Article

Not peer-reviewed version

The Finite–Time Turnpike Property in Machine Learning

Martin Gugat *

Posted Date: 20 August 2024

doi: 10.20944/preprints202408.1283.v1

Keywords: neural ode; turnpike property; finite-time turnpike; non-smooth loss function; tracking term

 doi:10.20944/preprints202408.1283.v1

Article

The Finite–Time Turnpike Property in Machine Learning

Martin Gugat

Lehrstuhl für dynamics, control, numerics and machine learning, (Alexander von Humboldt-Professur), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department Mathematik, Cauerstr. 11, 91058 Erlangen, Germany; martin.gugat@fau.de

Abstract: The finite-time turnpike property describes the situation in an optimal control problem where an optimal trajectory reaches the desired state before the end of the time interval and remains there. We consider a machine learning problem with a neural ordinary differential equation that can be seen as a homogenization of a deep ResNet. We show that with appropriate scaling of the quadratic control cost and the non-smooth tracking term the optimal control problem has the finite-time turnpike property, that is the desired state is reached in the interior of the time interval and the optimal state remains there until the terminal time T. This property is useful to achieve a compromise between the depth of the network and the size of the optimal system parameters which we hope will be useful to determine optimal depths for neural network architectures in the future.

Keywords: ResNet; neural ODE; finite-time turnpike property; turnpike phenomenon; non-smooth tracking term

1. Introduction

We consider a system that is governed by a neural ODE that can be considered as a continuous-time ResNet. The system **S** is defined as follows:

$$\mathbf{S} \begin{cases} x(0) = x_0 \in \mathbb{R}^d, \\ x'(t) = \sum_{i=1}^{p} \sigma(a_i(t)^\top x(t) + b_i(t))\, w_i(t) \end{cases}$$

(see for example [1,2]). For $i \in \{1, ..., p\}$ we have $w_i(t) \in \mathbb{R}^d$. The $w_i(t)$ are the columns of the matrix $W(t) \in \mathbb{R}^{d\times p}$. We have $a_i(t) \in \mathbb{R}^d$ and the $a_i(t)$ are the columns of the matrix $A(t) \in \mathbb{R}^{d\times p}$. The bias vector $b(t)$ is in $\mathbb{R}^p$ and has the components $b_i(t)$.

The motivation to study **S** is that a time-discrete version can be considered as a residual neural networks (ResNets) that has been very used in many applications, see [3] for example in image registration and classification problems. A time-discrete version can be obtained for example by an explicit Euler discretization of **S**.

The activation function σ is assumed to be Lipschitz continuous with a Lipschitz constant that is less than or equal to 1 and differentiable, for example

$$\sigma(z) = \tanh(z),$$

$\sigma(z) = \frac{1}{1+\exp(-z)}$. It acts on vectors componentwise.

For a given time horizon $T > 0$, we study an optimal control problem on the time interval $[0, T]$, where the desired state at T is prescribed by the terminal condition $x(T) = x_T$, where $x_T \in \mathbb{R}^d$ denotes the given desired output of the system. Let $t_0 \in (0, T)$ be given. For the training of the system, we study the loss function with a tracking term

$$Q(W, A, b) = \int_{t_0}^{T} |x(t) - x_T| + |x'(t)|\, dt.$$

 doi:10.20944/preprints202408.1283.v1

with the non–smooth norm $|z| = \sum_{i=1}^{d} |z_i|$.

We define the control cost (regularization term)

$$R(W, A, b) = \int_0^T \tfrac{1}{2}\|W(t)\|^2 + \tfrac{1}{2}\|A(t)\|^2 + \tfrac{1}{2}\|b(t)\|^2\, dt.$$

Here $(\|W(t)\|)$ denotes the Frobenius norm of $W(t)$. We introduce the space

$$X(T) = \{\text{measurable functions } (W(t), A(t), b(t)) \text{ defined on } (0,\, T)$$
$$\text{such that } \textstyle\int_0^T \|W(t)\|^2 + \|A(t)\|^2 + \|b(t)\|^2\, dt < \infty\}.$$

Lemma 10 in [4] states that system **S** is exactly controllable, that is the terminal condition

$$x(t_0) = z \tag{1}$$

can be satisfied for all $t_0 > 0$. To be precise, for all $t_0 > 0$ there exists a constant $C_e > 0$ such that for all $z \in \mathbb{R}^d$ we can find a control u_{exact} such that for the state $\tilde{x}$ that is generated by **S** with the initial condition $\tilde{x}(0) = x_0$ we have $\tilde{x}(t_0) = z$ and

$$\|u_{exact}\|_{L^2(0,\, t_0)} \le C_e\, \|z - x_0\|. \tag{2}$$

Also the linearized system is exactly controllable in the sense that for all $t_0 > 0$ there exists a constant $C_e > 0$ such that for all $z \in \mathbb{R}^d$ we can find a control $\tilde{v}$ such that for the state $\tilde{x}$ that is generated by the linearized system that is stated below with the initial condition $\tilde{x}(0) = 0$ we have $\tilde{x}(t_0) = z$ and

$$\|\tilde{v}\|_{L^2(0,\, t_0)} \le C_e\, \|z\|. \tag{3}$$

The linearized system at a given $u = (W,\, A,\, b)$ for the variation δx of the state that is generated by a variation $\delta u = (\delta W,\, \delta A,\, \delta b)$ of the control is

$$\delta x'(t) = \sum_{i=1}^{p} \sigma(a_i(t)^\top x(t) + b_i(t))\, \delta w_i(t) + \sum_{i=1}^{p} \sigma'(a_i(t)^\top x(t) + b_i(t)) w_i(t)\, x(t)^\top \delta a_i(t)$$

$$+ \sum_{i=1}^{p} \sigma'(a_i(t)^\top x(t) + b_i(t))\, w_i(t)\, \delta b_i(t) + \sum_{i=1}^{p} \sigma'(a_i(t)^\top x(t) + b_i(t)) w_i(t)\, a_i(t)^\top \delta x(t)$$

with the initial condition $\delta x(0) = 0$.

A universal approximation theorem for the corresponding time-discrete case with recurrent neural networks can be found in the seminal paper [5] by Cybenko, see also [6], [7–9].

For a parameter $\gamma > 0$ define

$$J(W, A, b) = \gamma\, Q(W, A, b) + R(W, A, b). \tag{4}$$

We study the minimization (training) problem

$$\mathbf{P}(T,\, \gamma) : \min_{(W,A,b) \in X(T)} J(W, A, b)$$

Our main result is that the optimal control problem $\mathbf{P}(T,\, \gamma)$ has the finite-time turnpike property, that is the desired state is already reached in the interior of the time-interval $[0,\, T]$ and remains there until the end of the time interval. The finite-time turnpike property has been studied for example in [10], [11] and [12]. In the first two references, the finite time-turnpike property is achieved by the non-smoothness of the objective functional. In this paper, we use a similar approach adapted to the framework of neural ordinary differential equations.

 doi:10.20944/preprints202408.1283.v1

The finite-time turnpike property is an extremal case of the celebrated turnpike property that has originally been studied in economics. The turnpike analysis investigates how the solutions of dynamic optimal control problems with a time evolution are related to the solutions of the corresponding static problems where the time-derivatives are set to zero and the initial conditions are cancelled. It turns out that often for large time horizons on large parts of the time interval the solution of the dynamic problems is very close to the solution of the corresponding static problem. For an overview about the turnpike property, see [13], [14], [15], [16] and the numerous references therein.

In the case of the finite-time turnpike property, after finite time the solution of the dynamic problem coincides with the solution of the static problem. The exponential turnpike property for ResNets and beyond has been studied for example in [17], but not the finite-time turnpike property.

2. The Finite-Time Turnpike Property

The following Theorem contains our main result, which states that the control cost entails the finite-time turnpike property.

Theorem 1. *For each sufficiently large $\gamma > 0$ each optimal trajectory for $\boldsymbol{P}(T, \gamma)$ satisfies*

$$x(t) = x_T \, , t \in [t_0, T],$$

that is $\boldsymbol{P}(T, \gamma)$ has the finite-time turnpike property. For $t \geq t_0$ for the optimal parameters we have $W(t) = 0$, $A(t) = 0$ and $b(t) = 0$. The optimal parameters remain unchanged if γ is further enlarged or if T is further enlarged.

For the proof of Theorem 1 we need a result about the embedding of the continuous functions in the Sobolev space $W^{1,1}$: Let

$$L^1(0, T) = \{f : [0, T] \to \mathbb{R}, \, f \text{ is measurable, i.e. } \int_0^T |f(t)| \, dt < \infty\}.$$

Consider the embedding of the space of continuous functions in the space

$$W^{1,1}(0, T) = \{f \in L^1(0, T) : f' \in L^1(0, T)\}.$$

Lemma 1. *Let $t_0 \in [0, T)$. For all $x \in W^{1,1}(t_0, T)$ we have*

$$\max_{t \in [t_0, T]} |x(t)| \leq \left(\frac{1}{T - t_0} + 1 \right) \int_{t_0}^T |x(t)| + |x'(t)| \, dt. \tag{5}$$

Proof of Lemma 1. For $t_1, t_2 \in [t_0, T]$ we have

$$\begin{aligned} |x(t_1) - x(t_2)| &= \left| \int_{t_1}^{t_2} x'(t) \, dt \right| \\ &\leq \left| \int_{t_1}^{t_2} |x'(t)| \, dt \right|. \end{aligned}$$

Thus x is continuous on $[t_0, T]$. Hence there exists a point $t_* \in [t_0, T]$ with

$$|x(t_*)| = \min_{t \in [t_0, T]} |x(t)| \leq \frac{1}{T - t_0} \int_{t_0}^T |x(t)| \, dt.$$

 doi:10.20944/preprints202408.1283.v1

Thus for all $\tau \in [t_0, T]$ the following inequality holds:

$$\begin{aligned} |x(\tau)| &\leq |x(t_*)| + |x(t_*) - x(\tau)| \\ &\leq \frac{1}{T-t_0}\int_{t_0}^{T} |x(t)|\,dt + \int_{t_0}^{T} |x'(t)|\,dt \\ &\leq \left(\frac{1}{T-t_0}+1\right)\int_{t_0}^{T} |x(t)| + |x'(t)|\,dt. \end{aligned}$$

□

Now we are prepared for the proof of Theorem 1.

Proof of Theorem 1. *Case 1:* If $x_0 = x_T$, the parameters $u_* = (W_*, A_*, b_*) = (0, 0, 0)$ generate the constant state $x(t) = x_T$. Hence $u_* = 0$ solves $\mathbf{P}(T, \gamma)$ and the assertion follows.

Case 2: Now we assume that $x_0 \neq x_T$. For $u = (W, A, b) \in X(T)$ define the cost

$$C_{(0,t_0)}(u) = \int_0^{t_0} \tfrac{1}{2}\|W(t)\|^2 + \tfrac{1}{2}\|A(t)\|^2 + \tfrac{1}{2}\|b(t)\|^2\,dt.$$

Define the non-smooth tracking term

$$I_{non}(u) = \int_{t_0}^{T} |x(t) - x_T| + |x'(t)|\,dt.$$

Define the objective functional

$$K_T(u) = C_{(0,t_0)}(u) + \gamma\, I_{non}(u).$$

We consider the auxiliary problem

$$\mathbf{Q}(T): \quad \min_{u\in X(T)} K_T(u).$$

We show that for solution u_* of $\mathbf{Q}(T)$ we have

$$I_{non}(u_*) = 0$$

by an indirect proof. Suppose that there exists a solution $u_* = (W_*, A_*, b_*)$ of $\mathbf{Q}(T)$ such that $I_{non}(u_*) > 0$. Then for the corresponding optimal state x_* that is generated by $\mathbf{S}$ we have $x_*(t_0) \neq x_T$; otherwise we could switch off the control at t_0 and continue with the zero control $(0, 0, 0)$ for $t \in (t_0, T]$ that generates the constant state x_T on $(t_0, T]$ to strictly improve the performance.

Define the auxiliary state

$$\tilde{x}(t_0) = x_T + \frac{1}{I_{non}(u_*)}\left(x_*(t_0) - x_T\right).$$

The exact controllability of the linearized system implies that we can find a control $\tilde{v} \in L^2(0, t_0)$ that due to (3) satisfies the inequality

$$\|\tilde{v}\|_{L^2(0,t_0)} \leq C_e\,\|\tilde{x}(t_0) - x_T\| = C_e\,\frac{1}{I_{non}(u_*)}\|x_*(t_0) - x_T\|$$

that generates the state $\tilde{V}$ with $\tilde{V}(0, \cdot) = 0$ and $\tilde{V}(t_0) = \tilde{x}(t_0) - x_T$.

Due to (5) we have

$$\|x_*(t_0) - x_T\| \leq \left(\frac{1}{T-t_0}+1\right)\int_{t_0}^{T} |x_*(t) - x_T| + |x_*'(t)|\,dt \tag{6}$$

 doi:10.20944/preprints202408.1283.v1

$$= \left(\frac{1}{T-t_0}+1\right) I_{non}(u_*).$$

Thus we have

$$\|\tilde{v}\|_{L^2(0,t_0)} \leq C_e \left(\frac{1}{T-t_0}+1\right).$$

For a step-size $\varepsilon \in (0,\, I_{non}(u_*))$ define

$$\lambda = 1 - \frac{\varepsilon}{I_{non}(u_*)} \in (0,1).$$

Consider the control u with

$$u(t) = u_*(t) - \varepsilon\tilde{v}(t)$$

for $t \in (0, t_0]$ and for $t \in (t_0,\, T)$ we defined $\tilde{v} = (\delta W,\, \delta A,\, \delta b)$ with $\delta W(t) = -\frac{\varepsilon}{I_{non}(u_*)} u_*(t)$, $\delta A(t) = -\frac{\varepsilon}{I_{non}(u_*)} A_*(t)$, $\delta b(t) = -\frac{\varepsilon}{I_{non}(u_*)} b_*(t)$.

Then if $\gamma > 0$ is sufficiently large, $-\tilde{v}$ is a descent direction in the sense that by a little step in the direction $-\tilde{v}$ we can improve the performance of the control u_*. This can be seen as follows.

For the state $x = x_* + \delta x$ that is generated with the solution δx of the linearized system with the initial condition $\delta x(0) = 0$ we have at t_0

$$x(t_0) - x_T = (x_*(t_0) - x_T) - \varepsilon(\tilde{x}(t_0) - x_T) = \left(1 - \frac{\varepsilon}{I_{non}(u_*)}\right)(x_*(t_0) - x_T)$$

$$= \lambda\,(x_*(t_0) - x_T).$$

Hence on $[t_0, T]$ the state $x = x_* + \delta x$ that is generated with the solution δx of the linearized system with the initial condition $\delta x(t_0) = -\frac{\varepsilon}{I_{non}(u_*)}\,(x_*(t_0) - x_T)$ is

$$x = x_T + \lambda\,(x_*(t) - x_T).$$

Thus for the tracking term we have the bound

$$I_{non}(u) = \lambda\, I_{non}(u_*) + \mathcal{O}(\|\delta u\|^2) = \left(1 - \frac{\varepsilon}{I_{non}(u_*)}\right) I_{non}(u_*) + \mathcal{O}(\|\delta u\|^2).$$

For the control cost we have

$$C_{(0,t_0)}(u) = \langle u_* - \varepsilon\tilde{v},\, u_* - \varepsilon\tilde{v}\rangle_{L^2(0,t_0)} = C_{(0,t_0)}(u_*) - 2\,\varepsilon\,\langle u_*,\, \tilde{v}\rangle_{L^2(0,t_0)} + \varepsilon^2\, C_{(0,t_0)}(\tilde{v}).$$

Define

$$p(\varepsilon) = K_T(u_* - \varepsilon\tilde{v})$$

$$= C_{(0,t_0)}(u_*) - 2\varepsilon\,\langle u_*,\, \tilde{v}\rangle_{L^2(0,t_0)} + \varepsilon^2\, C_{(0,t_0)}(\tilde{v}) + \gamma\left(1 - \frac{\varepsilon}{I_{non}(u_*)}\right) I_{non}(u_*) + \mathcal{O}(\|\delta u\|^2).$$

Then we have

$$p'(\varepsilon) = -2\,\langle u_*,\, \tilde{v}\rangle_{L^2(0,t_0)} + 2\,\varepsilon\, C_{(0,t_0)}(\tilde{v}) - \gamma + \mathcal{O}(\varepsilon).$$

This yields

$$p'(0) = -2\,\langle u_*,\, \tilde{v}\rangle_{L^2(0,t_0)} - \gamma.$$

The exact controllability of **S** implies that there is a control $u_{exact} \in L^2(0,t_0)$ with (due to (2))

$$\|u_{exact}\|_{L^2(0,t_0)} \leq C_e\,\|\tilde{x}_0 - x_T\|$$

 doi:10.20944/preprints202408.1283.v1

that generates the state V_{exact} with $V_{exact}(0,\cdot) = x_0$ and $V_{exact}(t_0,\cdot) = x_T$. For $t > t_0$, let $u_{exact}(t) = 0$. Since u_{exact} is feasible for $\mathbf{Q}(T)$, this yields the inequality

$$C_{(t_0,T)}(u_*) \le K_T(u_*) \le K_T(u_{exact}) = \|u_{exact}\|^2_{L^2(0,t_0)} \le C_e^2 \|x_0 - x_T\|^2_{L^2(0,L)}.$$

Hence we have

$$\begin{aligned}\langle u_*, \tilde{v}\rangle_{L^2(0,t_0)} &\le C_e \|x_0 - x_T\|_{L^2(0,L)} \|\tilde{v}\|_{L^2(0,t_0)} \\ &\le \|x_0 - x_T\|_{L^2(0,L)} C_e^2 \left(\frac{1}{T-t_0} + 1\right).\end{aligned}$$

Thus if

$$\gamma > 2\|x_0 - x_T\|_{L^2(0,L)} C_e^2 \left(\frac{1}{T-t_0} + 1\right),$$

we have $p'(0) \le -\gamma + 2\|x_0 - x_T\|_{L^2(0,L)} C_e^2 \left(\frac{1}{T-t_0} + 1\right) < 0$. This implies that for $\varepsilon > 0$ sufficiently small we have

$$K_T(u_* - \varepsilon\tilde{v}) < K_T(u_*),$$

which is a contradiction to the optimality of u^*.

Hence for any optimal control of $\mathbf{Q}(T)$ we have $I_{non}(u_*) = 0$. With inequality (6) this implies that for the optimal state we have $x_*(t_0) = x_T$.

Now we come back to problem

$$\mathbf{P}(T,\gamma): \quad \min_u J(u)$$

with J defined in (4). Let $v_P(T)$ denote the optimal value of $\mathbf{P}(T,\gamma)$ and $v_Q(T)$ denote the optimal value of $\mathbf{Q}(T)$. Since $K_T(u) \le J(u)$, we have $v_Q(T) \le v_P(T)$.

Moreover, any optimal control u_* for $\mathbf{Q}(T)$ is feasible for $\mathbf{P}(T,\gamma)$. Since $x_*(t_0) = x_T$, we have $C_{(t_0,T)}(u_*) = 0$. Hence $v_P(T) \le J(u_*) = K_T(u_*) = v_Q(T)$, and thus $v_P(T) \le v_Q(T)$. Therefore we have

$$v_P(T) = v_Q(T).$$

This implies that parameters that are optimal for $\mathbf{P}(T)$ are also optimal for $\mathbf{Q}(T)$ and the assertion follows. Thus we have proved Theorem 1. □

3. Existence of Solutions of $\mathbf{P}(T,\gamma)$ for Fixed A

For the sake of completeness of the analysis, we also state an existence result. However we can only prove the existence of a solution for the problem where the matrix A is fixed an not an optimization parameter for $\mathbf{P}(T,\gamma)$. Thus for a given matrix-valued function $A(t)$, we consider the problem

$$\mathbf{P}(T,\gamma,A): \min_{(\cdot,A,\cdot)\in X(T)} J(\cdot,A,\cdot)$$

In order to show the existence of a solution of $\mathbf{P}(T,\gamma,A)$, we assume that there exists a number $M > 0$ such that for $t \in [0,T]$ almost everywhere we have $\max_{i\in\{1,\dots,p\}} \|(a_i)(t)\| \le M$. This is the case if the a_i are elements of the function space $L^\infty(0,T)$, for example if they are step functions over $(0,T)$.

Theorem 2. *Assume that* $\sup_x |\sigma(x)| \le 1$ *and the Lipschitz constant of* σ *is less than or equal to* 1. *Assume that* $A(t)$ *is given such that we have*

$$\operatorname*{ess\,sup}_{i\in\{1,\dots,p\}, s\in[0,T]} \|(a_i)(s)\| < \infty.$$

Then each $T > 0$ *and* $\gamma > 0$, *problem* $\mathbf{P}(T,\gamma,A)$ *has a solution* W, b *such that in* $(W,A,b) \in X(T)$.

 doi:10.20944/preprints202408.1283.v1

If $A(t) = 0$ for $t \geq t_0$, for sufficiently large γ each solution has the finite-time turnpike property stated in Theorem 1.

The proof of Theorem 2 uses Gronwall's Lemma (see for example [18]). For the convenience of the reader we state it here:

Lemma 2 (Gronwall's Lemma)**.** *Let $L > 0$, $U_0 \geq 0$, $\varepsilon \geq 0$ and an integrable function U on $[0,\ T]$ be given. Assume that for $t \in [0,\ T]$ almost everywhere the integral inequality*

$$0 \leq U(t) \leq U_0 + \int_0^t L\, U(\tau) + \varepsilon \, d\tau$$

hold. Then for $t \in [0,\ T]$ almost everywhere the function U satisfies the inequality

$$U(t) \leq U_0 \, \mathrm{e}^{Lt} + \varepsilon \, \tfrac{\mathrm{e}^{Lt}-1}{L}.$$

Now we present the proof of Theorem 2.

Proof of Theorem 2. Consider a minimizing sequence $(u_n)_{n=1}^{\infty}$ with $u_n = (W_n,\ A,\ b_n) \in X(T)$ for all $n \in \{1, 2, 3, ...\}$. Define the norm

$$\|u_n\|_{X(T)} = \sqrt{\int_0^T \|W_n(t)\|^2 + \|A(t)\|^2 + \|b_n(t)\|^2 \, dt}$$

and the corresponding inner product that gives a Hilbert space structure to $X(T)$. Due to the definition of J, there exists a number $M > 0$ such that for all $n \in \{1, 2, ...\}$ we have

$$\|u_n\|_{X(T)} \leq M, \tag{7}$$

that is the sequence is bounded in $X(T)$.

Hence there exists a weakly-converging subsequence with a limit

$$u_* = (W_*,\ A,\ b_*) \in X(T).$$

Let x_* denote the state generated by u_*. For the states x_n generated by the u_n as a solution of **S** due to the definition of the tracking term R we can assume by increasing M if necessary that we have

$$\sup_{s \in [0,\ T],\, n \in \{1,2,3,...\}} \|x_n(s)\| \leq M.$$

Due to Mazur's Lemma (see for example [19], [20]), there exists a subsequence of convex combinations that converges strongly. To be precise, there exist convex combinations

$$v_k = \sum_{m=k}^{N(k)} \lambda_m^{(k)} u_m, \ \text{ with } \lambda_m^{(k)} \geq 0,\ k \leq m \leq N(k) \text{ and } \sum_{m=k}^{N(k)} \lambda_m^{(k)} = 1$$

such that

$$\lim_{k \to \infty} \|v_k - u_*\|_X(T) = 0.$$

This implies

$$\lim_{k \to \infty} \int_0^T \|W_n(t) - W_*(t)\| + \|b_n(t) - b_*(t)\| \, dt = 0.$$

 doi:10.20944/preprints202408.1283.v1

Since σ is Lipschitz continuous with a Lipschitz constant that is less than or equal to 1, this implies for $i \in \{1, ..., p\}$

$$\left|\sigma(\sum_{m=k}^{N(k)} \lambda_m^{(k)}[(a_i)(t)^\top x_m(t) + (b_i)_m(t))] - \sigma((a_i)(t)^\top x_*(t) + (b_i)_*(t))\right|$$

$$\leq \left|\sum_{m=k}^{N(k)} \lambda_m^{(k)}[(a_i)(t)^\top x_m(t) + (b_i)_m(t)))] - ((a_i)(t)^\top x_*(t) + (b_i)_*(t))\right|. \tag{8}$$

Thus for $t \in [0,\ T]$ almost everywhere we have

$$\left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t)\right\|$$

$$\leq \sum_{i=1}^{p} \int_0^t \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s) - (w_i)_*(s)\right\| \left|\sigma((a_i)(s)^\top x_*(s) + (b_i)_*(s))\right|$$

$$+ \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s)\right\| \left|\sigma(\sum_{m=k}^{N(k)} \lambda_m^{(k)}(a_i)(s)^\top x_m(s) + (b_i)_m(s))) - \sigma((a_i)(s)^\top x_*(s) + (b_i)_*(s))\right| ds.$$

Then the fact that $\sup_x |\sigma(x)| \leq 1$, the Cauchy-Schwarz inequality , (7) and (8) yield

$$\left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t)\right\| \leq \sum_{i=1}^{p} \int_0^t \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s) - (w_i)_*(s)\right\| ds$$

$$+\sqrt{\int_0^t \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s)\right\|^2 ds} \sqrt{\int_0^t \left|\sigma(\sum_{m=k}^{N(k)} \lambda_m^{(k)}(a_i)(s)^\top x_m(s) + (b_i)_m(s))) - \sigma((a_i)(s)^\top x_*(s) + (b_i)_*(s))\right|^2 ds}$$

$$\leq \sum_{i=1}^{p} \int_0^t \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s) - (w_i)_*(s)\right\| ds$$

$$+M\sqrt{\int_0^t \left|(\sum_{m=k}^{N(k)} \lambda_m^{(k)}(a_i)(s)^\top x_m(s) + (b_i)_m(s)) - ((a_i)(s)^\top x_*(s) + (b_i)_*(s))\right|^2 ds}$$

$$\leq \sum_{i=1}^{p} \int_0^t \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(w_i)_m(s) - (w_i)_*(s)\right\| ds$$

$$+M\sqrt{\int_0^t \left|(a_i)(s)^\top \left[\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(s) - x_*(s)\right]\right|^2 ds} + M\sqrt{\int_0^t \left|\sum_{m=k}^{N(k)} \lambda_m^{(k)}(b_i)_m(s) - (b_i)_*(s)\right|^2 ds}.$$

Due to Mazur's Lemma, this yields the existence of a sequence $(\epsilon_k)_k$ with $\epsilon_k \geq 0$ and $\lim_{k\to\infty} \epsilon_k = 0$ such that

$$\left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t)\right\| \leq \varepsilon_k + \sum_{i=1}^{p} M\sqrt{\int_0^t \operatorname*{ess\,sup}_{s\in(0,T)} \|(a_i)(s)\|^2 \left\|\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t)\right\|^2 dt}.$$

 doi:10.20944/preprints202408.1283.v1

Thus by increasing the value of M if necessary, we obtain for $t \in [0,T]$ almost everywhere

$$\left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right\| \leq \varepsilon_k + \sum_{i=1}^{p} M \sqrt{\int_0^t M^2 \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(s) - x_*(s) \right\|^2 ds}.$$

Since $(|u|+|v|)^2 \leq 2|u|^2 + 2|v|^2$ and
$\left(\sum_{i=1}^{p} \sqrt{|z_i|} \right)^2 \leq p \sum_{i=1}^{p} |z_i|$ this yields the integral inequality

$$\left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right\|^2 \leq 2\,(\varepsilon_k)^2 + 2\,p\,M^4 \sum_{i=1}^{p} \int_0^t \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(s) - x_*(s) \right\|^2 ds.$$

Now Gronwall's Lemma yields for $t \in [0,T]$ almost everywhere

$$\left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right\| = \mathcal{O}(\varepsilon_k).$$

This yields

$$\lim_{k\to\infty} \max_{t\in[0\,T]} \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right\| = 0.$$

For the time derivatives we obtain again by increasing the value of M if necessary

$$\int_0^T \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m'(t) - x_*'(t) \right\| dt$$

$$\leq \sum_{i=1}^{p} \int_0^T \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (w_i)_m(t) - (w_i)_*(t) \right\| \left| \sigma((a_i)(t)^\top x_*(t) + (b_i)_*(t)) \right|$$

$$+ \left\| \textstyle\sum_{m=k}^{N(k)} \lambda_m^{(k)} (w_i)_m(t) \right\| \left| \sigma(\textstyle\sum_{m=k}^{N(k)} \lambda_m^{(k)} (a_i)(t)^\top x_m(t) + (b_i)_m(t))) - \sigma((a_i)(t)^\top x_*(t) + (b_i)_*(t)) \right| dt$$

$$\leq \sum_{i=1}^{p} \int_0^T \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (w_i)_m(t) - (w_i)_*(t) \right\| dt$$

$$+ \sqrt{\textstyle\int_0^T \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (w_i)_m(t) \right\|^2 dt} \sqrt{\textstyle\int_0^T \left| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (a_i)(t)^\top x_m(t) + (b_i)_m(t) - ((a_i)(t)^\top x_*(t) + (b_i)_*(t)) \right|^2 dt}$$

$$\leq \varepsilon_k + \sum_{i=1}^{p} M \sqrt{\int_0^T \left| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (a_i)(t)^\top x_m(t) - (a_i)(t)^\top x_*(t) \right|^2 dt}$$

$$+ M \sqrt{\int_0^T \left| \sum_{m=k}^{N(k)} \lambda_m^{(k)} (b_i)_m(t) - (b_i)_*(t) \right|^2 dt}$$

$$\leq \varepsilon_k (1+M) + M \sum_{i=1}^{p} \sqrt{\int_0^T \left| (a_i)_m(t)^\top \left[\sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right] \right|^2 dt}$$

$$\leq \varepsilon_k (1+M) + M \sum_{i=1}^{p} \sqrt{\int_0^T \|(a_i)(t)\|^2 \| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_* \|^2 dt}$$

 doi:10.20944/preprints202408.1283.v1

$$\leq \varepsilon_k(1+M)$$
$$+M\sum_{i=1}^{p} \operatorname{ess}\sup_{s\in[0,T]} \|(a_i)(s)\| \sqrt{\int_0^T \left\| \sum_{m=k}^{N(k)} \lambda_m^{(k)} x_m(t) - x_*(t) \right\|^2 dt}$$
$$\leq \varepsilon_k(1+M) + M\sum_{i=1}^{p} \operatorname{ess}\sup_{s\in[0,T]} \|(a_i)(s)\|\, \varepsilon_k = \mathcal{O}(\varepsilon_k).$$

Thus we have

$$\liminf_{k\to\infty} Q(v_k) \geq Q(u_*),\ \liminf_{k\to\infty} R(v_k) \geq R(u_*).$$

This yields

$$\liminf_{k\to\infty} J(u_k) \geq J(u_*).$$

Hence u_* is a solution of $\mathbf{P}(T,\gamma,A)$. This shows that solution of $\mathbf{P}(T,\gamma,A)$ exist.

The exact controllability properties that have been used for the construction in the proof of Theorem 1 still hold if the matrix A is fixed. Hence the assertion follows. □

4. Discussion

We have shown that with a suitable non-smooth loss function each solution of a learning problem has the finite-time turnpike property which means that it reaches the desired state exactly after finite time. Since the finite time t_0 can be considered as a problem parameter, this situation allows to choose t_0 in a convenient way. Thus t_0 arises as an additional design parameter in the design of optimal neural networks, that corresponds to the number of layers. Since for $t \in [t_0, T]$ the optimal parameters are zero, System **S** does not change the state on $[t_0, T]$ and thus the time horizon can be cut off at t_0.

Hence the problem to find the optimal number of layers in a neural network corresponds in the setting of neural ODEs to the problem of time-optimal control where the task is to find a minimal value of t_0 subject to the constraint that $x(t_0) = x_T$ and for the optimal parameters $u(t)$ the constraint $\|u(t)\|^2_{X(t_0)} \leq \rho$ is satisfied. Here the number ρ is prescribed as a problem parameter. Let $\omega(T,\gamma)$ denote the optimal value of $\mathbf{P}(T,\gamma)$. Then for optimal parameters $u(t)$ that solve $\mathbf{P}(T,\gamma)$ we have $\|u(t)\|^2_{X(t_0)} \leq \omega(T,\gamma)$. Since Theorem 1 implies that for the optimal state we have $x(t_0) = x_T$, we conclude that optimal parameters for $\mathbf{P}(T,\gamma)$ also solve the time-optimal control problem with parameter $\rho = \omega(T,\gamma)$ and the optimal time is t_0.

We have shown the existence of a solution of the nonlinear optimization problem for the case that one of the parameters, namely the matrix $A(t)$ is fixed. In order to show that a solution also exists with A as an additional optimization parameter, we expect that an additional regularization term in the objective functional (for example $\int_0^T \|A'(t)\|^2\, dt$) is necessary. This is a topic for future research. We expect that the finite-time turnpike property also holds in the case $t_0 = 0$. However, the proof that is presented here does not apply to this case so this is another topic for future research. As a possible application of our results we have in mind the numerical solution of shape inverse problems as described in [21].

Funding: This research was funded by DFG in the framework of the Collaborative Research Centre CRC/Transregio 154, Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks, project C03 and project C05, Projektnummer 239904186 and the Bundesministerium für Bildung und Forschung (BMBF) and the Croatian Ministry of Science and Education under DAAD grant 57654073 'Uncertain data in control of PDE systems'.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Marion, P. Generalization bounds for neural ordinary differential equations and deep residual networks. *Advances in Neural Information Processing Systems* **2024**, *36*.

 doi:10.20944/preprints202408.1283.v1

2. Dupont, E.; Doucet, A.; Teh, Y.W. Augmented Neural ODEs. Advances in Neural Information Processing Systems; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, Vol. 32.
3. Thorpe, M.; van Gennip, Y. Deep limits of residual neural networks. *Research in the Mathematical Sciences* **2023**, *10*, 6.
4. Álvarez López, A.; Slimane, A.H.; Zuazua, E. Interplay between depth and width for interpolation in neural ODEs, 2024, [arXiv:math.OC/2401.09902].
5. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* **1989**, *2*, 303–314.
6. Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta numerica* **1999**, *8*, 143–195.
7. Schäfer, A.M.; Zimmermann, H.G. Recurrent neural networks are universal approximators. Artificial Neural Networks–ICANN 2006: 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part I 16. Springer, 2006, pp. 632–640.
8. Schäfer, A.M.; Udluft, S.; Zimmermann, H.G. Learning long term dependencies with recurrent neural networks. Artificial Neural Networks–ICANN 2006: 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part I 16. Springer, 2006, pp. 71–80.
9. Schaefer, A.M.; Udluft, S.; Zimmermann, H.G. Learning long-term dependencies with recurrent neural networks. *Neurocomputing* **2008**, *71*, 2481–2488.
10. Gugat, M.; Schuster, M.; Zuazua, E. The finite-time turnpike phenomenon for optimal control problems: Stabilization by non-smooth tracking terms. Stabilization of distributed parameter systems: design methods and applications. Springer, 2021, pp. 17–41.
11. Gugat, M.; Schuster, M. Optimal Neumann control of the wave equation with L 1-control cost: the finite-time turnpike property. *Optimization* **2024**, pp. 1–28.
12. Gugat, M. Optimal boundary control of the wave equation: The finite-time turnpike phenomenon. *Mathematical Reports* **2022**.
13. Zaslavski, A.J. *Turnpike Phenomenon in Metric Spaces*; Vol. 201, Springer Nature, 2023.
14. Grüne, L.; Faulwasser, T. Turnpike properties in optimal control: An overview of discrete-time and continuous-time results. Handbook of Numerical Analysis; Trelat, E.; Zuazua, E., Eds., 2022. doi:10.1016/bs.hna.2021.12.011.
15. Grüne, L.; Guglielmi, R. Turnpike properties and strict dissipativity for discrete time linear quadratic optimal control problems. *SIAM J. Control Optim.* **2018**, *56*, 1282–1302. doi:10.1137/17M112350X.
16. Trélat, E.; Zuazua, E. The turnpike property in finite-dimensional nonlinear optimal control. *Journal of Differential Equations* **2015**, *258*, 81–114.
17. Geshkovski, B.; Zuazua, E. Turnpike in optimal control of PDEs, ResNets, and beyond. *Acta Numerica* **2022**, *31*, 135–263.
18. Gugat, M. *Optimal boundary control and boundary stabilization of hyperbolic systems*; Birkhäuser, 2015.
19. Ciarlet, P.G. *Mathematical elasticity: Three-dimensional elasticity*; SIAM, 2021.
20. Heuser, H.G. Functional analysis. Transl. by John Horvath. A Wiley-Interscience Publication. Chichester etc.: John Wiley & Sons, 1982.
21. Jackowska-Strumillo, L.; Sokolowski, J.; Żochowski, A.; Henrot, A. On numerical solution of shape inverse problems. *Computational Optimization and Applications* **2002**, *23*, 231–255.