\cdot; \mathbf{x}, r, w, b)$ denote the solutions of (1.2), which is a well-posed system by the Cauchy-Lipschitz Theorem.

As stated in the first paragraph, we study the properties of any optimal control in a long time horizon. The main contribution of our paper is that, if the data can be interpolated and the error can be taken to 0 with a cost proportional to the current error, we improve the asymptotic bound $\mathcal{O}(1/T)$ for the error of the final state of the optimal trajectory obtained in [EGPZ20] and prove that it is exactly 0 for a sufficiently large time. In fact, in the simulations presented in [EGPZ20, Examples 4.2 and 4.4] the final errors seem to be 0, so we want to determine theoretically if, as their simulation suggests, the error is taken exactly to 0. Even if approximate controllability is usually enough for practical purposes, obtaining null controllability is interesting to broaden the perspective of the field. We work in an abstract setting, though we give concrete examples of problems that satisfy our assumptions, notably the ensemble controllability and classification problems. In ensemble controllability we aim to control two or more independent equations by applying the same control. The study of ensemble controllability dates back to [Rus86] and [Lio88, Chapter 5], and relevant papers on this topic include [TW00, LZ16, WZL17, AB19, RBZ21, RBAZ22]. The main difference of this paper and [RBZ21, RBAZ22] with the previous ones is that in our papers the trajectories satisfy the same differential equation, whereas in the

other papers they satisfy different differential equations (i.e. differential equations which at least do not have the same coefficients). As for the classification problem, it is a simplified version of the ensemble controllability problem, where the objective is to split the data into two sets, for instance, $\{x_1 \leq 1\}$ and $\{x_1 \geq 1\}$. An additional contribution of our paper is an example of neural ODE and loss functions where the error can be taken to 0, but for all time $T > 0$ the error at time T of the optimal trajectories is strictly positive. This illustrates that the results are far from being trivial.

This paper follows a well-established research line that studies the properties of the optimal control and trajectories in a long time horizon. This allows, for instance, that when doing numerical simulations, one may identify when a local minimum is not an optimal control. The *turnpike property* is a notion developed since the 1950s which means that when minimizing certain functionals all the optimal trajectories are most of the time near some specific state (the *turnpike*) independently of the initial value and the target (see, for instance, [McK76] and [DSS87]). An important recent paper regarding the study of the turnpike property is [PZ13], the first work which provides rigorous mathematical proof and a framework for the turnpike property for linear quadratic optimal control problems. Also, interesting recent studies on the turnpike property include discrete optimal control problems in [DGSW14] and [GG18], finite-dimensional nonlinear control problems in [TZ15] and [Tré20], optimal control problems for hyperbolic systems in [GTZ16] and [SPZ19], general Hilbert spaces in [TZZ18] and [BP20], boundary optimal control problems in [GH19], Navier-Stokes equation in [Zam18], fractional parabolic equation in [WZ21], Hamilton-Jacobi in [EKPZ20], the finite time turnpike phenomena in [GSZ21], hands-off controls in [SN20] and in deep neural networks in [FHS21], and Lipschitz nonlinear functions in [EGPZ22]. Finally, the optimal control for neural ODE is also studied in [EYG21], where the authors consider the cost of the L^1-norm of the control instead of the L^2-norm and obtain that for that norm the optimal control satisfies some sparsity properties and that the error of the final state belongs to $\mathcal{O}(1/T)$.

2 Main results

2.1 Optimal trajectories for control-affine neural ODE

As stated in the introduction, we study the optimal control of a data set ruled by a neural ODE. To measure how far the data is from the objective we introduce the *error function* (also referred in the literature of Machine Learning as *loss function*) $\mathcal{E} : (\mathbb{R}^d)^N \mapsto \mathbb{R}^+ := [0, \infty)$. We assume that $\mathcal{E}$ is continuous and satisfies the Hypotheses 1 and 2, which are later introduced in this section.

This allows to define the *empirical risk minimization functional for a target time* T:

$$J_T(w, b) := \mathcal{E}(y(T; \mathbf{x}, w, b)) + \int_0^T |(w(t), b(t))|^2 dt, \tag{2.1}$$

where y denotes a solution of (1.1) and $|\cdot|$ denotes the Frobenius norm. In this paper we denote any minimizer of J_T by (w_T, b_T). Moreover, the trajectories induced by such minimizers, called *optimal trajectories*, are denoted by $y_T(t; \mathbf{x}) := y(t; \mathbf{x}, w_T, b_T)$.

Example 2.1. One usual definition for the error function is

$$\mathcal{E}(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} E_i(x^i), \quad \forall \mathbf{x} \in (\mathbb{R}^d)^N, \tag{2.2}$$

for $E_i(x) = d(x, A_i)$, for d the euclidean distance and for given sets $A_i \subset \mathbb{R}$ (that might consist of a single element).

First of all, we recall that the functional J_T has at least a minimizer:

Proposition 2.2 (Existence of minimizers)**.** *The functional J_T given in* (2.1) *has at least one minimizer in $L^2(0,T;\mathcal{U})$.*

Proposition 2.2 is classical, and the proof can be found, for instance, in [Tré05, Proposition 6.2.3]. The main idea of the proof is that J_T is a sum of a positive weakly continuous functional and a positive continuous convex functional. For this result the continuity of $\mathcal{E}$ is essential.

For having J_T minimizers which take the error to 0 the first thing that we need, of course, is that the error can be taken to 0, a property known in Machine Learning as *interpolation* (see [EGPZ20]):

Hypothesis 1 (Interpolation). For the data set $\mathbf{x}$ there are controls:

$$(w_*, b_*) \in L^2(0,1;\mathcal{U}),$$

such that $\mathcal{E}(y(1;\mathbf{x},w_*,b_*)) = 0$.

Hypothesis 1 is used to show that the error of the final state of the optimal trajectories decays as $T \to \infty$ (see Lemma 3.3 below).

In addition, we assume that the error can be taken to 0 with a cost proportional to the error:

Hypothesis 2 (Local controllability of the system). Let the data set be $\mathbf{x} \in (\mathbb{R}^d)^N_*$. Then, there are $C, \tilde{\varepsilon} > 0$ both just depending on $\mathcal{E}$ and $\mathbf{x}$ such that for all $\overline{\mathbf{x}} = (\overline{x}_1, \dots, \overline{x}_N) \in (\mathbb{R}^d)^N_*$ satisfying $\mathcal{E}(\overline{\mathbf{x}}, \mathbf{x}) < \tilde{\varepsilon}$, there are some controls (w,b) satisfying:

$$\|(w,b)\|_{L^\infty(0,1;\mathcal{U})} < C\mathcal{E}(\overline{\mathbf{x}}),$$

such that:

$$\mathcal{E}(y(1;\overline{\mathbf{x}},w,b)) = 0.$$

As shown in Appendix A, an additional assumption to Hypothesis 1 is necessary to take the error to 0. Without Hypothesis 2 the cost to take the error to 0 may be considerably higher than obtaining some small error, so it might not compensate to take the error exactly to 0 (see Proposition A.1).

Now we have all the tools to state the first main result of this paper:

Theorem 2.3 (Annihilation of the error in a long time horizon)**.** *Let $\mathbf{x} \in (\mathbb{R}^d)^N_*$, σ be a Lipschitz activation function, $\mathcal{E}$ be an error function such that Hypotheses 1 and 2 are satisfied and J_T given in* (2.1)*. Then, for $T > 0$ large enough depending on σ, $\mathbf{x}$ and $\mathcal{E}$, and for all $\varepsilon > 0$ there is $\delta > 0$ such that $J_T(w,b) < \inf J_T + \delta$ implies:*

$$\mathcal{E}(y(T;\mathbf{x},w,b)) < \varepsilon. \tag{2.3}$$

Moreover, for $T > 0$ large enough the following equality holds for any optimal trajectory:

$$\mathcal{E}(y_T(T;\mathbf{x})) = 0. \tag{2.4}$$

Theorem 2.3 is proved by showing that if T is sufficiently large and if $\mathcal{E}(y(T;\mathbf{x},w,b))$ is small and strictly positive, we can construct with Hypothesis 2 a control $(\tilde{w},\tilde{b})$ such that:

$$J_T(\tilde{w},\tilde{b}) \leq J_T(w,b) - \frac{1}{2}\mathcal{E}(y(T;\mathbf{x},w,b)).$$

For that, we show in Lemma 3.1 that the trajectories may be preserved when we perform a diffeomorphism in the time variable, then in Lemma 3.4 given a control with a non-constant norm we construct a more efficient one

and in Proposition 3.6 we use this to construct a control for which the value of the empirical risk minimization functional is smaller for all controls with a non-constant norm.

The construction of such control is far from trivial and, as illustrated in Appendix A, the hypotheses are rather sharp. As explained in the first part of the introduction, Theorem 2.3 improves the results presented in [EGPZ20], where the authors prove that the error of the final state of the optimal trajectory is of size $\mathcal{O}(1/T)$.

Example 2.4 (Application of Theorem 2.3 to the classification problem). Let us consider:

$$\mathbf{x} = \left(x^1, \dots, x^M, x^{M+1}, \dots, x^N\right) \in (\mathbb{R}^d)^N_*,$$

the error function given by (2.2), for

$$E_i(x) = \begin{cases} (x_1+1)1_{x_1>-1}(x_1), & i = 1, \dots, M, \\ (x_1-1)1_{x_1>1}(x_1), & i = M+1, \dots, N, \end{cases}$$

and any neural function σ of the type $\sigma(x) = (\tilde{\sigma}(x_1), \dots, \tilde{\sigma}(x_d))$ such that there is $c > 0$ such that $cs \leq \tilde{\sigma}(s)$ for all $s \geq 0$ and $\tilde{\sigma}(s) \leq cs$ for all $s \leq 0$. Hypothesis 2 is clearly satisfied, as it suffices to consider $\tilde{\varepsilon} = 1/(2N)$, $b = 0$ and $w(t)x = (2Nc^{-1}\mathcal{E}(\overline{\mathbf{x}})x_1, 0, \dots, 0)$. Thus, Theorem 2.3 implies that if the data can be classified (i.e. if Hypothesis 1 is satisfied), then by computing the optimal control for a sufficiently large time, the data is sent to the sets $\{x_1 \leq -1\}$ and $\{x_1 \geq 1\}$.

2.2 Optimal trajectories for neural ODE with a homogeneous activation function

In this section we present the analogous results to those in Section 2.1 for the neural ODE (1.2) with activation functions which satisfy (1.4). Let us reformulate Hypotheses 1 and 2 in the context of (1.2):

Hypothesis 3 (Interpolation). For the data set $\mathbf{x}$ there are controls:

$$(r_*, w_*, b_*) \in L^2(0, 1; \tilde{\mathcal{U}}),$$

such that $\mathcal{E}(y(1; \mathbf{x}, r_*, w_*, b_*)) = 0$.

Hypothesis 4 (Local controllability of the system). Let the data set be $\mathbf{x} \in (\mathbb{R}^d)^N_*$. Then, there are $C > 0$ and $\tilde{\varepsilon} > 0$ both just depending on $\mathcal{E}$ and $\mathbf{x}$ such that for all $\overline{\mathbf{x}} = (\overline{x}_1, \dots, \overline{x}_N)$ satisfying $\mathcal{E}(\overline{\mathbf{x}}) < \tilde{\varepsilon}$, there are some controls (r, w, b) satisfying:

$$\|(w, b)\|_{L^\infty(0,1;\mathcal{U})} < C\mathcal{E}(\overline{\mathbf{x}}),$$

such that:

$$\mathcal{E}(y(1; \overline{\mathbf{x}}, r, w, b)) = 0.$$

Again, we seek to get sufficient conditions so that the optimal trajectories induced by:

$$\tilde{J}_T(r, w, b) := \mathcal{E}(y(T; \mathbf{x}, r, w, b)) + \int_0^T |(w(t), b(t))|^2 dt, \tag{2.5}$$

satisfy $\mathcal{E}(y_T(T; \mathbf{x})) = 0$. Since $|r|$ is constant (see (1.3)), it makes no sense to include it in the definition of $\tilde{J}_T$. For the functional $\tilde{J}_T$ the following result holds:

Theorem 2.5 (Annihilation of the error for a sufficiently large time)**.** *Let σ be a Lipschitz activation function satisfying* (1.4) *and E an error function satisfying Hypothesis 3 and 4. Then, for $T > 0$ large enough depending on σ, $\mathbf{x}$ and E, and all $\varepsilon > 0$ there is $\delta > 0$ such that if $J_T(r, w, b) < \inf J_T + \delta$:*

$$\mathcal{E}(y(T; \mathbf{x}, r, w, b)) < \varepsilon. \tag{2.6}$$

Moreover, if T is large enough and if $\tilde{J}_T$ has an optimal trajectory:

$$\mathcal{E}(y_T(T; \mathbf{x})) = 0. \tag{2.7}$$

The proof of Theorem 2.5 is analogous to that of Theorem 2.3, so we just give some brief explanations in the first comment of Section 4. As with Theorem 2.3, Theorem 2.5 improves the results presented in [EGPZ20], where the authors prove that the error of the optimal trajectory at a final time T is of magnitude $\mathcal{O}(1/T)$ also for the solutions of (1.2) with an activation functions satisfying (1.4).

Remark 2.6 (Existence of minimizers of $\tilde{J}_T$). We have stated "if $\tilde{J}_T$ has an optimal trajectory" in Theorem 2.5 because, as far as we know, it is an open question to see if $\tilde{J}_T$ admits a minimizer. The main obstacle to adapt the proof of Proposition 2.2 is that nonlinear functions and weak limits may not commute. However, as we see in the first comment of Section 4, we can improve Theorem 2.5 and obtain that for T large enough and all $\varepsilon > 0$ there are controls (r, w, b) such that $J_T(r, w, b) < \inf J_T + \varepsilon$ and $\mathcal{E}(y(T; \mathbf{x}, r, w, b)) = 0$.

Example 2.7 (Application of Theorem 2.5 to ensemble controllability). Theorem 2.5 can be applied to the ensemble controllability problem. Let $\mathbf{x} \in (\mathbb{R}^d)^N_*$ for $d \geq 2$, $\mathcal{M}$ given in (1.3):

$$\sigma(x) = (\max\{x_1, 0\}, \dots, \max\{x_d, 0\}), \tag{2.8}$$

the activation function, $\mathbf{z} = (z^1, \dots, z^N) \in (\mathbb{R}^d)^N_*$ the targets, and $\mathcal{E}$ given by (2.2) for $E_i(x) = |x - z^i|$ the error function. Then, it is proved in [RBZ21, Theorem 2] that Hypothesis 3 is satisfied. Moreover, as we prove in Appendix C, Hypothesis 4 also holds. We present the proof because the bounds for the cost of the control is not a straight consequence of the computations in [RBZ21]. Consequently, Theorem 2.5 (and all the auxiliary results and corollaries) can be applied to this neural problem.

Remark 2.8 (Functionals allowing expensive controls). As in [EGPZ20], we can consider the functional:

$$J_{T,\delta}(w, b) := \mathcal{E}(y(T; \mathbf{x}, w, b)) + \delta \int_0^T |(w(t), b(t))|^2 dt,$$

instead of J_T for (1.1), and:

$$J_{T,\delta}(r, w, b) := \mathcal{E}(y(T; \mathbf{x}, r, w, b)) + \delta \int_0^T |(w(t), b(t))|^2 dt,$$

instead of J_T for (1.2)-(1.4). By linearity (see Remark 3.2) it holds that:

$$J_{T,\delta}(w, b) = J_{T\delta^{-1},1}(\delta w(t\delta), \delta b(t\delta)),$$

and:

$$\tilde{J}_{T,\delta}(r, w, b) = \tilde{J}_{T\delta^{-1},1}(r(t\delta), \delta w(t\delta), \delta b(t\delta)),$$

respectively. A straight consequence is that (w, b) is a minimizer of $J_{T,\delta}$ if and only if $(\delta w(t\delta), \delta b(t\delta))$ is a minimizer of $J_{T\delta^{-1},1}$. Similarly, (r, w, b) is a minimizer of $\tilde{J}_{T,\delta}$ if and only if $(r(t\delta), \delta w(t\delta), \delta b(t\delta))$ is a minimizer of $J_{T\delta^{-1},1}$. Thus, analogous results to Theorems 2.3 and 2.5 and all the auxiliary results hold true for $J_{T,\delta}$ and $\tilde{J}_{T,\delta}$ when T is fixed and $\delta > 0$ is small enough depending on σ, $\mathcal{E}$, $\mathbf{x}$ and T.

3 Optimal control for control-affine neural ODE

In this section we work in the control problem described by (1.1) and the risk minimization functional J_T given by (2.1). In this section $C > 0$ denotes an arbitrary constant that may change from line to line and depends only on σ, $\mathcal{E}$ and $\mathbf{x}$. Similarly, when we assume that T is large enough we mean with respect to σ, $\mathcal{E}$ and $\mathbf{x}$. We first present some technical results in Section 3.1, then conclude the proof of Theorem 2.3 in Section 3.2 by a proof by contradiction, and finally provide additional properties of the optimal controls in Section 3.3.

3.1 Preliminaries

We first construct controls to have the same trajectory but at a different velocity:

Lemma 3.1 (A technical result regarding the time variable)**.** *Let* $\mathbf{x} \in (\mathbb{R}^d)^N$ *and*

$$\phi \in L^1_{loc}(0, \infty; \mathbb{R}^+).$$

Then:

$$y(\Gamma(t); \mathbf{x}, w, b) = y\Big(t; \mathbf{x}, \phi(\Gamma(s))w(\Gamma(s)), \phi(\Gamma(s))b(\Gamma(s))\Big), \quad \forall t \in [0, T^*], \tag{3.1}$$

for $T^* > 0$ *and* Γ *any solution of:*

$$\begin{cases} \dot{\Gamma}(s) = \phi(\Gamma(s)), \quad s \in [0, T^*), \\ \Gamma(0) = 0. \end{cases} \tag{3.2}$$

Remark 3.2 (Invariance of trajectories when ϕ is constant). An important application of Lemma 3.1 is the case $\phi(t) = \lambda \in \mathbb{R}^+$; that is, when ϕ is constant. Then, (3.1) becomes:

$$y(\lambda t; \mathbf{x}, w, b) = y(t; \mathbf{x}, \lambda w(\lambda s), \lambda b(\lambda s)). \tag{3.3}$$

Proof of Lemma 3.1. It suffices to see that for all i the function $t \mapsto y(\Gamma(t); x^i, w, b)$ is a solution of (1.1) with initial value x^i and controls $\phi(\Gamma(t))w(\Gamma(t))$ and $\phi(\Gamma(t))b(\Gamma(t))$, since (1.1) has a unique solution by the Cauchy-Lipschitz Theorem. From the initial condition on (3.2) we obtain that:

$$y(\Gamma(0); x^i, w, b) = y(0; x^i, w, b) = x^i.$$

Moreover, from the first equation of (3.2) and the chain rule:

$$\begin{aligned} \frac{d}{dt}\Big(y(\Gamma(t); x^i, w, b)\Big) &= \phi(\Gamma(t))\dot{y}(\Gamma(t); x^i, w, b) \\ &= \phi(\Gamma(t))\Big(w(\Gamma(t))\sigma(y(\Gamma(t); x^i, w, b)) + b(\Gamma(t))\Big) \\ &= \Big(\phi(\Gamma(t))w(\Gamma(t))\Big)\sigma\Big(y(\Gamma(t); x^i, w, b)\Big) + \Big(\phi(\Gamma(t))b(\Gamma(t))\Big). \end{aligned} \tag{3.4}$$

□

Next, we recall that Hypothesis 1 implies that the error is at most of size $\mathcal{O}(1/T)$:

Lemma 3.3 (Boundedness of the error with respect to T)**.** *Let* $T > 0$*,* σ *be an activation function and* $\mathcal{E}$ *an error function satisfying Hypothesis 1. Then:*

$$\mathcal{E}(y_T(T; \mathbf{x})) \leq \frac{C}{T}. \tag{3.5}$$

Lemma 3.3 is proven in [EGPZ20]. Briefly, it is a consequence of the definition of y_T as the optimal trajectory and that, by Remark 3.2, $\frac{1}{T}w_*(\frac{\cdot}{T})$ and $\frac{1}{T}b_*(\frac{\cdot}{T})$ are controls that take the error to 0 (see Hypothesis 1 for the definition of (w_*, b_*)). In fact:

$$\begin{aligned} \mathcal{E}(y_T(T; \mathbf{x})) \leq J_T(w_T, b_T) &\leq J_T\left(\frac{1}{T}w_*\left(\frac{\cdot}{T}\right), \frac{1}{T}b_*\left(\frac{\cdot}{T}\right)\right) \\ &= \frac{1}{T}\int_0^1 |(w_*(t), b_*(t))|^2 dt. \end{aligned} \tag{3.6}$$

Finally, we prove that we may assume that the norm of the optimal control is constant if it does not get null or explode:

Lemma 3.4 (Construction of more efficient controls)**.** *Let:*

$$(w,b) \in C^1([0,T];\mathcal{U} \setminus \{(0,0)\}),$$

be such that $t \mapsto |(w(t),b(t))|$ is not constant. Then, there is a control $(\tilde{w},\tilde{b})$ such that $t \mapsto |(\tilde{w}(t),\tilde{b}(t))|$ is constant, such that:

$$|(\tilde{w}(t),\tilde{b}(t))| \in \Big(\min_{[0,T]} |(w,b)|, \max_{[0,T]} |(w,b)|\Big) \quad \text{in } [0,T], \tag{3.7}$$

and such that:

$$y(T;\mathbf{x},\tilde{w},\tilde{b}) = y(T;\mathbf{x},w,b), \quad \int_0^T |(\tilde{w}(t),\tilde{b}(t))|^2 dt < \int_0^T |(w(t),b(t))|^2 dt. \tag{3.8}$$

Remark 3.5. Lemma 3.4 implies that $t \mapsto |(w(t),b(t))|$ is constant. However, it does not imply that $t \mapsto (w(t),b(t))$ is constant, just that its Frobenius norm is constant.

The proof consists on constructing new controls with a time-transformation that will allow us to arrive to the same target with a smaller cost:

Proof of Lemma 3.4. Let us consider the auxiliary function:

$$\phi_\gamma(t) = \frac{\gamma}{|(w(t),b(t))|} 1_{[0,T]}(t), \tag{3.9}$$

for $\gamma > 0$ to be fixed later, and Γ_γ given by:

$$\begin{cases} \dot{\Gamma}_\gamma(s) = \phi_\gamma(\Gamma_\gamma(s)), \quad s \in [0,T_\gamma], \\ \Gamma_\gamma(0) = 0, \end{cases} \tag{3.10}$$

for:

$$T_\gamma := \sup\{t : \Gamma_\gamma(t) < T\}.$$

Note that from the definition of T_γ it follows that:

$$\Gamma_\gamma(T_\gamma) = T. \tag{3.11}$$

Since ϕ_γ is C^1 (as $(w,b) \neq 0$, by compactness $\min_{[0,T]} |(w,b)| > 0$), (3.10) has a unique solution by the Cauchy-Lipschitz Theorem. Moreover, $\gamma \mapsto T_\gamma$ is continuous and decreasing, $\lim_{\gamma\to 0} T_\gamma = \infty$, and $\lim_{\gamma\to\infty} T_\gamma = 0$, so there is $\gamma^* > 0$ such that:

$$T_{\gamma^*} = T. \tag{3.12}$$

For that value γ^* it holds:

$$\gamma^* = \frac{1}{\frac{1}{T}\int_0^T \frac{dt}{|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t)))|}}, \tag{3.13}$$

because by (3.11) and (3.12):

$$\int_0^T \frac{\gamma^* dt}{|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t)))|} = \int_0^T \dot{\Gamma}_{\gamma^*}(t)dt = \Gamma_{\gamma^*}(T) = \Gamma_{\gamma^*}(T_{\gamma^*}) = T.$$

Considering (3.1) we obtain that:

$$\begin{aligned} y(T;\mathbf{x},\phi_{\gamma^*}(\Gamma_{\gamma^*}(s))w(\Gamma_{\gamma^*}(s)),\phi_{\gamma^*}(\Gamma_{\gamma^*}(s))b(\Gamma_{\gamma^*}(s))) &= y(\Gamma_{\gamma^*}(T);\mathbf{x},w,b) \\ &= y(T;\mathbf{x},w,b). \end{aligned} \tag{3.14}$$

Moreover, considering the strict inequality between the harmonic and arithmetic means (see, for instance, [Kom11]), (3.10)-(3.13), and the Change of Variables Theorem we get that:

$$\begin{aligned}\int_0^T \phi^2_{\gamma^*}(\Gamma_{\gamma^*}(t))|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t)))|^2dt &= \int_0^T (\gamma^*)^2dt = (\gamma^*)^2T \\ &= \frac{\gamma^*T}{\frac{1}{T}\int_0^T \frac{dt}{|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t)))|}} \\ &< \frac{1}{T}\int_0^T \gamma^*T|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t))|dt \\ &= \int_0^T \phi_{\gamma^*}(\Gamma_{\gamma^*}(t))|(w(\Gamma_{\gamma^*}(t)),b(\Gamma_{\gamma^*}(t)))|^2dt \\ &= \int_0^T |(w(t),b(t))|^2dt.\end{aligned} \tag{3.15}$$

Therefore, combining (3.14) and (3.15) we obtain (3.8) for:

$$\tilde{w}(t) = \phi_{\gamma^*}(\Gamma_{\gamma^*}(t))w(\Gamma_{\gamma^*}(t)), \quad \tilde{b}(t) = \phi_{\gamma^*}(\Gamma_{\gamma^*}(t))b(\Gamma_{\gamma^*}(t)).$$

Finally, since γ^* is the harmonic mean of values in:

$$\Big[\min_{[0,T]}|(w,b)|, \max_{[0,T]}|(w,b)|\Big],$$

we obtain (3.7). □

3.2 Construction of controls which take the error to zero

Let us state the properties of the controls that we construct in this section:

Proposition 3.6. *Let σ be an activation function and E an error function that satisfy Hypotheses 1 and 2. Let $T > 0$ be large enough and (w, b) be such that:*

$$J_T(w,b) \leq 2\inf J_T. \tag{3.16}$$

Then, there is a control $(\hat{w}, \hat{b})$ such that:

$$\mathcal{E}(y(T;\mathbf{x},\hat{w},\hat{b})) = 0, \tag{3.17}$$

and:

$$J_T(\hat{w},\hat{b}) \leq J_T(w,b) - \frac{1}{2}\mathcal{E}(y(T;\mathbf{x},w,b)). \tag{3.18}$$

The first step is to remark that $\mathcal{E}(y(T;\mathbf{x},w,b))$ is small for large T. The second step is to approximate (w, b) by some control $(\tilde{w}, \tilde{b})$ satisfying the hypothesis of Lemma 3.4. The third step is to show that if (2.4) is false, we may prolong for some $\tau > 0$ the controls $\tilde{w}$ and $\tilde{b}$ in $[T, T + \tau]$ so that:

$$\tilde{y}(T+\tau;\mathbf{x},\tilde{w},\tilde{b}) = 0. \tag{3.19}$$

The fourth step is to take those trajectories to $[0, T]$ with (3.3). The fifth and last step is to check that the new control satisfies (3.18).

Proof of Proposition 3.6. **Step 1: estimate of $\mathcal{E}(y(T;\mathbf{x},w,b))$.** If $\mathcal{E}(y(T;\mathbf{x},w,b)) = 0$, then it suffices to consider $(\hat{w},\hat{b}) = 0$, so we suppose from now on that $\mathcal{E}(y(T;\mathbf{x},w,b)) > 0$. Moreover, from (3.6) and (3.16) we obtain for T large enough that:

$$\mathcal{E}(y(T;\mathbf{x},w,b)) \in (0,\tilde{\varepsilon}/2), \tag{3.20}$$

for $\tilde{\varepsilon}$ the value in Hypothesis 2.

Step 2: approximating the control. Clearly, $C^1([0,T];\mathcal{U}\setminus\{0\})$ is dense is $L^2(0,T;\mathcal{U})$. Moreover,

$$(w,b)\mapsto \mathcal{E}(y(T;\mathbf{x},w,b)),$$

is continuous from $L^2(0,T;\mathcal{U})$ to $\mathbb{R}$. Thus, there is $(\tilde{w},\tilde{b})\in C^1([0,T];\mathcal{U}\setminus\{0\})$ such that:

$$\|(\tilde{w},\tilde{b})\|_{L^2(0,T;\mathcal{U})}\leq \|(w,b)\|_{L^2(0,T;\mathcal{U})}, \tag{3.21}$$

and:

$$\mathcal{E}(y(T;\mathbf{x},\tilde{w},\tilde{b}))\leq 2\mathcal{E}(y(T;\mathbf{x},w,b)). \tag{3.22}$$

Moreover, by Lemma 3.4 we can suppose that $t\mapsto|(\tilde{w}(t),\tilde{b}(t))|$ is constant. In addition, from (3.6), (3.16) and (3.21) it follows that:

$$\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})}\leq\frac{C}{\sqrt{T}}. \tag{3.23}$$

Step 3: taking the error to 0**.** From Hypothesis 2, (3.20) and (3.22) we obtain a control $(\overline{w},\overline{b})\in L^\infty(0,T;\mathcal{U})$ that takes the solution from $y(T;\mathbf{x},\tilde{w},\tilde{b})$ to a state $\tilde{\mathbf{x}}$ such that $\mathcal{E}(\tilde{\mathbf{x}})=0$. Moreover,

$$\|(\overline{w},\overline{b})\|_{L^\infty(0,1;\mathcal{U})}\leq C\mathcal{E}(y(T;\mathbf{x},\tilde{w},\tilde{b}))\leq C\mathcal{E}(y(T;\mathbf{x},w,b)).$$

Consequently, by Remark 3.2, for some:

$$\tau\leq C\frac{\mathcal{E}(y(T;\mathbf{x},\tilde{w},\tilde{b}))}{\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})}}, \tag{3.24}$$

the control $(\tilde{w},\tilde{b})$ can be prolonged to $[0,T+\tau]$ so that both:

$$\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T+\tau;\mathcal{U})}=\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})}, \tag{3.25}$$

and (3.19) are satisfied.

Step 4: taking the trajectory to $[0,T]$**.** We consider:

$$\begin{aligned}\hat{w}(t)&:=\frac{T+\tau}{T}\tilde{w}\left(\frac{T+\tau}{T}t\right),\\ \hat{b}(t)&:=\frac{T+\tau}{T}\tilde{b}\left(\frac{T+\tau}{T}t\right).\end{aligned} \tag{3.26}$$

Then, (3.17) is true. In fact, the equation (3.3) with $\lambda=\frac{T+\tau}{T}$ implies:

$$y(T;\mathbf{x},\hat{w},\hat{b})=y(T+\tau;\mathbf{x},\tilde{w},\tilde{b}).$$

Step 5: efficiency of the new control. First, we realize that:

$$\begin{aligned}J_T(w,b)-J_T(\hat{w},\hat{b})=\mathcal{E}(y(T;\mathbf{x},w,b))+\int_0^T|(w(t),b(t))|^2dt\\ -\left(\frac{T+\tau}{T}\right)^2\int_0^T\left|\left(\tilde{w}\left(\frac{T+\tau}{T}t\right),\tilde{b}\left(\frac{T+\tau}{T}t\right)\right)\right|^2dt.\end{aligned} \tag{3.27}$$

Considering that $t \mapsto (\tilde{w}(t), \tilde{b}(t))$ is constant in $[0,T]$, and that (3.21) and (3.25) are satisfied we deduce that:

$$\int_0^T |(w(t), b(t))|^2 dt - \int_0^T \left| \left(\tilde{w}\left(\frac{T+\tau}{T}t\right), \tilde{b}\left(\frac{T+\tau}{T}t\right)\right)\right|^2 dt \geq 0. \tag{3.28}$$

Consequently, we obtain from (3.20), (3.23)-(3.25) and (3.27)-(3.28) that:

$$\begin{aligned} J_T(w,b) - J_T(\hat{w},\hat{b}) &\geq \mathcal{E}(y_T(T;\mathbf{x},w,b),\mathbf{x}) - \left(\frac{2\tau}{T} + \frac{\tau^2}{T^2}\right) \int_0^T \left| \left(\tilde{w}\left(\frac{T+\tau}{T}t\right), \tilde{b}\left(\frac{T+\tau}{T}t\right)\right)\right|^2 dt \\ &\geq \left(1 - C\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})} - C\tilde{\varepsilon}T^{-1}\right) \mathcal{E}(y(T;\mathbf{x},w,b)) \\ &\geq \left(1 - CT^{-1/2}\right) \mathcal{E}(y(T;\mathbf{x},w,b)), \end{aligned}$$

which implies (3.18) for T large enough. □

Now we may conclude the proof of Theorem 2.3 by a proof by contradiction:

Conclusion of the proof of Theorem 2.3. Let $\varepsilon > 0$. It suffices to consider $\delta = \varepsilon/3$. If (w,b) are such that $J_T(w,b) \leq \inf J_T + \varepsilon/3$, then $\mathcal{E}(y(T;\mathbf{x},w,b)) < \varepsilon$. Otherwise, by Proposition 3.6 there are $(\hat{w},\hat{b})$ such that:

$$J_T(\hat{w},\hat{b}) \leq J_T(w,b) - \frac{\varepsilon}{2} \leq \inf J_T - \frac{\varepsilon}{6},$$

which is absurd. Similarly, if (w_T, b_T) is a minimizer of J_T and (2.4) is not satisfied, then the control $(\hat{w},\hat{b})$ of Proposition 3.6 satisfies $J_T(\hat{w},\hat{b}) < J_T(w_T,b_T)$, contradicting the definition of minimizer. □

3.3 Additional properties of the optimal control

As a consequence of Remark 3.2, we can easily prove that, assuming Hypotheses 1 and 2, for a sufficiently large time the optimal controls are of the form:

$$\left(\frac{1}{T} w_*\left(\frac{t}{T}\right), \frac{1}{T} b_*\left(\frac{t}{T}\right)\right),$$

for (w_*, b_*) the minimizers of the functional:

$$t \mapsto \int_0^T |(w(t), b(t))|^2 dt,$$

considered in the domain:

$$\{(w,b) : \mathcal{E}(y(1;\mathbf{x},w,b)) = 0\}.$$

In addition, we can prove that such minimizers belong to $L^\infty(0,T)$ and satisfy that $t \mapsto |(w(t), b(t))|$ is constant, which follows from:

Lemma 3.7 (A more efficient control)**.** *Let (w,b) a control in $L^2(0,1)$ such that $t \mapsto |(w(t),b(t))|$ is not constant. Then, there is a control $(\tilde{w},\tilde{b})$ such that:*

$$y(1;\mathbf{x},\tilde{w},\tilde{b}) = y(1;\mathbf{x},w,b),$$

$$\|(\tilde{w},\tilde{b})\|_{L^2(0,T;\mathcal{U})} < \|(w,b)\|_{L^2(0,T;\mathcal{U})},$$

and, if $(w,b) \in L^\infty(0,T;\mathcal{U})$,

$$\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})} \leq 2\|(w,b)\|_{L^\infty(0,T;\mathcal{U})}. \tag{3.29}$$

The proof of Lemma 3.7 is based on classical results from Measure Theory and is postponed to Appendix B. Lemma 3.7, compared to Lemma 3.4, has the advantage of having less restrictive hypothesis. However, it has the disadvantage that we do not obtain neither a contraction for the L^∞ norm (see Remark B.2) nor a control with constant norm, which is needed for proving Proposition 3.6.

4 Further comments and open problems

- **Analogous results for neural ODE whose dynamics are described by** (1.2)**.** Clearly Lemmas 3.1, 3.3, 3.4, and 3.7 and Proposition 3.6 can be proved for system (1.2) with σ satisfying (1.4) as in Section 3. The key lemma is Lemma 3.1, since the other results use the homogeneity of the system via Lemma 3.1. The analogous of Lemma 3.1 can be proved by replacing (3.4) by:

$$\begin{aligned}\frac{d}{dt}(y(\Gamma(t);x^i,w,b,r)) &= \phi(\Gamma(t))r(\Gamma(t))\dot{y}(\Gamma(t);x^i,w,b)\\ &= \phi(\Gamma(t))r(\Gamma(t))\sigma\Big(w(\Gamma(t))y(\Gamma(t);x^i,w,b)+b(\Gamma(t))\Big)\\ &= r(\Gamma(t))\sigma\Big(\phi(\Gamma(t))w(\Gamma(t))y(\Gamma(t);x^i,w,b)+\phi(\Gamma(t))b(\Gamma(t))\Big).\end{aligned}$$

  The last equality follows from (1.4). Finally, Theorem 2.5 and the analogous of Proposition 3.6 imply that for all $\delta>0$ there is a control (r,w,b) such that $\tilde{J}_T(r,w,b)<\inf\tilde{J}_T-\delta$ and $\mathcal{E}(y(T;\mathbf{x},r,w,b))=0$.

- **Optimal control for non-homogenous activation functions.** It remains an open problem to determine if similar results to Theorem 2.5 hold for non-homogeneous activation functions satisfying $\sigma(0)=0$ such as the hyperbolic tangent,

$$\sigma(x)=(\tanh(x_1),\dots,\tanh(x_d)),$$

  see [FS18]. We may wonder whether similar results hold with more general activation functions if we replace $\mathcal{M}$ (see (1.3)) by the unitary matrices or by $\mathcal{M}_{d\times d}$ (of course, the cost of r must also be included in the risk minimization functional). This would include, for instance, sigmoid,

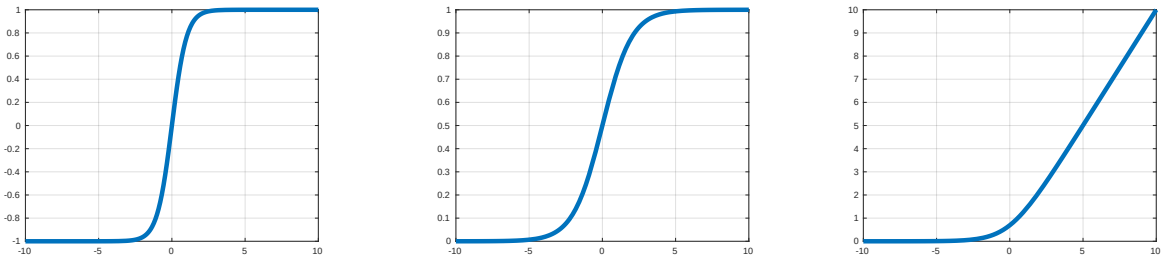
$$\sigma(x)=((1+e^{-x_1})^{-1},\dots,(1+e^{-x_d})^{-1}),$$

  see [MS95]; softplus,

$$\sigma(x)=(\log(1+e^{x_1}),\dots,\log(1+e^{x_d})),$$

  see [GBB11] (see Figure 2 for their graph in one dimension), and others like logistic and cross-entropy function. The main difficulty is that the analogue of Lemma 3.1 cease to be true, so another tool is needed to prove the main result, probably a local inverse theorem result.

- **Optimal control with other norms.** It is a relevant problem to determine if similar results to Theorems 2.3 and 2.5 hold for any other Lebesgue or Sobolev norms. In particular, the most interesting scenario is to replace both in J_T and $\tilde{J}_T$ the terms $\|(w,b)\|^2_{L^2(0,T)}$ by $\|(w,b)\|^2_{H^1(0,T)}$ and adding the restriction that the component of r can only change signs if $(w,b)=0$ or to measure the H^1 norm of r if the space $\mathcal{M}$ is connected. The interest of this is double: thinking in potential applications it makes sense to also try to bound the variations in the time variable, which can be obtained by minimizing the time derivative. Moreover, if we consider the H^1-norm we can prove as in Proposition 2.2 that $\tilde{J}_T$ admits a minimizer. The main difficulties when studying these norms are that Lemmas 3.4 and 3.7 may not be proved as easily (if they are true) because we need to keep track of the time derivative and because we cannot define the control on $[T,T+\tau]$ independently to the controls on $[0,T]$ due to the time derivative.



(a) Hyperbolic tangent (b) Sigmoid (c) Softplus

Figure 2: Some activation functions for $d = 1$

A A pathological case

In this section we prove that without Hypothesis 2 the error may not be taken exactly to 0 if the ratio between the cost of correcting the error and the error explodes as the error vanishes. We present an example for the sake of simplicity, though the proof can be replicated whenever the gradient of the error is null on all the points where the error is null, which is the key impediment for taking the error exactly to zero.

Proposition A.1 (Necessity of local controllability)**.** *Let us consider* $d = 1$, $\mathbf{x} = x_1 = 1$, $\mathcal{E}(x) = x^2$, $\sigma(s) = s$ *and* J_T *given by* (2.1)*. Then,* $y_T(T) > 0$ *for all* $T > 0$*.*

Proof of Proposition A.1. Let (w_T, b_T) be a minimizer of J_T. Clearly $w_T, b_T \leq 0$. Let us prove by contradiction that $y_T(T) > 0$. For that, we suppose that $y_T(T) = 0$. By Lemma 3.7, $t \mapsto |(w_T(t), b_T(t))|$ is a constant function equal to some constant $\mathfrak{c}$. In particular, for $\delta > 0$ small enough the following inequality is satisfied:

$$(y_T(T-\delta))^2 - \int_{T-\delta}^{T} |(w_T(t), b_T(t))|^2 = (y_T(T-\delta))^2 - \mathfrak{c}^2\delta \leq (C\delta)^2 - \mathfrak{c}^2\delta < 0. \tag{A.1}$$

The estimate $|y_T(T-\delta)| \leq C\delta$ follows from the formula:

$$y_T(T-\delta) = -\int_{T-\delta}^{T} b_T(s) \exp\left(-\int_{T-\delta}^{s} w_T(z)dz\right) ds,$$

which follows from $y_T(T) = 0$. Consequently, we obtain from (A.1) that:

$$J_T(w_T 1_{(0,T-\delta)}, b_T 1_{(0,T-\delta)}) - J_T(w_T, b_T) < 0, \tag{A.2}$$

which contradicts that (w_T, b_T) is a minimizer of J_T. □

Remark A.2 (On Hypothesis 1). It is trivial that Hypothesis 1 is satisfied by the activation and error function introduced in Propositions A.1.

B Proof of Lemma 3.7

In this section we prove Lemma 3.7. Here μ denotes the Lebesgue measure. In order to prove Lemma 3.7 we need the following classical result of measure theory, whose proof can be found in [Yeh06, Thm. 3.25]:

Lemma B.1 (Comparison between sets of positive measure and open sets)**.** *Let* $S \subset [0, T]$ *be a measurable set such that* $\mu(S) > 0$*. Then, for all* $\varepsilon > 0$ *there is an open set* $\mathcal{O}^\varepsilon = \bigcup_{i=1}^{n^\varepsilon} (a_i^\varepsilon, b_i^\varepsilon)$ *such that* $\mu(\mathcal{O}^\varepsilon \triangle S) < \varepsilon$*.*

Proof of Lemma 3.7. Since $|(w,b)|$ is not constant, there are some sets S_1 and S_2 and some constants $C_1, C_2 > 0$ such that $C_1 < C_2$, $|(w,b)| < C_1$ on S_1, $|(w,b)| > C_2$ on S_2 and:

$$\inf\{|x^2 - x^1| : x^1 \in S_1, x^2 \in S_2\} > 0. \tag{B.1}$$

From Lemma B.1 we get that for $\varepsilon > 0$ small enough there are two sets $\mathcal{O}_1^\varepsilon = \bigcup_{i=1}^{n_1^\varepsilon}(a_{1,i}^\varepsilon, b_{1,i}^\varepsilon)$ and $\mathcal{O}_2^\varepsilon = \bigcup_{i=1}^{n_2^\varepsilon}(a_{2,i}^\varepsilon, b_{2,i}^\varepsilon)$ satisfying:

$$\mu(\mathcal{O}_1^\varepsilon \setminus S_1) < \varepsilon, \quad \mu(\mathcal{O}_2^\varepsilon \setminus S_2) < \varepsilon, \tag{B.2}$$

and:

$$\mu(\mathcal{O}_1^\varepsilon) = \mu(\mathcal{O}_2^\varepsilon) = \frac{\min\{\mu(S_1), \mu(S_2)\}}{2}. \tag{B.3}$$

If ε is small enough, because of (B.1) we may also assume that:

$$\mathcal{O}_1^\varepsilon \cap \mathcal{O}_2^\varepsilon = \emptyset. \tag{B.4}$$

Let us consider the auxiliary function:

$$\phi_\gamma(t) = \begin{cases} 1 & t \in [0,T] \setminus (\mathcal{O}_1^\varepsilon \cap \mathcal{O}_2^\varepsilon), \\ 1+\gamma & t \in \mathcal{O}_1^\varepsilon, \\ \frac{1+\gamma}{1+2\gamma} & t \in \mathcal{O}_2^\varepsilon, \\ 0 & t \geq T, \end{cases} \tag{B.5}$$

for $\gamma > 0$ to be fixed later, and Γ_γ given by:

$$\begin{cases} \dot{\Gamma}_\gamma(s) = \phi_\gamma(\Gamma_\gamma(s)), \quad \forall s \geq 0, \\ \Gamma_\gamma(0) = 0. \end{cases} \tag{B.6}$$

We remark that:

$$\Gamma_\gamma(T) = T. \tag{B.7}$$

Indeed, it can be proved that if $\Gamma_\gamma(T_*) = a$ and $\phi_\gamma(t) = c$ on $[a,b]$, then $\Gamma_\gamma\left(T_* + \frac{b-a}{c}\right) = b$. Hence:

$$\Gamma_\gamma\left(\mu\big([0,T] \setminus (\mathcal{O}_1^\varepsilon \cap \mathcal{O}_2^\varepsilon)\big) + \frac{1}{1+\gamma}\mu(\mathcal{O}_1^\varepsilon) + \frac{1+2\gamma}{1+\gamma}\mu(\mathcal{O}_2^\varepsilon)\right) = T,$$

which considering (B.3), (B.4) and (B.5) implies (B.7).

Consequently, the following controls satisfy the conclusions of Lemma 3.7:

$$(\tilde{w}, \tilde{b}) = \phi_\gamma(\Gamma_\gamma(t))(w(\Gamma_\gamma(t)), b(\Gamma_\gamma(t))). \tag{B.8}$$

Indeed, from (3.1) and (B.1) it holds that:

$$\begin{aligned} y(T; \mathbf{x}, \phi_\gamma(s)w(\Gamma_\gamma(s)), \phi_\gamma(s)b(\Gamma_\gamma(s))) &= y(\Gamma_\gamma(T); \mathbf{x}, w, b) \\ &= y(T; \mathbf{x}, w, b). \end{aligned}$$

In addition, if γ and ε are small enough:

$$\begin{aligned}\int_0^T |(w(t),b(t))|^2dt - \int_0^T |\phi_\gamma(\Gamma_\gamma(t))(w(\Gamma_\gamma(t)),b(\Gamma_\gamma(t)))|^2dt \\ &= \int_{\mathcal{O}_1^\varepsilon\cup\mathcal{O}_2^\varepsilon} |(w(t),b(t))|^2dt \\ &- \int_{\Gamma_\gamma^{-1}(\mathcal{O}_1^\varepsilon)\cup\Gamma_\gamma^{-1}(\mathcal{O}_2^\varepsilon)} \phi_\gamma^2(\Gamma_\gamma(t))|(w(\Gamma_\gamma(t)),b(\Gamma_\gamma(t)))|^2dt \\ &= -\gamma\int_{\mathcal{O}_1^\varepsilon} |(w(t),b(t))|^2dt + \frac{\gamma}{1+2\gamma}\int_{\mathcal{O}_2^\varepsilon} |(w(t),b(t))|^2dt \\ &\geq \frac{\gamma}{1+2\gamma}C_2\left(\frac{\min\{\mu(S_1),\mu(S_2)\}}{2}-\varepsilon\right) \\ &- \gamma C_1\frac{\min\{\mu(S_1),\mu(S_2)\}}{2} - \|(w,b)\|^2_{L^2(\mathcal{O}_1^\varepsilon\setminus S_1)} > 0.\end{aligned}$$

The second equality follows from the change of variable $s=\Gamma_\gamma(t)$, the first inequality from the definitions of S_1, S_2, $\mathcal{O}_1^\varepsilon$, $\mathcal{O}_2^\varepsilon$ and (B.3), and the last inequality from $C_2 > C_1$, (B.2), being γ and ε small enough, and the well known identity:

$$\lim_{c\to 0}\sup_{\mu(A)=c}\|g\|_{L^2(A,dx)} = 0, \quad \forall g\in L^2(0,T).$$

Finally, if $(w,b)\in L^\infty(0,T;\mathcal{U})$ the estimate (3.29) follows from (B.5) and (B.8) by taking $\gamma\leq 1$. □

Remark B.2 (Sharpness of the estimate (3.29)). The construction provided in the previous proof may not ensure us that:

$$\|(\tilde{w},\tilde{b})\|_{L^\infty(0,T;\mathcal{U})} \leq \|(w,b)\|_{L^\infty(0,T;\mathcal{U})};$$

for instance if $|(w,b)| = 1_\Omega$, for $\Omega\subset[0,T]$ a set such that $\mu(\Omega)\in(0,T)$ and which contains an open neighbourhood of every rational number in $[0,T]$. However, we can replace in the estimate (3.29) the constant 2 by any constant strictly greater than 1.

C Local ensemble controllability

In this section we prove the following result:

Lemma C.1 (Local ensemble controllability result)**.** *Let σ be the activation function defined by* (2.8) *and $\mathcal{E}$ defined in Example 2.7. Then σ and $\mathcal{E}$ satisfy Hypothesis 4.*

The main contribution with respect to the result on [RBZ21] is that we keep track of the cost and continuity of the control. The controls that we construct are different to those in [RBZ21], in which w and b have a single non-zero component at any time, since we do not search for a sparse property, but to obtain the continuity of the controls with respect to the initial data. We recall that C is a positive constant sufficiently large changing from line to line which depends on the target set $\mathbf{z}$. Moreover, for $d=2$ we denote:

$$r=\begin{bmatrix} r_1 & 0\\ 0 & r_2\end{bmatrix}, \quad w=\begin{bmatrix} w_1\\ w_2\end{bmatrix}=\begin{bmatrix} w_{11} & w_{12}\\ w_{21} & w_{22}\end{bmatrix}, \quad b=\begin{bmatrix} b_1\\ b_2\end{bmatrix}.$$

Proof of Lemma C.1. In order to simplify the notation we prove Lemma C.1 for the case $d=2$, though the proof is analogous for any $d\geq 2$. We prove Lemma C.1 by induction on N.

Step 1: the base case. Let us begin with the case $N = 1$. We may take $x = (x_1, x_2)$ to $z = (z_1, z_2)$ with a force proportional to $|z - x|$ by applying the controls $w = 0$, $b_1 = |z_1 - x_1|$, $b_2 = |z_2 - x_2|$, $r_1 = \text{sign}(z_1 - x_1)$ and $r_2 = \text{sign}(z_2 - x_2)$.

Step 2: the inductive case. Step 2.1: rearranging the points. We may suppose by rearranging the indexes that $|z^N| = \max_{i=1,\dots,N} |z^i|$. For the rest of the proof we define $e := \frac{z^N}{|z^N|}$ and:

$$\delta := \min\{|z^N| - \max_{i=1,\dots,N-1} z^i \cdot e, 1\}. \tag{C.1}$$

Then, $\delta > 0$ since, for $i = 1, \dots, N-1$, either $|z^i| < |z^N|$ or $|z^i| = |z^N|$ but $z^i \neq z^N$, so $\cos(z^i, e) < 1$.

Step 2.2: controlling $(x^1, \dots, x^{N-1})$ in $[0, 1/2]$. By the induction hypothesis and linearity we know that for $\hat{\varepsilon}$ small enough, if $\sum_{i=1}^{N-1} |z^i - x^i| < \hat{\varepsilon}$ there are some controls (r, w, b) satisfying:

$$\|(w, b)\|_{L^\infty(0,1/2)} < C \sum_{i=1}^{N-1} |z^i - x^i|, \tag{C.2}$$

and such that:

$$y(1/2; (x^1, \dots, x^{N-1}), r, w, b) = (z^1, \dots, z^{N-1}). \tag{C.3}$$

We fix:

$$\tilde{\varepsilon} = \min\left\{\hat{\varepsilon}, \frac{\delta}{C(|z^N| + 1)}\right\}.$$

If:

$$\sum_{i=1}^{N} |z^i - x^i| < \tilde{\varepsilon}, \tag{C.4}$$

then:

$$|y(t; x^N, r, w, b) - z^N| < \frac{\delta}{2} \quad \forall t \in \left[0, \frac{1}{2}\right]. \tag{C.5}$$

Indeed, $|z^N - x^N| < \frac{\delta}{4}$ by (C.4) and, if $|y(\cdot; x^N, w, b)| < |z^N| + \frac{\delta}{2}$ on $[0, t)$, for $t \leq 1/2$, then:

$$\int_0^t \sigma(w(s)y(s; x^N, r, w, b) + b(s))ds \leq \frac{\|w\|_{L^\infty(0,t)} \left(|z^N| + \frac{\delta}{2}\right) + \|b\|_{L^\infty(0,t)}}{2} < \frac{\delta}{4},$$

considering (C.2) and that $\tilde{\varepsilon} < \frac{\delta}{C(|z^N|+1)}$. In a similar way, we can prove that:

$$|y(1/2; x^N, r, w, b) - z^N| \leq C \sum_{i=1}^{N} |z^i - x^i|. \tag{C.6}$$

Indeed, for $t \in [0, 1/2]$ by (C.2):

$$|\sigma(wy(1/2; x^N, r, w, b) + b)| \leq \|w\|_{L^\infty(0,1/2)}(|z^N| + \delta) + \|b\|_{L^\infty(0,1/2)} \leq C \sum_{i=1}^{N-1} |z^i - x^i|.$$

Step 2.3: controlling $y(1/2; x^N, r, w, b)_1$ in $[1/2, 3/4]$. We seek to obtain that:

$$y(3/4; x^N, r, w, b)_1 = z_1^N. \tag{C.7}$$

If $y(1/2; x^N, w, b)_1 = z_1^N$, it suffices to consider $r_1 = 1$, $r_2 = 1$, $w = 0$ and $b = 0$, so we may restrict to the case $y(1/2; x^N, w, b) \neq z_1^N$. To obtain (C.7) we consider the controls $r_1 = \text{sign}(z_1 - x_1)$, $r_2 = 1$, $w_1 = \mathfrak{c} \sum_{i=1}^N |z^i - x^i| e$, $b_1 = \mathfrak{c} \sum_{i=1}^N |z^i - x^i|(-|z^N| + \delta)$, $w_2 = 0$, $b_2 = 0$ in $[1/2, 3/4]$, for $\mathfrak{c}$ to be fixed later. These controls are constant in $[1/2, 3/4]$. First, we remark that $\sigma(w_1 \cdot x + b) = 0$ for all x such that $x \cdot e \leq |z^N| - \delta$. In particular, from (C.1) and (C.3) we derive:

$$y(3/4; (x^1, \dots, x^{N-1}), r, w, b) = y(1/2; (x^1, \dots, x^{N-1}), r, w, b) = (z^1, \dots, z^{N-1}). \tag{C.8}$$

Moreover, $|y(t; x^N, r, w, b)_1 - z_1^N|$ is decreasing on $[1/2, T_*]$, for:

$$T_* := \inf\{T_* \geq 1/2 : y(T_*; x^N, r, w, b)_1 = z_1^N\}.$$

In addition, thanks to (C.5) in $[1/2, T_*]$ the following inequality is satisfied:

$$\left|\dot{y}(t; x^N, r, w, b)_1\right| = |w_1 \cdot y(t; x^N, r, w, b) - b_1| \geq \mathfrak{c} \sum_{i=1}^N |z^i - x^i| \frac{\delta}{2}.$$

Combining this with (C.6) we obtain that $T_* < 3/4$ if $\mathfrak{c} \geq C$; i.e. if $\mathfrak{c}$ is sufficiently large just with respect to $\mathbf{z}$ (recall that δ is a fixed parameter depending only on $\mathbf{z}$). Consequently, since T_* is continuous with $\mathfrak{c}$ and $\lim_{\mathfrak{c} \to 0} T_* = \infty$ there is some $\mathfrak{c} \in (0, C]$ such that $T_* = 3/4$. In particular, there are some controls (r, w, b) such that (C.7), (C.8) hold, and such that:

$$\|(w, b)\|_{L^\infty(0,3/4)} < C \sum_{i=1}^N |z^i - x^i|.$$

Finally, arguing as in the final part of Step 2.3 we obtain for C large enough that:

$$|y(3/4; x^N, r, w, b)_2 - z_2^N| \leq \min\left\{ C \sum_{i=1}^N |z^i - x^i|, \frac{3\delta}{4} \right\}.$$

Step 2.4: controlling $y(\cdot; x^N, r, w, b)$ in $[3/4, 1]$. In a similar way, we can prolong the controls in $[3/4, 1]$ so that $y(1; \mathbf{x}, r, w, b) = \mathbf{z}$ and:

$$\|(w, b)\|_{L^\infty(0,1)} < C \sum_{i=1}^N |z^i - x^i|,$$

are satisfied. This can be done as in Step 2.3 by fixing $r_1 = 1$, $r_2 = \text{sign}(z_2 - x_2)$, $w_1 = 0$, $b_1 = 0$, $w_2 = \mathfrak{c} \sum_{i=1}^N |z^i - x^i| e$, $b_2 = \mathfrak{c} \sum_{i=1}^N |z^i - x^i|(-|z^N| + \delta)$ for some $\mathfrak{c} > 0$. Indeed, since $w_1 = 0$ and $b_1 = 0$, the function $t \mapsto y(t; x^N, r, w, b)_1$ is constant in $[3/4, 1]$. □

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement NO: 694126-DyCon). I would like to thank Carlos Esteve, Borjan Geshkovski and Domènec Ruiz i Balet for fruitful discussion. In particular, I would like to thank Carlos Esteve for proposing Proposition A.1.

References

[AB19] J. B. Amara and E. Beldi. Simultaneous controllability of two vibrating strings with variable coefficients. *Evol. Equ. Control The.*, 8(4):687–694, 2019.

[BP20] T. Breiten and L. Pfeiffer. On the turnpike property and the receding-horizon method for linear-quadratic optimal control problems. *SIAM J. Control Optim.*, 58(2):1077–1102, 2020.

[Ces12] L. Cesari. *Optimization—theory and applications: problems with ordinary differential equations*, volume 17. Springer Science & Business Media, 2012.

[DGSW14] T. Damm, L. Grüne, M. Stieler, and K. Worthmann. An exponential turnpike theorem for dissipative discrete time optimal control problems. *SIAM J. Control Optim.*, 52(3):1935–1957, 2014.

[DSS87] Robert Dorfman, Paul Anthony Samuelson, and Robert M Solow. *Linear programming and economic analysis*. Courier Corporation, 1987.

[EGPZ20] C. Esteve, B. Geshkovski, D. Pighin, and E. Zuazua. Large-time asymptotics in deep learning. *arXiv preprint arXiv:2008.02491*, 2020.

[EGPZ22] C. Esteve, B. Geshkovski, D. Pighin, and E. Zuazua. Turnpike in lipschitz-nonlinear optimal control. *Nonlinearity*, 35(4):1652, 2022.

[EKPZ20] C. Esteve, H. Kouhkouh, D. Pighin, and E. Zuazua. The turnpike property and the long-time behavior of the Hamilton-Jacobi equation. *arXiv preprint arXiv:2006.10430*, 2020.

[EYG21] C. Esteve-Yagüe and B. Geshkovski. Sparse approximation in learning via neural odes. *arXiv preprint arXiv:2102.13566*, 2021.

[FHS21] T. Faulwasser, A.-J. Hempel, and S. Streif. On the turnpike to design of deep neural nets: Explicit depth bounds. *arXiv preprint arXiv:2101.03000*, 2021.

[FS18] E. Fathi and B. M. Shoja. Deep neural networks for natural language processing. In *Handbook of statistics*, volume 38, pages 229–316. Elsevier, 2018.

[GBB11] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[GG18] L. Gruüne and R. Guglielmi. Turnpike properties and strict dissipativity for discrete time linear quadratic optimal control problems. *SIAM J. Control Optim.*, 56(2):1282–1302, 2018.

[GH19] M. Gugat and F. M. Hante. On the turnpike phenomenon for optimal boundary control problems with hyperbolic systems. *SIAM J. Control Optim.*, 57(1):264–289, 2019.

[GSZ21] M. Gugat, M. Schuster, and E. Zuazua. The Finite-Time Turnpike Phenomenon for Optimal Control Problems: Stabilization by Non-smooth Tracking Terms. In G. Sklyar and A. Zuyev, editors, *Stabilization of Distributed Parameter Systems: Design Methods and Applications*, pages 17–41, Cham, 2021. Springer International Publishing.

[GTZ16] M. Gugat, R. Trélat, and E. Zuazua. Optimal Neumann control for the 1D wave equation: Finite horizon, infinite horizon, boundary tracking terms and the turnpike property. *Syst. Control Lett.*, 90:61–70, 2016.

[HR17] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Probl.*, 34(1):014004, 2017.

[HSV95] R. F Hartl, S. P. Sethi, and R. G. Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Rev.*, 37(2):181–218, 1995.

[HZRS15] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[Kom11] J. Komić. *International Encyclopedia of Statistical Science. Harmonic Mean.*, pages 622–624. Springer, Heidelberg, 2011.

[Lio88] J-L Lions. Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome 1. *RMA*, 8, 1988.

[LZ16] 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.

[McK76] Lionel W McKenzie. Turnpike theory. *Econometrica: Journal of the Econometric Society*, pages 841–865, 1976.

[MS95] J. Mira and F. Sandoval. *From Natural to Artificial Neural Computation: International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, June 7-9, 1995: Proceedings*, volume 930. Springer Science & Business Media, 1995.

[NH10] V. Nair and G. E Hinton. Rectified linear units improve restricted boltzmann machines. In *27th International Conference on International Conference on Machine Learning, ICML 10*, pages 807–814, 2010.

[PZ13] A. Porretta and E. Zuazua. Long time versus steady state optimal control. *SIAM J. Control Optim.*, 51(6):4242–4273, 2013.

[RBAZ22] D. Ruiz-Balet, E. Affili, and E. Zuazua. Interpolation and approximation via momentum resnets and neural odes. *Systems & Control Letters*, 162:105182, 2022.

[RBZ21] D. Ruiz-Balet and E. Zuazua. Neural ODE control for classification, approximation and transport. *arXiv preprint arXiv:2104.05278*, 2021.

[Rus86] D. L Russell. The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in a cylindrical region. *SIAM J. Control Optim.*, 24(2):199–229, 1986.

[SN20] N. Sakamoto and M. Nagahara. The turnpike property in the maximum hands-off control. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2350–2355. IEEE, 2020.

[Son92] E. D. Sontag. Neural nets as systems models and controllers. In *Proc. Seventh Yale Workshop on Adaptive and Learning Systems*, pages 73–79, 1992.

[SPZ19] N. Sakamoto, D. Pighin, and E. Zuazua. The turnpike property in nonlinear optimal control—a geometric approach. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2422–2427. IEEE, 2019.

[SS97] E. Sontag and H. Sussmann. Complete controllability of continuous-time recurrent neural networks. *Syst. Control Lett.*, 30(4):177–183, 1997.

[Tré05] E. Trélat. *Contrôle optimal: théorie & applications*. Vuibert Paris, 2005.

[Tré20] E. Trélat. Linear turnpike theorem. *arXiv preprint arXiv:2010.13605*, 2020.

[TW00] M. Tucsnak and G. Weiss. Simultaneous exact controllability and some applications. *SIAM J. Control Optim.*, 38(5):1408–1427, 2000.

[TZ15] E. Trélat and E. Zuazua. The turnpike property in finite-dimensional nonlinear optimal control. *J. Differ. Equations*, 258(1):81–114, 2015.

[TZZ18] E. Trélat, C. Zhang, and E. Zuazua. Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces. *SIAM J. Control Optim.*, 56(2):1222–1252, 2018.

[Wei17] E Weinan. A proposal on machine learning via dynamical systems. *Commun. Math. Stat.*, 5(1):1–11, 2017.

[WZ21] M. Warma and S. Zamorano. Exponential turnpike property for fractional parabolic equations with non-zero exterior data. *ESAIM:COCV*, 27(1):1–35, 2021.

[WZL17] J. Wu, X. Zhu, and S. Li. Simultaneous controllability of damped wave equations. *Math. Method Appl. Sci.*, 40(1):319–324, 2017.

[Yeh06] J. Yeh. *Real analysis: theory of measure and integration second edition*. World Scientific Publishing Company, 2006.

[Zam18] S. Zamorano. Turnpike property for two-dimensional Navier–Stokes equations. *J. Math. Fluid Mech.*, 20(3):869–888, 2018.

[Zbi93] R. Zbikowski. Lie algebra of recurrent neural networks and identifiability. In *1993 American Control Conference*, pages 2900–2901. IEEE, 1993.