SPARSE APPROXIMATION IN LEARNING VIA NEURAL ODES

CARLOS ESTEVE-YAGÜE AND BORJAN GESHKOVSKI

Abstract. We consider the neural ODE and optimal control perspective of supervised learning with $L^1(0,T;\mathbb{R}^{d_u})$ control penalties, where rather than only minimizing a final cost for the state, we integrate this cost over the entire time horizon. Under natural homogeneity assumptions on the nonlinear dynamics, we prove that any optimal control (for this cost) is sparse, in the sense that it vanishes beyond some positive stopping time. We also provide a polynomial stability estimate for the running cost of the state with respect to the time horizon. This can be seen as a *turnpike property* result, for nonsmooth functionals and dynamics, and without any smallness assumptions on the data, both of which are new in the literature. In practical terms, the temporal sparsity and stability results could then be used to discard unnecessary layers in the corresponding residual neural network (ResNet), without removing relevant information.

Contents

Keywords. Deep Learning; Neural ODEs; Supervised Learning; Sparsity; Optimal control; Turnpike property, Stabilization.
AMS Subject Classification. 49J15; 49M15; 49J20; 49K20; 93C20; 49N05.

1. Introduction

1.1. **Motivation.** *Sparsity* is a highly desirable property in many machine learning and optimization tasks due to the inherent reduction of computational complexity. Typically induced by ℓ^1 penalties/regularizations, it has been used extensively for simplifying machine learning tasks by selecting, in an automatized manner, a strict subset of the available features to be used. This is exemplified by the well-known Lasso (least absolute shrinkage and selection operator, [Santosa and Symes, 1986; Tibshirani, 1996]), which consists in minimizing a least squares cost function and an ℓ^1 parameter penalty for an affine parametric model $y = wx + b$. As the ℓ^1 penalty enforces a subset of the optimizable parameters (w, b) to become zero, the associated features may be discarded safely.

Date: October 20, 2021.

With such insights in mind, in this work we analyze supervised learning problems viewed from the lens of optimal control and neural ODEs, and demonstrate the appearance of sparsity patterns for global minimizers in the context of $L^1(0,T;\mathbb{R}^{d_u})$ control penalties. Rather than typical sparsity in which, at a given time t, all but few of the components of a control $u(t) \in \mathbb{R}^{d_u}$ are zero, we shall demonstrate a *temporal sparsity property*, namely that an optimal control $u(t)$ concentrates all its value within an interval $[0, T^*]$, and vanishes beyond time $t \geqslant T^*$. We motivate our setting and main result in what follows, and refer the reader to Section 1.6 for a roadmap of the paper.

1.2. **Supervised learning.** To put the above discussion into context, we recall that *supervised learning* addresses the problem of predicting from labeled data, which consists in approximating an unknown function $f : \mathcal{X} \to \mathcal{Y}$ from known samples

$$\left\{x^{(i)}, y^{(i)}\right\}_{i\in[n]} \subset \mathcal{X} \times \mathcal{Y}.$$

Here and henceforth, $[n] := \{1, \dots, n\}$ and $\mathcal{X} \subset \mathbb{R}^d$. Depending on the nature of the label space $\mathcal{Y}$, one distinguishes two types of supervised learning tasks: *classification*, when labels take values in a finite set of $m \geqslant 2$ classes, e.g. $\mathcal{Y} = [m]$, and *regression*, when the labels take continuous values in $\mathcal{Y} \subset \mathbb{R}^m$ with $m \geqslant 1$. To solve a supervised learning problem, one seeks to construct a map $f_{\text{approx}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, which, desirably, is such that for any $x \in \mathcal{X}$ and for any Borel measurable $A \subset \mathcal{Y}$, $f_{\text{approx}}(x)(A) \simeq 1$ whenever $f(x) \in A$, and $f_{\text{approx}}(x)(A) \simeq 0$ whenever $f(x) \notin A$; here, $\mathcal{P}(\mathcal{Y})$ denotes the space of probability measures on $\mathcal{Y}$. In other words, one looks for a map f_{approx} which approximates the map $x \to \delta_{f(x)}$ where δ_z denotes the Dirac measure centered at z. Ultimately, this translates to simultaneously interpolating the above dataset through f_{approx}, whilst ensuring generalization/extrapolation, namely reliable prediction on points in $\mathcal{X}$ which are outside of said dataset ([Mallat, 2016]).

1.3. **An optimal control perspective.** There are various ways in which one can construct such an approximation f_{approx}, with different degrees of empirical and theoretical guarantees. In this paper, following a recent trend started with the works [E, 2017; Haber and Ruthotto, 2017; Chen et al., 2018], we shall focus on parametrizing f_{approx} by the flow of neural ODEs, such as

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w(t)\sigma(\mathbf{x}_i(t)) + b(t) & \text{for } t \in (0,T), \\ \mathbf{x}_i(0) = x^{(i)}, \end{cases} \tag{1.1}$$

for $i \in [n]$ and $T > 0$, with σ being a scalar, globally Lipschitz function defined componentwise in (1.1). The matrix $w(t) \in \mathbb{R}^{d\times d}$ and vector $b(t) \in \mathbb{R}^d$ play the role of controls (called *parameters* in machine learning jargon), which in practice are found by solving an empirical risk minimization problem of the form

$$\inf_{\substack{u=(w,b)\in\mathfrak{U} \\ \mathbf{x}_i \text{ solves } (1.1)}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \text{loss}\left(P\mathbf{x}_i(T), y^{(i)}\right)}_{:=\mathscr{E}(\mathbf{x}(T))} + \int_0^T \|u(t)\|_1 \,\mathrm{d}t. \tag{1.2}$$

Here, $\mathfrak{U}$ is an appropriate Banach subspace of $L^1(0,T;\mathbb{R}^{d_u})$, $P:\mathbb{R}^d\to\mathbb{R}^m$ is an affine map which we suppose to be given[1], and which serves to match the states $\mathbf{x}_i(T)$ with the labels $y^{(i)}$ (typically of different dimensions), while

$$\mathrm{loss}(\cdot,\cdot):\mathbb{R}^m\times\mathcal{Y}\to\mathbb{R}_+$$

is such that $x\mapsto\mathrm{loss}(x,y)$ is continuous for all $y\in\mathcal{Y}$, $\mathrm{loss}(x,y)\neq 0$ whenever $\mu(x)\neq\delta_y$, and $\mathrm{loss}(x,y)\to 0$ when $\mu(x)\to\delta_y$ in an appropriate sense of measures (e.g., for some Wasserstein distance, or for the Kullback-Liebler divergence). A canonical example is given by the square of the euclidean distance (*least squares error*). But more tailored loss functions may be used, including positive and non-coercive ones, such as the *cross-entropy* loss commonly used for classification tasks

$$\mathrm{loss}\big(x,y\big):=-\log\left(\frac{e^{x_y}}{\sum_{j=1}^m e^{x_j}}\right)\qquad\text{for } x\in\mathbb{R}^m,\ y\in[m]. \tag{1.3}$$

Once a solution $u=(w,b)$ to (1.2) is found, one may construct the approximation f_{approx} by setting $f_{\mathrm{approx}}(x)=\mu(\mathbf{x}(T))$ for $x\in\mathcal{X}\subset\mathbb{R}^d$, where $\mathbf{x}(T)$ solves (1.1) with $\mathbf{x}(0)=x$ and control u. The choice of $\mu:\mathcal{X}\to\mathcal{P}(\mathcal{Y})$ depends on the loss function and task at hand; for the least squares error loss for instance, one sets $\mu(x):=\delta_{Px}$, while for the cross-entropy loss, one sets $\mu:=\mathrm{softmax}\circ P$, with $\mathrm{softmax}(z)_\ell=e^{z_\ell}/\sum_{j=1}^m e^{z_j}$ for $\ell\in[m]$ and $z\in\mathbb{R}^m$, as in (1.3) (designating a smooth approximation of the argmax).

The above presentation thus leads one to note that, in the neural ODE setting, supervised learning is a particular optimal control problem, wherein one looks to find a single pair of controls $u=(w,b)$ which steer n trajectories of a nonlinear ODE such as (1.1), corresponding to n different initial data, to n different targets.

1.4. **The role of T.** Let us motivate our reason for considering the neural ODE and optimal control interpretation of supervised learning. In practice, one typically considers some discrete-time analog of (1.1), e.g. a forward Euler scheme of the form

$$\begin{cases}\mathbf{x}_i^{k+1}=\mathbf{x}_i^k+\triangle t\left(w^k\sigma\big(\mathbf{x}_i^k\big)+b^k\right) & \text{for } k\in\{0,\dots,n_t-1\},\\ \mathbf{x}_i^0=x^{(i)},\end{cases} \tag{1.4}$$

for $i\in[n]$, where $n_t\geqslant 2$ and $\triangle t=T/n_t$. The scheme (1.4) is an example of a *residual neural network* (ResNet), a popular neural network architecture introduced in [He et al., 2016]. As shown in [He et al., 2016], such neural networks provide, empirically, remarkable interpolation *and* extrapolation performance when n_t is large (of the orders of hundreds or thousands). Here, n_t is referred to as the *depth* of the network (1.4) and each time-step k is called a *layer*. When n_t is large, one is said to do *deep learning*. However, the theory supporting these empirical results is not completely mature ([Zhang et al., 2021]).

We observe that when $\triangle t>0$ is fixed, the time horizon T can be used to estimate the depth n_t. This warrants the study of the behavior of optimal control problems for neural ODEs when T is increased. On another hand, for many problems in optimal control, tracking the control and the trajectory over the entire time interval yields quantitative stability estimates for both when T is large enough. This is for instance

[1] In practice, P is either an optimizable variable, or its coefficients may be chosen at random. While we fix P for technical purposes, our numerical experiments indicate that the results presented in what follows persist when P is optimized as well.

the case in turnpike theory for linear quadratic (LQ) problems ([Porretta and Zuazua, 2013]). Consequently, in this work, rather than (1.2), we are led to consider

$$\inf_{\substack{u=(w,b)\in\mathfrak{U}\\ \mathbf{x}_i \text{ solves (1.1)}}} \int_0^T \mathscr{E}(\mathbf{x}(t))\,\mathrm{d}t + \int_0^T \|u(t)\|_1\,\mathrm{d}t, \tag{1.5}$$

where $\mathscr{E}$ is defined in (1.2), and where we set $\mathbf{x}(t) = \{\mathbf{x}_i(t)\}_{i\in[n]}$. Our goal in this work is to provide a rather complete picture of the behavior of solutions to (1.5) and (1.1) as functions of T.

1.5. **Our result: temporal sparsity.** As insinuated in above discussions, penalizing the $L^1(0,T;\mathbb{R}^{d_u})$ norm promotes the control $u(t)$ to be sparse in time. This can already be confirmed through numerical experiments[2] before proceeding with theoretical setups and proofs. In Figure 1 (see Figure 2–Figure 3 for further illustrations), we depict a solution of (1.5) for a binary classification task ($\mathcal{Y} := \{-1, 1\}$), with $\sigma \equiv \tanh$, using the cross-entropy loss defined in (1.3), with $T = 5$, $\triangle t = 1/3$ with a midpoint scheme, and $n = 3000$. We also impose the constraint $\|u(t)\|_1 \leqslant M$ with $M = 8$, to avoid concentration near $t = 0$. (See Remark 1.) The numerical results show that optimal controls $u(t) = (w(t), b(t))$ concentrate within an interval $[0, T^*]$, and vanish beyond time T^*. Furthermore, the corresponding states $\{\mathbf{x}_i(t)\}_{i\in[n]}$ are, naturally, stationary for $t \geqslant T^*$, but actually in the regime in which $\mathscr{E}(\mathbf{x}(t))$ is near 0, as desired. In the following section, we shall mathematically formalize these results and provide rigorous proofs ensuring their validity in a wide array of functional settings.

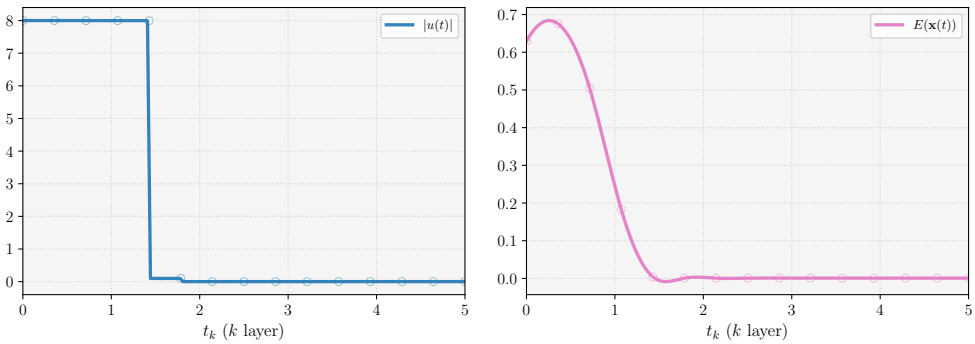

FIGURE 1. (*Left*) Sparsity in time for optimal controls $u(t)$ solving (1.5). (*Right*) Decay of the error $\mathscr{E}(\mathbf{x}(t))$ of the optimal states $\{\mathbf{x}_i(t)\}_{i\in[n]}$ to zero. Both vanish in a finite time $T^* \sim 1.5$, which corresponds to 5 layers. More layers are hence not necessary.

1.6. **Outline.** The remainder of this work is structured as follows. In **Section 2**, we provide the functional setting and our main result (Theorem 2.1), which corroborates the numerical experiment presented just above. Further numerical visualizations of the same experiment may also be found therein. The proof of Theorem 2.1 may be found in **Section 3**. We conclude with a selection of open problems in **Section 4**.

[2]Codes may be found at `https://github.com/borjanG/dynamical.systems`. Experiments were done using `PyTorch` [Paszke et al., 2017]. Minimization was done with Adam [Kingma and Ba, 2014].

1.7. **Notation.** We denote $\mathbb{R}_+ := [0, +\infty)$. For any tensor $u = (u_1, \dots, u_{d_u}) \in \mathbb{R}^{d_u}$ and any $p \in [1, +\infty)$, we denote by $\|u\|_p$ the p-Frobenius norm of u. We focus on $p = 1$, but our results hold for any p.

2. Main result

2.1. **Setup.** We henceforth suppose we are given a dataset

$$\left\{x^{(i)}, y^{(i)}\right\}_{i\in[n]} \subset \mathcal{X} \times \mathcal{Y} \tag{2.1}$$

with $\mathcal{X} \subset \mathbb{R}^d$ and $x^{(i)} \neq x^{(j)}$ for $i \neq j$. The label space $\mathcal{Y}$ may either be a finite subset of $\mathbb{N}$, or a subset of $\mathbb{R}^m$. To have a more coherent presentation and simplify the technical details, we shall stack all of the trajectories $\mathbf{x}_i(t)$ appearing in neural ODEs as (1.1), in order, into one single vector $\mathbf{x}(t) \in \mathbb{R}^{dn}$. Namely, we set

$$\mathbf{x}(t) := \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_n(t) \end{bmatrix} \in \mathbb{R}^{d_x}, \qquad \mathbf{x}^0 := \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix} \in \mathbb{R}^{d_x}$$

for $i \in [n]$ and $t \geqslant 0$, where $d_x := dn$, and consider stacked neural ODEs in the general form

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), u(t)) & \text{for } t \in (0, T), \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases} \tag{2.2}$$

where $u(t) := (w(t), b(t)) \in \mathbb{R}^{d^2+d}$. We provide some important comments on the choice of initial datum $\mathbf{x}^0$ to Remark 3. As presented in (1.1), for the stacked system the nonlinearity $\mathbf{f} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_x}$ may take the form

$$\mathbf{f}(\mathbf{x}, u) = \begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix} \sigma(\mathbf{x}) + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix} \tag{2.3}$$

for $\mathbf{x} \in \mathbb{R}^{d_x}$ and $u = (w, b) \in \mathbb{R}^{d_u}$, with $d_u := d^2 + d$. Once again, $\sigma \in \mathrm{Lip}(\mathbb{R})$ is defined componentwise, so that each component of $\mathbf{f}$ coincides with the canonical neural ODE given in (1.1). Permutations may also be considered, such as

$$\mathbf{f}(\mathbf{x}, u) = \sigma\left(\begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix} \mathbf{x} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}\right), \tag{2.4}$$

as in the original paper [E, 2017]. Actually the key assumption we shall henceforth make regarding $\mathbf{f}$ is the following.

Assumption 1 (Homogeneous dynamics)**.** *We suppose that $\sigma \in \mathrm{Lip}(\mathbb{R})$. We suppose that $\mathbf{f}$ is 1–homogeneous with respect to the controls u, in the sense that*

$$\mathbf{f}(\mathbf{x}, \alpha u) = \alpha\, \mathbf{f}(\mathbf{x}, u)$$

for all $(\mathbf{x}, u) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_u}$ and for all $\alpha > 0$.

This is clearly the case for dynamics $\mathbf{f}$ parametrized as in (2.3), whilst for (2.4), we shall moreover assume that σ is 1–homogeneous – a canonical example is the ReLU $\sigma(x) = \max\{x, 0\}$, or more general variants such as $\sigma(x) = \max\{ax, x\}$ for $a \in [0, 1)$. Such homogeneity assumptions are not rare in neural network theory, in which one

commonly makes use of scaling arguments, see [Chizat and Bach, 2018] for instance. Now, as seen in (1.5), given $T > 0$ we shall consider the following minimization problem

$$\inf_{\substack{u \in \mathfrak{U}_{\mathrm{ad},T} \\ \mathbf{x} \text{ solves } (2.2)}} \underbrace{\int_0^T \mathscr{E}(\mathbf{x}(t))\,\mathrm{d}t + \int_0^T \|u(t)\|_1\,\mathrm{d}t}_{:=\mathscr{J}_T(u)}, \tag{2.5}$$

where $\mathscr{E}$ is defined in (1.2), and

$$\mathfrak{U}_{\mathrm{ad},T} := \left\{ u \in L^1(0,T;\mathbb{R}^{d_u}) \colon \ \|u(t)\|_1 \leqslant M \ \text{ a.e. in } (0,T) \right\}$$

for a fixed thresholding constant $M > 0$. Note that for such controls, (2.2) admits a unique solution $\mathbf{x} \in C^0([0,T];\mathbb{R}^{d_x})$ by the Cauchy-Lipschitz theorem. We postpone commenting the need of having an L^∞ constraint in $\mathfrak{U}_{\mathrm{ad},T}$ to Remark 1. Before doing so, we make precise the exact assumptions we shall henceforth make regarding the loss function inducing the error $\mathscr{E}$, defined in (1.2), appearing in (2.5).

Assumption 2 (The loss function)**.** *We suppose that* $\mathrm{loss}(\cdot,\cdot) : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}_+$ *appearing in* (1.2) *satisfies*

$$\mathrm{loss}(\cdot, y) \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^m;\mathbb{R}_+) \qquad \textit{and} \qquad \inf_{x \in \mathbb{R}^m} \mathrm{loss}(x,y) = 0$$

for all $y \in \mathcal{Y}$.

This assumption is generic among most losses considered in practice, including all those induced by a distance (e.g., least squares error) and the cross-entropy loss (1.3).

2.2. **Main result.** Throughout the paper, we will assume that the neural ODE can interpolate the dataset defined in (2.1), either in finite or in infinite time. This is an exact controllability assumption, as we shall suppose that there exist controls for which the corresponding stacked trajectory $\mathbf{x}(t)$ makes $\mathscr{E}(\mathbf{x}(\cdot))$ (defined in (1.2)) vanish in finite or in infinite time respectively.

Definition 1 (Interpolation)**.** *We say that*

(i) (2.2) interpolates *the dataset* (2.1) *in some time* $T > 0$ *if there exists* $T > 0$ *and* $u \in L^\infty(0,T;\mathbb{R}^{d_u})$ *such that the solution* $\mathbf{x} \in C^0([0,T];\mathbb{R}^{d_x})$ *to* (2.2) *satisfies*

$$\mathscr{E}(\mathbf{x}(T)) = 0.$$

(ii) (2.2) asymptotically interpolates *the dataset* (2.1) *if there exist* $T > 0$, *some function* $h \in C^\infty([T,+\infty);\mathbb{R}_+)$ *satisfying*

$$\dot{h} < 0 \qquad \textit{and} \qquad \lim_{t \to +\infty} h(t) = 0,$$

and some $u \in L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})$ *such that the solution* $\mathbf{x} \in C^0(\mathbb{R}_+;\mathbb{R}^{d_x})$ *to* (2.2) *set on* $\mathbb{R}_+$ *satisfies*

$$\mathscr{E}(\mathbf{x}(t)) \leqslant h(t)$$

for $t \geqslant T$.

These conditions actually hold for the dynamics $\mathbf{f}$ and many of the errors $\mathscr{E}$ we consider here – we postpone this discussion to Remark 2. We may now state our main result.

Theorem 2.1. *Suppose $T > 0$ and $M > 0$ are fixed. Let $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ be any (should it exist[3]) solution to (2.5). Let $\mathbf{x}_T \in C^0([0,T];\mathbb{R}^{d_x})$ denote the corresponding solution to (2.2). Then, there exists some time $T^* \in (0,T]$ such that*

$$\begin{aligned} \|u_T(t)\|_1 &= M \qquad \text{for a.e. } t \in (0,T^*), \\ \|u_T(t)\|_1 &= 0 \qquad \text{for a.e. } t \in (T^*,T). \end{aligned} \tag{2.6}$$

Moreover, T^* is such that

$$\mathscr{E}(\mathbf{x}_T(T^*)) \leqslant \mathscr{E}(\mathbf{x}_T(t)) \qquad \text{for } t \in [0,T], \tag{2.7}$$

and, furthermore,

(i) If system (2.2) interpolates the dataset in some time $T_0 > 0$ as per Definition 1, then there exists a constant $\mathfrak{C} > 0$ independent of both T and M, such that

$$T^* \leqslant \mathfrak{C}\left(\frac{1}{M} + \frac{1}{M^2}\right)$$

and

$$\mathscr{E}(\mathbf{x}_T(T^*)) \leqslant \frac{\mathfrak{C}}{T}\left(\frac{1}{M} + 1\right).$$

(ii) If system (2.2) asymptotically interpolates the dataset as per Definition 1, then there exists a constant $\mathfrak{C}(M) > 0$ independent of T such that

$$T^* \leqslant \frac{\mathfrak{C}(M)}{M} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{M}$$

and

$$\mathscr{E}(\mathbf{x}_T(T^*)) \leqslant \frac{\mathfrak{C}(M)}{T} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{T},$$

where h^{-1} denotes the inverse function of h.

Sketch of the proof. In the proof of the theorem, which may be found in Section 3, the stopping time $T^* > 0$ is precisely defined as

$$T^* := \min\left\{t \in [0,T] \colon \mathscr{E}(\mathbf{x}_T(t)) = \min_{s\in[0,T]} \mathscr{E}(\mathbf{x}_T(s))\right\}.$$

This implies (2.7) by definition. One then shows that the temporal sparsity in equations (2.6) holds. This is done by a contradiction argument: one supposes that either of both conclusions doesn't hold, and in both cases, constructs auxiliary controls which are strict minimizers for $\mathscr{J}_T$ defined in (2.5). This is quite transparent in the case in which $\|u_T(t)\| \neq 0$ for $t \geqslant T^*$, in which case, one can simply use a zero extension of $u_T(t)$ for $t \geqslant T^*$ to conclude. On the other hand, if $\|u_T(t)\| < M$ for $t \in (0,T^*)$, the construction is more delicate and technical, and makes crucial use of the scaling provided provided by the homogeneous dynamics, and the invariance of the $L^1(0,T;\mathbb{R}^{d_u})$ by this scaling. The estimates on the stopping time T^* and on the error evaluated at the stopping time can then be obtained by making use of the interpolation assumptions and the mentioned scaling, for constructing suboptimal controls which can be estimated appropriately. In particular, our arguments do not rely on studying the

[3]One can show that a minimizer exists when $\mathbf{f}$ is as in (2.3) by means of the direct method in the calculus of variations. However, for $\mathbf{f}$ as in (2.4), it's not clear if there is enough compactness to convert weak convergences into pointwise ones for passing to the limit inside σ.

first-order optimality system, and is specifically tailored to the particular ODEs in question. This allows us to avoid smallness assumptions on the data, and smoothness assumptions on the nonlinearity.

2.3. **Turnpike property.** The behavior displayed in Theorem 2.1 and Figure 1 – Figure 5 can, in some contexts, be seen as a novel manifestation of the *turnpike property* in optimal control: over long time horizons, the optimal pair $(u_T(t), \mathbf{x}_T(t))$ should be "near" an optimal steady pair $(\overline{u}, \overline{\mathbf{x}})$, namely a solution to the problem

$$\inf_{\substack{(u,\mathbf{x})\in\mathbb{R}^{d_u}\times\mathbb{R}^{d_x}\\ \mathbf{f}(\mathbf{x},u)=0}} \mathscr{E}(\mathbf{x}) + \|u\|_1. \tag{2.8}$$

(See [Porretta and Zuazua, 2013; Trélat and Zuazua, 2015].) Let us suppose that $\mathrm{loss}(x,y) = \|x-y\|_2^2$ (but the discussion remains true for any distance) and drop the subscript T, hence

$$\mathscr{E}(\mathbf{x}(t)) = \frac{1}{n}\sum_{i=1}^{n} \left\| P\mathbf{x}_i(t) - y^{(i)} \right\|_2^2.$$

Theorem 2.1 then implies that

$$\left\| P\mathbf{x}_i(t) - y^{(i)} \right\|_2^2 \leqslant \frac{C(M)}{T} \tag{2.9}$$

for all $t \geqslant T^*$ and $i \in [n]$. Now note that $\mathbf{f}(\overline{\mathbf{x}}, 0) = 0$ for any $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$. In particular, if $P : \mathbb{R}^d \to \mathbb{R}^m$ is surjective, then taking $\overline{\mathbf{x}}_i \in P^{-1}\left(y^{(i)}\right)$ for $i \in [n]$, we see that there exists some $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$, with $\overline{\mathbf{x}}_i \in P^{-1}\left(y^{(i)}\right)$ such that $(0, \overline{\mathbf{x}})$ is the unique solution to the steady problem (2.8). Now, on one hand, the sparsity in time result already ensures a finite-time turnpike property for the optimal controls $u_T(t)$ to the steady correspondent $\overline{u} \equiv 0$. On the other hand, (2.9) can be seen as

$$\left\| P\Big(\mathbf{x}_i(t) - \overline{\mathbf{x}}_i\Big) \right\|_2^2 \leqslant \frac{C(M)}{T}$$

for all $t \geqslant T^*$, $i \in [n]$ and for some $\overline{\mathbf{x}}_i \in P^{-1}\left(y^{(i)}\right)$. This is a turnpike property for (a projection of) the state $\mathbf{x}(t)$.

Actually, one can see that the above artifact is not bound to machine learning, and applies to more classical optimal control problems of the form

$$\inf_{\substack{u\in\mathfrak{U}_{\mathrm{ad},T}\\ \mathbf{x} \text{ solves (2.11)}}} \int_0^T \|\mathbf{x}(t) - \overline{\mathbf{x}}\|_p^p + \int_0^T \|u(t)\|_1 \,\mathrm{d}t, \tag{2.10}$$

where $p \in [1, +\infty)$, $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$ is fixed, and the underlying system is of *driftless control-affine* form

$$\begin{cases} \dot{\mathbf{x}}(t) = \displaystyle\sum_{j=1}^{d_u} u_j(t) f_j(\mathbf{x}(t)) & \text{in } (0,T), \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases} \tag{2.11}$$

with $f_j : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ for $j \in [d_u]$. Then $(\overline{u}, \overline{\mathbf{x}}) = (0, \overline{\mathbf{x}})$ is the optimal steady pair, namely the unique solution to

$$\inf_{\substack{(u,\mathbf{x})\in\mathbb{R}^{d_u}\times\mathbb{R}^{d_x}\\ \sum_{j=1}^{d_u} u_j f_j(\mathbf{x})=0}} \|\mathbf{x} - \overline{\mathbf{x}}\|_p^p + \|u\|_1,$$

and we have the following corollary of Theorem 2.1.

Corollary 2.1 (Turnpike property). *Suppose $\mathbf{x}_0, \overline{\mathbf{x}} \in \mathbb{R}^{d_x}$ are given, and let $T > 0$, $M > 0$ be fixed. Suppose $f_j \in \mathrm{Lip}(\mathbb{R}^{d_x}; \mathbb{R}^{d_x})$ for $j \in [d_u]$. Let $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ be any solution to (2.10). Let $\mathbf{x}_T$ denote the corresponding solution to (2.11). Then there exists some time $T^* \in (0, T]$ and some constant $\mathfrak{C} > 0$ independent of both T and M such that*

$$\|u_T(t)\|_1 = M 1_{[0,T^*]}(t)$$

holds for a.e. $t \in (0, T)$, and

$$\|\mathbf{x}_T(t) - \overline{\mathbf{x}}\|_p^p \leqslant \frac{\mathfrak{C}}{T}\left(\frac{1}{M} + 1\right).$$

holds for all $t \in [T^*, T]$.

Theorem 2.1 and Corollary 2.1 can then be seen as a new result in the turnpike literature: they provide a finite-time, exact turnpike for any optimal control u_T solving (2.10) (new on its own, due to the L^1 penalty of the controls), and a polynomial turnpike for the corresponding optimal state $\mathbf{x}_T(t)$ for $t \in [T^*, T]$, without any smallness assumptions on the initial data $\mathbf{x}^0$, on the target $\overline{\mathbf{x}}$, or smoothness assumptions on the dynamics f. The latter are deemed necessary for arguments which make use of the Pontryagin Maximum Principle and linearization ([Trélat and Zuazua, 2015]). A final arc near $t = T$ doesn't appear as the running cost is at its minimal value for $t \in [T^*, T]$. Similar results have been obtained for L^2 penalties in [Esteve et al., 2020a,b] (see also [Faulwasser et al., 2021; Effland et al., 2020; Gugat et al., 2021]).

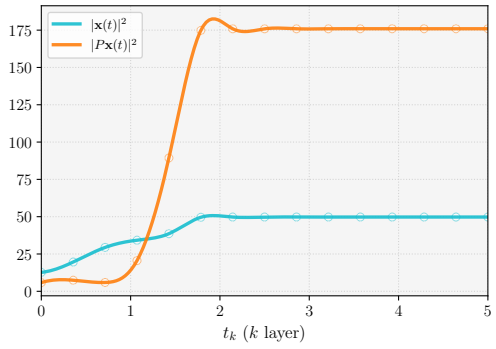

FIGURE 2. For the experiment of Figure 1, we see that not only the error $\mathscr{E}(\mathbf{x}(t))$ decays (at least polynomially), but the trajectories $\mathbf{x}(t)$ also reach some stationary point which ought to be near argmin $\mathscr{E}$. (See Section 2.3.)

It is curious that in Figure 2, we actually see this phenomenon for the trajectories when $\mathscr{E}$ is given by the cross-entropy loss (1.3). In this case, $\mathscr{E}$ is not coercive: $\mathscr{E}(\mathbf{x}(t))$ approaches 0 only if the margin $\gamma(\mathbf{x}_T(T))$ defined in (2.12) goes to $+\infty$. Namely, every trajectory $\mathbf{x}_i(T)$ for $i \in [n]$ ought to grow to $+\infty$ in an appropriate direction in $\mathbb{R}^d$. Thus, in this non-coercive case, we do not interpret the graph of Figure 2 as a turnpike property, since the turnpike would depend on (and increase with) T. Rather, the trajectories $\mathbf{x}(t)$ become stationary beyond time $t \geqslant T^*$ to some point $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$, which is polynomially "sliding" to $+\infty$ (the "argmin" of $\mathscr{E}$) as $T \to +\infty$.

2.4. **Discussion.** Let us provide a structured commentary regarding the different assumptions surrounding the above result, possible extensions, and novelty with respect to past literature on both neural ODEs and optimal control.

Remark 1 (L^∞ constraint)**.** *Penalizing the L^1 norm in (2.5) enforces the use of sparse controls, which without an L^∞ constraint, would a priori concentrate near $t = 0$ as a Dirac mass. We include the L^∞ constraint in the definition of $\mathfrak{U}_{\mathrm{ad},T}$ in order to prevent such degeneracy. One can then recover a Dirac mass centered at $t = 0$ when $M \to +\infty$.*

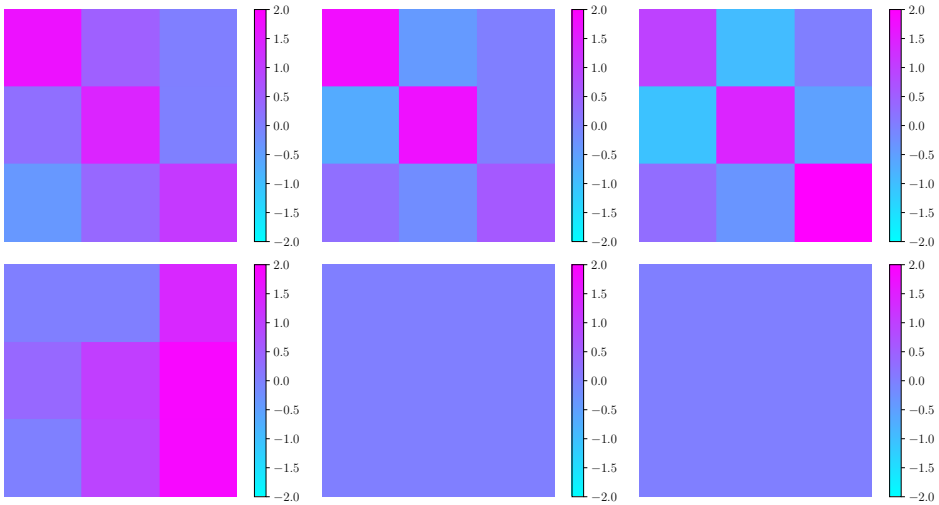

FIGURE 3. The matrix $w_T(t) \in \mathbb{R}^{3\times 3}$ part of $u_T(t) = (w_T(t), b_T(t))$ for the experiment of Figure 1 at $t \in \{0.33, 0.67, 1, 1.33, 1.67, 4.33\}$, seen from left to right, indicating the temporal sparsity (here, beyond $T^* \sim 1.5$) shown in Theorem 2.1 and seen in Figure 1. One sees the lack of coordinate-wise sparsity ($u_i(t)u_j(t) = 0$ for $i \neq j$ and all t), for which different penalties should be used ([Kalise et al., 2020]).

Remark 2 (Interpolation)**.** *In the case where ℓ attains its infimum (here 0), (finite-time) interpolation as per Definition 1, which can be seen as simultaneous or ensemble controllability, has been shown to hold for the dynamics $\mathbf{f}$ as considered here in several recent works [Cuchiero et al., 2020; Li et al., 2019; Esteve et al., 2020a; Agrachev and Sarychev, 2021; Ruiz-Balet and Zuazua, 2021; Ruiz-Balet et al., 2021; Bárcena-Petisco, 2021]. We have stated it as an assumption in Theorem 2.1 to make transparent the ingredients used in the proof. On another hand, as our setting includes losses which do not attain their infimum, one cannot expect exact interpolation to always hold. This is exemplified by the cross-entropy defined in (1.3), which motivates the asymptotic interpolation hypothesis. Under the assumption that there exits a control $u \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ for which the* margin $\gamma = \gamma(\mathbf{x}(T_0))$ *defined as*

$$\gamma(\mathbf{x}(T_0)) := \min_{i\in[n]} \left\{ \left(P\mathbf{x}_i(T_0)\right)_{y^{(i)}} - \max_{\substack{j\in[m]\\ j\neq y^{(i)}}} \left(P\mathbf{x}_i(T_0)\right)_j \right\} \tag{2.12}$$

is positive in some $T_0 > 0$, in [Geshkovski, 2021, Proposition 7.4.2] asymptotic interpolation is shown to hold for the cross-entropy (1.3) with

$$h(t) = \log\left(1 + (m-1)e^{-\gamma e^t}\right).$$

Remark 3 (Initial data in (2.2))**.** *In binary classification tasks ($\mathcal{Y} = \{-1, 1\}$) for instance, we are looking to approximate a characteristic function of some set $A \subset \mathbb{R}^d$. If the dataset is not linearly separable, in the sense that there exists $\mathbf{w} \in \mathbb{R}^d$ such that*

$$\min_{i\in[n]} \left(y^{(i)}\mathbf{w}^\top x^{(i)}\right) > 0,$$

then solving such a problem would entail separating, over time, the dataset by means of the controlled flow of an ODE. This cannot always be done due to the backward uniqueness of ODEs, as trajectories cannot cross in the state space $\mathbb{R}^d$. As noted in [Dupont et al., 2019], a simple remedy is to embed every datum $x^{(i)} \in \mathbb{R}^d$ into $\mathbb{R}^{d+1}$ by appending a 0 to its tail, namely considering

$$\mathbf{x}_i^0 = \begin{bmatrix} x^{(i)} \\ 0 \end{bmatrix}.$$

This is seen in Figure 4. By abuse of notation, whenever the dataset is not linearly separable, we shall suppose that $x^{(i)}$ are already in an augmented form, and keep the same notation for simplicity.

Remark 4 (The dynamics)**.**

- *While there are several works in the literature which prove sparsity in time for controls found by minimizing some functional, even for systems with drifts (unlike ours), the theory is either done for linear systems ([Zuazua, 2010; Alt and Schneider, 2015; Geshkovski and Zuazua, 2021]), or nonlinear ones for specific regression functionals and/or differentiable dynamics and/or infinite time horizons ([Kalise et al., 2017, 2020; Vossen and Maurer, 2006]). Similar considerations can be found in the literature on optimal control of multi-agent/mean-field systems ([Caponigro et al., 2013; Fornasier et al., 2014; Caponigro et al., 2015]). The setting we presented herein makes no such assumptions, and our results can then be seen as complementary to these works. Our consideration of divergences instead of distances in the optimization problem can be seen as a novelty in the optimal control context.*
- *More complicated neural ODEs of the form*

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w^2(t)\sigma\left(w^1(t)\mathbf{x}_i(t)\right) & \text{in } (0,T) \\ \mathbf{x}_i(0) = \mathbf{x}^0 \end{cases} \tag{2.13}$$

for $i \in [n]$, where $w^2(t) \in \mathbb{R}^{d\times d_{hid}}$ and $w^1(t) \in \mathbb{R}^{d_{hid}\times d}$ (we omit the translation control for simplicity), tend to perform well in experiments due to the higher number of controls. When σ is 1–homogeneous, and $w^2(t) = \pm 1$ or is an orthogonal matrix for all t, Theorem 2.1 still holds due to the fact that Lemma 3.1 applies for such dynamics. When we remove such assumptions on $w^2(t)$, the technical impediment we encounter is the lack of invariance of the $L^1(0,T;\mathbb{R}^{d_u})$ norm with respect to the canonical scaling induced by the equation (Lemma 3.1). Indeed, if one sets $w_1^1(t) := T^\alpha w^1(tT)$ and

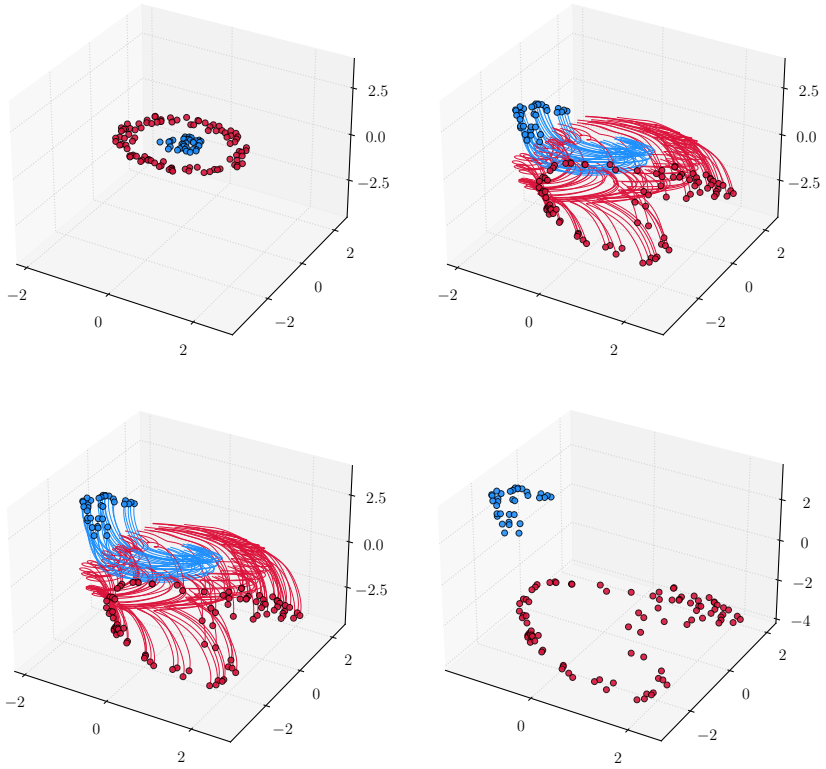

FIGURE 4. The evolution of (a part of) the states $\{\mathbf{x}_i(t)\}_{i\in[n]}$ solving (1.1), for the experiment of Figure 1. Clockwise from top to bottom: $t = 0$, $t \leqslant 1.33$, $t \leqslant 5$, $t = 5$. The states are stationary in a separation regime beyond $t \geqslant T^*$, as indicated by Figure 1.

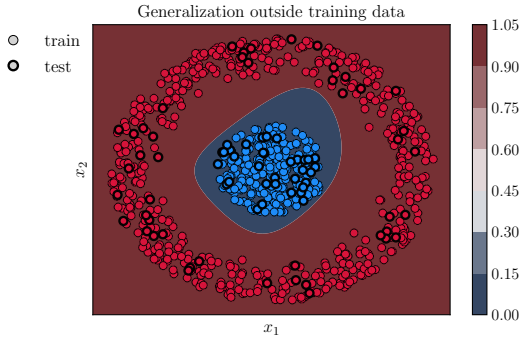

FIGURE 5. The learned predictor f_{approx} through the neural ODE flow. It captures the shape of the dataset given by f, accurately classifies the test data, thus ensuring satisfactory generalization.

$w_1^2(t) := T^{1-\alpha} w^2(tT)$ *for* $t \in [0,1]$ *and some* $\alpha \in (0,1)$, *then it can be seen*

that $\mathbf{x}_i^1(t) := \mathbf{x}_i(tT)$ *solves* (2.13) *on* $[0,1]$. *Yet,*

$$\int_0^T \left\|w^1(t)\right\|_1 \,\mathrm{d}t + \int_0^T \left\|w^2(t)\right\|_1 \,\mathrm{d}t = T^{\alpha-1}\int_0^1 \left\|w_1^1(s)\right\|_1 \,\mathrm{d}s + T^{-\alpha}\int_0^1 \left\|w_1^2(s)\right\|_1 \,\mathrm{d}s.$$

This is incompatible with our proof strategy. However, noting the above identity, one could investigate the applicability of our techniques to (2.13) *and parameter regularizations of the form*

$$\int_0^T \left\|w^1(t)\right\|_1^{1/\alpha} \,\mathrm{d}t + \int_0^T \left\|w^2(t)\right\|_1^{1/1\,-\,\alpha} \,\mathrm{d}t,$$

which would be invariant by the above scaling. In such a case, the sparsity pattern should be defined with respect to the regularization one considers. Due to the likely nontrivial nature of the proof, we leave it open.

3. Proofs

In this section we provide the proof of Theorem 2.1. We shall split the proof into two parts. We first state and prove Proposition 3.1, which contains the first part of Theorem 2.1, concerning the temporal sparsity of optimal controls. The proof of the latter is done throughout Section 3.1. We then provide the remainder of the proof in Section 3.2.

3.1. Preliminary results.

The main goal of this subsection is to state and prove Proposition 3.1, ensuring the temporal sparsity of optimal controls. A cornerstone of our forthcoming arguments is the possibility of rescaling any trajectory of (2.2) set in $[0,T_0]$ to obtain the same trajectory set on $[0,T]$.

Lemma 3.1. *Let* $\mathbf{x}^0 \in \mathbb{R}^{d_x}$, $T_0 > 0$, $u_{T_0} \in L^1(0,T_0;\mathbb{R}^{d_u})$, *and let* $\mathbf{x}_{T_0}$ *be the unique solution to* (2.2) *set on* $[0,T_0]$, *with control* u_{T_0}. *Let* $T > 0$, *and define*

$$u_T(t) := \frac{T_0}{T} u_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t \in [0,T],$$

and

$$\mathbf{x}_T(t) := \mathbf{x}_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t \in [0,T].$$

Then $\mathbf{x}_T$ *is the unique solution to* (2.2) *with control* u_T.

Such time-scaling arguments are standard in the context of driftless control-affine systems (see [Coron, 2007, Chapter 3, Section 3.3]). It is here that the homogeneity of the dynamics with respect to the controls plays a crucial role. We omit the proof, which is straightforward. We also summarize the notion of temporal sparsity through the following definition.

Definition 2 (Sparse controls)**.** *Let* $M > 0$ *and* $0 < T^* \leqslant T$ *be fixed. We say that* $u \in \mathfrak{U}_{\mathrm{ad},T}$ *is* sparse *in* (T^*,T) *if*

$$\|u(t)\|_1 = M \qquad \text{a.e. } t \in (0,T^*), \tag{3.1}$$

$$\|u(t)\|_1 = 0 \qquad \text{a.e. } t \in (T^*,T). \tag{3.2}$$

For any $T^* > 0$, *we shall denote by* $\mathfrak{U}_{\mathrm{sp},T^*}$ *the set consisting of all* $u \in \mathfrak{U}_{\mathrm{ad},T}$ *which are sparse in* (T^*,T), *namely which satisfy* (3.1) – (3.2).

Proposition 3.1. *Let $T > 0$ and $M > 0$ be fixed. Let $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ be a global minimizer of $\mathscr{J}_T$ defined in (2.5), and let $\mathbf{x}_T$ be the corresponding unique solution to (2.2). Then $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$, where T^* is defined as*

$$T^* := \min\left\{ t \in [0,T] \colon \mathscr{E}(\mathbf{x}_T(t)) = \min_{s\in[0,T]} \mathscr{E}(\mathbf{x}_T(s)) \right\}. \tag{3.3}$$

Note that the T^* is clearly well defined, as the set over which the min is taken is clearly bounded, and is also closed as the preimage of the singleton

$$\left\{ \min_{s\in[0,T]} \mathscr{E}(\mathbf{x}_T(s)) \right\}$$

under the continuous map $t \longmapsto \mathscr{E}(\mathbf{x}(t))$. The core of the proof of Proposition 3.1 lies in the following lemma, which ensures that if a control $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ does not saturate the L^∞–constraint before some time T^*, then u_T is not optimal for $\mathscr{J}_T$ and can always be "improved" through the scaling of Lemma 3.1.

Lemma 3.2. *Let $T > 0$ and $M > 0$ be fixed. Let $u_T \in \mathfrak{U}_{\mathrm{ad},\mathfrak{T}}$ be any admissible (but not necessarily optimal) control, and let $T^* > 0$ be defined as in (3.3). Assume that, for some $\theta \in (0,1)$, there exists a finite collection of disjoint non-empty intervals $\{(a_j,b_j)\}_{j\in[\mathfrak{J}]}$ with $(a_j,b_j) \subset (0,T^*)$ for which*

$$\|u_T(t)\|_1 \leqslant (1-\theta)M \qquad \text{for a.e. } t \in \mathbf{O}_{\mathfrak{J}}, \tag{3.4}$$

and

$$\mathscr{E}(\mathbf{x}_T(t)) - \mathscr{E}(\mathbf{x}_T(T^*)) \geqslant \theta \qquad \text{for all } t \in \mathbf{O}_{\mathfrak{J}} \tag{3.5}$$

hold, where

$$\mathbf{O}_{\mathfrak{J}} := \bigcup_{j=1}^{\mathfrak{J}} (a_j,b_j)\,.$$

Then there exists some $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ satisfying

$$\overline{u}(t) = 0 \qquad \text{for a.e. } t \in (T^*-\tau, T), \tag{3.6}$$

and

$$\mathscr{J}_T(\overline{u}) \leqslant \mathscr{J}_T(u_T) - \theta\tau,$$

where

$$\tau := \theta\,\mathrm{meas}(\mathbf{O}_{\mathfrak{J}}) = \theta \sum_{j=1}^{\mathfrak{J}} (b_j - a_j).$$

We may now provide the proof to Proposition 3.1.

Proof of Proposition 3.1. We argue by contradiction. Suppose that $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ is a global minimizer of $\mathscr{J}_T$ such that $u_T \notin \mathfrak{U}_{\mathrm{sp},T^*}$, where $T^* > 0$ is defined as in the statement. Hence, either condition (3.1) or condition (3.2) does not hold.

Case 1: (3.2) does not hold. Let us thus suppose that

$$\|u(t)\|_1 \neq 0 \qquad \text{a.e. } t \in (T^*,T). \tag{3.7}$$

Consider

$$\overline{u}(t) = \begin{cases} u_T(t) & \text{for } t \in [0,T^*] \\ 0 & \text{for } t \in (T^*,T]. \end{cases}$$

Clearly $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$. Furthermore, we have

$$\overline{\mathbf{x}}(t) = \mathbf{x}_T(t) \qquad \text{for } t \in [0, T^*],$$

and since $\mathbf{f}(\cdot, 0) \equiv 0$, also

$$\overline{\mathbf{x}}(t) = \overline{\mathbf{x}}(T^*) = \mathbf{x}_T(T^*), \qquad \text{for } t \in [T^*, T].$$

Combining these facts with the definition (3.3) of T^*, we are lead to

$$\int_0^T \mathscr{E}(\overline{\mathbf{x}}(t))\,\mathrm{d}t = \int_0^{T^*} \mathscr{E}(\mathbf{x}_T(t))\,\mathrm{d}t + \int_{T^*}^T \mathscr{E}(\mathbf{x}_T(T^*))\,\mathrm{d}t \leqslant \int_0^T \mathscr{E}(\mathbf{x}_T(t))\,\mathrm{d}t.$$

By virtue of (3.7) we also find

$$\begin{aligned}\int_0^T \|\overline{u}(t)\|_1\,\mathrm{d}t &= \int_0^{T^*} \|u_T(t)\|_1\,\mathrm{d}t \\ &< \int_0^{T^*} \|u_T(t)\|_1\,\mathrm{d}t + \int_{T^*}^T \|u_T(t)\|_1\,\mathrm{d}t = \int_0^T \|u_T(t)\|_1\,\mathrm{d}t.\end{aligned}$$

Combining the two previous inequalities, we deduce that $\mathscr{J}_T(\overline{u}) < \mathscr{J}_T(u_T)$, which contradicts the optimality of u_T.

Case 2: (3.1) does not hold. The idea is to again construct an auxiliary control which improves u_T to deduce a contradiction. We now split the proof in three steps.

Step 1. If (3.1) is not fulfilled, then there must exist some $\theta \in (0, 1)$ such that the set

$$\mathbf{A}_\theta := \Big\{ t \in (0, T^*) : \ \|u_T(t)\|_1 \leqslant (1-\theta)M \Big\}$$

has positive Lebesgue measure, namely $\mathrm{meas}(\mathbf{A}_\theta) > 0$. Now set $\omega := \frac{\mathrm{meas}(\mathbf{A}_\theta)}{2}$, and using elementary set theory we find

$$\mathbf{A}_\theta \cap (0, T^* - \omega) = \mathbf{A}_\theta \setminus \Big((0, T^*) \setminus (0, T^* - \omega) \Big) = \mathbf{A}_\theta \setminus [T^* - \omega, T^*),$$

whence the set

$$\mathbf{B}_\theta := \mathbf{A}_\theta \cap (0, T^* - \omega)$$

also has positive Lebesgue measure: $\mathrm{meas}(\mathbf{B}_\theta) > 0$. By classical results in Lebesgue measure theory (see [Yeh, 2006, Thm. 3.25]), for all $\varepsilon > 0$ there exists a finite collection of disjoint nonempty intervals $\{(a_j, b_j)\}_{j \in [n(\varepsilon)]}$, with $(a_j, b_j) \subset (0, T^* - \omega)$, such that the set

$$\mathbf{O}_\varepsilon := \bigcup_{j=1}^{n(\varepsilon)} (a_j, b_j)$$

satisfies

$$\mathrm{meas}\left(\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta\right) < \varepsilon \quad \text{ and } \quad \mathrm{meas}\left(\mathbf{B}_\theta \setminus \mathbf{O}_\varepsilon\right) < \varepsilon. \tag{3.8}$$

In particular,

$$\mathrm{meas}\left(\mathbf{O}_\varepsilon\right) > \mathrm{meas}(\mathbf{B}_\theta) - \varepsilon. \tag{3.9}$$

Step 2. Let $\varepsilon \in (0, \mathrm{meas}(\mathbf{B}_\theta))$ be arbitrary and to be chosen later, and let $\{(a_j, b_j)\}_{j \in [n(\varepsilon)]}$ be the corresponding collection of disjoint intervals satisfying (3.8), with $\mathbf{O}_\varepsilon$ denoting the union of these intervals as defined above. We now look to construct a control $u^\varepsilon \in \mathfrak{U}_{\mathrm{ad},T}$ such that

$$\|u^\varepsilon(t)\|_1 \leqslant (1 - \theta^*)M$$

and

$$\mathscr{E}(\mathbf{x}^\varepsilon(t)) - \mathscr{E}(\mathbf{x}^\varepsilon(T_\bullet)) \geqslant \theta^*$$

for some $\theta^* > 0$ and for all $t \in \mathbf{O}_\varepsilon$, where

$$T_\bullet := \min\left\{ t \in [0,T] \ : \quad \mathscr{E}(\mathbf{x}^\varepsilon(t)) = \min_{s\in[0,T]} \mathscr{E}(\mathbf{x}^\varepsilon(s)) \right\}$$

should also satisfy $T_\bullet \geqslant T^* - \omega$. To this end, set

$$u^\varepsilon(t) := \begin{cases} u_T(t) & \text{for } t \in (0,T) \setminus (\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta) \\ 0 & \text{for } t \in \mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta. \end{cases}$$

Since $u_T \in \mathfrak{U}_{\mathrm{ad},T}$, it may readily be seen that

$$\|u^\varepsilon(t)\|_1 \leqslant M \qquad \text{for a.e. } t \in (0,T).$$

Hence $u^\varepsilon \in \mathfrak{U}_{\mathrm{ad},T}$. Now let $\mathbf{x}^\varepsilon$ denote the solution to (2.2) associated to u^ε. By virtue of the specific form of $\mathbf{f}$, the Lipschitz continuity of σ, and the Grönwall inequality, we may readily deduce that there exists a constant $C_1 = C_1(T, M, \sigma) > 0$ independent of ε such that

$$\left\|\mathbf{x}^\varepsilon(t) - \mathbf{x}_T(t)\right\|_1 \leqslant C_1 \int_0^T \left\|u^\varepsilon(s) - u_T(s)\right\|_1 \mathrm{d}s \tag{3.10}$$

for all $t \in [0,T]$. On the other hand, by using (3.8), we also deduce that

$$\int_0^T \left\|u^\varepsilon(s) - u_T(s)\right\|_1 \mathrm{d}s \leqslant M \mathrm{meas}\left(\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta\right) < M\varepsilon. \tag{3.11}$$

Combining (3.10) and (3.11) leads us to

$$\left\|\mathbf{x}^\varepsilon(t) - \mathbf{x}_T(t)\right\|_1 < C_1 M \varepsilon$$

for $t \in [0,T]$. Now since $\mathbf{x}_T \in C^0([0,T];\mathbb{R}^{d_x})$, the stacked trajectory $\mathbf{x}_T(t)$ remains in a compact subset of $\mathbb{R}^{d_x}$ for all $t \in [0,T]$. Due to (3.1), and since $\varepsilon \leqslant \mathrm{meas}(\mathbf{B}_\theta)$, we also find that $\mathbf{x}_\varepsilon$ remains in a slightly larger compact subset, independent of ε. Hence, by the locally Lipschitz character of $\mathrm{loss}(\cdot, y)$, implying that of $\mathscr{E}$, the estimate

$$\left|\mathscr{E}\big(\mathbf{x}^\varepsilon(t)\big) - \mathscr{E}\big(\mathbf{x}_T(t)\big)\right| \leqslant C_2 M \varepsilon, \tag{3.12}$$

holds for some $C_2 = C_2(T, M, \sigma, \mathscr{E}) > 0$ independent of ε, and for all $t \in [0,T]$. On the other hand, using only the definition (3.3) of T^*, we find that there exists some $\lambda > 0$ such that

$$\mathscr{E}(\mathbf{x}_T(t)) \geqslant \mathscr{E}(\mathbf{x}_T(T^*)) + \lambda \tag{3.13}$$

for all $t \in [0, T^* - \omega]$. Estimate (3.12) combined with (3.13) yields

$$\begin{aligned} \mathscr{E}(\mathbf{x}^\varepsilon(T^*)) &\leqslant \mathscr{E}(\mathbf{x}_T(T^*)) + C_2 M\varepsilon \leqslant \mathscr{E}(\mathbf{x}_T(t)) - \lambda + C_2 M\varepsilon \\ &\leqslant \mathscr{E}(\mathbf{x}^\varepsilon(t)) - \lambda + 2C_2 M\varepsilon, \end{aligned} \tag{3.14}$$

for all $t \in [0, T^* - \omega]$, which, by choosing $\varepsilon < \lambda/2C_2M$, implies that $T_\bullet \geqslant T^* - \omega$, as desired. The computations done in (3.14) also yield

$$\begin{aligned} \mathscr{E}\big(\mathbf{x}^\varepsilon(t)\big) &\geqslant \mathscr{E}\big(\mathbf{x}^\varepsilon(T^*)\big) + \lambda - C_2\, M\varepsilon \\ &\geqslant \mathscr{E}\big(\mathbf{x}^\varepsilon(T_\bullet)\big) + \lambda - 2C_2 M\varepsilon \end{aligned} \tag{3.15}$$

for all $t \in [0, T^* - \omega]$. As we chose $\varepsilon < \lambda/2C_2M$, we have that $\lambda - 2C_2M\varepsilon > 0$, and may then set

$$\theta^* := \min\left\{\theta, \lambda - 2C_2 M\varepsilon\right\},$$

so that $\theta^* > 0$. By virtue of (3.15),

$$\mathscr{E}\big(\mathbf{x}^\varepsilon(t)\big) - \mathscr{E}\big(\mathbf{x}^\varepsilon(T_\bullet)\big) \geqslant \theta^*$$

holds for all $t \in \mathbf{O}_\varepsilon$. Now, observe that u^ε also satisfies

$$\|u^\varepsilon(t)\|_1 \leqslant (1-\theta^*)\, M$$

for a.e. $t \in \mathbf{O}_\varepsilon$. Indeed, if $t \in \mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta$, then $u^\varepsilon(t) = 0$ by definition, so the inequality clearly holds. On the other hand, if $t \in \mathbf{O}_\varepsilon \cap \mathbf{B}_\theta$, then $t \in \mathbf{A}_\theta$, and since $\theta^* \geqslant \theta$, the conclusion follows.

Step 3. We may now apply Lemma 3.2, which ensures the existence of some $\overline{u}^\varepsilon \in \mathfrak{U}_{\mathrm{ad},T}$ for which

$$\mathscr{J}_T\left(\overline{u}^\varepsilon\right) \leqslant \mathscr{J}_T\left(u^\varepsilon\right) - (\theta^*)^2\, \mathrm{meas}\left(\mathbf{O}_\varepsilon\right) \tag{3.16}$$

holds. As a consequence of (3.11) and (3.12), we have

$$\mathscr{J}_T\left(u^\varepsilon\right) \leqslant \mathscr{J}_T\left(u_T\right) + (1 + C_2 T)\, M\varepsilon,$$

which, when combined with (3.16) and (3.9), yields

$$\mathscr{J}_T\left(\overline{u}^\varepsilon\right) < \mathscr{J}_T(u_T) + (1 + C_2 T) M\varepsilon - (\theta^*)^2(\mathrm{meas}(\mathbf{B}_\theta) - \varepsilon).$$

Looking at the above inequality, we may note that, by choosing $\varepsilon > 0$ even smaller (namely taking

$$\varepsilon \leqslant \frac{(\theta^*)^2\, \mathrm{meas}(\mathbf{B}_\theta)}{(1 + C_2 T) M}$$

would do), we may ensure that

$$\mathscr{J}_T(\overline{u}^\varepsilon) < \mathscr{J}_T(u_T),$$

which contradicts the optimality of u_T. This concludes the proof. □

We conclude this section with a proof of Lemma 3.2.

Proof of Lemma 3.2. We will argue by induction over the number of intervals $\mathfrak{I} \geqslant 1$, constructing appropriately the control $\overline{u}$ explicitly in each step via affine transformations of u_T – the desired estimates will follow by using the time-scaling invariance of the L^1–norm of the controls.

Step 1). Initialization. Let us first assume that $\mathfrak{I} = 1$. Consider

$$\overline{u}(t) := \begin{cases} u_T(t) & \text{for } \ t \in (0, a_1) \\ \dfrac{b_1 - a_1}{c_1 - a_1} u_T\left((t - a_1)\dfrac{b_1 - a_1}{c_1 - a_1} + a_1\right) & \text{for } \ t \in [a_1,\, c_1) \\ u_T\left(t + b_1 - c_1\right) & \text{for } t \in [c_1,\, T^* - (b_1 - c_1)), \\ 0 & \text{for } t \in [T^* - (b_1 - c_1),\, T), \end{cases}$$

where $c_1 \in (a_1, b_1)$ is chosen so that

$$\frac{b_1 - a_1}{c_1 - a_1}(1 - \theta) = 1,$$

which is equivalent to

$$b_1 - c_1 = \theta(b_1 - a_1) =: \tau.$$

Observe that as a consequence of (3.4), we clearly have $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$. In addition, by virtue of the choice of c_1, and the definition of τ, $\overline{u}(t)$ also satisfies (3.6). Now, making

use of the scaling provided by Lemma 3.1, and the fact that $\mathbf{f}(\cdot,0)\equiv 0$, one can check that the state trajectory $\overline{\mathbf{x}}(t)$ associated to $\overline{u}(t)$ is exactly given by

$$\overline{\mathbf{x}}(t)=\begin{cases}\mathbf{x}_T(t) & \text{for } \ t\in[0,a_1)\\ \mathbf{x}_T\left((t-a_1)\dfrac{b_1-a_1}{c_1-a_1}+a_1\right) & \text{for } \ t\in[a_1,\,c_1)\\ \mathbf{x}_T\left(t+b_1-c_1\right) & \text{for } \ t\in[c_1,\,T^*-(b_1-c_1)),\\ \mathbf{x}_T(T^*) & \text{for } \ t\in[T^*-(b_1-c_1),\,T].\end{cases}$$

Moreover, observe that since $\tau:=b_1-c_1$,

$$\mathscr{E}(\overline{\mathbf{x}}(t))=\mathscr{E}(\mathbf{x}_T(T^*)) \qquad \text{for } t\in[T^*-\tau,T]. \tag{3.17}$$

Let us now evaluate the functional $\mathscr{J}_T$ along $\overline{u}$. We start by computing the L^1 norm of $\overline{u}$:

$$\begin{aligned}\|\overline{u}\|_{L^1(0,T;\mathbb{R}^{d_u})}&=\int_0^{a_1}\|u_T(t)\|_1\,\mathrm{d}t+\int_{c_1}^{T^*-(b_1-c_1)}\|u_T(t+b_1-c_1)\|_1\,\mathrm{d}t\\&\quad+\frac{b_1-a_1}{c_1-a_1}\int_{a_1}^{c_1}\left\|u_T\left((t-t_1)\frac{b_1-a_1}{c_1-a_1}+a_1\right)\right\|_1\mathrm{d}t\\&=\int_0^{b_1}\|u_T(s)\|_1\,\mathrm{d}s+\int_{T^*-b_1}^{T^*}\|u_T(s)\|_1\,\mathrm{d}s\\&\leqslant\|u_T\|_{L^1(0,T;\mathbb{R}^{d_u})}.\end{aligned} \tag{3.18}$$

On the other hand, by virtue of (3.17), (3.5), the definition (3.3) of T^*, and the same changes of variable used to deduce (3.18), we find

$$\begin{aligned}\int_0^T\Big(\mathscr{E}(\overline{\mathbf{x}}(t))-\mathscr{E}\big(\mathbf{x}_T(T^*)\big)\Big)\,\mathrm{d}t&=\int_0^{a_1}\Big(\mathscr{E}(\mathbf{x}_T(t))-\mathscr{E}\big(\mathbf{x}_T(T^*)\big)\Big)\,\mathrm{d}t\\&\quad+\underbrace{\frac{c_1-a_1}{b_1-a_1}}_{1-\theta}\int_{a_1}^{b_1}\Big(\mathscr{E}(\mathbf{x}_T(t))-\mathscr{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}t\\&\quad+\int_{b_1}^{T^*}\Big(\mathscr{E}(\mathbf{x}_T(t))-\mathscr{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}t\\&\leqslant\int_0^T\Big(\mathscr{E}(\mathbf{x}_T(t))-\mathscr{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}t-\theta^2(b_1-a_1).\end{aligned}$$

By combining the above inequality with (3.18), it follows that

$$\mathscr{J}_T\left(\overline{u}\right)\leqslant\mathscr{J}_T\left(u_T\right)-\theta^2(b_1-a_1).$$

The statement of the Lemma thus holds for $\mathfrak{I}=1$.

Step 2). Heredity. Let us suppose that, for some $n\geqslant 1$, the statement of the lemma holds whenever $\mathfrak{I}=n$, and let u_T satisfy (3.4) and (3.5) with $\mathfrak{I}=n+1$. Assume without loss of generality that $a_1>a_j$ for all $j\in\{2,\dots,\mathfrak{I}\}$. Using precisely the same argument as in Step 1, we can construct a control $\overline{u}_1$ satisfying

$$\overline{u}_1(t)=0 \qquad \text{for a.e. } t\in(T^*-\tau_1,T)$$

with $\tau_1=\theta(b_1-a_1)$, and

$$\mathscr{J}_T(\overline{u}_1)\leqslant\mathscr{J}_T\left(u_T\right)-\theta^2(b_1-a_1),$$

and which is such that $\overline{u}_1(t) = u_T(t)$ for all $t \in (0, t_1)$. Now observe that, since $a_1 > a_j$ for all $j \geqslant 2$, and in view of (3.17), it follows that $\overline{u}_1$ satisfies (3.4) and (3.5) with $\mathfrak{I} - 1 = n$ number of intervals and with $T_1^* = T^* - \tau_1$ instead of T^*. By the induction hypothesis, we conclude that there exists some control $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ such that

$$\overline{u}(t) = 0 \qquad \text{for a.e. } t \in (T_1^* - \tau, T)$$

with $\tau = \theta \sum_{j=2}^{\mathfrak{I}} (b_j - a_j)$, and

$$\mathscr{J}_T(\overline{u}) \leqslant \mathscr{J}_T(\overline{u}_1) - \theta^2 \sum_{j=2}^{\mathfrak{I}} (b_j - a_j) \leqslant \mathscr{J}_T(u_T) - \theta^2 \sum_{j=1}^{\mathfrak{I}} (b_j - a_j).$$

The statement of the Lemma thus also holds for $\mathfrak{I} = n + 1$. This concludes the proof. □

3.2. **Proof of Theorem 2.1.**

Proof of Theorem 2.1. Properties (2.6) and (2.7) for the minimizers of $\mathscr{J}_T$ follow directly from Proposition 3.1. Let us give the proof of the statements *(i)* and *(ii)* in Theorem 2.1.

Proof of *(i)*. If the interpolation property holds, then there exist $T_0 > 0$ and some control $u_{T_0} \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ such that the associated solution $\mathbf{x}_{T_0} \in C^0([0, T_0]; \mathbb{R}^{d_x})$ to (2.2) satisfies $\mathscr{E}(\mathbf{x}_{T_0}(T_0)) = 0$. Set

$$T_1 := \frac{T_0 \, \|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}}{M}, \tag{3.19}$$

and consider

$$u_{T_1}(t) := \frac{M}{\|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}} u_{T_0}\left(t \frac{T_0}{T_1}\right) \qquad \text{for } t \in (0, T_1).$$

Observe that $u_{T_1} \in \mathfrak{U}_{\mathrm{ad},T_1}$. Furthermore, in view of Lemma 3.1, the associated solution $\mathbf{x}_{T_1}$ to (2.2), is given by

$$\mathbf{x}_{T_1}(t) = \mathbf{x}_{T_0}\left(t \frac{T_0}{T_1}\right) \qquad \text{for } t \in (0, T_1),$$

and hence,

$$\mathscr{E}(\mathbf{x}_{T_1}(T_1)) = \mathscr{E}(\mathbf{x}_{T_0}(T_0)) = 0.$$

Now for any $T > 0$, we define

$$\overline{u}(t) = \begin{cases} u_{T_1}(t) & \text{for } t \in (0, T) \cap (0, T_1) \\ 0 & \text{for } t \in (0, T) \setminus (0, T_1). \end{cases}$$

Clearly $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$. By a simple change of variable, and using (3.19), one sees that

$$\begin{aligned} \mathscr{J}_T(\overline{u}) &\leqslant \int_0^{T_1} \mathscr{E}(\mathbf{x}_{T_1}(t)) \, \mathrm{d}t + M \, T_1 \\ &= \frac{\|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}}{M} \int_0^{T_0} \mathscr{E}(\mathbf{x}_{T_0}(t)) \, \mathrm{d}t + \|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})} \, T_0 \\ &= \frac{C_1}{M} + C_2, \end{aligned} \tag{3.20}$$

holds, where $C_1 > 0$ and $C_2 > 0$ are independent of both T and M. In view of (2.6), any minimizer u_T of $\mathscr{J}_T$ satisfies $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$ for some $T^* \in (0,T]$. Since $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$, using (3.20), we obtain

$$\mathscr{J}_T(u_T) = \int_0^T \mathscr{E}(\mathbf{x}_T(t))\,\mathrm{d}t + M\,T^* \leqslant \mathscr{J}_T(\overline{u}) \leqslant \frac{C_1}{M} + C_2. \tag{3.21}$$

Since $\mathscr{E} \geqslant 0$, using (3.21) we deduce that

$$T^* \leqslant \frac{C_1}{M^2} + \frac{C_2}{M}.$$

Moreover, using (2.7) in (3.21), we also deduce that

$$T\mathscr{E}(\mathbf{x}_T(T^*)) \leqslant \mathscr{J}_T(u_T) \leqslant \frac{C_1}{M} + C_2.$$

The last two estimates imply *(i)* in the statement of Theorem 2.1, as desired.

Proof of *(ii)*. If the asymptotic interpolation property holds, then there exist $T_0 > 0$, a function h as in Definition 1, and some control $u^\infty \in L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})$ such that the corresponding solution $\mathbf{x}^\infty$ to (2.2) set on $\mathbb{R}_+$ satisfies

$$\mathscr{E}(\mathbf{x}^\infty(t)) \leqslant h(t) \tag{3.22}$$

for all $t \geqslant T_0$. Combining this with the continuity of the map $t \longmapsto \mathscr{E}(\mathbf{x}^\infty(t))$, we can readily deduce that there exists a constant $C_0 > 0$ depending only on $T_0 > 0$ such that

$$\mathscr{E}(\mathbf{x}^\infty(t)) \leqslant C_0 \tag{3.23}$$

for all $t \geqslant 0$. Let us henceforth set

$$\lambda := \frac{M}{\left\|u^\infty\right\|_{L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})}}.$$

For any $T_1 > 0$, we also define

$$u_{T_1}(t) = \begin{cases} \lambda u^\infty(\lambda t) & \text{for } t \in (0,T_1] \\ 0 & \text{for } t > T_1. \end{cases}$$

Observe that, by definition of λ, one has $u_{T_1} \in \mathfrak{U}_{\mathrm{ad},T}$ for any $T > 0$. By virtue of Lemma 3.1, the state associated to u_{T_1} is precisely

$$\mathbf{x}_{T_1}(t) = \begin{cases} \mathbf{x}^\infty(\lambda t) & \text{for } \ t \in (0,T_1) \\ \mathbf{x}^\infty(\lambda T_1) & \text{for } t \geqslant T_1. \end{cases}$$

Now, by virtue of the definition of u_{T_1}, for any $T > 0$, we have

$$\begin{aligned} \mathscr{J}_T(u_{T_1}) &\leqslant \int_0^{T_1} \mathscr{E}\left(\mathbf{x}^\infty(\lambda t)\right)\,\mathrm{d}t + \max\Big\{0, T - T_1\Big\}\mathscr{E}\Big(\mathbf{x}^\infty(\lambda T_1)\Big) + M\,T_1 \\ &\leqslant (C_0 + M)\,T_1 + T\,\mathscr{E}\Big(\mathbf{x}^\infty(\lambda T_1)\Big). \end{aligned} \tag{3.24}$$

We now distinguish two cases. If $T \leqslant 1/h(T_0)$, then using (3.23), the optimality of u_T as well as the fact that $u_{T_1} \in \mathfrak{U}_{\mathrm{ad},T}$, along with $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$, and the definition (3.3) of T^*, through (3.24) we find

$$T\mathscr{E}(\mathbf{x}_T(T^*)) + MT^* \leqslant (C_0 + M)T_1 + \frac{C_0}{h(T_0)},$$

and choosing $T_1 = 1$ leads us to the conclusion. Now suppose that $T > 1/h(T_0)$. By Definition 1, the decreasing function h is a bijection from $(T_0, +\infty)$ onto its range $(0, h(T_0))$, and so $h^{-1}(1/T)$ is well defined precisely for $T > 1/h(T_0)$. We set

$$T_1 := \frac{1}{\lambda} h^{-1}\left(\frac{1}{T}\right).$$

Combining the optimality of u_T with (3.24), and using the fact that $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$, we find

$$\begin{aligned}\mathscr{J}_T(u_T) = M\,T^* + \int_0^T \mathscr{E}(\mathbf{x}_T(t))\,\mathrm{d}t &\leqslant \mathscr{J}_T(u_{T_1}) \\ &\leqslant \mathfrak{C}(M)\,h^{-1}\left(\frac{1}{T}\right) + T\,\mathscr{E}\left(\mathbf{x}^\infty\left(h^{-1}\left(\frac{1}{T}\right)\right)\right),\end{aligned} \tag{3.25}$$

where the constant

$$\mathfrak{C}(M) := \frac{(C_0 + M)}{\lambda}$$

is independent of T. Now since $h^{-1} : (0, h(T_0)) \to (0, +\infty)$ is non-decreasing, and $T > 1/h(T_0)$, we have that $h^{-1}(1/T) \geqslant T_0$. Using this fact, along with (3.22) in (3.25), combined with the definition (3.3) of T^*, allows us to deduce that

$$T\mathscr{E}(\mathbf{x}_T(T^*)) + MT^* \leqslant \mathfrak{C}(M)\,h^{-1}\left(\frac{1}{T}\right) + 1.$$

The desired statement *(ii)* then follows also for $T > 1/h(T_0)$. This concludes the proof. □

4. Concluding remarks

4.1. **Epilogue.** We have presented a manifestation of temporal sparsity and approximation/stability properties for supervised learning problems for neural ODEs with $L^1(0, T; \mathbb{R}^{d_u})$ penalties. Our main result ensures that any solution u_T to (2.5) is sparse in time, in the sense that $u_T \equiv 0$ on (T^*, T) for some $T^* \in (0, T]$. Under appropriate controllability assumptions, we also provide estimates on the stopping time T^*, and on the error $\mathscr{E}(\mathbf{x}_T(t))$ for $t \geqslant T^*$. The impact of this result, corroborated by numerical experiments, is (at least) twofold:

(i) When extrapolated to the discrete-time, ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as (safely) considering a shallower ResNet, namely a ResNet with less layers n_t, which could naturally lower the computational cost of the optimization process. In computing terms, the stabilization result further indicates that, perhaps, a model predictive control (MPC)-type strategy is warranted for an optimal choice of the stopping time (see [Grüne et al., 2019; Esteve et al., 2020a] for similar considerations).

(ii) Our result also applies for more classical optimal control problems, and provides a polynomial turnpike property for optimal states, and an exact turnpike property for optimal controls, minimizing a functional with a running cost for the state and an L^1 penalty for the control. This conclusion also comes along without any smallness conditions on the data, targets, or smoothness assumptions on the functional and/or dynamics (which, albeit, ought to be

homogeneous with respect to the control), which is new in the literature. We believe that the setting presented in this paper also invites the consideration of more unconventional functionals in control theory, e.g. minimizing divergences, which are more common in the literature on machine learning and inverse problems.

4.2. **Outlook.** We comment some questions that remain regarding our study.

1. The existence of minimizers for (2.5)– (2.4) remains unclear. It can be ensured if one replaces the L^1 penalty by a BV one, for which compactness of minimizing sequences holds. BV controls fit in the setting of temporal sparsity, unlike $W^{1,1}$ ones, which are continuous. The BV norm is also invariant with respect to the scaling of Lemma 3.1. But a complete extension of our arguments to this case would require further work.
2. It is curious that, when seen in the classical L^2 tracking context (i.e. the loss is the squared ℓ^2 distance) with an L^1 penalty for the controls, Theorem 2.1 only provides a polynomial turnpike estimate for the state. This is different to the L^2 penalty context, presented in [Esteve et al., 2020a,b], in which an exponential turnpike/stabilization estimate for the state is shown. There is reason to believe that for more specific loss functions, our stability results can be sharpened.
3. As a matter of fact, since $u_T(t) = 0$ for $t \geqslant T^*$, and our numerical experiments show that the state is stable in a regime in which the error $\mathscr{E}$ is 0, one could also stipulate that a result of the mould $\mathscr{E}(\mathbf{x}_T(t)) = 0$ for $t \geqslant T^*$ holds. Such an exact turnpike property for the state has been obtained in the linear setting in [Gugat et al., 2021]. However, the transfer of the techniques of the latter paper to our setting does not appear straightforward.

Acknowledgments. We thank Dario Pighin and Enrique Zuazua for insightful discussions.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No.765579-ConFlex and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement NO. 694126-DyCon).

References

Agrachev, A. and Sarychev, A. (2021). Control on the manifolds of mappings with a view to the deep learning. *Journal of Dynamical and Control Systems*, pages 1–20.

Alt, W. and Schneider, C. (2015). Linear-quadratic control problems with L^1-control cost. *Optimal Control Appl. Methods*, 36(4):512–534.

Bárcena-Petisco, J. A. (2021). Optimal control for neural ODE in a long time horizon and applications to the classification and simultaneous controllability problems.

Caponigro, M., Fornasier, M., Piccoli, B., and Trélat, E. (2013). Sparse stabilization and optimal control of the Cucker-Smale model. *Mathematical Control and Related Fields*, 3(4):447–466.

Caponigro, M., Fornasier, M., Piccoli, B., and Trélat, E. (2015). Sparse stabilization and control of alignment models. *Mathematical Models and Methods in Applied Sciences*, 25(03):521–564.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046.

Coron, J.-M. (2007). *Control and nonlinearity*. Number 136. American Mathematical Soc.

Cuchiero, C., Larsson, M., and Teichmann, J. (2020). Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM J. Math. Data Sci.*, 2(3):901–919.

Dupont, E., Doucet, A., and Teh, Y. W. (2019). Augmented Neural ODEs. In *Advances in Neural Information Processing Systems*, pages 3134–3144.

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

Effland, A., Kobler, E., Kunisch, K., and Pock, T. (2020). Variational networks: An optimal control approach to early stopping variational methods for image restoration. *J. Math. Imaging Vision*, pages 1–21.

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

Esteve, C., Geshkovski, B., Pighin, D., and Zuazua, E. (2020b). Turnpike in Lipschitz-nonlinear optimal control. *arXiv preprint arXiv:2011.11091*.

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

Fornasier, M., Piccoli, B., and Rossi, F. (2014). Mean-field sparse optimal control. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130400.

Geshkovski, B. (2021). Control in moving interfaces and deep learning.

Geshkovski, B. and Zuazua, E. (2021). Optimal actuator design via Brunovsky's normal form. *arXiv preprint arXiv:2108.05629*.

Grüne, L., Schaller, M., and Schiela, A. (2019). Sensitivity analysis of optimal control for a class of parabolic PDEs motivated by model predictive control. *SIAM J. Control Optim.*, 57(4):2753–2774.

Gugat, M., Schuster, M., and Zuazua, E. (2021). The finite-time turnpike phenomenon for optimal control problems: Stabilization by non-smooth tracking terms. In *Stabilization of Distributed Parameter Systems: Design Methods and Applications*, pages 17–41. Springer.

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

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

Kalise, D., Kunisch, K., and Rao, Z. (2017). Infinite horizon sparse optimal control. *J. Optim. Theory Appl.*, 172(2):481–517.

Kalise, D., Kunisch, K., and Rao, Z. (2020). Sparse and switching infinite horizon optimal controls with mixed-norm penalizations. *ESAIM Control Optim. Calc. Var.*, 26:61.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch.

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

Ruiz-Balet, D., Affili, E., and Zuazua, E. (2021). Interpolation and approximation via Momentum ResNets and Neural ODEs.

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

Santosa, F. and Symes, W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7:1307–1330.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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

Vossen, G. and Maurer, H. (2006). On L^1-minimization in optimal control and applications to robotics. *Optimal Control Applications and Methods*, 27(6):301–321.

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zuazua, E. (2010). Switching control. *J. Eur. Math. Soc.*, 13(1):85–117.

Carlos Esteve-Yagüe, Borjan Geshkovski
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, Spain
and
Chair of Computational Mathematics
Fundación Deusto
Av. de las Universidades, 24
48007 Bilbao, Basque Country, Spain
Email address: {carlos.esteve, borjan.geshkovski}@deusto.es