Applied Mathematics & Optimization (2023) 88:84
https://doi.org/10.1007/s00245-023-10059-5

Strong Stationarity for Optimal Control Problems with Non-smooth Integral Equation Constraints: Application to a Continuous DNN

Harbir Antil[1] · **Livia Betz**[2] · **Daniel Wachsmuth**[2]

Accepted: 25 August 2023 / Published online: 26 September 2023

Abstract
Motivated by the residual type neural networks (ResNet), this paper studies optimal control problems constrained by a non-smooth integral equation associated to a fractional differential equation. Such non-smooth equations, for instance, arise in the continuous representation of fractional deep neural networks (DNNs). Here the underlying non-differentiable function is the ReLU or max function. The control enters in a nonlinear and multiplicative manner and we additionally impose control constraints. Because of the presence of the non-differentiable mapping, the application of standard adjoint calculus is excluded. We derive strong stationary conditions by relying on the limited differentiability properties of the non-smooth map. While traditional approaches smoothen the non-differentiable function, no such smoothness is retained in our final strong stationarity system. Thus, this work also closes a gap which currently exists in continuous neural networks with ReLU type activation function.

Keywords Non-smooth optimization · Optimal control of fractional ODEs · Integral equations · Strong stationarity · Caputo derivative · Deep neural networks

Mathematics Subject Classification 34A08 · 45D05 · 49J21

✉ Livia Betz
livia.betz@uni-wuerzburg.de

[1] The Center for Mathematics and Artificial Intelligence, Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA

[2] Institut für Mathematik, Universität Würzburg, 97074 Würzburg, Germany

1 Introduction

In this paper, we establish strong stationary optimality conditions for the following control constrained optimization problem

$$\left.\begin{aligned} \min_{(a,\ell)\in H^1(0,T;\mathbb{R}^{n\times n})\times H^1(0,T;\mathbb{R}^n)} \quad & J(y,a,\ell) \\ \text{s.t.} \quad & \partial^\gamma y(t) = f(a(t)y(t)+\ell(t)) \quad \text{a.e. in } (0,T), \\ & y(0) = y_0, \\ & \ell \in \mathcal{K}, \end{aligned}\right\} \tag{P}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a *non-smooth* non-linearity. This is assumed to be Lipschitz continuous and directionally differentiable only. An important example falling into this class is the ReLU (max) function used in the description of fractional DNNs. In this case, a and l, respectively indicate the weights and biases. The objective functional is given by

$$J(y,a,\ell) := g(y(T)) + \frac{1}{2}\|a\|^2_{H^1(0,T;\mathbb{R}^{n\times n})} + \frac{1}{2}\|\ell\|^2_{H^1(0,T;\mathbb{R}^n)},$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. The values $\gamma \in (0,1)$ and $y_0 \in \mathbb{R}^n$ are fixed and the set $\mathcal{K} \subset H^1(0,T;\mathbb{R}^n)$ is convex and closed. The symbol ∂^γ denotes the fractional time derivative, more details are provided in the forthcoming sections. Notice that the entire discussion in this paper also extends (and is new) for the case $\gamma = 1$, i.e., the standard time derivative. This has been substantiated with the help of several remarks throughout the paper. Recently, optimal control of fractional ODEs/PDEs have received a significant interest, we refer to the articles [1, 7] and the references therein. The most generic framework is considered in [5]. However, none of these articles deal with the non-smooth setting presented in this paper.

The essential feature of the problem under consideration is that the mapping f is not necessarily differentiable, i.e., its directional derivative is non-linear with respect to the direction. Thus, standard methods for the derivation of qualified optimality conditions are not applicable here. In view of our goal to establish strong stationarity, the main novelties in this paper arise from:

- the presence of the fractional time derivative;
- the fact that the controls appear in the argument of the non-smooth non-linearity f;
- the presence of control constraints (in this context, we are able to prove strong stationarity without resorting to unverifiable "constraint qualifications").

All these challenges appear in applications concerned with the control of neural networks. The non-smooth and nonlinear function f encompasses functions such as max or ReLU arising in deep neural networks (DNNs). The objective function J encompasses a generic class of functionals such as cross entropy and least squares. In fact, the optimal control problem (P) is motivated by residual neural networks [24, 33]

and fractional deep neural networks [3, 4, 6]. The control constraints can capture the bias ordering notion recently introduced in [2]. All existing approaches in the neural network setting assume differentiability of f in deriving the gradients via backpropagation. No such smoothness conditions are assumed in this paper.

Deriving necessary optimality conditions is a challenging issue even in finite dimensions, where a special attention is given to MPCCs (mathematical programs with complementarity constraints). In [34] a detailed overview of various optimality conditions of different strength was introduced, see also [27] for the infinite-dimensional case. The most rigorous stationarity concept is strong stationarity. Roughly speaking, the strong stationarity conditions involve an optimality system, which is equivalent to the purely primal conditions saying that the directional derivative of the reduced objective in feasible directions is nonnegative (which is referred to as B-stationarity).

While there are plenty of contributions in the field of optimal control of smooth problems, see e.g. [38] and the references therein, fewer papers are dealing with non-smooth problems. Most of these works resort to regularization or relaxation techniques to smooth the problem, see e.g. [8, 28] and the references therein. The optimality systems derived in this way are of intermediate strength and are not expected to be of strong stationary type, since one always loses information when passing to the limit in the regularization scheme. Thus, proving strong stationarity for optimal control of non-smooth problems requires direct approaches, which employ the limited differentiability properties of the control-to-state map. In this context, there are even less contributions. Based on the pioneering work [31] (strong stationarity for optimal control of elliptic VIs of obstacle type), most of them focus on elliptic VIs [12, 18, 26, 32, 39, 40]; see also [14] (parabolic VIs of the first kind) and the more recent contribution [15] (evolutionary VIs). Regarding strong stationarity for optimal control of non-smooth PDEs, the literature is rather scarce and the only papers known to the authors addressing this issue so far are [30] (parabolic PDE), [11, 16, 17] (elliptic PDEs) and [10] (coupled PDE system). We point out that, in contrast to our problem, all the above mentioned works feature controls which appears outside the non-smooth mapping. Moreover, none of these contributions deals with a fractional time derivative.

Let us give an overview of the structure and the main results in this paper. After introducing the notation, we present in Sect. 2 some fractional calculus results which are needed throughout the paper.

Section 3 focuses on the analysis of the state equation in (P). Here we address the existence and uniqueness of so-called mild solutions, i.e., solutions of the associated integral Volterra equation (Sect. 3.1). The properties of the respective control-to-state operator are investigated in Sect. 3.2. In particular, we are concerned with the *directional differentiability* of the solution mapping of the non-smooth integral equation associated to the fractional differential equation in (P). While optimal control of nonlinear (and smooth) integral equations attracted much attention, see, e.g., [13, 23, 41], to the best of our knowledge, the sensitivity analysis of non-smooth integral equations has not been yet investigated in the literature. In Sect. 3.3 we show that the previously found mild solution is in fact strong. That is, the unique solution to the state equation in (P) is absolutely continuous, and it thus possesses a so-called *Caputo-derivative*. We underline that, the only paper known to the authors which deals with optimal control and proves the existence of strong solutions in the framework of fractional differential

equations is [5]. In [5], the absolute continuity of the mild solution of a fractional in time PDE (state equation) is shown by imposing pointwise (time-dependent) bounds on the time derivative of the control which then carry over to the time derivative of the state. We point out that we do not need such bounds in our case. Moreover, the result in this subsection stands by its own and it adds to the key novelties of the present paper.

Section 4 focuses on the main contribution, namely the strong stationarity for the optimal control of (P). Via a classical smoothening technique, we first prove an auxiliary result (Lemma 4.1) which will serve as an essential tool in the context of establishing strong stationarity. Our main Theorem 4.7 is then shown by extending the "surjectivity" trick from [10, 30]. In this context, we resort to a verifiable "constraint qualification" (CQ), cf. Assumption 4.3 below. The CQ requires that one of the components of the optimal state is non-zero at all times. We underline that this assumption is satisfied by state systems describing neural networks with the max or ReLu function. In addition, there are many other settings where the CQ can be a priori checked, as pointed out in Remark 4.4 below. In a more general case, this CQ is the price to pay for imposing constraints on the control ℓ (and not on the control a), see Remark 4.12. As already emphasized in contributions where strong stationarity is investigated, CQs are to be expected when examining control constrained problems [11, 39] or, they may be required by the complex nature of the state system [10]. At the end of Sect. 4 we gather some important remarks regarding the main result. A fundamental aspect resulting from the findings in this paper is that, when it comes to strong stationarity, the presence of *more than one control* allows us to impose control constraints without having to resort to unverifiable CQs, see Remark 4.13.

In Sect. 5 we state the strong stationarity conditions associated to the control of a continuous deep neural network. Finally, we include in Appendix A the proof of Lemma 4.1, for convenience of the reader.

Notation Throughout the paper, $T > 0$ is a fixed final time and $n \in \mathbb{N}$ is a fixed dimension. By $\|\cdot\|$ we denote the Frobenius norm. If X and Y are linear normed spaces, $X \hookrightarrow\hookrightarrow Y$ means that X is compactly embedded in Y, while $X \overset{d}{\hookrightarrow} Y$ means that X is densely embedded in Y. The dual space of X will be denoted by X^*. For the dual pairing between X and X^* we write $\langle ., . \rangle_X$. If X is a Hilbert space, $(\cdot, \cdot)_X$ stands for the associated scalar product. The closed ball in X around $x \in X$ with radius $\alpha > 0$ is denoted by $B_X(x, \alpha)$. The conical hull of a set $\mathcal{M} \subset X$ is defined as $\text{cone}\, \mathcal{M} := \cap\{A \subset X : \mathcal{M} \subset A,\ A \text{ is a convex cone}\}$. With a little abuse of notation, the Nemytskii-operators associated with the mappings considered in this paper will be denoted by the same symbol, even when considered with different domains and ranges. We use sometimes the notation $h \lesssim g$ to denote $h \leq Cg$, for some constant $C > 0$, when the dependence of the constant C on some physical parameters is not relevant.

2 Preliminaries

In this section we gather some fractional calculus tools that are needed for our analysis.

Definition 2.1 (Left and right Riemann-Liouville fractional integrals) For $\phi \in L^1(0,T;\mathbb{R}^n)$, we define

$$I_{0+}^{\gamma}(\phi)(t) := \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)}\phi(s)\,\mathrm{d}s, \qquad I_{T-}^{\gamma}(\phi)(t) := \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)}\phi(s)\,\mathrm{d}s$$

for all $t \in [0,T]$. Here Γ is the Euler-Gamma function.

Definition 2.2 (The Caputo fractional derivative) Let $y \in W^{1,1}(0,T;\mathbb{R}^n)$. The (strong) Caputo fractional derivative of order $\gamma \in (0,1)$ is given by

$$\partial^{\gamma} y := I_{0+}^{1-\gamma} y'.$$

Lemma 2.3 *(Fractional integration by parts,* [29, Lemma 2.7a]*) If $\phi \in L^{\varrho}(0,T;\mathbb{R}^n)$ and $\psi \in L^{\zeta}(0,T;\mathbb{R}^n)$ with $\varrho, \zeta \in (1,\infty]$, $1/\varrho + 1/\zeta \le 1+\gamma$, then*

$$\int_0^T \psi^{\top} I_{0+}^{\gamma}(\phi)\,\mathrm{d}t = \int_0^T \phi^{\top} I_{T-}^{\gamma}(\psi)\,\mathrm{d}t.$$

Remark 2.4 Note that the identity in Lemma 2.3 implies that I_{T-}^{γ} is the adjoint operator of I_{0+}^{γ}.

Lemma 2.5 *(Boundedness of fractional integrals,* [29, Lemma 2.1a]*) The operators I_{0+}^{γ}, I_{T-}^{γ} map $L^{\infty}(0,T;\mathbb{R}^n)$ to $C([0,T];\mathbb{R}^n)$. Moreover, it holds*

$$\|I_{0+}^{\gamma}\|_{\mathcal{L}(L^{\varrho}(0,t;\mathbb{R}^n),L^{\varrho}(0,t;\mathbb{R}^n))} \le \frac{t^{\gamma}}{\gamma\Gamma(\gamma)} \quad \forall\, t \in [0,T] \tag{2.1}$$

for all $\varrho \in [1,\infty]$. The same estimate is true for I_{T-}^{γ}, cf. also Remark 2.4.

Lemma 2.6 *(Gronwall's inequality,* [20, Lemma 6.3]*) Let $\varphi \in C([0,T];\mathbb{R}^n)$ with $\varphi \ge 0$. If*

$$\varphi(t) \le c_1 t^{\alpha-1} + c_2 \int_0^t (t-s)^{\beta-1}\varphi(s)\,\mathrm{d}s \quad \forall\, t \in [0,T],$$

where $c_1, c_2 \ge 0$ are some constants and $\alpha, \beta > 0$, then there is a positive constant $C = C(\alpha,\beta,T,c_2)$ such that

$$\varphi(t) \le C c_1 t^{\alpha-1} \quad \forall\, t \in [0,T].$$

Finally, let us state a result that will be very useful throughout the entire paper.

Lemma 2.7 *Let $r \in [1, 1/(1-\gamma))$ be given for $\gamma \in (0,1)$. Then for each $t \in [0,T]$, we have*

$$\Big(\int_0^t (t-s)^{(\gamma-1)r}\,\mathrm{d}s\Big)^{1/r} \le \frac{t^{(\gamma-1)r+1}}{(\gamma-1)r+1} \le \frac{T^{(\gamma-1)r+1}}{(\gamma-1)r+1}.$$

Proof By assumption, we have $(\gamma - 1)r + 1 > 0$ and

$$\int_0^t (t-s)^{(\gamma-1)r} \, \mathrm{d}s = \frac{t^{(\gamma-1)r+1}}{(\gamma-1)r+1} \leq \frac{T^{(\gamma-1)r+1}}{(\gamma-1)r+1}$$

follows from elementary calculations. The proof is complete. □

3 The State Equation

In this section we address the properties of the solution operator of the state equation

$$\partial^\gamma y(t) = f(a(t)y(t) + \ell(t)) \quad \text{a.e. in } (0,T), \qquad y(0) = y_0. \tag{3.1}$$

Throughout the paper, $\gamma \in (0,1)$, unless otherwise specified, and $y_0 \in \mathbb{R}^n$ is fixed. For all $z \in \mathbb{R}^n$, the non-linearity $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies

$$f(z)_i = \widetilde{f}(z_i), \quad i = 1, ..., n,$$

where $\widetilde{f} : \mathbb{R} \to \mathbb{R}$ is a non-smooth nonlinear function. For convenience we will denote both non-smooth functions by f; from the context it will always be clear which one is meant.

Assumption 3.1 For the non-smooth mapping appearing in (P) we require:

1. The non-linearity $f : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous with constant $L > 0$, i.e.,

$$\|f(z_1) - f(z_2)\| \leq L\|z_1 - z_2\| \quad \forall\, z_1, z_2 \in \mathbb{R}^n.$$

2. The function f is directionally differentiable at every point, i.e.,

$$\lim_{\tau \searrow 0} \left\| \frac{f(z + \tau\,\delta z) - f(z)}{\tau} - f'(z;\delta z) \right\| = 0 \quad \forall\, z, \delta z \in \mathbb{R}^n.$$

As a consequence of Assumption 3.1 we have

$$\|f'(z;\delta z_1) - f'(z;\delta z_2)\| \leq L\|\delta z_1 - \delta z_2\| \quad \forall\, z, \delta z_1, \delta z_2 \in \mathbb{R}^n. \tag{3.2}$$

3.1 Mild Solutions

Definition 3.2 Let $(a, \ell) \in L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n)$ be given. We say that $y \in C([0,T];\mathbb{R}^n)$ is a mild solution of the state Eq. (3.1) if it satisfies the following integral equation

$$y(t) = y_0 + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f(a(s)y(s) + \ell(s)) \, \mathrm{d}s \quad \forall\, t \in [0,T]. \tag{3.3}$$

Remark 3.3 One sees immediately that if $\gamma = 1$ then $y \in W^{1,\infty}(0, T; \mathbb{R}^n)$ and the mild solution is in fact a strong solution of the following ODE

$$y'(t) = f(a(t)y(t) + \ell(t)) \quad \text{a.e. in } (0, T), \qquad y(0) = y_0.$$

Proposition 3.4 *For every $(a, \ell) \in L^\infty(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n)$ there exists a unique mild solution $y \in C([0, T]; \mathbb{R}^n)$ to the state equation (3.1).*

Proof To show the existence of a mild solution in the general case $\gamma \in (0, 1)$ we define the operator

$$F : C([0, t^*]; \mathbb{R}^n) \ni z \mapsto I^\gamma_{0+}(f(az + \ell)) \in C([0, t^*]; \mathbb{R}^n),$$

where t^* will be computed such that F is a contraction. Indeed, according to Lemma 2.5, F is well-defined (since f maps bounded sets to bounded sets). Moreover, by applying (2.1) with $\varrho = \infty$, and using Assumption 3.1, we see that

$$\begin{aligned}\|F(z_1) - F(z_2)\|_{C([0,t^*];\mathbb{R}^n)} &\le \frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)}\|f(az_1 + \ell) - f(az_2 + \ell)\|_{C([0,t^*];\mathbb{R}^n)} \\ &\le \frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\|z_1 - z_2\|_{C([0,t^*];\mathbb{R}^n)}\end{aligned}$$

for all $z_1, z_2 \in C([0, t^*]; \mathbb{R}^n)$. Thus, $F : C([0, t^*]; \mathbb{R}^n) \to C([0, t^*]; \mathbb{R}^n)$ is a contraction provided that $\frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} < 1$. If $\frac{T^\gamma}{\gamma\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} < 1$, the proof is complete. Otherwise we fix t^* as above and conclude that $z = F(z)$ admits a unique solution in $C([0, t^*]; \mathbb{R}^n)$, which for later purposes, is denoted by $\widetilde{y}$. To prove that this solution can be extended on the whole given interval $[0, T]$, we use a concatenation argument. We define

$$\begin{aligned}\widehat{F} : C([t^*, 2t^*]; \mathbb{R}^n) \ni z \mapsto\ & y_0 + \int_{t^*}^{\cdot} \frac{(\cdot - s)^{\gamma-1}}{\Gamma(\gamma)} f(a(s)z(s) + \ell(s))\,\mathrm{d}s \\ &+ \int_0^{t^*} \frac{(\cdot - s)^{\gamma-1}}{\Gamma(\gamma)} f(a(s)\widetilde{y}(s) + \ell(s))\,\mathrm{d}s \in C([t^*, 2t^*]; \mathbb{R}^n).\end{aligned}$$

Using a simple coordinate transform, we can apply again (2.1) with $\varrho = \infty$ on the interval $(t^*, 2t^*)$, and we have

$$\begin{aligned}&\|\widehat{F}(z_1) - \widehat{F}(z_2)\|_{C([t^*,2t^*];\mathbb{R}^n)} \\ &\le \sup_{t\in[t^*,2t^*]} \int_{t^*}^t \frac{(t - s)^{\gamma-1}}{\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\|(z_1 - z_2)(s)\|_{\mathbb{R}^n}\,\mathrm{d}s \\ &\le \frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\|z_1 - z_2\|_{C([t^*,2t^*];\mathbb{R}^n)}\end{aligned}$$

for all $z_1, z_2 \in C([t^*, 2t^*]; \mathbb{R}^n)$. Since t^* was fixed so that $\frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)} L\,\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} < 1$, we deduce that $z = \widehat{F}(z)$ admits a unique solution $\widehat{y}$ in $C([t^*, 2t^*]; \mathbb{R}^n)$. By con-

catenating the local solutions found on the intervals $[0, t^*]$ and $[t^*, 2t^*]$ one obtains a unique continuous function on $[0, 2t^*]$ which satisfies the integral equation (3.3). Proceeding further in the exact same way, one finds that (3.3) has a unique solution in $C([0, T]; \mathbb{R}^n)$. □

3.2 Control-to-State Operator

Next, we investigate the properties of the solution operator associated to (3.1)

$$S : L^\infty(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n) \ni (a, \ell) \mapsto y \in C([0, T]; \mathbb{R}^n).$$

Proposition 3.5 *(S is Locally Lipschitz) For every $M > 0$ there exists a constant $L_M > 0$ such that*

$$\begin{aligned}\|S(a_1, \ell_1) - S(a_2, \ell_2)\|_{C([0,T];\mathbb{R}^n)} &\le L_M(\|a_1 - a_2\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \\ &+\|\ell_1 - \ell_2\|_{L^\infty(0,T;\mathbb{R}^n)}),\end{aligned} \tag{3.4}$$

for all $(a_1, \ell_1), (a_2, \ell_2) \in B_{L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)}(0, M)$.

Proof First we show that S maps bounded sets to bounded sets. To this end, let $M > 0$ and $(a, \ell) \in L^\infty(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n)$ be arbitrary but fixed such that $\|(a, \ell)\|_{L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \le M$. From (3.3) we have

$$\begin{aligned}\|y(t)\| &\le \|y_0\| + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)}(\|f(0)\| + L(\|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\|y(s)\| + \|\ell\|_{L^\infty(0,T;\mathbb{R}^n)}))\,\mathrm{d}s \\ &\le c_1 + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L\,M\|y(s)\|\,\mathrm{d}s \quad \forall\, t \in [0, T],\end{aligned}$$

with

$$c_1 = \|y_0\| + \frac{T^\gamma}{\gamma}(\|f(0)\| + L\|\ell\|_{L^\infty(0,T;\mathbb{R}^n)}),$$

where we used Assumption 3.1 and Lemma 2.7 with $r = 1$. By means of Lemma 2.6 we deduce

$$\|y\|_{C([0,T];\mathbb{R}^n)} \le c(\|y_0\|, \|f(0)\|, L, M, \gamma, T) =: c_M. \tag{3.5}$$

Now, let $M > 0$ be further arbitrary but fixed. Define $y_k := S(a_k, \ell_k)$, $k = 1, 2$ and consider $\|(a_k, \ell_k)\|_{L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \le M$, $k = 1, 2$. Subtracting the integral

formulations associated to each k and using the Lipschitz continuity of f with constant L yields for $t \in [0, T]$

$$\begin{aligned}\|(y_1 - y_2)(t)\| &\le \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L \, \|a_1(s)y_1(s) - a_2(s)y_2(s) + \ell_1(s) - \ell_2(s)\| \, \mathrm{d}s \\ &\le \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L \, \|a_1\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \|y_1(s) - y_2(s)\| \, \mathrm{d}s \\ &\quad + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L \|y_2\|_{L^\infty(0,T;\mathbb{R}^n)} (\|a_1(s) - a_2(s)\| + \|\ell_1(s) - \ell_2(s)\|) \, \mathrm{d}s \\ &\le \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L \, M \, \|y_1(s) - y_2(s)\| \, \mathrm{d}s \\ &\quad + \frac{t^\gamma}{\gamma} L \, (c_M \|a_1 - a_2\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} + \|\ell_1 - \ell_2\|_{L^\infty(0,T;\mathbb{R}^n)}),\end{aligned}$$

where in the last inequality we used (3.5) and Lemma 2.7 with $r = 1$. Now, Lemma 2.6 implies that

$$\|y_1 - y_2\|_{C([0,T];\mathbb{R}^n)} \lesssim c_M (\|a_1 - a_2\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} + \|\ell_1 - \ell_2\|_{L^\infty(0,T;\mathbb{R}^n)}),$$

which completes the proof. □

Theorem 3.6 *(S is directionally differentiable) The control to state operator*

$$S : L^\infty(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n) \ni (a, \ell) \mapsto y \in C([0, T]; \mathbb{R}^n)$$

is directionally differentiable with directional derivative given by the unique solution $\delta y \in C([0, T]; \mathbb{R}^n)$ *of the following integral equation*

$$\begin{aligned}\delta y(t) = \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f'(a(s)y(s) + \ell(s); a(s)\delta y(s) + \delta a(s)y(s) \\ + \delta\ell(s)) \, \mathrm{d}s \quad \forall\, t \in [0, T],\end{aligned} \tag{3.6}$$

i.e., $\delta y = S'((a, \ell); (\delta a, \delta\ell))$ *for all* $(a, \ell), (\delta a, \delta\ell) \in L^\infty(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n)$.

Proof We first show that (3.6) is uniquely solvable. To this end, we argue as in the proof of Proposition 3.4. From Lemma 2.5 we know that the operator

$$F : C([0, t^*]; \mathbb{R}^n) \ni z \mapsto I_{0+}^\gamma (f'(ay + \ell; az + \delta ay + \delta\ell)) \in C([0, t^*]; \mathbb{R}^n),$$

is well defined since $f'(ay+\ell; az+\delta ay+\delta\ell) \in L^\infty(0, T; \mathbb{R}^n)$ for $z \in L^\infty(0, T; \mathbb{R}^n)$ [see (3.2)]. By employing the Lipschitz continuity of $f'(ay+\ell; \cdot)$ with constant L, one obtains the exact same estimate as in the proof of Proposition 3.4 and the remaining arguments stay the same.

Next we focus on proving that δy is the directional derivative of S at (a, ℓ) in direction $(\delta a, \delta\ell)$. For $\tau \in (0, 1]$ we define $y^\tau := S(a + \tau\delta a, \ell + \tau\delta\ell)$, $(a^\tau, \ell^\tau) :=$

$(a+\tau\delta a, \ell+\tau\delta\ell)$. From (3.3) we have

$$\begin{aligned}
&\Big(\frac{y^\tau - y}{\tau} - \delta y\Big)(t)\\
&= \frac{1}{\tau}\int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)}\big(f(a^\tau(s)y^\tau(s)+\ell^\tau(s)) - f(a(s)y(s)+\ell(s))\big)\,\mathrm{d}s\\
&\quad - \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f'\big(\underbrace{a(s)y(s)+\ell(s)}_{=:h(s)}; \underbrace{a(s)\delta y(s)+\delta a(s)y(s)+\delta\ell(s)}_{=:\delta h(s)}\big)\,\mathrm{d}s\\
&= \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)}\frac{1}{\tau}\big[f(a^\tau(s)y^\tau(s)+\ell^\tau(s)) - f((h+\tau\delta h)(s))\big]\,\mathrm{d}s\\
&\quad + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)}\Big(\underbrace{\frac{1}{\tau}[f((h+\tau\delta h)(s)) - f(h(s))] - f'(h(s);\delta h(s))}_{=:B_\tau(s)}\Big)\,\mathrm{d}s
\end{aligned}$$

for all $t\in[0,T]$. Since f is Lipschitz continuous with constant L we get

$$\begin{aligned}
\Big\|\Big(\frac{y^\tau - y}{\tau} - \delta y\Big)(t)\Big\| &\le \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} L\\
\Big(\frac{1}{\tau}\|a^\tau(s)y^\tau(s)+\ell^\tau(s) - h(s) - \tau\delta h(s)\|\Big)\,\mathrm{d}s &+ \|B_\tau\|_{L^q(0,t;\mathbb{R}^n)} \quad \forall\, t\in[0,T],
\end{aligned} \tag{3.7}$$

where $q = r' < \infty$, with r given by Lemma 2.7. Note that, in view of the directional differentiability of f combined with Lebesgue dominated convergence theorem it holds

$$B_\tau \to 0 \quad \text{in } L^q(0,T) \quad \text{as } \tau \searrow 0. \tag{3.8}$$

Now, let us take a closer look at the term

$$\begin{aligned}
&\frac{1}{\tau}\|a^\tau(s)y^\tau(s)+\ell^\tau(s) - h(s) - \tau\delta h(s)\|\\
&= \frac{1}{\tau}\big\|(a+\tau\delta a)(s)y^\tau(s) - a(s)y(s) - \tau(a\delta y + \delta a y)(s)\big\|\\
&= \Big\|a(s)\Big(\frac{y^\tau - y}{\tau} - \delta y\Big)(s) + \delta a(s)(y^\tau(s) - y(s))\Big\|\\
&\le \|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\Big\|\Big(\frac{y^\tau - y}{\tau} - \delta y\Big)(s)\Big\|\\
&\quad + \underbrace{\tau L_M \|\delta a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}(\|\delta a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} + \|\delta\ell\|_{L^\infty(0,T;\mathbb{R}^n)})}_{=:b_\tau} \quad \forall\, s\in[0,T],
\end{aligned}$$

where in the last inequality we used the Lipschitz continuity of S, cf. Proposition 3.5, with $M := \|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} + \|\delta a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} + \|\ell\|_{L^\infty(0,T;\mathbb{R}^n)} + \|\delta\ell\|_{L^\infty(0,T;\mathbb{R}^n)}$.

Going back to (3.7), we see that

$$\left\| \left(\frac{y^\tau - y}{\tau} - \delta y \right)(t) \right\| \leq \|B_\tau\|_{L^q(0,T;\mathbb{R}^n)} + L\, b_\tau + L\, \|a\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} \left\| \left(\frac{y^\tau - y}{\tau} - \delta y \right)(s) \right\| \mathrm{d}s \quad \forall\, t \in [0,T],$$

where we relied again on Lemma 2.7 with $r = 1$. In light of (3.8), Lemma 2.6 finally implies

$$\left\| \frac{y^\tau - y}{\tau} - \delta y \right\|_{C([0,T];\mathbb{R}^n)} \lesssim \|B_\tau\|_{L^q(0,T;\mathbb{R}^n)} + L\, b_\tau \to 0 \qquad \text{as } \tau \searrow 0.$$

The proof is now complete. □

Remark 3.7 Note that in the case $\gamma = 1$, one obtains by arguing exactly as in the proof of Theorem 3.6 that

$$S : L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n) \ni (a,\ell) \mapsto y \in W^{1,\infty}(0,T;\mathbb{R}^n)$$

is directionally differentiable with directional derivative given by the unique solution $\delta y \in W^{1,\infty}(0,T;\mathbb{R}^n)$ of the following ODE

$$\begin{aligned} &\delta y'(t) = f'(a(t)y(t) + \ell(t); a(t)\delta y(t) + \delta a(t) y(t) + \delta\ell(t)) \\ &\quad \text{f.a.a. } t \in (0,T), \quad \delta y(0) = 0. \end{aligned}$$

Proposition 3.8 (*Existence of optimal solutions for* (P)) *The optimal control problem* (P) *admits at least one solution in* $H^1(0,T;\mathbb{R}^{n\times n}) \times \mathcal{K}$.

Proof The assertion follows by standard arguments which rely on the direct method of the calculus of variations combined with the radial unboundedness of the reduced objective

$$H^1(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n) \ni (a,\ell) \mapsto J(S(a,\ell),a,\ell) \in \mathbb{R},$$

the compact embedding $H^1(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n) \hookrightarrow\hookrightarrow L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n)$, the continuity of

$$S : L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n) \to C([0,T];\mathbb{R}^n),$$

cf. Proposition 3.5, and of $g : \mathbb{R}^n \to \mathbb{R}$, and the weak lower semicontinuity of the norm. □

3.3 Strong Solutions

Next we prove that the state equation

$$\partial^\gamma y(t) = f(a(t)y(t) + \ell(t)) \quad \text{a.e. in } (0,T), \qquad y(0) = y_0, \tag{3.9}$$

admits in fact strong solutions, i.e., solutions that possess a so-called Caputo derivative, see Definition 2.2.

Definition 3.9 We say that $y \in W^{1,1}(0,T;\mathbb{R}^n)$ is a strong solution to (3.9) if

$$I_{0+}^{1-\gamma} y' = f(ay + \ell) \quad \text{a.e. in } (0,T), \qquad y(0) = y_0.$$

The following well known result is a consequence of the identity $I_{0+}^{\gamma} I_{0+}^{1-\gamma} y' = I_{0+}^{1} y' = y$, which is implied by the semigroup property of the fractional integrals cf. e.g. [29, Lemma 2.3] and Definition 2.1.

Lemma 3.10 *A function $y \in W^{1,1}(0,T;\mathbb{R}^n)$ is a strong solution of (3.9) if and only if it satisfies the integral formulation (3.3).*

Theorem 3.11 *Let $\gamma \in (0,1)$ and $r \in [1, \frac{1}{1-\gamma})$ be given. For each $(a,\ell) \in W^{1,\varrho}(0,T;\mathbb{R}^{n\times n}) \times W^{1,\varrho}(0,T;\mathbb{R}^n)$, $\varrho > 1$, (3.9) admits a unique strong solution $y \in W^{1,\zeta}(0,T;\mathbb{R}^n)$, where $\zeta = \min\{r,\varrho\} > 1$.*

Proof Let $t \in [0,T]$ and $h \in (0,1]$ be arbitrary but fixed. Note that the existence of a unique solution $y \in C([0,T+1];\mathbb{R}^n)$ is guaranteed by Proposition 3.4; this solution coincides with the mild solution of (3.1) on the interval $[0,T]$. From (3.3) we have

$$\begin{aligned} y(t+h) - y(t) = &\int_0^{t+h} \frac{s^{\gamma-1}}{\Gamma(\gamma)} f(ay+\ell)(t+h-s)\,\mathrm{d}s \\ &- \int_0^{t} \frac{s^{\gamma-1}}{\Gamma(\gamma)} f(ay+\ell)(t-s)\,\mathrm{d}s, \end{aligned}$$

which implies

$$\begin{aligned} \|y(t+h) - y(t)\| &\le \int_t^{t+h} \frac{s^{\gamma-1}}{\Gamma(\gamma)} f(ay+\ell)(t+h-s)\,\mathrm{d}s \\ &\quad + \int_0^{t} \frac{s^{\gamma-1}}{\Gamma(\gamma)} L\,\|(ay+\ell)(t+h-s) - (ay+\ell)(t-s)\|\,\mathrm{d}s \\ &:= z_1(t,h) + z_2(t,h). \end{aligned}$$

For z_1 and z_2 we find the following estimates

$$\begin{aligned}
\frac{z_1(t,h)}{h} &\lesssim \frac{1}{h}\int_t^{t+h} s^{\gamma-1}\,\mathrm{d}s \le t^{\gamma-1},\\
z_2(t,h) &\lesssim \int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|(a(t+h-s)-a(t-s))y(t+h-s)\|\,\mathrm{d}s\\
&\quad+\int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|a(t-s)(y(t+h-s)-y(t-s))\|\,\mathrm{d}s\\
&\quad+\int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|\ell(t+h-s)-\ell(t-s)\|\,\mathrm{d}s\\
&\lesssim \int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|(a(t+h-s)-a(t-s))\|\,\mathrm{d}s\\
&\quad+\int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|y(t+h-s)-y(t-s)\|\,\mathrm{d}s\\
&\quad+\int_0^t \frac{s^{\gamma-1}}{\Gamma(\gamma)}\|\ell(t+h-s)-\ell(t-s)\|\,\mathrm{d}s,
\end{aligned}$$

where we relied on the fact that $y, a,$ and ℓ are essentially bounded. Altogether we have

$$\begin{aligned}
\left\|\frac{y(t+h)-y(t)}{h}\right\| &\lesssim t^{\gamma-1}+\int_0^t (t-s)^{\gamma-1}\left\|\frac{\ell(s+h)-\ell(s)}{h}\right\|\mathrm{d}s\\
&\quad+\int_0^t (t-s)^{\gamma-1}\left\|\frac{a(s+h)-a(s)}{h}\right\|\mathrm{d}s\\
&\quad+\int_0^t (t-s)^{\gamma-1}\left\|\frac{y(s+h)-y(s)}{h}\right\|\mathrm{d}s
\end{aligned}\tag{3.10}$$

for all $t \in [0,T]$ and all $h \in (0,1]$. Let us define

$$B_h(t) := t^{\gamma-1}+\int_0^t (t-s)^{\gamma-1}\left\|\frac{\ell(s+h)-\ell(s)}{h}\right\|\mathrm{d}s+\int_0^t (t-s)^{\gamma-1}\left\|\frac{a(s+h)-a(s)}{h}\right\|\mathrm{d}s.$$

Since $(a,\ell) \in W^{1,\varrho}(0,T;\mathbb{R}^{n\times n})\times W^{1,\varrho}(0,T;\mathbb{R}^n)$, and $\zeta = \min\{r,\varrho\}$, where r is given by Lemma 2.7, we can estimate B_h as follows

$$\begin{aligned}
\|B_h\|_{L^\zeta(0,T;\mathbb{R})} &\le \frac{T^{(\gamma-1)r+1}}{(\gamma-1)r+1}+\frac{T^\gamma}{\gamma\Gamma(\gamma)}\left\|\frac{\ell(\cdot+h)-\ell(\cdot)}{h}\right\|_{L^\zeta(0,T;\mathbb{R}^n)}\\
&\quad+\frac{T^\gamma}{\gamma\Gamma(\gamma)}\left\|\frac{a(\cdot+h)-a(\cdot)}{h}\right\|_{L^\zeta(0,T;\mathbb{R}^{n\times n})}\\
&\le \frac{T^{(\gamma-1)r+1}}{(\gamma-1)r+1}+\frac{T^{\gamma+\zeta^{-1}-\varrho^{-1}}}{\gamma\Gamma(\gamma)}(\|\ell'\|_{L^\varrho(0,T;\mathbb{R}^n)}+\|a'\|_{L^\varrho(0,T;\mathbb{R}^{n\times n})}),
\end{aligned}\tag{3.11}$$

where we relied on Lemmas 2.7, 2.5 and [21, Theorem 3, p. 277]. Hence, $\{B_h\}$ is uniformly bounded in $L^\zeta(0,T;\mathbb{R})$ with respect to $h \in (0,1]$. Further, the generalized Gronwall inequality of [25, Lemma 7.1.1], see also [42, Corollary 1], applied to (3.10) yields

$$\left\| \frac{y(t+h)-y(t)}{h} \right\| \lesssim B_h(t) + \int_0^t \Big[\sum_{n=0}^{\infty} \frac{\Gamma(\gamma)^n}{\Gamma(n\gamma)} (t-s)^{n\gamma-1} B_h(s) \Big] \mathrm{d}s$$

for all $t \in [0,T]$ and all $h \in (0,1]$. Using monotone convergence theorem, we can exchange the order of integration and summation to get

$$\left\| \frac{y(t+h)-y(t)}{h} \right\| \lesssim B_h(t) + \sum_{n=0}^{\infty} \Gamma(\gamma)^n (I_{0+}^{n\gamma} B_h)(t),$$

where we used the definition of $I_{0+}^{n\gamma}$ from Definition 2.1. Applying Lemma 2.5 we obtain

$$\begin{aligned} \left\| \frac{y(\cdot+h)-y(\cdot)}{h} \right\|_{L^\zeta(0,T;\mathbb{R}^n)} &\lesssim \|B_h\|_{L^\zeta(0,T;\mathbb{R})} + \sum_{n=0}^{\infty} \Gamma(\gamma)^n \|I_{0+}^{n\gamma} B_h\|_{L^\zeta(0,T;\mathbb{R})} \\ &\leq \|B_h\|_{L^\zeta(0,T;\mathbb{R})} + \sum_{n=0}^{\infty} \Gamma(\gamma)^n \frac{T^{n\gamma}}{n\gamma\Gamma(n\gamma)} \|B_h\|_{L^\zeta(0,T;\mathbb{R})} \\ &= \|B_h\|_{L^\zeta(0,T;\mathbb{R})} [1 + E_{\gamma,1}(\Gamma(\gamma)T^\gamma)] \quad \forall\, h \in (0,1], \end{aligned}$$

where $E_{\gamma,1}(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(n\gamma+1)} < \infty$ is the celebrated Mittag-Leffler function; note that here we used $n\gamma\Gamma(n\gamma) = \Gamma(n\gamma+1)$. Since $\{B_h\}$ is uniformly bounded in $L^\zeta(0,T;\mathbb{R})$, see (3.11), we obtain that the difference quotients of y are uniformly bounded in $L^\zeta(0,T;\mathbb{R})$ with respect to $h \in (0,1]$. Hence, y has a weak derivative in $L^\zeta(0,T;\mathbb{R})$ by [21, Theorem 3, p. 277]. The proof is now complete. □

Remark 3.12 We remark that the degree of smoothness of the right-hand sides a, ℓ does not necessarily carry over to the strong solution y (unless a certain compatibility condition is satisfied, see Remark 3.13 below). This is in accordance with observations made in literature, see e.g., [19, Example 6.4, Remark 6.13, Theorem 6.27] (fractional ODEs) and [36, Corollary 2.2] (fractional in time PDEs). Indeed, for large values for ϱ and small values of γ tending to 0, the strong solution $y \in W^{1,\zeta}(0,T;\mathbb{R}^n)$, where $\zeta = r \in (1, 1/(1-\gamma))$ is close to 1. However, as γ approaches the value 1, one can expect the strong solutions to become as regular as their right-hand sides. This can be seen in the case $\gamma = 1$, where the smoothness of the strong solution improves as the smoothness of a, ℓ does so. Note that in this particular situation the solution of (3.9) is in fact far more regular than as in the statement in Theorem 3.11, see Remark 3.3.

Remark 3.13 (Compatibility condition) If $f(a(0)y_0 + \ell(0)) = 0$, then the regularity of the strong solution to (3.9) can be improved by looking at the equation satisfied by the weak derivative y' and inspecting its smoothness. Since the focus of this paper lies

on the optimal control and not on the analysis of fractional equations, we do not give a proof here. We just remark that the requirement $f(a(0)y_0 + \ell(0)) = 0$ corresponds to the one in e.g. [19, Theorem 6.26], cf., also [36, Corollary 2.2], where it is proven that the smoothness of the derivative of the strong solution improves if and only if such a compatibility condition is true.

4 Strong Stationarity

The first result in this section will be an essential tool for establishing the strong stationarity in Theorem 4.7 below, as it guarantees the existence of a multiplier satisfying both a gradient equation and an inequality associated to a local minimizer of (P).

Lemma 4.1 *Let $(\bar{a}, \bar{\ell})$ be a given local optimum of* (P)*. Then there exists a multiplier $\lambda \in L^r(0, T; \mathbb{R}^n)$ with r as in Lemma 2.7 such that*

$$(\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda, \delta a\, \bar{y}\rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall\, \delta a \in H^1(0, T; \mathbb{R}^{n\times n}), \tag{4.1a}$$

$$(\bar{\ell}, \delta\ell)_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda, \delta\ell\rangle_{L^r(0,T;\mathbb{R}^n)} \geq 0 \quad \forall\, \delta\ell \in \operatorname{cone}(\mathcal{K} - \bar{\ell}), \tag{4.1b}$$

where we abbreviate $\bar{y} := S(\bar{a}, \bar{\ell})$.

Proof The technical proof can be found in Appendix A.. □

The next step towards the derivation of our strong stationary system is to write the first order necessary optimality conditions in primal form.

Lemma 4.2 *(B-stationarity) If $(\bar{a}, \bar{\ell})$ is locally optimal for* (P)*, then there holds*

$$\begin{aligned} &\nabla g(\bar{y}(T))^\top\, S'\left((\bar{a}, \bar{\ell}); (\delta a, \delta\ell)\right)(T) \\ &\quad + (\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + (\bar{\ell}, \delta\ell)_{H^1(0,T;\mathbb{R}^n)} \geq 0 \\ &\quad \forall\, (\delta a, \delta\ell) \in H^1(0, T; \mathbb{R}^{n\times n}) \times \operatorname{cone}(\mathcal{K} - \bar{\ell}), \end{aligned} \tag{4.2}$$

where we abbreviate $\bar{y} := S(\bar{a}, \bar{\ell})$.

Proof The result follows from the continuous differentiability of g combined with the directional differentiability of S, see Theorem 3.6, and the local optimality of $(\bar{a}, \bar{\ell})$. □

Assumption 4.3 ('Constraint Qualification') There exists some index $m \in \{1, ..., n\}$ such that the optimal state satisfies $\bar{y}_m(t) \neq 0$ for all $t \in [0, T]$.

Remark 4.4 Let us underline that there is a zoo of situations where the requirement in Assumption 4.3 is fulfilled. We just enumerate a few in what follows.

- If there exists some index $m \in \{1, ..., n\}$ such that $y_{0,m} > 0$ and $f(z) \geq 0 \quad \forall\, z \in \mathbb{R}$ then the optimal state satisfies $\bar{y}_m(t) \geq y_{0,m} > 0$ for all $t \in [0, T]$, in view of (3.3).

In particular, our 'constraint qualification' is fulfilled by continuous fractional deep neural networks (DNNs) with ReLU activation function, since $f = \max\{0, \cdot\}$ in this case, while an additional initial datum can be chosen so that $y_{0,m} > 0$.

- Similarly, if there exists some index $m \in \{1, ..., n\}$ such that $y_{0,m} < 0$ and $f(z) \leq 0 \quad \forall z \in \mathbb{R}$, then the optimal state satisfies $\bar{y}_m(t) \leq y_{0,m} < 0$ for all $t \in [0, T]$. In both situations, the CQ in Assumption 4.3 is satisfied.
- If there exists some index $m \in \{1, ..., n\}$ such that $y_{0,m} \neq 0$ and $f(\bar{\ell}_m(t)) = 0$ for all $t \in [0, T]$, then, according to [19, Theorem 6.14] the optimal state satisfies $\bar{y}_m(t) \neq 0$ for all $t \in [0, T]$. This is the case if e.g. $f = \max\{0, \cdot\}$ and $\mathcal{K} \subset \{v \in H^1(0, T; \mathbb{R}^n) : v_m(t) \leq 0 \, \forall t \in [0, T]\}$.

Remark 4.5 We point out that Assumption 4.3 is due to the structure of the state equation and due to the fact that constraints are imposed on the control ℓ (and not on the control a), see Remark 4.12 below for more details. This assumption is essential for using the purely primal optimality condition of Lemma 4.2 to derive a formulation involving adjoint (or dual) quantities in (4.6) below. In this sense, Assumption 4.3 plays a role similar to constraint qualifications in nonlinear differentiable programming, which are used to prove existence of Lagrange multipliers such that the Karush-Kuhn-Tucker system (KKT) is satisfied, see e.g. [22, Sect. 2]. In Remark 4.9 below we will state the KKT conditions associated to our control problem in the case that f is differentiable.

The following result describes the density of the set of arguments into which the non-smoothness is derived in the "linearized" state equation (3.6). This aspect is crucial in the context of proving strong stationarity for the control of non-smooth equations, cf. [10, 30] and Remark 4.12 below.

Lemma 4.6 *(Density of the set of arguments of $f'((\bar{a}\bar{y}+\bar{\ell})_i(t); \cdot)$) Let $(\bar{a}, \bar{\ell})$ be a given local optimum for (P) with associated state $\bar{y} := S(\bar{a}, \bar{\ell})$. Under Assumption 4.3, it holds*

$$\{\bar{a}S'((\bar{a}, \bar{\ell}); (\delta a, 0)) + \delta a \bar{y} : \delta a \in H^1(0, T; \mathbb{R}^{n\times n})\} \overset{d}{\hookrightarrow} C([0, T]; \mathbb{R}^n).$$

Proof Let $\rho \in C([0, T]; \mathbb{R}^n)$ be arbitrary, but fixed and define the function

$$\widehat{\delta y}(t) := \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f'(\bar{a}(s)\bar{y}(s) + \bar{\ell}(s); \rho(s)) \, \mathrm{d}s \quad \forall t \in [0, T].$$

Note that $\widehat{\delta y} \in C([0, T]; \mathbb{R}^n)$, in view of Lemma 2.5. We will now construct $\widehat{\delta a}$ such that

$$\bar{a}\widehat{\delta y} + \widehat{\delta a}\bar{y} = \rho. \tag{4.3}$$

This is possible due to Assumption 4.3. Indeed, for $j = 1, ..., n$ and $t \in [0, T]$ we can define

$$\widehat{\delta a}_{jm}(t) := \frac{(\rho(t) - \bar{a}(t)\widehat{\delta y}(t))_j}{\bar{y}_m(t)}, \qquad \widehat{\delta a}_{ji}(t) := 0, \quad \text{for } i \neq m.$$

Note that $\widehat{\delta a}_{jm} \in C[0,T]$. Due to (4.3), $\widehat{\delta y}$ satisfies the integral equation

$$\widehat{\delta y}(t) = \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f'(\bar a(s)\bar y(s) + \bar\ell(s); \bar a(s)\widehat{\delta y}(s) + \widehat{\delta a}(s)\bar y(s))\,\mathrm{d}s \quad \forall\, t \in [0,T].$$

By Theorem 3.6, the integral equation is equivalent to

$$\widehat{\delta y} = S'((\bar a, \bar\ell); (\widehat{\delta a}, 0)). \tag{4.4}$$

Now, let us consider a sequence $\delta a_k \in H^1(0,T;\mathbb{R}^{n\times n})$ with

$$\delta a_k \to \widehat{\delta a} \quad \text{in } C([0,T];\mathbb{R}^{n\times n}). \tag{4.5}$$

In view of Proposition 3.5 and Theorem 3.6, the mapping S is locally Lipschitz continuous and directionally differentiable from $L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)$ to $C([0,T];\mathbb{R}^n)$. Hence, the mapping $S'((\bar a,\bar\ell);\cdot) : L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n) \to C([0,T];\mathbb{R}^n)$ is continuous, see, e.g., [35, Lemmas 3.1.2 and 3.1.3]. Thus, the convergence (4.5) implies that

$$\delta y_k := S'((\bar a,\bar\ell);(\delta a_k, 0)) \to \widehat{\delta y} \quad \text{in } C([0,T];\mathbb{R}^n),$$

where we recall (4.4). This gives in turn

$$\bar a \delta y_k + \delta a_k \bar y \to \rho \quad \text{in } C([0,T];\mathbb{R}^n),$$

in view of (4.3). Since $\rho \in C([0,T];\mathbb{R}^n)$ was arbitrary, the proof is now complete. □

The main finding of this paper is stated in the following result.

Theorem 4.7 *(Strong stationarity) Let $r \in [1, \frac{1}{1-\gamma})$. Suppose that Assumption 4.3 is satisfied and let $(\bar a, \bar\ell)$ be a given local optimum for (P) with associated state*

$$\bar y := S(\bar a, \bar\ell) \in W^{1,\zeta}(0,T;\mathbb{R}^n),$$

where $\zeta = \min\{r, 2\}$. Then there exists a multiplier $\lambda \in L^r(0,T;\mathbb{R}^n)$ and an adjoint state $p \in L^r(0,T;\mathbb{R}^n)$, such that

$$p(t) = \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)}\nabla g(\bar y(T)) + \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)} \bar a^\top(s)\lambda(s)\,\mathrm{d}s \quad \forall\, t \in [0,T), \tag{4.6a}$$

$$\lambda_i(t) \in [p_i(t) f'_-((\bar a\bar y + \bar\ell)_i(t)),\ p_i(t) f'_+((\bar a\bar y + \bar\ell)_i(t))] \quad \text{for a.a. } t \in (0,T),\ i = 1,\dots,n, \tag{4.6b}$$

$$(\bar a, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda, \delta a\,\bar y\rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall\, \delta a \in H^1(0,T;\mathbb{R}^{n\times n}), \tag{4.6c}$$

$$(\bar\ell, \delta\ell)_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda, \delta\ell\rangle_{L^r(0,T;\mathbb{R}^n)} \geq 0 \quad \forall\, \delta\ell \in \mathrm{cone}(\mathcal{K} - \bar\ell), \tag{4.6d}$$

where, for an arbitrary $z \in \mathbb{R}$, the left and right-sided derivative of $f : \mathbb{R} \to \mathbb{R}$ are defined through $f'_+(z) := f'(z; 1)$ and $f'_-(z) := -f'(z; -1)$, respectively.

Remark 4.8 The adjoint (integral) equation (4.6a) describes the mild solution of a differential equation featuring the so-called right Riemann-Liouville operator [19, Chap. 5]:

$$D^{\gamma}_{T-}(p)(t) = \bar{a}^{\top}(t)\lambda(t) \quad \text{f.a.a. } t \in (0, T), \quad \lim_{t \to T} I^{1-\gamma}_{T-}(p)(t) = \nabla g(\bar{y}(T)), \tag{4.7}$$

where

$$D^{\gamma}_{T-}(\phi) := -\frac{d}{dt} I^{1-\gamma}_{T-}(\phi).$$

Here we recall Definition 2.1. If p is absolutely continuous, then, together with $\delta y = S'((\bar{a}, \bar{\ell}); (\delta a, \delta \ell))$, it satisfies the relation in [5, Proposition 2.5], which says that the right Riemann-Liouville operator is the adjoint of the Caputo fractional derivative (Definition 2.2). Note that $\delta y \in W^{1,1}(0, T; \mathbb{R}^n)$; this can be shown by arguing as in the proof of Theorem 3.11. If p has enough regularity, then $I^{1-\gamma}_{T-} p \in C([0, T]; \mathbb{R}^n)$ and thus, $I^{1-\gamma}_{T-}(p)(T) = \nabla g(\bar{y}(T))$, in view of (4.7).

Proof of Theorem 4.7 We begin by noticing that the regularity of the state is a consequence of $(\bar{a}, \bar{\ell}) \in H^1(0, T; \mathbb{R}^{n \times n}) \times H^1(0, T; \mathbb{R}^n)$ in combination with Theorem 3.11. From Lemma 4.1, we get the existence of $\lambda \in L^r(0, T; \mathbb{R}^n)$ satisfying (4.1). This allows us to define an adjoint state $p \in L^r(0, T; \mathbb{R}^n)$ such that (4.6a), (4.6c) and (4.6d) are satisfied. Note that the $L^r(0, T; \mathbb{R}^n)$ regularity of p is a result of Lemmas 2.7 and 2.5. Thus, it remains to show that (4.6b) is true. Let $(\delta a, \delta \ell) \in H^1(0, T; \mathbb{R}^{n \times n}) \times \operatorname{cone}(\mathcal{K} - \bar{\ell})$ be arbitrary but fixed and abbreviate $\widetilde{\delta y} := S'((\bar{a}, \bar{\ell}); (\delta a, \delta \ell))$ and $f'(\cdot; \cdot) := f'(\bar{a}\bar{y} + \bar{\ell}; \bar{a}\widetilde{\delta y} + \delta a \bar{y} + \delta \ell)$. Note that

$$\widetilde{\delta y} = I^{\gamma}_{0+}(f'(\cdot; \cdot)), \tag{4.8}$$

see (3.6). Now, using (4.6a) and (4.8) in Lemma 2.3 leads to

$$\int_0^T (\bar{a}^{\top}\lambda)(t)^{\top} \widetilde{\delta y}(t) \, \mathrm{d}t = \int_0^T f'(\cdot; \cdot)(t)^{\top} \left[p(t) - \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} \nabla g(\bar{y}(T)) \right] \mathrm{d}t.$$

Thus,

$$\int_0^T f'(\cdot; \cdot)(t)^{\top} p(t) - (\bar{a}^{\top}\lambda)(t)^{\top} \widetilde{\delta y}(t) \, \mathrm{d}t = \nabla g(\bar{y}(T))^{\top} \underbrace{\int_0^T \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} f'(\cdot; \cdot)(t) \, \mathrm{d}t}_{= \widetilde{\delta y}(T), \text{ see (4.8)}}. \tag{4.9}$$

By inserting (4.9) in (4.2), we arrive at

$$\begin{aligned}&\int_0^T f'(\cdot;\cdot)(t)^\top p(t) - (\bar{a}^\top\lambda)(t)^\top\widetilde{\delta y}(t)\,\mathrm{d}t\\&\quad+(\bar{a},\delta a)_{H^1(0,T;\mathbb{R}^{n\times n})}+(\bar{\ell},\delta\ell)_{H^1(0,T;\mathbb{R}^n)}\geq 0\\&\quad\forall\,(\delta a,\delta\ell)\in H^1(0,T;\mathbb{R}^{n\times n})\times\mathrm{cone}(\mathcal{K}-\bar{\ell}).\end{aligned}\tag{4.10}$$

Setting $\delta\ell := 0$, taking into account $\widetilde{\delta y} = S'((\bar{a},\bar{\ell});(\delta a,0))$ and the definition of $f'(\cdot;\cdot)$, and making use of (4.6c), results in

$$\begin{aligned}&\int_0^T f'(\cdot;\cdot)(t)^\top p(t) - (\bar{a}^\top\lambda)(t)^\top\widetilde{\delta y}(t)\,\mathrm{d}t + (\bar{a},\delta a)_{H^1(0,T;\mathbb{R}^{n\times n})}\\&\overset{(4.6c)}{=}\int_0^T p(t)^\top f'(\bar{a}(t)\bar{y}(t)+\bar{\ell}(t);\bar{a}(t)\widetilde{\delta y}(t)+\delta a(t)\bar{y}(t))\,\mathrm{d}t\\&-\int_0^T\lambda(t)^\top(\bar{a}(t)\widetilde{\delta y}(t)+\delta a(t)\bar{y}(t))\,\mathrm{d}t\geq 0\quad\forall\,\delta a\in H^1(0,T;\mathbb{R}^{n\times n}).\end{aligned}\tag{4.11}$$

Now let $\rho\in C([0,T];\mathbb{R}^n)$ be arbitrary but fixed. According to Lemma 4.6 there exists a sequence $\{\delta a_n\}\subset H^1(0,T;\mathbb{R}^{n\times n})$ such that

$$\bar{a}S'((\bar{a},\bar{\ell});(\delta a_n,0))+\delta a_n\bar{y}\to\rho\quad\text{in } C([0,T];\mathbb{R}^n).$$

Thus, testing with $\delta a_n\in H^1(0,T;\mathbb{R}^{n\times n})$ in (4.11) and passing to the limit $n\to\infty$ leads to

$$\int_0^T p(t)^\top f'(\bar{a}(t)\bar{y}(t)+\bar{\ell}(t);\rho(t))-\lambda(t)^\top\rho(t)\,\mathrm{d}t\geq 0\quad\forall\,\rho\in C([0,T];\mathbb{R}^n),$$

where we relied on the continuity of $f'(\bar{a}\bar{y}+\bar{\ell};\cdot):L^\infty(0,T;\mathbb{R}^n)\to L^\infty(0,T;\mathbb{R}^n)$, cf. (3.2), and on the fact that $\lambda, p\in L^r(0,T;\mathbb{R}^n)$. Now, by testing with $\rho\geq 0$ and by employing the fundamental lemma of calculus of variations combined with the positive homogeneity of the directional derivative with respect to the direction we deduce

$$p_i(t)f'((\bar{a}\bar{y}+\bar{\ell})_i(t);1)-\lambda_i(t)\geq 0\qquad\text{a.e. in }(0,T),\quad i=1,\ldots n.$$

In an analogous way, testing with $\rho\leq 0$ implies

$$p_i(t)f'((\bar{a}\bar{y}+\bar{\ell})_i(t);-1)+\lambda_i(t)\geq 0\qquad\text{a.e. in }(0,T),\quad i=1,\ldots n,$$

from which (4.6b) follows. □

Remark 4.9 (Correspondence to KKT conditions) If $(\bar{a}\bar{y}+\bar{\ell})_i(t)\notin\mathcal{N}$ f.a.a. $t\in(0,T)$ and for all $i=1,\ldots,n$, where $\mathcal{N}$ denotes the set of non-smooth points of f, then $\lambda_i(t)=p_i(t)f'((\bar{a}\bar{y}+\bar{\ell})_i(t))$ f.a.a. $t\in(0,T)$ and for all $i=1,\ldots,n$, cf. (4.6b). In

this case, (4.1) is equivalent to the optimality system or standard KKT-conditions, which one would obtain if one would assume f to be continuously differentiable. These conditions are given by:

$$p(t) = \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} \nabla g(\bar{y}(T)) + \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)} \bar{a}^\top(s)\lambda(s)\,\mathrm{d}s \quad \forall\, t \in [0,T), \tag{4.12a}$$

$$\lambda_i(t) = p_i(t) f'((\bar{a}\bar{y} + \bar{\ell})_i(t)) \quad \text{for a.a. } t \in (0,T),\ i = 1, ..., n, \tag{4.12b}$$

$$(\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda, \delta a\, \bar{y}\rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall\, \delta a \in H^1(0,T;\mathbb{R}^{n\times n}), \tag{4.12c}$$

$$(\bar{\ell}, \delta\ell)_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda, \delta\ell\rangle_{L^r(0,T;\mathbb{R}^n)} \geq 0 \quad \forall\, \delta\ell \in \mathrm{cone}(\mathcal{K} - \bar{\ell}). \tag{4.12d}$$

The optimality system in Theorem 4.7 is indeed of strong stationary type, as the next result shows:

Theorem 4.10 *(Equivalence between B- and strong stationarity) Let $(\bar{a}, \bar{\ell}) \in H^1(0,T;\mathbb{R}^{n\times n}) \times \mathcal{K}$ be given and let $\bar{y} := S(\bar{a}, \bar{\ell})$ be its associated state. If there exists a multiplier $\lambda \in L^r(0,T;\mathbb{R}^n)$ and an adjoint state $p \in L^r(0,T;\mathbb{R}^n)$, where $r \in [1, \frac{1}{1-\gamma})$, such that (4.6) is satisfied, then $(\bar{a}, \bar{\ell})$ also satisfies the variational inequality (4.2). Moreover, if Assumption 4.3 is satisfied, then (4.2) is equivalent to (4.6).*

Proof We first show that (4.6b) implies

$$\int_0^T p(t)^\top f'(\bar{a}(t)\bar{y}(t) + \bar{\ell}(t); \rho(t)) - \lambda(t)^\top \rho(t)\,\mathrm{d}t \geq 0 \quad \forall\, \rho \in C([0,T];\mathbb{R}^n). \tag{4.13}$$

To this end, let $\rho \in C([0,T];\mathbb{R}^n)$ and $i = 1, ..., n$ be arbitrary, but fixed. We denote by $\mathcal{N}$ the set of non-differentiable points of f. From (4.6b), we deduce that

$$\lambda_i(t)\rho_i(t) = p_i(t) f'((\bar{a}\bar{y} + \bar{\ell})_i(t))\rho_i(t) \quad \text{a.e. where } (\bar{a}\bar{y} + \bar{\ell})_i \notin \mathcal{N}. \tag{4.14}$$

Further, we define $\mathcal{N}_i^+ := \{t \in [0,T] : (\bar{a}\bar{y} + \bar{\ell})_i(t) \in \mathcal{N} \text{ and } \rho_i(t) > 0\}$ and $\mathcal{N}_i^- := \{t \in [0,T] : (\bar{a}\bar{y} + \bar{\ell})_i(t) \in \mathcal{N} \text{ and } \rho_i(t) \leq 0\}$. Then, (4.6b) and the positive homogeneity of the directional derivative with respect to the direction yield

$$\begin{aligned} \lambda_i(t)\rho_i(t) &\leq \begin{cases} p_i(t) f'_+((\bar{a}\bar{y} + \bar{\ell})_i(t))\rho_i(t) & \text{a.e. in } \mathcal{N}_i^+ \\ p_i(t) f'_-((\bar{a}\bar{y} + \bar{\ell})_i(t))\rho_i(t) & \text{a.e. in } \mathcal{N}_i^- \end{cases} \\ &= p_i(t) f'((\bar{a}\bar{y} + \bar{\ell})_i(t); \rho_i(t)) \quad \text{a.e. where } (\bar{a}\bar{y} + \bar{\ell})_i \in \mathcal{N}. \end{aligned} \tag{4.15}$$

Now, (4.13) follows from (4.14) and (4.15).

Next, let $(\delta a, \delta\ell) \in H^1(0,T;\mathbb{R}^{n\times n}) \times \mathrm{cone}(\mathcal{K} - \bar{\ell})$ be arbitrary but fixed and test (4.13) with $\bar{a}\widetilde{\delta y} + \delta a\bar{y} + \delta\ell$, where we abbreviate $\widetilde{\delta y} := S'((\bar{a}, \bar{\ell}); (\delta a, \delta\ell))$. This results in

$$\int_0^T p(t)^\top f'(\bar{a}(t)\bar{y}(t)+\bar{\ell}(t);(\bar{a}\widetilde{\delta y}+\delta a\bar{y}+\delta\ell)(t))\,\mathrm{d}t$$
$$-\int_0^T \lambda(t)^\top(\bar{a}\widetilde{\delta y}+\delta a\bar{y}+\delta\ell)(t)\,\mathrm{d}t\geq 0. \tag{4.16}$$

Then, by using (4.9) one sees that (4.16) implies

$$\nabla g(\bar{y}(T))^\top\widetilde{\delta y}(T)-\langle\lambda,\delta a\bar{y}+\delta\ell\rangle_{L^r(0,T;\mathbb{R}^n)}\geq 0$$
$$\forall\,(\delta a,\delta\ell)\in H^1(0,T;\mathbb{R}^{n\times n})\times\mathrm{cone}(\mathcal{K}-\bar{\ell}). \tag{4.17}$$

Finally (4.6c)–(4.6d) in combination with (4.17) yield that (4.2) is true. Moreover, if Assumption 4.3 is satisfied, then (4.2) implies (4.6), see the proof of Theorem 4.7. We underline that the only information about the local minimizer that is used in the proof of Theorem 4.7 is contained in (4.2). □

Remark 4.11 (Strong stationarity in the case $\gamma=1$) If $\gamma=1$, then the state $\bar{y}$ associated to the local optimum $(\bar{a},\bar{\ell})\in H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)$ belongs to $W^{2,2}(0,T;\mathbb{R}^n)$; this is a consequence of the statement in Remark 3.3 combined with the fact that $f(\bar{a}\bar{y}+\bar{\ell})\in H^1(0,T;\mathbb{R}^n)$, since $f\in W^{1,\infty}(\mathbb{R}^n;\mathbb{R}^n)$, as a result of Assumption 3.1. Moreover, by taking a look at (4.6a) we see that the adjoint equation reads

$$-p'(t)=\bar{a}^\top(t)\lambda(t)\quad\forall\,t\in[0,T],\quad p(T)=\nabla g(\bar{y}(T))$$

for $\gamma=1$. A close inspection of step (III) in the proof of Lemma 4.1 shows that $p\in W^{2,2}(0,T;\mathbb{R}^n)$ and $\lambda\in L^\infty(0,T;\mathbb{R}^n)$, see (4.6b).

4.1 Some Comments Regarding the Main Result

We end this section by collecting some important remarks concerning Theorem 4.7.

Remark 4.12 (Density of the set of arguments of $f'((\bar{a}\bar{y}+\bar{\ell})_i(t);\cdot)$) The proof of Theorem 4.7 shows that it is essential that the set of directions into which the non-smooth mapping f is differentiated—in the 'linearized' state equation associated to $(\bar{a},\bar{\ell})$—is dense in a (suitable) Bochner space (which is the assertion in Lemma 4.6). This has also been pointed out in [10, Remark 2.12], where strong stationarity for a coupled non-smooth system is proven.

Let us underline that the 'constraint qualification' in Assumption 4.3 is not only due to the structure of the state equation, but also due to the presence of constraints on ℓ. If constraints were imposed on a instead of ℓ, then there would be no need for a CQ in the sense of Assumption 4.3. An inspection of the proof of Theorem 4.7 shows that in this case one needs to show that

$$\{\bar{a}S'((\bar{a},\bar{\ell});(0,\delta\ell))+\delta\ell:\delta\ell\in H^1(0,T;\mathbb{R}^n)\}\overset{d}{\hookrightarrow}C([0,T];\mathbb{R}^n).$$

This is done by arguing as in the proof of Lemma 4.6, where this time, one defines $\widehat{\delta\ell}:=\rho-\bar{a}\widehat{\delta y}$.

Thus, depending on the setting, the 'constraint qualification' may vanish or may read completely differently [10, Assumption 2.6], but it should imply that the set of directions into which f is differentiated -in the "linearized" state equation-is dense in an entire space [10, Lemma 2.8], see also [10, Remark 2.12].

These observations are also consistent with the result in [30]. Therein, the direction into which one differentiates the non-smoothness—in the 'linearized' state equation—is the 'linearized' solution operator, such that the counterpart of our Lemma 4.6 is [30, Lemma 5.2]. In [30], there is no constraint qualification in the sense of Assumption 4.3; however, the density assumption [30, Assumption 2.1.6] can be regarded as such. In [16, Remark 4.15] the authors also acknowledge the necessity of a density condition similar to that described above in order to ensure strong stationarity.

Remark 4.13 (Control constraints) We point out that we deal with controls (a, ℓ) mapping to $(\mathbb{R}^n)^{n+1}$, whereas the space of functions we want to cover in Lemma 4.6 consists of functions that map to $\mathbb{R}^n$ only. This allows us to restrict n controls by constraints (if we look at (P) as having $n+1$ controls mapping to $\mathbb{R}^n$.) Indeed, a closer inspection of the proof of Lemma 4.6 shows that one can impose control constraints on all columns of the control a except the m-th column. This still implies that the set of directions into which f is differentiated—in the 'linearized' state equation—is dense in an entire space. The fact that two or more controls provide advantages in the context of strong stationarity has already been observed in [26, Sect. 4]. Therein, an additional control has to be considered on the right-hand side of the VI under consideration in order to be able to prove strong stationarity, see [26, Sect. 4] for more details.

The situation changes when, in addition to asking that $\ell \in \mathcal{K}$, control constraints are imposed on all columns of a. In this case, we deal with a fully control constrained problem. By looking at the proof of Lemma 4.6 we see that the arguments cannot be applied in this case, see also [10, 16, 30, 32] where the same observation was made. This calls for a different approach in the proof of Theorem 4.7 and additional "constraint qualifications" [11, 39].

Remark 4.14 (Sign condition on the adjoint state. Optimality conditions obtained via smoothening)

(i) An essential information contained in the strong stationary system (4.6) is the fact that

$$p_i(t)\, f'_-((\bar{a}\,\bar{y} + \bar{\ell})_i(t)) \le p_i(t)\, f'_+((\bar{a}\,\bar{y} + \bar{\ell})_i(t)) \tag{4.18}$$

f.a.a. $t \in (0, T)$, $i = 1, ..., n$, see (4.6b). This is crucial for showing the implication (4.6) $\Rightarrow$ (4.2), which ultimately yields that (4.6) is indeed of strong stationary type (see the proof of Theorem 4.10).

If f is convex or concave around its non-smooth points, this translates into a sign condition for the adjoint state. Indeed, if $f : \mathbb{R} \to \mathbb{R}$ is convex around a non-smooth point z, this means that $f'_-(z) < f'_+(z)$, and from (4.18) we have

$$p_i(t) \ge 0 \quad \text{a.e. where } (\bar{a}\,\bar{y} + \bar{\ell})_i = z, \quad i = 1, ..., n.$$

Similarly, in the concave case, p is negative for those pairs (i,t) for which $(\bar{a}\bar{y}+\bar{\ell})_i(t)$ is a non-differentiable point of f.

In addition, we note that, if f is piecewise continuously differentiable, (4.18) implies the regularity (cf. [35, Definition 7.4.1]) of the mapping $p_i(t)f : \mathbb{R} \to \mathbb{R}$ at $(\bar{a}\bar{y}+\bar{\ell})_i(t)$ f.a.a. $t \in (0,T)$ and for all $i = 1, ..., n$, in view of [30, Lemma C.1]. See also [30, Remark 6.9] and [10] for similar situations.

(ii) By contrast, optimality systems derived by classical smoothening techniques often lack a sign for the adjoint state (and the above mentioned regularity in the sense of [35, Definition 7.4.1]) eventually along with other information which gets lost in the limit analysis. See e.g. [11, Proposition 2.17], [30, Sect. 4], [16, Theorem 4.4], [37, Theorem 2.4] (optimal control of non-smooth PDEs) and [32] (optimal control of VIs). Generally speaking, a sign condition for the adjoint state in those points (i,t) where the argument of the non-smoothness f in the state equation, in our case $(\bar{a}\bar{y}+\bar{\ell})_i(t)$, is such that f is not differentiable at $(\bar{a}\bar{y}+\bar{\ell})_i(t)$, is what ultimately distinguishes a strong stationary optimality system from very 'good' optimality systems obtained by smoothening procedures, cf. [11, Proposition 2.17] and [30, Sect. 7.2], see also [16, Remark 4.15].

Remark 4.15 (The multi-data case) Let us assume that the number of input data is larger than one and at the same time not larger than n. In this case our optimal control problem is replaced by

$$\left.\begin{aligned} \min_{(a,\ell)} \quad & \sum_{j=1}^{m} g(y^{(j)}(T)) + \frac{1}{2}\|a\|^2_{H^1(0,T;\mathbb{R}^{n\times n})} + \frac{1}{2}\|\ell\|^2_{H^1(0,T;\mathbb{R}^n)} \\ \text{s.t.} \quad & \partial^\gamma y^{(j)}(t) = f(a(t)y^{(j)}(t) + \ell(t)) \quad \text{a.e. in } (0,T), \\ & y^{(j)}(0) = y_0^{(j)}, \quad j = 1, ..., m \\ & \ell \in \mathcal{K}, \end{aligned}\right\} \tag{4.19}$$

where $m \in \{2, ..., n\}$ is fixed. Then, an inspection of the proof of Lemma 4.6 shows that the strong stationarity result remains true provided that the system

$$\widehat{\delta a}(t)\bar{y}(t) = \psi(t) \quad \forall\, t \in [0,T]$$

admits at least one solution $\widehat{\delta a} \in C([0,T];\mathbb{R}^{n\times n})$ for each $\psi \in C([0,T];\mathbb{R}^{n\times m})$; here $\bar{y} : [0,T] \to \mathbb{R}^{n\times m}$ denotes the state associated to a local optimum. The strong stationary optimality conditions in this particular case are given by

$$p^{(j)}(t) = \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)}\nabla g(\bar{y}^{(j)}(T)) + \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)}\bar{a}^\top(s)\lambda^{(j)}(s)\,\mathrm{d}s \quad \forall\, t \in [0,T), \tag{4.20a}$$

$$\lambda_i^{(j)}(t) \in [p_i^{(j)}(t)\, f'_-((\bar{a}\bar{y}^{(j)}+\bar{\ell})_i(t)),\, p_i^{(j)}(t)\, f'_+((\bar{a}\bar{y}^{(j)}+\bar{\ell})_i(t))] \quad \text{for a.a. } t \in (0,T), \tag{4.20b}$$

$$(\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \sum_{j=1}^{m} \langle \lambda^{(j)}, \delta a\, \bar{y}^{(j)} \rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall\, \delta a \in H^1(0,T;\mathbb{R}^{n\times n}), \tag{4.20c}$$

$$(\bar{\ell}, \delta \ell)_{H^1(0,T;\mathbb{R}^n)} + \sum_{j=1}^{m} \langle \lambda^{(j)}, \delta \ell \rangle_{L^r(0,T;\mathbb{R}^n)} \geq 0 \quad \forall\, \delta \ell \in \mathrm{cone}(\mathcal{K} - \bar{\ell}), \tag{4.20d}$$

$j = 1, ..., m$.

5 Application to a Continuous DNN

This section is concerned with the application of our main result to the following optimal control problem:

$$\left.\begin{aligned} \min \quad & \frac{1}{2}\|y(T) - y_d\|^2_{\mathbb{R}^n} + \frac{1}{2}\|a\|^2_{H^1(0,T;\mathbb{R}^{n\times n})} + \frac{1}{2}\|\ell\|^2_{H^1(0,T;\mathbb{R}^n)} \\ \text{s.t.} \quad & \partial^\gamma y(t) = \max(a(t)y(t) + \ell(t), 0) \quad \text{a.e. in } (0,T), \\ & y_1(0) = 1, \quad y_i(0) = 0, \ i = 2, ..., n. \\ & (a, \ell) \in H^1(0,T;\mathbb{R}^{n\times n}) \times \mathcal{K}, \end{aligned}\right\} \tag{Q}$$

where $y_d \in \mathbb{R}^n$ is fixed and the set $\mathcal{K}$ captures the fixed bias ordering [2], i.e.,

$$\mathcal{K} := \{\ell \in H^1(0,T;\mathbb{R}^n) : \ell_i(t) \leq \ell_{i+1}(t) \quad \forall\, i = 1, ..., n-1, \ \forall\, t \in [0,T]\}.$$

The state equation in (Q) describes a continuous deep neural network. For a discrete representation of this network, we refer to [4, 6].

We note that $\mathcal{K} \subset H^1(0,T;\mathbb{R}^n)$ is a convex, closed cone. Thus, all the quantities in (Q) fit in our general setting, cf. also Assumption 3.1. To see that the 'constraint qualification' in Assumption 4.3 is fulfilled, we refer to Remark 4.4.

Hence, we can apply Theorem 4.7, which in this particular case reads as follows:

Theorem 5.1 *(Strong stationarity for the control of continuous DNNs with fixed bias ordering) Let $(\bar{a}, \bar{\ell})$ be a given local optimum for (Q) with associated state $\bar{y}$. Then there exists a multiplier $\lambda \in L^r(0,T;\mathbb{R}^n)$ and an adjoint state $p \in L^r(0,T;\mathbb{R}^n)$, where $r \in [1, \frac{1}{1-\gamma})$, such that*

$$p(t) = \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)}(\bar{y}(T) - y_d) + \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)} \bar{a}^\top(s)\lambda(s)\,\mathrm{d}s \quad \forall\, t \in [0,T). \tag{5.1a}$$

$$\left.\begin{aligned} & \lambda_i(t) = p_i(t), \quad \text{a.e. where } (\bar{a}\bar{y} + \bar{\ell})_i > 0, \\ & \lambda_i(t) \in [0, p_i(t)] \quad \text{a.e. where } (\bar{a}\bar{y} + \bar{\ell})_i = 0, \\ & \lambda_i(t) = 0 \quad \text{a.e. where } (\bar{a}\bar{y} + \bar{\ell})_i < 0, \qquad i = 1, ..., n, \end{aligned}\right\} \tag{5.1b}$$

$$(\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda, \delta a\, \bar{y}\rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall \delta a \in H^1(0, T; \mathbb{R}^{n\times n}), \tag{5.1c}$$

$$(\bar{\ell}, \delta \ell)_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda, \delta \ell\rangle_{L^r(0,T;\mathbb{R}^n)} \geq 0 \quad \forall \delta \ell \in \mathcal{K} - \bar{\ell}. \tag{5.1d}$$

Moreover, (5.1) *is equivalent to*

$$\begin{aligned} &(\bar{y}(T) - y_d)^\top S'\left((\bar{a}, \bar{\ell}); (\delta a, \delta \ell)\right)(T) \\ &\quad + (\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + (\bar{\ell}, \delta \ell)_{H^1(0,T;\mathbb{R}^n)} \geq 0 \\ &\forall\, (\delta a, \delta \ell) \in H^1(0, T; \mathbb{R}^{n\times n}) \times \operatorname{cone}(\mathcal{K} - \bar{\ell}), \end{aligned} \tag{5.2}$$

where S is the control-to-state operator.

Proof The result follows from Theorems 4.7 and 4.10, by taking into account that (4.6b) is equivalent to (5.1b) if $f = \max\{\cdot, 0\}$. Note that (5.2) is just (4.2) in this particular setting. □

Funding Open Access funding enabled and organized by Projekt DEAL. Harbir Antil is partially supported by NSF Grant DMS-2110263 and the AirForce Office of Scientific Research under Award NO: FA9550-22-1-0248. Livia Betz is supported by the DFG Grant BE 7178/3-1. Daniel Wachsmuth is partially supported by the DFG Grant WA 3626/5-1.

Declarations

Conflict of interest The authors have no conflicts of interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix A: Proof of Lemma 4.1

Proof of Lemma 4.1 We associate a state equation to a smooth approximation of the non-differentiable function f, such that the respective solution mapping is Gâteaux-differentiable [step (I)]. Then, by arguments inspired by e.g. [9], it follows that $(\bar{a}, \bar{\ell})$ can be approximated by a sequence of local minimizers of an optimal control problem governed by the regularized state equation [step (II)]. Passing to the limit in the adjoint system associated to the regularized optimal control problem finally yields the desired assertion [step (III)]. Although many of the arguments are well-known, we give a detailed proof, for completeness and for convenience of the reader.

(I) Let $\varepsilon > 0$ be arbitrary, but fixed. We begin by investigating the smooth integral equation

$$y_\varepsilon(t) = y_0 + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} f_\varepsilon(a(s)y_\varepsilon(s) + \ell(s))\,\mathrm{d}s \quad \forall\, t \in [0,T], \tag{A.1}$$

where the differentiable function $f_\varepsilon : \mathbb{R} \to \mathbb{R}$ is defined as

$$f_\varepsilon(z) := \int_{-\infty}^{\infty} f(z - \varepsilon s)\varphi(s)\,\mathrm{d}s,$$

where $\varphi \in C_c^\infty(\mathbb{R})$, $\varphi \geq 0$, $\operatorname{supp}\varphi \subset [-1,1]$ and $\int_{-\infty}^{\infty} \varphi(s)\,\mathrm{d}s = 1$. Once again, we do not distinguish between $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ and $f_\varepsilon : \mathbb{R} \to \mathbb{R}$. As in the case of its non-smooth counterpart, $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ is assumed to satisfy for all $z \in \mathbb{R}^n$

$$f_\varepsilon(z)_i = \widetilde{f}_\varepsilon(z_i), \quad i = 1, ..., n$$

where $\widetilde{f}_\varepsilon : \mathbb{R} \to \mathbb{R}$ is a smooth function. We observe that for all $z \in \mathbb{R}$ it holds

$$f_\varepsilon(z) \to f(z) \quad \text{as } \varepsilon \searrow 0. \tag{A.2}$$

Moreover,

$$\|f_\varepsilon(z_1) - f_\varepsilon(z_2)\| \leq L\,\|z_1 - z_2\| \quad \forall\, z_1, z_2 \in \mathbb{R}^n, \tag{A.3}$$

where $L > 0$ is the Lipschitz constant of f.

By employing the exact same arguments as in the proof of Proposition 3.4, one infers that (A.1) admits a unique solution $y_\varepsilon \in C([0,T];\mathbb{R}^n)$ for every $(a,\ell) \in L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n)$, which allows us to define the smooth solution mapping

$$S_\varepsilon : L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n) \ni (a,\ell) \mapsto y_\varepsilon \in C([0,T];\mathbb{R}^n).$$

The operator S_ε is Gâteaux-differentiable and its derivative is the unique solution of

$$\delta y_\varepsilon(t) = \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} {f_\varepsilon}'(a(s)y_\varepsilon(s) + \ell(s))(a(s)\delta y_\varepsilon(s) + \delta a(s)y_\varepsilon(s) + \delta\ell(s))\,\mathrm{d}s \tag{A.4}$$

for all $t \in [0,T]$, i.e., $\delta y_\varepsilon = S_\varepsilon'(a,\ell)(\delta a, \delta\ell)$; note that here we use the notation ${f_\varepsilon}'$ for the Jacobi-matrix of $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$. By using the integral formulations (3.3) and (A.1), Lemma 2.6 and (A.2), we obtain the convergence $S_\varepsilon(a,\ell) - S(a,\ell) \to 0$ in $C([0,T];\mathbb{R}^n)$ as $\varepsilon \searrow 0$. On the other hand, by arguing as in the proof of Proposition 3.5 we deduce that S_ε is Lipschitz-continuous in the sense of (3.4) (with constant independent of ε). As a result, we have

$$S_\varepsilon(a_\varepsilon, \ell_\varepsilon) - S(a,\ell) \to 0 \quad \text{in } C([0,T];\mathbb{R}^n), \tag{A.5}$$

when $(a_\varepsilon, \ell_\varepsilon) \to (a,\ell)$ in $L^\infty(0,T;\mathbb{R}^{n\times n} \times \mathbb{R}^n)$.

(II) Next, we focus on proving that $(\bar{a}, \bar{\ell})$ can be approximated via local minimizers of optimal control problems governed by (A.1). To this end, let

$$B_\rho := B_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)}((\bar{a},\bar{\ell}),\rho) \cap (H^1(0,T;\mathbb{R}^{n\times n})\times\mathcal{K}),\ \ \rho>0,$$

be the ball of local optimality of $(\bar{a}, \bar{\ell})$ and consider the smooth (reduced) optimal control problem

$$\left.\begin{aligned} \min_{(a,\ell)\in H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \quad & J(S_\varepsilon(a,\ell),a,\ell)+\frac{1}{2}\|(a,\ell)-(\bar{a},\bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \\ \text{s.t.} \quad & (a,\ell)\in B_\rho. \end{aligned}\right\} \tag{A.6}$$

Let us recall here that

$$J(y,a,\ell)=g(y(T))+\frac{1}{2}\|a\|^2_{H^1(0,T;\mathbb{R}^{n\times n})}+\frac{1}{2}\|\ell\|^2_{H^1(0,T;\mathbb{R}^n)},$$

where $g:\mathbb{R}^n\to\mathbb{R}$ is a differentiable, and thus, continuous function. By the Lipschitz continuity of $S_\varepsilon : L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)\to C([0,T];\mathbb{R}^n)$ and the compact embedding $H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)\hookrightarrow\hookrightarrow L^\infty(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)$, we see that (A.6) admits a global solution $(a_\varepsilon,\ell_\varepsilon)\in B_\rho$. Since B_ρ is weakly closed in $H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)$ we can extract a weakly convergent subsequence

$$(a_\varepsilon,\ell_\varepsilon)\rightharpoonup(\widetilde{a},\widetilde{\ell})\in B_\rho \quad \text{in } H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n).$$

For simplicity, we abbreviate in the following

$$\mathcal{J}(a,\ell) := J(S(a,\ell),a,\ell), \tag{A.7a}$$

$$\mathcal{J}_\varepsilon(a,\ell) := J(S_\varepsilon(a,\ell),a,\ell)+\frac{1}{2}\|(a,\ell)-(\bar{a},\bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)}. \tag{A.7b}$$

Due to (A.5) combined with the continuity of $g:\mathbb{R}^n\to\mathbb{R}$, it holds

$$\begin{aligned} \mathcal{J}(\bar{a},\bar{\ell}) &\overset{(A.7a)}{=} J(S(\bar{a},\bar{\ell}),\bar{a},\bar{\ell}) = \lim_{\varepsilon\to0} J(S_\varepsilon(\bar{a},\bar{\ell}),\bar{a},\bar{\ell}) \\ &\overset{(A.7b)}{=} \lim_{\varepsilon\to0}\mathcal{J}_\varepsilon(\bar{a},\bar{\ell}) \geq \limsup_{\varepsilon\to0}\mathcal{J}_\varepsilon(a_\varepsilon,\ell_\varepsilon), \end{aligned} \tag{A.8}$$

where in the last inequality we used the fact that $(a_\varepsilon, \ell_\varepsilon)$ is a global minimizer of (A.6) and that $(\bar{a}, \bar{\ell})$ is admissible for (A.6). In view of (A.7b), (A.8) can be continued as

$$\begin{aligned}\mathcal{J}(\bar{a}, \bar{\ell}) &\geq \limsup_{\varepsilon\to 0} J(S_\varepsilon(a_\varepsilon, \ell_\varepsilon), a_\varepsilon, \ell_\varepsilon) + \frac{1}{2}\|(a_\varepsilon, \ell_\varepsilon) - (\bar{a}, \bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \\ &\geq \liminf_{\varepsilon\to 0} J(S_\varepsilon(a_\varepsilon, \ell_\varepsilon), a_\varepsilon, \ell_\varepsilon) + \frac{1}{2}\|(a_\varepsilon, \ell_\varepsilon) - (\bar{a}, \bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \\ &\geq J(S(\widetilde{a}, \widetilde{\ell}), \widetilde{a}, \widetilde{\ell}) + \frac{1}{2}\|(\widetilde{a}, \widetilde{\ell}) - (\bar{a}, \bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} \geq \mathcal{J}(\bar{a}, \bar{\ell}),\end{aligned} \tag{A.9}$$

where we used again $H^1(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n) \hookrightarrow\hookrightarrow L^\infty(0, T; \mathbb{R}^{n\times n})$ and (A.5) combined with the continuity of $g : \mathbb{R}^n \to \mathbb{R}$; note that for the last inequality in (A.9) we employed the fact that $(\widetilde{a}, \widetilde{\ell}) \in B_\rho$. From (A.9) we conclude

$$\lim_{\varepsilon\to 0} J(S_\varepsilon(a_\varepsilon, \ell_\varepsilon), a_\varepsilon, \ell_\varepsilon) + \frac{1}{2}\|(a_\varepsilon, \ell_\varepsilon) - (\bar{a}, \bar{\ell})\|^2_{H^1(0,T;\mathbb{R}^{n\times n}\times\mathbb{R}^n)} = \mathcal{J}(\bar{a}, \bar{\ell}).$$

By arguing as above we also get

$$\lim_{\varepsilon\to 0} J(S_\varepsilon(a_\varepsilon, \ell_\varepsilon), a_\varepsilon, \ell_\varepsilon) = \mathcal{J}(\bar{a}, \bar{\ell}),$$

which implies

$$(a_\varepsilon, \ell_\varepsilon) \to (\bar{a}, \bar{\ell}) \quad \text{in } H^1(0, T; \mathbb{R}^{n\times n} \times \mathbb{R}^n). \tag{A.10}$$

As a consequence, (A.5) yields

$$y_\varepsilon \to \bar{y} \quad \text{in } C([0, T]; \mathbb{R}^n), \tag{A.11}$$

where we abbreviate $y_\varepsilon := S_\varepsilon(a_\varepsilon, \ell_\varepsilon)$. By classical arguments one then obtains that $(a_\varepsilon, \ell_\varepsilon)$ is a local minimizer for

$$\min_{H^1(0,T;\mathbb{R}^{n\times n})\times\mathcal{K}} \mathcal{J}_\varepsilon(a, \ell).$$

(III) Due to the above established local optimality of $(a_\varepsilon, \ell_\varepsilon)$ and on account of the differentiability properties of S_ε, cf. step (I), we can write down the following necessary optimality condition

$$\begin{aligned}&\nabla g(y_\varepsilon(T))^\top S'_\varepsilon(a_\varepsilon, \ell_\varepsilon)((a, \ell) - (a_\varepsilon, \ell_\varepsilon))(T) \\ &\qquad + (2a_\varepsilon - \bar{a}, a - a_\varepsilon)_{H^1(0,T;\mathbb{R}^{n\times n})} + (2\ell_\varepsilon - \bar{\ell}, \ell - \ell_\varepsilon)_{H^1(0,T;\mathbb{R}^n)} \geq 0\end{aligned} \tag{A.12}$$

for all $(a, \ell) \in H^1(0, T; \mathbb{R}^{n\times n}) \times \mathcal{K}$. Now, let us consider for $t \in [0, T)$

$$p_\varepsilon(t) = \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} \nabla g(y_\varepsilon(T)) + \int_t^T \frac{(s-t)^{\gamma-1}}{\Gamma(\gamma)} a_\varepsilon^\top(s) \underbrace{f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)(s) p_\varepsilon(s)}_{=:\lambda_\varepsilon(s)} \, \mathrm{d}s. \tag{A.13}$$

To see that (A.13) admits a unique solution, we argue as in the proof of Proposition 3.4. From Lemma 2.5 we know that the operator $z \mapsto I^\gamma_{T-}(a_\varepsilon^\top f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon) z)$ maps continuous functions to continuous functions, since $a_\varepsilon^\top$, $f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon) \in L^\infty(0, T; \mathbb{R}^{n\times n})$. However, the first term in (A.13) is only L^r- integrable with r given by Lemma 2.7. This means that (no matter how smooth z is) the fix point operator associated to (A.13), namely

$$z \mapsto \frac{(T-\cdot)^{\gamma-1}}{\Gamma(\gamma)} \nabla g(y_\varepsilon(T)) + I^\gamma_{T-}(a_\varepsilon^\top f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon) z)$$

maps only to $L^r(0, T; \mathbb{R}^n)$ with r given by Lemma 2.7. Due to Lemma 2.5 we have for all $z_1, z_2 \in L^r(0, t^*; \mathbb{R}^n)$ the estimate

$$\begin{aligned} &\| I^\gamma_{T-}(a_\varepsilon^\top f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)(z_1 - z_2)) \|_{L^r(0,t^*;\mathbb{R}^n)} \\ &\qquad \le \frac{(t^*)^\gamma}{\gamma\Gamma(\gamma)} L \, \|a_\varepsilon\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \|z_1 - z_2\|_{L^r(0,t^*;\mathbb{R}^n)} \end{aligned}$$

and by arguing exactly as in the proof of Proposition 3.4 one obtains that (A.13) admits a unique solution $p_\varepsilon \in L^r(0, T; \mathbb{R}^n)$ with r given by Lemma 2.7. This immediately implies that $\lambda_\varepsilon \in L^r(0, T; \mathbb{R}^n)$, since $f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon) \in L^\infty(0, T; \mathbb{R}^{n\times n})$.

Next, let $(a, \ell) \in H^1(0, T; \mathbb{R}^{n\times n}) \times \mathcal{K}$ be arbitrary but fixed. We abbreviate

$$\delta y_\varepsilon := S'_\varepsilon(a_\varepsilon, \ell_\varepsilon)((a, \ell) - (a_\varepsilon, \ell_\varepsilon))$$

and

$$f_\varepsilon{}'(\cdot)(\cdot) := f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)(a_\varepsilon \delta y_\varepsilon + (a - a_\varepsilon) y_\varepsilon + \ell - \ell_\varepsilon),$$

which implies

$$\delta y_\varepsilon = I_{0+}(f_\varepsilon{}'(\cdot)(\cdot)), \tag{A.14}$$

on account of (A.4). Now, in the light of Lemma 2.3 combined with the identities (A.13) and (A.14) we have

$$\int_0^T (a_\varepsilon^\top \lambda_\varepsilon)(t)^\top \delta y_\varepsilon(t) \, \mathrm{d}t = \int_0^T f_\varepsilon{}'(\cdot)(\cdot)(t)^\top \Big[p_\varepsilon(t) - \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} \nabla g(y_\varepsilon(T)) \Big] \mathrm{d}t.$$

Thus, we obtain

$$\begin{aligned}&\int_0^T f_\varepsilon{}'(\cdot)(\cdot)(t)^\top p_\varepsilon(t) - (a_\varepsilon^\top \lambda_\varepsilon)(t)^\top \delta y_\varepsilon(t)\,\mathrm{d}t \\ &\quad = \nabla g(y_\varepsilon(T))^\top \underbrace{\int_0^T \frac{(T-t)^{\gamma-1}}{\Gamma(\gamma)} f'(\cdot)(\cdot)(t)\,\mathrm{d}t}_{=\delta y_\varepsilon(T),\ \text{see (A.14)}}.\end{aligned} \tag{A.15}$$

Since $\lambda_\varepsilon = f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)p_\varepsilon$, we can simplify the left-hand side of the above equation as

$$\begin{aligned}&\int_0^T f_\varepsilon{}'(\cdot)(\cdot)(t)^\top p_\varepsilon(t) - (a_\varepsilon^\top \lambda_\varepsilon)(t)^\top \delta y_\varepsilon(t)\,\mathrm{d}t \\ &\quad = \int_0^T \lambda_\varepsilon(t)^\top ((a-a_\varepsilon)y_\varepsilon + \ell - \ell_\varepsilon)(t)\,\mathrm{d}t.\end{aligned} \tag{A.16}$$

Inserting (A.15) and (A.16) in (A.12) leads to

$$\begin{aligned}0 \le &\int_0^T f_\varepsilon{}'(\cdot)(\cdot)(t)^\top p_\varepsilon(t) - (a_\varepsilon^\top \lambda_\varepsilon)(t)^\top \delta y_\varepsilon(t)\,\mathrm{d}t \\ &+ (2a_\varepsilon - \bar{a}, a - a_\varepsilon)_{H^1(0,T;\mathbb{R}^{n\times n})} + (2\ell_\varepsilon - \bar{\ell}, \ell - \ell_\varepsilon)_{H^1(0,T;\mathbb{R}^n)} \\ = &\langle \lambda_\varepsilon, (a-a_\varepsilon)y_\varepsilon + \ell - \ell_\varepsilon\rangle_{L^r(0,T;\mathbb{R}^n)} \\ &+ (2a_\varepsilon - \bar{a}, a - a_\varepsilon)_{H^1(0,T;\mathbb{R}^{n\times n})} + (2\ell_\varepsilon - \bar{\ell}, \ell - \ell_\varepsilon)_{H^1(0,T;\mathbb{R}^n)}.\end{aligned} \tag{A.17}$$

Setting (a, ℓ) in (A.17) to $(a_\varepsilon \pm \delta a, \ell_\varepsilon)$, $\delta a \in H^1(0,T;\mathbb{R}^{n\times n})$, and (a_ε, ℓ), $\ell \in \mathcal{K}$, respectively, yields

$$(2a_\varepsilon - \bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda_\varepsilon, \delta a\, y_\varepsilon\rangle_{L^r(0,T;\mathbb{R}^n)} = 0 \quad \forall \delta a \in H^1(0,T;\mathbb{R}^{n\times n}), \tag{A.18a}$$

$$(2\ell_\varepsilon - \bar{\ell}, \ell - \ell_\varepsilon)_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda_\varepsilon, \ell - \ell_\varepsilon\rangle_{L^r(0,T;\mathbb{R}^n)} \ge 0 \quad \forall \ell \in \mathcal{K}. \tag{A.18b}$$

The next step is to show that p_ε and hence λ_ε are bounded independently of ε. From (A.13) we further obtain

$$\begin{aligned}p_\varepsilon(T-t) = &\frac{t^{\gamma-1}}{\Gamma(\gamma)}\nabla g(y_\varepsilon(T)) \\ &+ \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} a_\varepsilon^\top (T-s) f_\varepsilon{}'(a_\varepsilon \bar{y}_\varepsilon + \ell_\varepsilon)(T-s) p_\varepsilon(T-s)\,\mathrm{d}s \quad \forall t \in (0,T].\end{aligned}$$

We abbreviate $\widetilde{p}_\varepsilon := p(T - \cdot)$. Since $\|f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \le L$, we have the estimate

$$\|\widetilde{p}_\varepsilon(t)\| \le c_1 t^{\gamma-1} + \int_0^t \frac{(t-s)^{\gamma-1}}{\Gamma(\gamma)} c_2\, L\, \|\widetilde{p}_\varepsilon(s)\|\, \mathrm{d}s, \quad t \in (0,T],$$

where $c_1, c_2 > 0$ are independent of ε. Here we also used that $\{\|a_\varepsilon\|_{L^\infty(0,T;\mathbb{R}^{n\times n})}\}$, $\{\|\nabla g(y_\varepsilon(T))\|_{\mathbb{R}^n}\}$ are uniformly bounded with respect to ε, cf., (A.10) and (A.11); recall that $g : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, by assumption. In view of Lemma 2.6, this implies

$$\|\widetilde{p}_\varepsilon(t)\| \le C\, c_1 t^{\gamma-1}, \quad t \in (0,T].$$

Thus, by employing again $\|f_\varepsilon{}'(a_\varepsilon y_\varepsilon + \ell_\varepsilon)\|_{L^\infty(0,T;\mathbb{R}^{n\times n})} \le 1$, we have $\|\lambda_\varepsilon\|_{L^r(0,T;\mathbb{R}^n)} \le \tilde{C}$ with $r \in (1, \frac{1}{1-\gamma})$ given by Lemma 2.7, and we can extract a weakly convergent subsequence

$$\lambda_\varepsilon \rightharpoonup^* \lambda \quad \text{in } L^r(0,T;\mathbb{R}^n).$$

Passing to the limit in (A.18) and using (A.10), (A.11) now yields

$$\begin{aligned}
&(\bar{a}, \delta a)_{H^1(0,T;\mathbb{R}^{n\times n})} + \langle \lambda, \delta a\, \bar{y}\rangle_{L^r(0,T;\mathbb{R}^{n\times n})} = 0 \quad \forall\, \delta a \in H^1(0,T;\mathbb{R}^{n\times n}),\\
&(\bar{\ell}, \ell - \bar{\ell})_{H^1(0,T;\mathbb{R}^n)} + \langle \lambda, \ell - \bar{\ell}\rangle_{L^r(0,T;\mathbb{R}^n)} \ge 0 \quad \forall\, \ell \in \mathcal{K}.
\end{aligned}$$

The proof is now complete. □

References

1. Agrawal, O.P.: A general formulation and solution scheme for fractional optimal control problems. Nonlinear Dyn. **38**(1–4), 323–337 (2004)
2. Antil, H., Brown, T.S., Lohner, R., Togashi, F., Verma, D.: Deep neural nets with fixed bias configuration. In: Numerical algebra, control and optimization. (2022)
3. Antil, H., Díaz, H., Herberg, E.: An optimal time variable learning framework for deep neural networks. Technical report. (2022). arXiv arXiv:2204.08528
4. Antil, H., Elman, H.C., Onwunta, A., Verma, D.: Novel deep neural networks for solving Bayesian statistical inverse problems. Technical report. (2021). arXiv arXiv:2102.03974
5. Antil, H., Gal, C.G., Warma, M.: A unified framework for optimal control of fractional in time subdiffusive semilinear PDEs. Discrete Contin. Dyn. Syst. Ser. S **15**(8), 1883–1918 (2022)
6. Antil, H., Khatri, R., Löhner, R., Verma, D.: Fractional deep neural network via constrained optimization. Mach. Learn.: Sci. Technol. **2**(1), 015003 (2020)
7. Antil, H., Otárola, E., Salgado, A.J.: A space-time fractional optimal control problem: analysis and discretization. SIAM J. Control Optim. **54**(3), 1295–1328 (2016)
8. Barbu, V.: Necessary conditions for distributed control problems governed by parabolic variational inequalities. SIAM J. Control Optim. **19**(1), 64–86 (1981)
9. Barbu, V.: Optimal control of variational inequalities. In: Research notes in mathematics, vol. 100. Pitman, Boston (1984)
10. Betz, L.: Strong stationarity for optimal control of a non-smooth coupled system: application to a viscous evolutionary VI coupled with an elliptic PDE. SIAM J. Optim. **29**(4), 3069–3099 (2019)

11. Betz, L.: Strong stationarity for a highly nonsmooth optimization problem with control constraints. Math. Control Relat. Fields (2022). https://doi.org/10.3934/mcrf.2022047
12. Betz, L., Yousept, I.: Optimal control of elliptic variational inequalities with bounded and unbounded operators. Math. Control Relat. Fields **11**(3), 479–498 (2021)
13. Bittner, L.: On optimal control of processes governed by abstract functional, integral and hyperbolic differential equations. Math. Operationsforsch. Statist. **6**(1), 107–134 (1975)
14. Christof, C.: Sensitivity analysis and optimal control of obstacle-type evolution variational inequalities. SIAM J. Control Optim. **57**(1), 192–218 (2019)
15. Christof, C., Brokate, M.: Strong stationarity conditions for optimal control problems governed by a rate-independent evolution variational inequality. (2022). arXiv:2205.01196
16. Christof, C., Clason, C., Meyer, C., Walther, S.: Optimal control of a non-smooth, semilinear elliptic equation. Math. Control Relat. Fields **8**(1), 247–276 (2018)
17. Clason, C., Nhu, V.H., Rösch, A.: Optimal control of a non-smooth quasilinear elliptic equation. Math. Control Relat. Fields **11**(3), 521–554 (2021)
18. De los Reyes, J.C., Meyer, C.: Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind. J. Optim. Theory Appl. **168**(2), 375–409 (2016)
19. Diethelm, K.: The analysis of fractional differential equations. Lecture Notes in Mathematics, Springer, Berlin (2004)
20. Elliott, C.M., Larsson, S.: Error estimates with smooth and nonsmooth data for a finite element method for the Cahn-Hilliard equation. Math. Comput. **58**(198), 603–630 (1992)
21. Evans, L.C.: Partial differential equations. American Mathematical Society, Providence (2010)
22. Geiger, C., Kanzow, C.: Theorie und Numerik restringierter Optimierungsaufgaben. Springer, Berlin (2002)
23. Goldberg, H., Tröltzsch, F.: Second order optimality conditions for a class of control problems governed by nonlinear integral equations with application to parabolic boundary control. Optimization **20**(5), 687–698 (1989)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 770–778. (2016)
25. Henry, D.: Geometric theory of semilinear parabolic equations. Springer, Berlin (1981)
26. Herzog, R., Meyer, C., Wachsmuth, G.: B- and strong stationarity for optimal control of static plasticity with hardening. SIAM J. Optim. **23**(1), 321–352 (2013)
27. Hintermüller, M., Kopacka, I.: Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. SIAM J. Optim. **20**(2), 868–902 (2009)
28. Ito, K., Kunisch, K.: Optimal control of parabolic variational inequalities. Journal de Mathématiques Pures et Appliqués **93**(4), 329–360 (2010)
29. Kilbas, A.A., Srivastava, H.M., Trujillo, J.J.: Theory and applications of fractional differential equations. In: North-Holland mathematics studies, vol. 204. Elsevier, Amsterdam (2006)
30. Meyer, C., Susu, L.M.: Optimal control of nonsmooth, semilinear parabolic equations. SIAM J. Control Optim. **55**(4), 2206–2234 (2017)
31. Mignot, F.: Contrôle dans les inéquations variationelles elliptiques. J. Funct. Anal. **22**(2), 130–185 (1976)
32. Mignot, F., Puel, J.-P.: Optimal control in some variational inequalities. SIAM J. Control Optim. **22**(3), 466–476 (1984)
33. Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. J. Math. Imaging Vis. **62**(3), 352–364 (2020)
34. Scheel, H., Scholtes, S.: Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. Math. Oper. Res. **25**(1), 1–22 (2000)
35. Schirotzek, W.: Nonsmooth analysis. Springer, Berlin (2007)
36. Stynes, M.: Too much regularity may force too much uniqueness. Fract. Calc. Appl. Anal. **19**(6), 1554–1562 (2016)
37. Tiba, D.: Optimal control of nonsmooth distributed parameter systems. Springer, Berlin (1990)
38. Tröltzsch, F.: Optimal control of partial differential equations. In: Graduate studies in mathematics, vol. 112, American Mathematical Society, Providence. (2010). Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels

39. Wachsmuth, G.: Strong stationarity for optimal control of the obstacle problem with control constraints. SIAM J. Optim. **24**(4), 1914–1932 (2014)
40. Wachsmuth, G.: Elliptic quasi-variational inequalities under a smallness assumption: uniqueness, differential stability and optimal control. Calc. Var. Partial Differ. Equ. **59**(2), 82 (2020)
41. Wolfersdorf, L.: Optimal control of a class of processes described by general integral equations of Hammerstein type. Math. Nachr. **71**, 115–141 (1976)
42. Ye, H., Gao, J., Ding, Y.: A generalized Gronwall inequality and its application to a fractional differential equation. J. Math. Anal. Appl. **328**(2), 1075–1081 (2007)