Noname manuscript No.
(will be inserted by the editor)

Finite Elements with Switch Detection for Direct Optimal Control of Nonsmooth Systems

Armin Nurkanović · Mario Sperl · Sebastian Albrecht · Moritz Diehl

Received: xxxx/ Accepted: xxxx

Abstract This paper introduces Finite Elements with Switch Detection (FESD), a numerical discretization method for nonsmooth differential equations. We consider the Filippov convexification of these systems and a transformation into dynamic complementarity systems introduced by Stewart [46]. FESD is based on solving nonlinear complementarity problems and can automatically detect nonsmooth events in time. If standard time-stepping Runge-Kutta (RK) methods are naively applied to a nonsmooth ODE, the accuracy is at best of order one. In FESD, we let the integrator step size be a degree of freedom. Additional complementarity conditions, which we call *cross complementarities*, enable *exact* switch detection, hence FESD can recover the high order accuracy that the RK methods enjoy for smooth ODE. Additional conditions called *step equilibration* allow the step size to change only when switches occur and thus avoid spurious degrees of freedom. Convergence results for the FESD method are derived, local uniqueness of the solution and convergence of numerical sensitivities are proven. The efficacy of FESD is demonstrated in several simulation and optimal control examples. In an optimal control problem benchmark with FESD, we achieve up to five orders of magnitude

A. Nurkanović,
Department of Microsystems Engineering (IMTEK), University of Freiburg, Germany, E-mail: armin.nurkanovic@imtek.uni-freiburg.de

M. Sperl,
Department of Mathematics, Chair of Applied Mathematics, University of Bayreuth, Germany, E-mail: Mario.Sperl@uni-bayreuth.de

S. Albrecht,
Siemens Technology, Munich, Germany, E-mail: sebastian.albrecht@siemens.com

M. Diehl,
Department of Microsystems Engineering (IMTEK) and Department of Mathematics, University of Freiburg, Germany, E-mail: moritz.diehl@imtek.uni-freiburg.de

more accurate solutions than a standard time-stepping approach for the same computational time.

Keywords switched systems · hybrid systems · nonsmooth ODE · numerical integration · optimal control · numerical methods
Mathematics Subject Classification 34A36, 49M25, 49Q12, 65L99, 49M37.

1 Introduction

The goal of this paper is to develop high-accuracy numerical simulation and optimal control methods for Ordinary Differential Equations (ODE) with a discontinuous vector field. We assume the following structure

$$\dot{x}(t) = f_i(x(t), u(t)), \text{ if } x(t) \in R_i \subset \mathbb{R}^{n_x},\ i \in \mathcal{J} \coloneqq \{1, \dots, n_f\}, \tag{1}$$

where R_i are disjoint open sets and $f_i(\cdot)$ are smooth functions on an open neighborhood of $\overline{R}_i$, n_f is a positive integer and $u(t)$ is an externally chosen control function.

This formulation of *piecewise smooth* ODE falls into the class of *hybrid systems* [14,50]. Many practical problems give rise to such ODE with a discontinuous right hand side (r.h.s.), e.g., in sliding mode control [4], mechanics problems with Coulomb friction [49], state constrained ODE derived from Pontryagin's maximum principle [41], electronic circuits [2], biological systems [5], vaccination strategies [15], transportation systems and traffic flow networks [9], constrained optimization algorithms viewed as dynamic systems [28] and many more. Systems with state jumps, including impact mechanics, robotics and hybrid systems with hysteresis can be transformed into systems matching the form of (1) via the *time-freezing reformulation* [35,37,40]. Consequently, efficient and accurate numerical optimal control algorithms for this class of systems are of great interest.

Related work: High-accuracy simulation of ODE with a discontinuous r.h.s. is numerically difficult since an accurate location of the nonsmooth events in time is needed. Standard time-stepping methods with a fixed step size applied to this class of ODE have at best first-order accuracy as they do not detect the switches [3]. Additionally, in the case of sliding modes the numerical solution obtained by explicit time-stepping methods tends to *chatter* around the discontinuity. Convergence of standard time-stepping discretization methods with order one is studied in [18,29,52]. Most high-accuracy methods include a root-finding procedure for accurate switch location and they usually assume that the trajectory crosses the discontinuity or only passes through two regions at a time [3]. An exception is Stewart's high accuracy method which can deal with almost all switching cases [46,48]. However, it also has an *external* switch location routine and is thus difficult to apply in direct optimal control, i.e., in *first discretize, then optimize* approaches to optimal control [44].

The parametric sensitivities of discontinuous ODE (1) have jump discontinuities [23,51] when the trajectory passes through or enters a surface of discontinuity, cf. Section 2.4. Therefore, the application of high-accuracy integrators even in a direct multiple shooting setting [13], which hides the external switch detection procedure from the optimizer, will be notoriously difficult, since derivative-based optimization algorithms will likely fail due to the non-Lipschitz sensitivities. Other fundamental difficulties within direct optimal control of non-smooth ODE are illuminated in the seminal paper by Stewart and Anitescu [51]. They show that direct methods based on time-stepping integration schemes with fixed step sizes are doomed to fail since the numerical sensitivities obtained by differentiating the results of a simulation are wrong no matter how small the step size is. We refer to this class of methods as standard methods because they are commonly used in optimal control of smooth dynamical systems. It is also shown that the numerical sensitivities of a smooth approximation of an ODE with a discontinuous r.h.s. are only correct if the step size approaches zero faster than the smoothing parameter, which makes accurate approximations computationally expensive. The same effects carry over to many Dynamic Complementarity Systems (DCS) [36].

On the theoretical side, necessary and/or sufficient conditions for optimality are provided in [16,24,45,53]. Guo and Ye [24] and Vieira et. al [53] study optimal control of a DCS with absolutely continuous solutions. The problems from this paper fall into this class. A broader overview can be found in [14].

On the practical side, many authors have developed methods to numerically treat discontinuous ODE in optimal control [10,11,12,30,31,36,42,51]. Kirches [31] develops a direct multiple shooting-based approach with a switch detecting integrator and sensitivity update formulae. Similarly, in [42] a method is developed that can treat sliding modes on a single switching surface. Katayama et al. [30] fix the switching sequence and optimize the lengths of the phases in a model predictive control loop, where at every sample the switching sequence is updated. Bemporad et al. [11] assign integer variables to every mode and solve mixed-integer optimization problems. In [10,12,36] the ODE is transformed into a DCS, resulting in a Mathematical Program with Complementarity Constraints (MPCC) to be solved after discretization. In [36] the authors use a standard discretization approach and suggest to use a homotopy to avoid spurious local minima due to wrong sensitivities [51]. Baumrucker and Biegler [10] consider systems with a single switching surface (or with multiple independent switching surfaces, cf. Section 2.3) and allow variable step sizes. This method yields exact switch detection, higher order integration accuracy, and correct numerical sensitivities. The step sizes are left to the optimizer as a degree of freedom, hence it can play with the discretization accuracy, possibly in an undesired way. Unfortunately, a formal proof of the appealing properties of the method is not provided in [10].

As it can be seen from the discussion above, most of the practical methods use only first-order accuracy methods with possibly incorrect sensitivities or treat the discontinuity in the integrator which complicates the use of derivative-based optimization algorithms (except [10]). They do not treat

sliding modes appropriately or handle only systems with a single switching surface, i.e., only two regions. The goal of this paper is to develop a method that resolves these issues and to provide a proper convergence theory. Note that the (easier) case of ODE with a continuous but nonsmooth r.h.s. fits into the structure of (1). Conversely, the *time-freezing reformulation* [35,37, 40] transforms many systems from the (more difficult) case of systems with state jumps into the form of (1). This enables us to treat many classes of nonsmooth systems with the same numerical method in a unified way. FESD uses the DCS representation introduced by Stewart [46,48] and is motivated by the variable step size ideas of Baumrucker and Biegler [10].

Contributions: In this paper, we develop the FESD method, which can be used both in simulation and optimal control problems. We start with a reformulation of (1) into dynamic complementary systems introduced by Stewart [46,48] and provide a constructive way to pass from more natural definitions of discontinuous ODE to Stewart's form. Discretization of the DCS results in a nonlinear complementary problem. We build on the ideas of varying the step size and allowing switches only to take place at the boundaries of the finite elements introduced in [10]. The FESD method can efficiently deal with multiple and simultaneous switches including sliding modes on higher co-dimension surfaces, and thus is more general than [10]. Moreover, in contrast to [10], where only Radau-IIA Implicit Runge-Kutta (IRK) methods were considered, in FESD one can use any Runge-Kutta method.

Additionally, we prove that FESD detects the switches *exactly* in time, recovers the high-order accuracy that RK methods enjoy for smooth ODE, and obtains the correct numerical sensitivities even when the solution crosses or stays on a discontinuity. To allow switching on the boundaries of the finite elements we introduce the *cross complementarity* formulation. Using FESD to discretize OCP with DCS results in mathematical programs with complementarity constraints which can be solved efficiently with smooth optimization techniques by solving few several Nonlinear Programming (NLP) problems. [8,27,32,33,43]. Thus, we avoid the nonsmooth difficulty encountered within direct multiple shooting with switch-detecting integrators.

Since the step sizes h_n are allowed to vary, if no switches occur, we would encounter spurious degrees of freedom which let the optimizer play with the integrator accuracy in a possibly undesired way. To avoid this problem we propose additional conditions called *step equilibration* which are based on an indicator function. This decouples the integrator accuracy from the optimizer and results in piecewise equidistant discretization grids between the switches. We illustrate the practicability of FESD on several simulation and optimal control examples and verify the theoretical findings.

The FESD method with its many variations and the remainder of the tool-chain for numerically solving optimal control problems with nonsmooth systems are implemented in the open source MATLAB toolbox `NOSNOC` [1,38].

Organization of the paper: Section 2 gives an introduction to Filippov systems and Stewart's reformulation. It provides a practical procedure to construct Stewart's indicator functions and discusses issues with discontinuous sensitivities. Section 3 introduces the FESD method and discusses the main ideas that lead to it. In Section 4 several relevant theoretical properties of the FESD are studied. The sections contain numerical examples that illustrate the theoretical and algorithmic developments. Section 5 shows how to use FESD in numerical optimal control. The section finishes with an optimal control example and a benchmark compassion of FESD to the standard approach. Finally, Section 6 summarizes the paper and outlines future research.

Notation: The complementary conditions for two vectors $a, b \in \mathbb{R}^n$ read as $0 \leq a \perp b \geq 0$, where $a \perp b$ means $a^\top b = 0$.

For two scalar variables a, b the so-called C-functions [20, Section 1.5.1] have the property $\phi(a,b) = 0 \iff a \geq 0, b \geq 0, ab = 0$. Examples are the natural residual functions $\phi_{\mathrm{NR}}(a,b) = \min(a,b)$ or the Fischer-Burmeister function $\phi_{\mathrm{FB}}(a,b) = a + b - \sqrt{a^2+b^2}$. If $a, b \in \mathbb{R}^n$, we use $\phi(\cdot)$ componentwise and define $\Phi(a,b) = (\phi(a_1,b_1), \ldots, \phi(a_n,b_n))$. All vector inequalities are to be understood element-wise, $\mathrm{diag}(x) \in \mathbb{R}^{n\times n}$ returns a diagonal matrix with $x \in \mathbb{R}^n$ containing the diagonal entries. The concatenation of two column vectors $a \in \mathbb{R}^{n_a}$, $b \in \mathbb{R}^{n_b}$ is denoted by $(a,b) \coloneqq [a^\top, b^\top]^\top$, the concatenation of several column vectors is defined analogously. The identity matrix is denoted by $I \in \mathbb{R}^{n\times n}$ and a column vector with all ones is denoted by $e = (1,1,\ldots,1) \in \mathbb{R}^n$, their dimension is clear from the context. The closure of a set C is denoted by $\overline{C}$, its boundary as ∂C and $\mathrm{conv}(C)$ is its convex hull. Given a matrix $M \in \mathbb{R}^{n\times m}$, its i-th row is denoted by $M_{i,\bullet}$ and its j-th column is denoted by $M_{\bullet,j}$. For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ we denote by $\mathrm{D}f(x) = \frac{\partial f}{\partial x}(x) \in \mathbb{R}^{m\times n}$ the Jacobian matrix and by $\nabla f(x) \coloneqq \frac{\partial f}{\partial x}(x)^\top$ its transpose.

For the left and the right limits, we use the notation $x(t_{\mathrm{s}}{}^+) = \lim_{t\to t_{\mathrm{s}},\ t>t_{\mathrm{s}}} x(t)$ and $x(t_{\mathrm{s}}{}^-) = \lim_{t\to t_{\mathrm{s}},\ t<t_{\mathrm{s}}} x(t)$, respectively. When clear from context, we often drop the dependency on time t.

2 Piecewise smooth differential equations

This section will introduce some necessary assumptions on the systems PSS (1), its Filippov convexification [22], and Stewart's reformulation into Dynamic Complementarity Systems (DCS) [46,48] to prepare the ground for the novel method presented in Section 3. We discuss some properties of the DCS for a fixed active set and active-set changes.

2.1 Filippov convexification

Initial value problems arising from the nonsmooth ODE (1) usually fail to have classic Carathéodory solutions, for a counterexample, see e.g., [46, Section 1].

To have a meaningful solution concept for this class of ODE, the main idea of Filippov was to replace the r.h.s. of (1) with a convex set and to obtain the following Differential Inclusion (DI):

$$\dot{x}(t) \in F_{\mathrm{F}}(x(t), u(t)) \coloneqq \bigcap_{\delta>0} \bigcap_{\mu(N)=0} \overline{\mathrm{conv}} f(x + \delta B(x) \setminus N, u(t)), \tag{2}$$

where $f(x,u) = f_i(x,u)$ if $x \in R_i$, $B(x)$ is the Euclidean unit ball at x in $\mathbb{R}^{n_x}$, $\mu(\cdot)$ is the Lebesgue measure on $\mathbb{R}^{n_x}$ and $\overline{\mathrm{conv}}(\cdot)$ maps a subset of $\mathbb{R}^{n_x}$ to its closed convex hull. Throughout the paper we assume that the regions R_i are disjoint, connected and open. They are assumed to be nonempty and to have piecewise-smooth boundaries ∂R_i. We assume that $\overline{\bigcup_{i\in\mathcal{J}} R_i} = \mathbb{R}^{n_x}$ and that $\mathbb{R}^{n_x} \setminus \bigcup_{i\in\mathcal{J}} R_i$ is a set of measure zero.

Let $\mathcal{I}(x) \coloneqq \{i \mid x \in \overline{R}_i\} \subseteq \mathcal{J}$ be the active set at $x \in \mathbb{R}^{n_x}$. Due to the special structure of (1) the Filippov DI (2) can be written as

$$\dot{x} \in \overline{\mathrm{conv}}\{f_i(x,u) \mid i \in \mathcal{I}(x)\}.$$

This means that in the interior of the regions, R_i the Filippov set $F_{\mathrm{F}}(x,u)$ is equal to $\{f_i(x,u)\}$ and on the boundary between regions we have a convex combination of the neighboring vector fields. If $\dot{x}$ exists, functions $\theta_i(\cdot)$ which serve as convex multipliers can be introduced and the Filippov DI can be written as

$$\dot{x} \in F_{\mathrm{F}}(x,u) = \Big\{ \sum_{i\in\mathcal{J}} f_i(x,u)\,\theta_i \mid \sum_{i\in\mathcal{J}} \theta_i = 1,\ \theta_i \geq 0,\ \theta_i = 0 \text{ if } x \notin \overline{R_i}, \forall i \in \mathcal{J} \Big\}. \tag{3}$$

We call the functions $\theta_i(\cdot)$ *Filippov multipliers*. As it will be seen later the functions $\theta_i(\cdot)$ lack any continuity properties. But it can be shown that they are at least measurable [21,46]. Given (3), we will compute *piecewise active* solutions [46], which are defined as follows.

Definition 1 (Piecewise active solution [46]) *For an initial value $x(0) = x_0$, a given measurable control function $u(t)$ and a compact interval $[0,T]$, a function $x : [0,T] \to \mathbb{R}^{n_x}$ is said to be a solution of (2), if $\dot{x}(t) \in F_{\mathrm{F}}(x(t),u(t))$ almost everywhere on $[0,T]$. This function is called a piecewise active solution if the active set $\mathcal{I}(x(t))$ is a piecewise constant function of time and it changes its value only finitely many times on $[0,T]$. A time point $t_{\mathrm{s}} \in [0,T]$ is called a switching point if $\mathcal{I}(x(t))$ is not constant in any sufficiently small neighborhood of t_{s}.*

Note that this definition assumes a finite number of switches and excludes so-called Zeno solutions, where infinitely many switches can occur in a finite time interval. Zeno solutions cannot be treated with event-detecting methods.

For a constant active set $\mathcal{I}(x)$ one can derive an ODE or Differential Algebraic Equation (DAE) (for sliding modes) from (3) and apply standard integration methods. The overall ODE solution $x(t)$ is continuous and consists of smooth pieces connected by nondifferentiable points ("kinks") at the switching times t_{s}.

2.2 Stewart's reformulation

We regard a specific representation of the sets R_i which was introduced by Stewart [46]. The main assumption is that the regions R_i are given as

$$R_i = \{x \in \mathbb{R}^{n_x} \mid g_i(x) < \min_{j \in \mathcal{J},\, j \neq i} g_j(x)\}. \tag{4}$$

It is assumed that the *indicator* functions $g_i(\cdot),\ i \in \mathcal{J}$, are smooth functions. Moreover, throughout the paper we assume additionally that $g_i(\cdot), f_i(\cdot)$ and $\nabla g_i(\cdot)$ are Lipschitz continuous.

Note that due to the definition of the sets R_i in (4), the active set can be defined as

$$\mathcal{I}(x(t)) \coloneqq \Big\{ i \in \mathcal{J} \mid g_i(x(t)) = \min_{j \in \mathcal{J}} g_j(x(t)) \Big\}. \tag{5}$$

We define the vectors $\theta = (\theta_1, \ldots, \theta_{n_f}) \in \mathbb{R}^{n_f}$, $g(x) = (g_1(x), \ldots, g_{n_f}(x)) \in \mathbb{R}^{n_f}$ and the matrix $F(x) = \big[f_1(x), \ldots, f_{n_f}(x)\big] \in \mathbb{R}^{n_x \times n_f}$. Using the specific representations (4), from [46] and equation (3) one can deduce that the Filippov DI can be written as

$$\dot{x} = F(x,u)\theta(x), \tag{6}$$

where the algebraic variables $\theta(x(t))$ are a solution of the parametric Linear Program (LP)

$$\mathrm{LP}(x): \quad \theta(x) \in \arg\min_{\tilde{\theta} \in \mathbb{R}^{n_f}} \quad g(x)^\top \tilde{\theta} \quad \text{s.t.} \quad e^\top \tilde{\theta} = 1,\ \tilde{\theta} \geq 0. \tag{7}$$

Using the Karush–Kuhn–Tucker (KKT) conditions of $\mathrm{LP}(x)$ and (6), we obtain the dynamic complementarity system

$$\dot{x} = F(x,u)\theta, \tag{8a}$$
$$0 = g(x) - \lambda - \mu e, \tag{8b}$$
$$1 = e^\top \theta, \tag{8c}$$
$$0 \leq \theta \perp \lambda \geq 0, \tag{8d}$$

where the algebraic variables $\lambda \in \mathbb{R}^{n_f}$ and $\mu \in \mathbb{R}$ are the Lagrange multipliers of the parametric LP (7). To have an even more compact representation we use a C-function Ψ for the complementarity conditions and rewrite the KKT conditions of the LP (7) as the nonsmooth equation:

$$G_{\mathrm{LP}}(x,\theta,\lambda,\mu) \coloneqq \begin{bmatrix} g(x) - \lambda - \mu e \\ 1 - e^\top \theta \\ \Psi(\theta,\lambda) \end{bmatrix} = 0. \tag{9}$$

It provides $2n_f+1$ conditions for the $2n_f+1$ algebraic variables $\theta,\ \lambda$ and μ. The DCS reads in compact form as a nonsmooth differential algebraic equation:

$$\dot{x} = F(x,u)\theta, \tag{10a}$$
$$0 = G_{\mathrm{LP}}(x,\theta,\lambda,\mu). \tag{10b}$$

Example 1 *We illustrate this formulation on the simple example of* $\dot{x} \in 2 - \text{sign}(x)$. *This ODE is characterized by the regions* $R_1 = \{x \in \mathbb{R} \mid x < 0\}$ *and* $R_2 = \{x \in \mathbb{R} \mid x > 0\}$, *with* $f_1(x) = 3$, $f_2(x) = 1$ *and* $F(x) = [3 \quad 1]$. *It can be verified that with the functions* $g_1(x) = x$ *and* $g_2(x) = -x$ *we have a representation of the regions as in* (4). *Moreover, we have the multipliers* $\theta, \lambda \in \mathbb{R}^2$ *and* $\mu \in \mathbb{R}$. *Thus, the corresponding DCS reads as:*

$$\dot{x} = \begin{bmatrix} 3 & 1 \end{bmatrix} (\theta_1, \theta_2), \tag{11a}$$
$$0 = x - \lambda_1 - \mu, \tag{11b}$$
$$0 = -x - \lambda_2 - \mu, \tag{11c}$$
$$1 = \theta_1 + \theta_2, \tag{11d}$$
$$0 = \theta \perp \lambda \geq 0. \tag{11e}$$

2.2.1 Remark on how to treat switching functions

Definition (4) might not be the most intuitive way to represent the sets R_i. In many practical examples some smooth scalar functions $c_i(\cdot)$, called *switching functions*, are given. Their zero-level sets define the boundaries of the regions R_i. For example, $R_1' = \{x \in \mathbb{R}^{n_x} \mid c_1(x) > 0, \ldots, c_m(x) > 0\}$, $R_2' = \{x \in \mathbb{R}^{n_x} \mid c_1(x) > 0, \ldots, c_{m-1}(x) > 0, c_m(x) < 0\}$ and so on. Let $c(x) = (c_1(x), \ldots, c_m(x)) \in \mathbb{R}^m$ and assume that $\nabla c(x) \in \mathbb{R}^{n \times m}$ has rank m. Thus, we can locally define up to $n_f = 2^m$ regions and encode them via a sign matrix $S \in \mathbb{R}^{2^m \times m}$ defined as

$$S = \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & \ldots & 1 & -1 \\ \vdots & \vdots & \ldots & \vdots & \\ -1 & -1 & \ldots & -1 & -1 \end{bmatrix}.$$

Note that the matrix S has no repeating rows. Moreover, we assume that this matrix has no zero entries. The sets R_i' can be compactly represented using the rows $S_{i,\bullet}$ as

$$R_i' = \{x \in \mathbb{R}^{n_x} \mid \text{diag}(S_{i,\bullet}) c(x) > 0\}. \tag{12}$$

The next proposition provides a constructive way to find the functions $g(\cdot)$ from the more intuitive representation of the regions via $c(\cdot)$.

Proposition 2 *Let the function* $g : \mathbb{R}^{n_x} \to \mathbb{R}^{n_f}$ *be defined as*

$$g(x) = -Sc(x), \tag{13}$$

then for all $x \in R_i'$ *the following statements are true:*

(i) $g_i(x) < g_j(x)$, *for* $i \neq j$,
(ii) the definitions (4) *and* (12) *define the same set, i.e.,* $R_i = R_i'$.

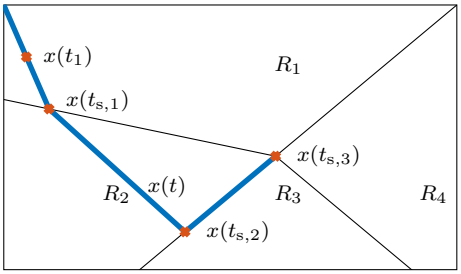

Fig. 1 Illustration of active sets at different points. It can be seen that $\mathcal{I}(x(t_1)) = \mathcal{I}_0 = \{1\}$. At $x(t_{\mathrm{s},1})$ the trajectory crosses the surface of discontinuity between R_1 and R_2, hence $\mathcal{I}(x(t_{\mathrm{s},1})) = \mathcal{I}_1^0 = \{1,2\}$ and later $\mathcal{I}_1 = \{2\}$. The segment between $x(t_{\mathrm{s},2})$ and $x(t_{\mathrm{s},3})$ is a sliding mode and we have $\mathcal{I}_2^0 = \{2,3\}$ and $\mathcal{I}_2 = \{2,3\}$. Finally we have at $x(t_{\mathrm{s},3})$ that $\mathcal{I}_3^0 = \{1,2,3,4\}$.

Proof. For (i), note that for $x \in R_i'$ all terms in the sum $g_i(x) = -S_{i,\bullet}c(x) = -\sum_k S_{i,k}c_k(x)$ are strictly positive. On the other hand, for any $g_j(x) = -S_{j,\bullet}c(x) = -\sum_k S_{j,k}c_k(x)$, $j \neq i$ and $x \in R_i'$, due to (12), all terms in the sum where $S_{j,k} \neq S_{i,k}$ are strictly negative. Therefore $S_{i,\bullet}c(x) > S_{j,\bullet}c(x)$, thus (i) holds.

For (ii), first regard the rows $S_{j,\bullet}$ that differ from $S_{i,\bullet}$ only in the k-th column. Then $g_i(x) - g_j(x) = -(S_{i,k} - S_{j,k})c_k(x) < 0$. If $S_{i,k} = 1$, then $g_i(x) - g_j(x) = -2c_k(x) < 0$. Likewise, for $S_{i,k} = -1$, then $g_i(x) - g_j(x) = 2c_k(x) < 0$. Therefore, from (4) we recover the definition of (12) by looking at the rows where $S_{i,k}$ and $S_{j,k}$ differ by one element. For all rows j that differ from $S_{i,\bullet}$ by more than one column, by similar reasoning, we obtain inequalities that do not tighten (12), since $g_i(x) - g_j(x)$ consists of a sum of the terms from the inequalities where only one component of $c(x)$ is left. Therefore, statement (ii) holds and this completes the proof. □

Example 2 *For our tutorial example* $\dot{x} \in 2 - \operatorname{sign}(x)$ *and the corresponding DCS* (11) *we have* $c(x) = x$ *and* $S = \begin{bmatrix}-1 & 1\end{bmatrix}^\top$ *and we obtain* $g(x) = -Sc(x) = (x, -x)$ *as used in Example 1.*

2.2.2 Fixed active set

For a given solution $x(\cdot)$ let us denote all switching points by $0 = t_{\mathrm{s},0} < t_{\mathrm{s},1} < \cdots < t_{\mathrm{s},N_{\mathrm{sw}}} = T$. The fixed active set between two switches is denoted by $\mathcal{I}_n := \mathcal{I}(x(t))$, $t \in (t_{\mathrm{s},n}, t_{\mathrm{s},n+1}) =: I_n$ and at a switching point $t_{\mathrm{s},n}$ by $\mathcal{I}_n^0 := \mathcal{I}(x(t_{\mathrm{s},n}))$. Note that $\mathcal{I}_n^0 = \mathcal{I}_n \cup \mathcal{I}_{n-1}$. These definitions are illustrated in Figure 1. In this subsection, we regard the DCS (8) for a single fixed $\mathcal{I}_n$. For ease of notation, we drop the subscripts in this subsection and denote the fixed active set by $\mathcal{I}$. Depending on the active set, the DCS (8) reduces either to an ODE or to a DAE.

To simplify our exposition we introduce the following notation. For a given vector $a \in \mathbb{R}^n$ and set $\mathcal{I} \subseteq \{1, \ldots, n\}$, we define the projection matrix $P_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times n}$ which has zeros or ones as entries. It selects all component $a_i, i \in \mathcal{I}$ from the vector a, i.e., $a_{\mathcal{I}} = P_{\mathcal{I}} a \in \mathbb{R}^{|\mathcal{I}|}$ and $a_{\mathcal{I}} = [a_i \mid i \in \mathcal{I}]$.

In the DAE case, x is on the boundary of one or more regions R_i, we speak of sliding modes [23], i.e., $|\mathcal{I}| > 1$ and typically obtain an index 2 differential algebraic equation. In this case, two or more equal entries of $g(x)$ are the smallest components of this vector, and the solution θ of the LP(x) is not unique and lies on a facet of the unit simplex. To compute the values of θ, we must treat the DCS as a DAE. We define $F_{\mathcal{I}}(x, u) \coloneqq F(x, u) P_{\mathcal{I}}^{\top}$, which selects the appropriate columns of $F(x, u)$. For $t \in I$, we have $\theta_i = 0, i \notin \mathcal{I}$ and $\lambda_i = 0, i \in \mathcal{I}$, thus the DCS (8) reduces to the DAE

$$\dot{x} = F_{\mathcal{I}}(x, u)\theta_{\mathcal{I}}, \tag{14a}$$

$$0 = g_{\mathcal{I}}(x) - \mu e, \tag{14b}$$

$$1 = e^{\top}\theta_{\mathcal{I}}. \tag{14c}$$

There are $|\mathcal{I}| + 1$ nontrivial algebraic equations and $|\mathcal{I}| + 1$ unknown algebraic variables, namely μ and θ_i for $i \in \mathcal{I}$, since we consider $\theta_i(t) = 0,\ i \notin \mathcal{I}$ as fixed.

In the ODE case, x is in the interior of some region R_i, we have $|\mathcal{I}| = 1$. The algebraic variables μ and θ_i can be computed explicitly from (14) and we have $\theta_i = 1$ and $\mu = g_i(x)$. Thus, the DCS reduces to the ODE $\dot{x} = f_i(x)$.

Next, we provide sufficient conditions for solution uniqueness of the DAE (14) for a given $|\mathcal{I}| \geq 1$. We define the matrix

$$M_{\mathcal{I}}(x) = \nabla g_{\mathcal{I}}(x)^{\top} F_{\mathcal{I}}(x, u) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}. \tag{15}$$

Note that entries of this matrix arise by taking the total time derivative of (14b).

Assumption 3 *Given a fixed active set $\mathcal{I}(x(t)) = \mathcal{I}$ for $t \in I$, it holds that the matrix $M_{\mathcal{I}}(x(t))$ is invertible and $e^{\top} M_{\mathcal{I}}(x(t))^{-1} e \neq 0$ for all $t \in I$.*

Proposition 4 *Suppose that Assumption 3 holds. Given the initial value $x(t_{\mathrm{s},n})$, then the DAE (14) has a unique solution for all $t \in I$.*

Proof. For a given $x(\cdot)$ we can differentiate equation (14b) w.r.t. t and obtain the following index 1 DAE

$$\dot{x} = F_{\mathcal{I}}(x, u)\theta_{\mathcal{I}},\ \dot{\mu} = -v, \tag{16a}$$

$$\begin{bmatrix} M_{\mathcal{I}}(x) & e \\ e^{\top} & 0 \end{bmatrix} \begin{bmatrix} \theta_{\mathcal{I}} \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{16b}$$

with the algebraic variables $\theta_{\mathcal{I}}$ and $v \in \mathbb{R}$. For a given initial condition $x(t_{\mathrm{s,n}})$, $\mu(t_{\mathrm{s,n}})$ can be directly computed from any component of (14b). Using the Schur complement and Assumption 3, we conclude that we can find unique $\theta_{\mathcal{I}}$ and v by solving the linear system (16b). Therefore, the DAE (14) can be

reduced to an ODE. Since the functions f_i are assumed to be Lipschitz the resulting ODE has a unique solution $x(t), t \in I$. □

A similar result, with a more complicated proof but different assumptions can be found in [46, Section 2]. Note that even though the DAE has a unique solution for a given active set $\mathcal{I}$, there might be multiple $\mathcal{I}$ that give a well-defined ODE, as we discuss in the subsequent sections. We do not know a priori whether we need to treat an ODE or a DAE, but for both cases, we will use Runge-Kutta methods within FESD to provide high-accuracy solutions. The crucial part of FESD is the automatic active set and switching time detection so that sliding modes and crossings of region boundaries can be treated in a unified way.

2.2.3 Active-set changes and continuity of λ and μ

Every active-set change in (8d) corresponds to crossing a discontinuity, entering or leaving a sliding mode, or a spontaneous leaving of a surface of discontinuity. These events in time are called *switches*.

From (3), Eq. (8b) and the complementarity conditions (8d) for $i \in \mathcal{I}(x)$ it follows that $\theta_i \geq 0$ and $\lambda_i = 0$. Likewise, for $i \notin \mathcal{I}(x)$ it follows that $\theta_i = 0$ and $\lambda_i \geq 0$. Hence, for $i \in \mathcal{I}(x)$ from (8b) and (5) we conclude that $\mu = \min_{j \in \mathcal{J}} g_j(x)$.

Lemma 5 *The functions $\lambda(t)$ and $\mu(t)$ in (8) are continuous in time.*

Proof: The function $\mu(t)$ is a minimum of continuous functions and is thus continuous. Therefore, continuity of $\lambda(t) = g(x(t)) - \mu(t)e$ follows from the continuity of $x(t)$ and $g(x)$ and Equation (8b). □

Remark 6 *Continuity of $\lambda(t)$ implies that at an active-set change of a component i at $t_{\mathrm{s},n+1}$ some $\lambda_i(t_{\mathrm{s},n+1})$ must be zero. Moreover, for some $i \notin \mathcal{I}_n$, in the case of crossing a discontinuity or entering a sliding mode with $i \in \mathcal{I}_{n+1}$, it holds that the left time derivative of λ_i is negative, i.e., $\dot{\lambda}_i(t^-_{\mathrm{s},n+1}) < 0$. Likewise, in the case of leaving a sliding mode or a spontaneous switch, with $i \in \mathcal{I}_n$ and $i \notin \mathcal{I}_{n+1}$, it follows that the right time derivative of λ_i is positive, i.e., $\dot{\lambda}_i(t^+_{\mathrm{s},n+1}) > 0$. If some of the first-order one-sided derivatives of $\lambda(\cdot)$ are zero at a switching point $t_{\mathrm{s},n+1}$, then one must look at higher-order derivatives to determine if it stays active or not.*

We exploit the continuity of $\lambda(\cdot)$ and $\mu(\cdot)$ later in the derivation of the FESD method. The next example discusses the difference between the possible switching cases.

Example 3 *There are four possible switching cases which we illustrate with the following examples:*

(a) crossing a surface of discontinuity, $\dot{x}(t) \in 2 - \mathrm{sign}(x(t))$,
(b) sliding mode, $\dot{x}(t) \in -\mathrm{sign}(x(t))$,
(c) leaving sliding mode $\dot{x}(t) \in -\mathrm{sign}(x(t)) + t$.
(d) spontaneous switch, $\dot{x}(t) \in \mathrm{sign}(x(t))$,

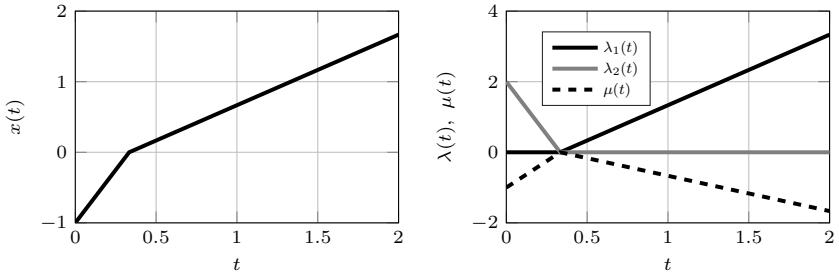

Fig. 2 Illustration of the arguments of Lemma 5 and Remark 6 on the Example 3 (a) and its corresponding DCS (11) for $t \in [0, 2]$ and $x(0) = -1$ with $t_\mathrm{s} = -\frac{1}{3}$. The functions $\mu = \min(-x, x)$, $\lambda_1 = x - \mu$ and $\lambda_2 = -x - \mu$ are continuous. At the switching point t_s we have $\dot{\lambda}_1(t_\mathrm{s}^-) = 0$, $\dot{\lambda}_1(t_\mathrm{s}^-) > 0$ and $\dot{\lambda}_2(t_\mathrm{s}^-) < 0$, $\dot{\lambda}_1(t_\mathrm{s}^-) = 0$.

In case (a), for $x(0) < 0$ the trajectory reaches $x = 0$ and crosses it. In example (b), for any finite $x(0)$, the trajectory reaches $x = 0$ and stays there. On the other hand, in example (c), for $x(0) = 0$ the DI has a unique solution and leaves $x = 0$ at $t = 1$. In the last example, for $x(0) = 0$, the DI has infinitely many solutions, and $x(t)$ can spontaneously leave $x = 0$ at any $t \geq 0$. Note that there is a qualitative difference between leaving a sliding mode (c) and spontaneous switch (d). The arguments of Lemma 5 and Remark 6 are illustrated in Figure 2 for the Example 3 (a).

2.2.4 Predicting the new active set

In this subsection, we restate a more technical result from [46] which is later needed in the convergence proof. The reader not interested in the proofs may skip this part.

As already noted in Remark 6, switches are characterized by the time derivative of $\lambda(\cdot)$. Note that, e.g., for crossing a discontinuity or entering a sliding mode at a switching point and for the subsequent interval it holds that $\mathcal{I}_n \subseteq \mathcal{I}_n^0$. Moreover, one can construct a Linear Complementarity Problem (LCP) with the data at $x(t_{\mathrm{s},n})$ and predict $\mathcal{I}_{n+1}$.

We define the vector $w_\mathcal{I}(t) \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}\lambda_\mathcal{I}(t) = M_\mathcal{I}(x(t))\theta_\mathcal{I}(t) - \dot{\mu}(t)e$. One can construct the following mixed LCP between $\dot{\lambda}_{\mathcal{I}_n^0}$ and $\theta_{\mathcal{I}_n^0}$ at $t_{\mathrm{s},n}$:

$$w_{\mathcal{I}_n^0} = M_{\mathcal{I}_n^0}(x)\theta_{\mathcal{I}_n^0} - \dot{\mu}e, \tag{17a}$$

$$1 = e^\top \theta_{\mathcal{I}_n^0}, \tag{17b}$$

$$0 \leq w_{\mathcal{I}_n^0} \perp \theta_{\mathcal{I}_n^0} \geq 0. \tag{17c}$$

For a sufficiently large $\alpha > 0$ all entries of the matrix $M_{\mathcal{I},\alpha}(x) = M_\mathcal{I}(x) + \alpha e e^\top$ are strictly positive. This means the matrix $M_{\mathcal{I},\alpha}(x)$ is strictly copositive, i.e., for any $a \geq 0, a \neq 0$ it holds that $a^\top M_{\mathcal{I},\alpha}(x)a > 0$ [20]. One can derive an LCP equivalent to (17) [46, Lemma 3.3]:

$$0 \leq \tilde{w}_{\mathcal{I}_n^0} = M_{\mathcal{I}_n^0,\alpha}(x)\tilde{\theta}_{\mathcal{I}_n^0} - e \perp \tilde{\theta}_{\mathcal{I}_n^0} \geq 0. \tag{18}$$

The motivation for rewriting (17) as (18) is twofold. It is both easier to prove solution existence and to compute a solution for an LCP with a strictly copositve matrix than for the initial mixed LCP [46]. The solution of the initial LCP $(w_{\mathcal{I}_n^0}, \theta_{\mathcal{I}_n^0})$ can be reconstructed via $\theta_{\mathcal{I}_n^0} = \tilde{\theta}_{\mathcal{I}_n^0}/e^\top \tilde{\theta}_{\mathcal{I}_n^0}$ and $w_{\mathcal{I}_n^0} = \tilde{w}_{\mathcal{I}_n^0}/e^\top \tilde{\theta}_{\mathcal{I}_n^0}$, for further details cf. [46, Lemma 3.3]. There is a one-to-one correspondence between the active set in a neighborhood of a switching point $t_{\mathrm{s},n}$ and the solutions of the LCP (18). This is summarized in the next theorem proved by Stewart [46].

Theorem 7 (Theorem 3.2 [46]) *Let $x(t)$ be a solution in the sense of Definition 1 for $t \in [t_a, t_b]$, with $\mathcal{I}^0 = \mathcal{I}(x(t_a))$ and $\mathcal{I} = \mathcal{I}(x(t))$ for all $t \in (t_a, t_b)$. Suppose Assumption 3 holds for all $t \in (t_a, t_b)$. Then for each $t \in (t_a, t_b)$ there is a solution of the LCP (18) such that*

$$\{i \mid \tilde{\theta}_i > 0\} \subseteq \mathcal{I} \subseteq \{i \mid \tilde{w}_i = 0\}.$$

Conversely, let $x_0 \in \mathbb{R}^{n_x}$ and t_a be given with $\mathcal{I}^0 = \mathcal{I}(x_0)$. Then if $(\tilde{w}_{\mathcal{I}^0}, \tilde{\theta}_{\mathcal{I}^0})$ is a solution of the LCP (18) such that

$$\{i \mid \tilde{\theta}_i > 0\} = \mathcal{I} = \{i \mid \tilde{w}_i = 0\}$$

and the conditions of Assumptions 3 are satisfied for $\nabla g_i(x)$ and $f_i(x,u)$, $i \in \mathcal{I}$, then there is a $t_b > t_a$ and a solution $x(\cdot)$ in the sense of Definition 1 on $[t_a, t_b]$ such that $x(t_a) = x_0$ and $\mathcal{I}(x(t)) = \mathcal{I}$ for all $t \in (t_a, t_b)$.

Regard an LCP

$$0 \leq M\theta + q \perp \theta \geq 0, \tag{19}$$

with $M \in \mathbb{R}^{l \times l}$ and $q \in \mathbb{R}^l$. The given LCP (19) is compactly denoted by $\mathrm{LCP}(M, q)$ and its set of solutions is denoted by $\mathrm{SOL}(M, q) \subseteq \mathbb{R}^l$. If a solution satisfies $(M\theta + q) + \theta > 0$, we say that strict complementarity holds.

To show convergence we will require the solutions of the LCP to be *strongly stable* [20,46]. A solution $(w^*, \theta^*) \in \mathrm{SOL}(M, q)$ of a given $\mathrm{LCP}(M, q)$ is said to be strongly stable if there is a neighborhood U of θ^* and a neighborhood V of the problem data $M \in \mathbb{R}^{l \times l}$ and $q \in \mathbb{R}^l$, such that the intersection of U with the solution set of an LCP constructed from the data from one point in V is a singleton. We state a regularity assumption about the LCP (18).

Assumption 8 *Consider a solution $x(t)$ in the sense of Definition 1 for $t \in [0, T]$, and let $\mathcal{S} = \{t_{\mathrm{s},0}, \ldots, t_{\mathrm{s},N_{\mathrm{sw}}}\}$ be the set of switching points. The solutions of the LCP (18) are strongly stable and satisfy strict complementarity for all $t \in [t_\mathrm{s} - \epsilon, t_\mathrm{s} + \epsilon] \cap [0, T]$, $t_\mathrm{s} \in \mathcal{S}$, for a sufficiently small $\epsilon > 0$.*

The strict complementarity assumption is needed to obtain a tight prediction of the next active set $\mathcal{I}$, cf. first part of Theorem 7. From the proof of [46, Theorem 3.2] it follows that the strict complementarity condition implies that the one-sided time derivatives of $\lambda_i(t), i \notin \mathcal{I}(x(t))$ are nonzero, see also Remark 6. Without this assumption, one can obtain only an over-approximation of $\mathcal{I}$.

However, it can be relaxed at the cost of looking at higher-order time derivatives of $\lambda_i(t_{\mathrm{s},n}), i \in \mathcal{I}_n^0$ and constructing an appropriate LCP for determining the active sets past some switching point, cf. [47, Section 4.2] for derivations. Note that the strict complementarity is needed only in a neighborhood of the switching points. Strong stability is assumed in order to apply some results on parametric LCPs. In our case, we will use it to draw the same conclusions from LCPs constructed at t and t', where t and t' are sufficiently close.

Example 4 *We briefly illustrate Theorem 7 on our example* $\dot{x} = 2 - \mathrm{sign}(x)$ *with* $x(0) = -1$*, cf. Figure 2. It is easy to see that* $t_{\mathrm{s},1} = -\frac{1}{3}$ *and that the relevant active sets are* $\mathcal{I}_0 = \{1\}$*,* $\mathcal{I}_1 = \{2\}$ *and* $\mathcal{I}_1^0(t_{\mathrm{s},1}) = \{1, 2\}$*. The LCP* (18) *for our example at* $t_{\mathrm{s},1}$ *reads as*

$$0 \leq \tilde{w}_{\mathcal{I}_1^0} = \begin{bmatrix} 3+\alpha & 1+\alpha \\ -3+\alpha & -1+\alpha \end{bmatrix} \tilde{\theta}_{\mathcal{I}_1^0} - e \perp \tilde{\theta}_{\mathcal{I}_1^0} \geq 0.$$

With $\alpha = 5$ *this LCP has the unique solution* $\tilde{\theta}_{\mathcal{I}_1^0} = (0, \frac{1}{4})$ *and* $\tilde{w}_{\mathcal{I}_1^0} = (\frac{1}{2}, 0)$ *and according to the last theorem it correctly predicts* $\mathcal{I}_1 = \{2\}$*.*

2.3 Remark on Cartesian products of Filippov systems

The reformulation from the last subsection given by the DCS (8) fails on some simple examples such as: $\dot{x}_1 \in -\mathrm{sign}(x_1),\ \dot{x}_2 \in -\mathrm{sign}(x_2)$, $x \in \mathbb{R}^2$. This example satisfies the one-sided Lipschitz condition and has a unique Filippov solution [23,48]. However, as shown in [48] at $(0, 0)$ the DAE arising from (8) fails to have a unique solution. One can see that $\dot{x}_1 \in -\mathrm{sign}(x_1)$ and $\dot{x}_2 \in -\mathrm{sign}(x_2)$ are completely independent and thus they should be treated in such a way.

Stewart introduced a generalization of his reformulation for such cases in [48]. One should identify the n_{sys} independent subsystems with index $k = 1, \ldots, n_{\mathrm{sys}}$, where each subsystem has n_f^k modes. We equip all variables related to the $k-$th subsystem with the superscript k. Instead of (3) one can write

$$\dot{x} \in \Big\{ \sum_{k=1}^{n_{\mathrm{sys}}} \sum_{i=1}^{n_f^k} \theta_i^k f_i^k(x, u) \ \mid \ \sum_{i=1}^{n_f^k} \theta_i^k = 1,\ \theta^k \geq 0, k = 1, \ldots, n_{\mathrm{sys}} \Big\}. \tag{20}$$

Finding the functions $g^k(\cdot) \in \mathbb{R}^{n_f^k}$ from $c^k(\cdot) \in \mathbb{R}^{n_f^k}$ for every subsystem works the same way as in Section 2.2.1. Thereby, the regions of every subsystem are defined via the matrix S^k and the switching functions $c^k(x) \in \mathbb{R}^{n_c^k}$. Every mode's convex combination is encoded by its parametric linear program (7), constructed with the k-th modes' switching functions $g^k(x) \in \mathbb{R}^{n_f^k}$. Thus, we can derive the DCS

$$\dot{x} = \sum_{k=1}^{n_{\mathrm{sys}}} F^k(x, u)\theta^k, \tag{21a}$$

$$0 = g^k(x) - \lambda^k - \mu^k e, \qquad \text{for all } k \in \{1, \dots n_{\text{sys}}\}, \tag{21b}$$
$$1 = e^\top \theta^k, \qquad \text{for all } k \in \{1, \dots n_{\text{sys}}\}, \tag{21c}$$
$$0 \leq \theta^k \perp \lambda^k \geq 0, \qquad \text{for all } k \in \{1, \dots n_{\text{sys}}\}, \tag{21d}$$

where $F^k(x,u) = [f_1^k(x,u), \dots, f_{n_f^k}^k(x,u)] \in \mathbb{R}^{n_x \times n_f^k}$ and $g^k(x) \in \mathbb{R}^{n_f^k}$, $\theta^k \in \mathbb{R}^{n_f^k}$, $\lambda^k \in \mathbb{R}^{n_f^k}$ and $\mu^k \in \mathbb{R}$, for all $k \in \{1, \dots, n_{\text{sys}}\}$. For ease of notation, in the remainder of the paper we treat the case with $n_{\text{sys}} = 1$, as all extensions are straightforward.

To the best of the authors' knowledge, there are no general conditions known which identify when the r.h.s. of (1) is partially separable as in (20) and there might even be multiple ways to write it in this form. However, in practice, it is usually easy to identify the structure of (20) by inspection. For example, this occurs if we have multiple surfaces with friction, or multiple objects touching the same frictional surface [48].

2.4 Sensitivities with respect to parameters and initial values

Correct calculation of derivatives of solutions w.r.t. parameters (e.g., discretized control functions) and initial values is crucial for efficient numerical optimal control algorithms and verifying the optimality of a solution. This is not straightforward for ODE with a discontinuous r.h.s., as the sensitivity usually exhibits jumps when switches occur. As any constant parameter $\hat{p}$ can be modeled via adding the state $\dot{p} = 0$ and $p(0) = \hat{p}$, we restrict our attention to sensitivities w.r.t. initial values.

Regard the DCS given by Eq. (8) on a time interval $[0, T]$ with the initial condition $x(0) = x_0$. Assume that the surface ∂R_j is reached at $t_s(x_0) \in (0, T)$ and that $x_0 \in R_i$. We consider the case where the solution crosses a co-dimension one surface of discontinuity ∂R_j. Other cases are where the trajectory: (a) slides on the surface of discontinuity after reaching it, (b) starts on a surface of discontinuity and stays on it or leaves it, or (c) goes from one surface to another. They can be analyzed with the same arguments as below, but we omit these cases here for brevity, cf. [23, Section 2.11].

In the case of crossing, we have for $t \in [0, t_s)$ that $\mathcal{I}(x(t)) = \{i\}$ and from (8) it follows that $\dot{x} = f_i(x)$. After crossing ∂R_j at t_s we have $\mathcal{I}(x(t)) = \{j\}$ for $t \in (t_s, T]$ and $\dot{x} = f_j(x)$. At t_s it holds that $\psi_{i,j}(x(t_s)) = 0$ with

$$\psi_{i,j}(x(t)) \coloneqq g_i(x(t)) - g_j(x(t)). \tag{22}$$

Thus, the system can be compactly represented by

$$\dot{x}(t) = \begin{cases} f_i(x(t)), & \psi_{i,j}(x(t)) < 0, \\ f_j(x(t)), & \psi_{i,j}(x(t)) \geq 0. \end{cases} \tag{23}$$

We are interested in the exact sensitivity matrix $X(t, 0; x_0) = \frac{\partial x(t; x_0)}{\partial x_0} \in \mathbb{R}^{n_x \times n_x}$ of a solution $x(t; x_0)$ of the system (23). The function $X(t, 0; x_0)$ obeys

smooth linear variational differential equations on both sides of t_s, but exhibits a jump at t_s [23]. The statement of the next proposition is adapted from [51, Section 3.3].

Proposition 9 *Regard the system* (23) *with* $x(0) = x_0 \in R_i$ *on an interval* $[0,T]$ *with a switch at* $t_\mathrm{s} \in (0,T)$. *Assume that the functions* $f_i(x)$, $f_j(x), \psi_{i,j}(x)$ *are continuously differentiable along* $x(t), t \in [0,T]$. *Assume the solution* $x(t)$ *reaches the surface of discontinuity transversally, i.e.,* $\nabla\psi_{i,j}(x(t_\mathrm{s}))^\top f_i(x(t_\mathrm{s})) > 0$. *Then the sensitivity* $X(T,0;x_0)$ *of a solution* $x(t;x_0)$ *of the system described by the ODE* (23) *is given by*

$$\begin{aligned} &X(T,0;x_0) = X(T,{t_\mathrm{s}}^+;x(t_\mathrm{s}))J(x(t_\mathrm{s};x_0))X({t_\mathrm{s}}^-,0;x_0) \text{ with} \\ &J(x(t_\mathrm{s};x_0)) \coloneqq I + \frac{(f_j(x(t_\mathrm{s};x_0)) - f_i(x(t_\mathrm{s};x_0)))\nabla\psi_{i,j}(x(t_\mathrm{s};x_0))^\top}{\nabla\psi_{i,j}(x(t_\mathrm{s};x_0))^\top f_i(x(t_\mathrm{s};x_0))}. \end{aligned} \tag{24}$$

This proposition can also be adapted to the case of sliding modes. We obtain similar expressions for the sensitivity jump formula as in (24). The only change needed to be made is to replace $f_j(x)$ with $f^*(x)$, where $f^*(x)$ defines the sliding vector field [22].

Since numerical sensitivities obtained via standard time-stepping methods fail to converge to their correct values (24) [40,51], artificial local minima arbitrarily close to the initialization point may arise in the context of optimization and impair the progress of the optimizer. This is resolved within FESD, where the convergence of the discrete-time sensitivities is recovered, cf. Section 4.5.

3 Finite Elements with Switch Detection

This section introduces the main algorithmic ingredients of the FESD method. The goal of the method is to: (a) detect exactly the time of reaching or leaving the region boundaries which is necessary for high accuracy of integration methods, (b) exactly compute the sensitivities across regions in order to correctly treat the nonsmoothness and (c) appropriately treat the possible evolution on the boundary that is present in sliding modes.

In this section, we regard a single control interval $[0,T]$ with a constant externally chosen control input $q \in \mathbb{R}^{n_u}$, i.e., we set $u(t) = q$ for $t \in [0,T]$. Extensions with more complex smooth parametrizations of the control function are straightforward.

3.1 Standard Runge-Kutta discretization

As a starting point in our analysis, we regard a standard Runge-Kutta (RK) discretization of the DCS (8). In the nonsmooth ODE community, these schemes are known as *time-stepping* methods. Opposed to *event-based/switch-detection* methods, they assume fixed step sizes h_n and do not try to detect the switches. As a consequence, they have in general only first-order accuracy [3].

The theoretical properties of RK methods for DI and DCS have been studied by many authors, e.g., [18,29,50,52].

Suppose the initial value $x(0) = s_0$ is given. We divide the control interval into N_{FE} *finite elements* (i.e., integration intervals) $[t_n, t_{n+1}]$ via the grid points $0 = t_0 < t_1 < \ldots < t_{N_{\mathrm{FE}}} = T$. On each of the finite elements we consider an n_{s}-stage Runge-Kutta method which is characterized by the Butcher tableau entries $a_{i,j}, b_i$ and c_i with $i, j \in \{1, \ldots, n_{\mathrm{s}}\}$ [26]. The fixed step size reads as $h_n = t_{n+1} - t_n,\ n = 0, \ldots, N_{\mathrm{FE}} - 1$. The approximation of the differential state at the grid points t_n is denoted by $x_n \approx x(t_n)$. We regard a *differential* representation of the Runge-Kutta method where the derivatives of states at the stage points $t_{n,i} \coloneqq t_n + c_i h_n,\ i = 1, \ldots, n_{\mathrm{s}}$, are degrees of freedom. For a single finite element, they are summarized in the vector $V_n \coloneqq (v_{n,1}, \ldots, v_{n,n_{\mathrm{s}}}) \in \mathbb{R}^{n_{\mathrm{s}} n_x}$. The stage values for the algebraic variables are collected in the vectors: $\Theta_n \coloneqq (\theta_{n,1}, \ldots, \theta_{n,n_{\mathrm{s}}}) \in \mathbb{R}^{n_{\mathrm{s}} \cdot n_f}$, $\Lambda_n \coloneqq (\lambda_{n,1}, \ldots, \lambda_{n,n_{\mathrm{s}}}) \in \mathbb{R}^{n_{\mathrm{s}} \cdot n_f}$ and $M_n \coloneqq (\mu_{n,1}, \ldots, \mu_{n,n_{\mathrm{s}}}) \in \mathbb{R}^{n_{\mathrm{s}}}$. We also define the vector $Z_n = (x_n, \Theta_n, \Lambda_n, M_n, V_n)$ which collects all internal variables. With x_n^{next} we denote the value at t_{n+1}, which is obtained after a single integration step. Finally, the RK equations for a single finite element for the DCS (8) are given by:

$$G_{\mathrm{rk}}(x_n^{\mathrm{next}}, Z_n, h_n, q) \coloneqq \begin{bmatrix} v_{n,1} - F(x_n + h_n \sum_{j=1}^{n_{\mathrm{s}}} a_{1,j} v_{n,j}, q)\theta_{n,1} \\ \vdots \\ v_{n,n_{\mathrm{s}}} - F(x_n + h_n \sum_{j=1}^{n_{\mathrm{s}}} a_{n_{\mathrm{s}},j} v_{n,j}, q)\theta_{n,n_{\mathrm{s}}} \\ G_{\mathrm{LP}}(x_n + h_n \sum_{j=1}^{n_{\mathrm{s}}} a_{1,j} v_{n,j}, \theta_{n,1}, \lambda_{n,1}, \mu_{n,1}) \\ \vdots \\ G_{\mathrm{LP}}(x_n + h_n \sum_{j=1}^{n_{\mathrm{s}}} a_{n_{\mathrm{s}},j} v_{n,j}, \theta_{n,n_{\mathrm{s}}}, \lambda_{n,n_{\mathrm{s}}}, \mu_{n,n_{\mathrm{s}}}) \\ x_n^{\mathrm{next}} - x_n - h_n \sum_{i=1}^{n_{\mathrm{s}}} b_i v_{n,i} \end{bmatrix} = 0. \tag{25}$$

Next, we summarize the equations for all N_{FE} finite elements over the whole interval $[0, T]$ in a discrete-time system manner. For this purpose, we introduce some additional shorthands. All variables of all finite elements for a single control interval are collected in the vectors $\mathbf{x} = (x_0, \ldots, x_{N_{\mathrm{FE}}}) \in \mathbb{R}^{(N_{\mathrm{FE}}+1) n_x}$, $\mathbf{V} = (V_0, \ldots, V_{N_{\mathrm{FE}}-1}) \in \mathbb{R}^{N_{\mathrm{FE}} n_{\mathrm{s}} n_x}$ and $\mathbf{h} \coloneqq (h_0, \ldots, h_{N_{\mathrm{FE}}-1}) \in \mathbb{R}^{N_{\mathrm{FE}}}$. Note that the simple continuity condition $x_{n+1} = x_n^{\mathrm{next}}$ holds. We collect all stage values of the Filippov multipliers in the vector $\mathbf{\Theta} = (\Theta_0, \ldots, \Theta_{N_{\mathrm{FE}}-1}) \in \mathbb{R}^{n_\theta}$ and $n_\theta = N_{\mathrm{FE}} n_{\mathrm{s}} n_f$. The vectors $\mathbf{\Lambda} \in \mathbb{R}^{n_\theta}$, $\mathbf{M} \in \mathbb{R}^{n_\mu}$ for the stage values of the Lagrange multipliers are defined accordingly, with $n_\mu = \frac{n_\theta}{n_f}$. The vector $\mathbf{Z} = (\mathbf{x}, \mathbf{V}, \mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{M}) \in \mathbb{R}^{n_{\mathbf{Z}}}$ collects all *internal* variables and $n_{\mathbf{Z}} = (N_{\mathrm{FE}} + 1) n_x + N_{\mathrm{FE}} n_{\mathrm{s}} n_x + 2 n_\theta + n_\mu$.

All computations over a single control interval which we call here the *standard discretization* are summarized in the following equations which resemble a discrete-time system:

$$s_1 = F_{\mathrm{std}}(\mathbf{Z}), \tag{26a}$$

$$0 = G_{\text{std}}(\mathbf{Z}, \mathbf{h}, s_0, q), \tag{26b}$$

where $s_1 \in \mathbb{R}^{n_x}$ is the approximation of $x(T)$ and

$$F_{\text{std}}(\mathbf{Z}) = x_{N_{\text{FE}}},$$
$$G_{\text{std}}(\mathbf{Z}, \mathbf{h}, s_0, q) \coloneqq \begin{bmatrix} x_0 - s_0 \\ G_{\text{rk}}(x_1, Z_0, h_0, q) \\ \vdots \\ G_{\text{rk}}(x_{N_{\text{FE}}}, Z_{N_{\text{FE}}-1}, h_{N_{\text{FE}}-1}, q) \end{bmatrix}.$$

Note that $\mathbf{h}$ are given parameters, implicitly fixed by the chosen discretization grid. It is usually impossible to obtain high-accuracy solutions with this method, as this can only happen if active-set changes occur coincidentally at t_n. Despite the high accuracy in this unlikely case, the numerical sensitivities would still be wrong [36,51]. When active-set changes happen within a finite element, the IRK method tries to approximate a nonsmooth trajectory by a smooth polynomial, cf. the left plot in Figure 3, which results in a poor approximation.

3.2 Algorithmic ingredients of the FESD method

To ensure high-accuracy solutions of FESD, we allow the optimization routine to vary the lengths h_n of the finite elements such that all switching points coincide with grid points t_n. Consequently, active-set changes cannot happen in the interior of each finite element, and smooth functions are approximated by smooth polynomials within a finite element, cf. the right plot in Figure 3. Thus, the active set $\mathcal{I}(x(t))$ changes its value only at some grid point t_n and is constant in the interior of all intervals (t_n, t_{n+1}). A key assumption in any event-based method is that there are finitely many switches in finite time. We also assume that there are enough finite elements to capture every switch that occurs in the time interval $[0, T]$.

3.2.1 The step sizes as degrees of freedom

To capture the switches with the discretization grid points t_n, the step sizes h_n are left to be degrees of freedom in the RK method in the remainder of this paper. Additionally, the condition $\sum_{n=0}^{N_{\text{FE}}-1} h_n = T$ ensures that we regard a time interval of unaltered length.

3.2.2 Cross complementarity

We want to prohibit active-set changes on stage points inside a finite element. To achieve this, next to the complementarity conditions for every stage point $0 = \Phi(\theta_{n,m}, \lambda_{n,m})$ we include additional conditions on the variables $\mathbf{\Theta}$ and

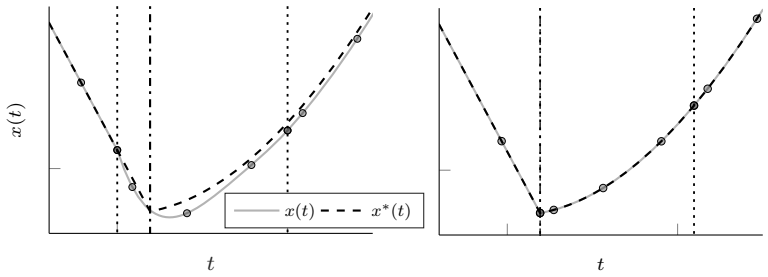

Fig. 3 Illustration of the analytic solution and a polynomial solution approximation to a PSS via an IRK Radau-IIA method of order 7. The left plot shows an approximation with a fixed step size where an active-set change happens on a stage point. The right plot shows an approximation obtained with FESD (based on the same IRK method) where the switch happens on the boundary. The circles represent the stage values, the vertical dotted lines the finite elements boundaries, and the vertical dashed line the switching time t_{s}.

Λ. These conditions ensure that the variable step size h_n adapts so that the switching times are indeed captured, as shown below.

For ease of exposition, we assume that the underlying RK scheme satisfies $c_{n_{\mathrm{s}}} = 1$ (e.g., Radau and Lobatto methods [26]). This means that the right boundary point of a finite element is a stage point, since $t_{n+1} = t_n + c_{n_{\mathrm{s}}} h_n$ for $c_{n_{\mathrm{s}}} = 1$. At the end of the section, we detail how to treat the case with $c_{n_{\mathrm{s}}} \neq 1$ (e.g., Gauss-Legendre methods).

Continuity of $\lambda(\cdot)$ and $\mu(\cdot)$. The boundary values of the approximation of $\lambda(\cdot)$ and $\mu(\cdot)$ on an interval $[t_n, t_{n+1}]$ play a crucial role in FESD. Therefore, we regard their values at t_n and t_{n+1} which are denoted by $\lambda_{n,0},\ \mu_{n,0}$ and $\lambda_{n,n_{\mathrm{s}}},\ \mu_{n,n_{\mathrm{s}}}$, respectively. We exploit the continuity of $\lambda(\cdot)$ and $\mu(\cdot)$ (cf. Lemma 5) and impose for their discrete-time counterparts for $n = 0, \ldots, N_{\mathrm{FE}} - 1$:

$$\lambda_{n,n_{\mathrm{s}}} = \lambda_{n+1,0},\ \mu_{n,n_{\mathrm{s}}} = \mu_{n+1,0}. \tag{27}$$

Therefore, in the sequel we use only the right boundary points $\lambda_{n,n_{\mathrm{s}}}$ and $\mu_{n,n_{\mathrm{s}}}$ which are degrees of freedom in the RK equations (26).

Moving the switching points to the boundary. Since $\lambda(\cdot)$ is continuous, on some interval (t_n, t_{n+1}) with a fixed active set $\mathcal{I}_n$, in the interior of the regarded interval its components are either zero or positive on the whole interval. The stage values $\lambda_{n,i}$ of the discrete-time counterpart should satisfy this property as well. This is achieved by the *cross complementarity* conditions, which read for all $n \in \{0, \ldots, N_{\mathrm{FE}} - 1\}$ as

$$0 = \mathrm{diag}(\theta_{n,m})\lambda_{n,m'},\ m \in \{1, \ldots, n_{\mathrm{s}}\},\ m' \in \{0, \ldots, n_{\mathrm{s}}\},\ m \neq m'. \tag{28}$$

Some of the appealing properties of the constraints (28) are given by the next lemma. In our notation $\theta_{n,m,i}$ is the i-th component of the vector $\theta_{n,m}$.

Lemma 10 *Regard a fixed $n \in \{0, \ldots, N_{\mathrm{FE}} - 1\}$ and a fixed $i \in \mathcal{J}$. If any $\theta_{n,m,i}$ with $m \in \{1, \ldots, n_{\mathrm{s}}\}$ is positive, then all $\lambda_{n,m',i}$ with $m' \in \{0, \ldots, n_{\mathrm{s}}\}$ must be zero. Conversely, if any $\lambda_{n,m',i}$ is positive, then all $\theta_{n,m,i}$ are zero.*

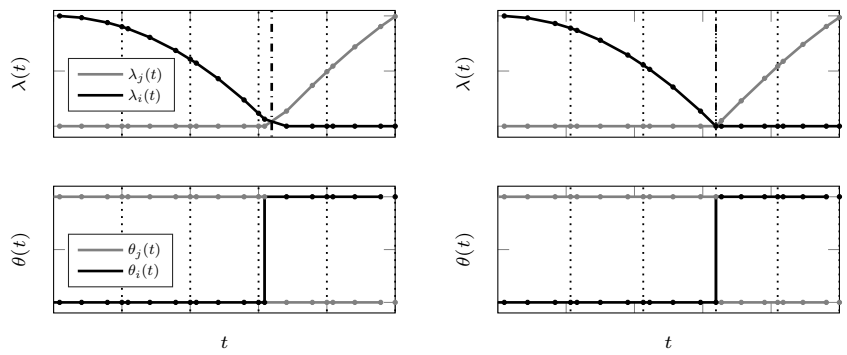

Fig. 4 An illustration of the standard complementarity conditions $\Psi(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = 0$ (left plot) and the standard complementarity conditions augmented by $0 = G_{\text{cross}}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$ (right plot). The dots represent the stage values. The vertical dotted line marks the finite element boundaries, and the vertical dashed line marks the switching time t_{s}. In the standard case (left plot), an active-set change can happen at any complementarity pair. With the cross complementarities (29) (right plot) an active-set change can only happen on the boundaries of a finite element.

Proof. Let $\theta_{n,m,i}$ be positive, and suppose $\lambda_{n,j,i} = 0$ and $\lambda_{n,k,i} > 0$ for some $k, j \in \{0, \dots, n_{\text{s}}\}, k \neq j$, then $\theta_{n,m,i}\lambda_{n,k,i} > 0$ which violates (28), thus all $\lambda_{n,m',i} = 0,\ m' \in \{0, \dots, n_{\text{s}}\}$. The converse is proven similarly. □

A consequence of this lemma is that at the boundary points t_{n+1} for active-set changes we have $\lambda_{n,n_{\text{s}},i} = \lambda_{n+1,n_{\text{s}},i} = 0$. This is important for the switch detection as we discuss below. The results of the last lemma is illustrated in Figure 4. Note that in contrast to the left plot illustrating the standard complementary conditions, in the right plot, all stage points inside a finite element have the same active set and on the finite element boundary we have $\lambda_{n,n_{\text{s}},i} = 0$.

Note that $\lambda_{0,0}$ and $\mu_{0,0}$ are not defined via Eq. (27), as we do not have a preceding finite element. However, they are crucial for the statement of the last lemma, especially, if the boundary point is the only stage point, as is the case for the implicit Euler method. This can be resolved by pre-computing $\lambda_{0,0}$ explicitly and using it in (28). Note that $\lambda_{0,0}$ is not a degree of freedom. Since x_0 is known, we obtain $\mu_{0,0} = \min_i g_i(x_0)$ and thus we have $\lambda_{0,0} = g(x_0) - \mu_{0,0}$.

The conditions (28) are given in their sparsest form. Due to the non-negativity of Λ_n and Θ_n there are many equivalent formulations of this condition, e.g., all conditions above can be summed up for a single finite element or even for all finite elements on the regarded control interval. Moreover, instead of the component-wise products in $\theta_{n,m}$ and $\lambda_{n,m'}$ we can use also inner products of these vectors. Thus, we use a more compact form of (28) where we combine the conditions for two neighboring finite elements. The motivation for this form is that we end up with the same number of new conditions as we

have new degrees of freedom by varying h_n. The conditions read as:

$$G_{\text{cross}}(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) := \begin{bmatrix} \sum_{m=1}^{n_\text{s}} \sum_{\substack{m'=0, \\ m' \neq m}}^{n_\text{s}} \theta_{0,m}^\top \lambda_{0,m'} + \sum_{m=1}^{n_\text{s}} \sum_{\substack{m'=0, \\ m' \neq m}}^{n_\text{s}} \theta_{1,m}^\top \lambda_{1,m'} \\ \vdots \\ \sum_{m=1}^{n_\text{s}} \sum_{\substack{m'=0, \\ m' \neq m}}^{n_\text{s}} \theta_{N_\text{FE}-2,m}^\top \lambda_{N_\text{FE}-2,m'} + \sum_{m=1}^{n_\text{s}} \sum_{\substack{m'=0, \\ m' \neq m}}^{n_\text{s}} \theta_{N_\text{FE}-1,m}^\top \lambda_{N_\text{FE}-1,m'} \end{bmatrix}. \tag{29}$$

Implicit switch detection. We briefly explain how the switch detection for the solution approximation is realized and formalize it later in Section 4. Note that for $x_n^{\text{next}} = x_{n+1}$ we have from the KKT conditions of the LP(x_{n+1}) (cf. Eq.(9)) that $\mu_{n,n_\text{s}} = \min_j g_j(x_{n+1})$. Moreover, if the active-set changes between the n-th and $n+1$-st finite element in the i-th component, then from Lemma 10 it follows that $\lambda_{n,n_\text{s},i} = 0$. Therefore, we obtain from (25) implicitly the condition

$$g_i(x_{n+1}) = \lambda_{n,n_\text{s},i} - \mu_{n,n_\text{s}},$$

which is equal to

$$0 = g_i(x_{n+1}) - g_j(x_{n+1}) = \psi_{i,j}(x_{n+1}), \tag{30}$$

where $\psi_{i,j}(x_{n+1}) = 0$ defines the switching surface between R_i and R_j. This condition forces h_n to adapt such that the switch is detected exactly. Note that the condition (30) appears only if active-set changes happen, hence the whole switch detection procedure is implicit.

3.2.3 Step equilibration

If no switches occur, i.e., the active sets $\mathcal{I}_n$ do not change between two neighboring finite elements, then the cross complementarity conditions in (29) are trivially satisfied. This yields spurious degrees of freedom in the step sizes h_n and the optimizer can adapt the grid in an undesirable way and harm the discretization accuracy. Also, the path-constraint discretization can be exploited unfavorably, just to decrease the objective value. To resolve this problem we introduce *step equilibration* conditions.

The step size should only change if a switch occurs and otherwise be constant. This results in a piecewise uniform discretization grid for the differential and algebraic states on the regarded control interval. To accomplish this, we derive an indicator function that is zero only if a switch occurs otherwise its value is strictly positive.

If some $\lambda_i(t_n)$ is equal to zero and its left or right time derivative is nonzero, then an active-set change has occurred. Instead of looking at the time derivatives, in the discrete-time case, we exploit the non-negativity of $\lambda_{n,m}$ and the fact that the active set is fixed for the whole finite element (due to cross

complementarity, cf. Lemma 10). For $n \in \{1, \dots, N_{\mathrm{FE}} - 1\}$, we define the following backward and forward sums of the stage values over the neighboring finite elements $[t_{n-1}, t_n]$ and $[t_n, t_{n+1}]$:

$$\sigma_n^{\lambda,\mathrm{B}} = \sum_{m=0}^{n_{\mathrm{s}}} \lambda_{n-1,m},\ \sigma_n^{\lambda,\mathrm{F}} = \sum_{m=0}^{n_{\mathrm{s}}} \lambda_{n,m}.$$

The components of $\sigma_n^{\lambda,\mathrm{B}}$ and $\sigma_n^{\lambda,\mathrm{F}}$ are zero if the left and right time derivatives of the corresponding components of $\lambda_{n,m}$ are zero. Likewise, they are positive when the left and right time derivatives are nonzero. Analogously, the sums for $\theta_{n,m}$ are defined as:

$$\sigma_n^{\theta,\mathrm{B}} = \sum_{m=1}^{n_{\mathrm{s}}} \theta_{n-1,m},\ \sigma_n^{\theta,\mathrm{F}} = \sum_{m=1}^{n_{\mathrm{s}}} \theta_{n,m}.$$

Additionally, we define the following vectors for all $n \in \{1, \dots, N_{\mathrm{FE}} - 1\}$:

$$\pi_n^{\lambda} = \mathrm{diag}(\sigma_n^{\lambda,\mathrm{B}})\sigma_n^{\lambda,\mathrm{F}},\ \pi_n^{\theta} = \mathrm{diag}(\sigma_n^{\theta,\mathrm{B}})\sigma_n^{\theta,\mathrm{F}}.$$

If there is an active-set change in the i-th complementarity pair, then at most one of the i-th components of $\sigma_n^{\lambda,\mathrm{B}}$ and $\sigma_n^{\lambda,\mathrm{F}}$ is nonzero, hence their product, i.e., the i-th component of π_n^{λ}, is zero. Due to complementarity, the same holds for π_n^{θ}. For sliding modes the corresponding components of π_n^{λ} are zero and of π_n^{θ} they are positive (due to complementarity). Thus, the i-th component of

$$\upsilon_n = \pi_n^{\lambda} + \pi_n^{\theta},$$

is only zero if there is an active-set change in the i-th complementarity pair at t_n. A function that has the desired properties is defined as:

$$\eta_n(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \coloneqq \prod_{i=1}^{n_f} (\upsilon_n)_i.$$

This scalar function summarizes the effects of all components. It is zero only if an active-set change happens at the boundary point t_n, otherwise, it is strictly positive. Finally, the constraints that remove possible spurious degrees of freedom in h_n read as:

$$0 = G_{\mathrm{eq}}(\mathbf{h}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \coloneqq \begin{bmatrix} (h_1 - h_0)\eta_1(\boldsymbol{\Theta},\boldsymbol{\Lambda}) \\ \vdots \\ (h_{N_{\mathrm{FE}}-1} - h_{N_{\mathrm{FE}}-2})\eta_{N_{\mathrm{FE}}-1}(\boldsymbol{\Theta},\boldsymbol{\Lambda}) \end{bmatrix}. \tag{31}$$

Since many products are involved in $\eta_n(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$, one can replace it by $\tilde{\eta}_n(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \coloneqq \tanh(\eta_n(\boldsymbol{\Theta}, \boldsymbol{\Lambda}))$ to have a better scaling. An example for step equilibration is studied in Subsection 5.3 numerically.

3.2.4 The FESD discretization

We have now all the ingredients to extend the standard RK discretization (26) to the *FESD discretization*. We use again the same discrete-time representation

$$s_1 = F_{\text{fesd}}(\mathbf{Z}), \tag{32a}$$
$$0 = G_{\text{fesd}}(\mathbf{Z}, \mathbf{h}, s_0, q, T), \tag{32b}$$

where $F_{\text{fesd}}(\mathbf{x}) = x_{N_{\text{FE}}}$ is the state transition map and $G_{\text{fesd}}(\mathbf{x}, \mathbf{h}, \mathbf{Z}, q, T)$ collects all other internal computations including all RK steps within the regarded control interval:

$$G_{\text{fesd}}(\mathbf{Z}, \mathbf{h}, s_0, q, T) := \begin{bmatrix} G_{\text{std}}(\mathbf{Z}, \mathbf{h}, s_0, q) \\ G_{\text{cross}}(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\ G_{\text{eq}}(\mathbf{h}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\ \sum_{n=0}^{N_{\text{FE}}-1} h_n - T \end{bmatrix}.$$

For a fixed control function q, horizon length T and initial value s_0, the formulation (32) can be used as an integrator with exact switch detection for PSS (1). Since Filippov DI does not always have unique solutions, one cannot expect uniqueness of solutions for their discrete-time counterparts (32) in all cases. In simulation methods, a common approach is to either pick one *local* solution obtained by the solver for the nonlinear complementarity problem (32) or to enumerate all possible solutions at an active-set change [5, 46]. In this paper, we consider only the first option. Note that in sliding modes, we implicitly obtain differential algebraic equations of index 2, cf. Section 2.2.2. To achieve good accuracy in practice it is usually required to use stiffly accurate methods, e.g., Radau-IIA methods [26].

3.2.5 Remark on RK methods with $c_{n_s} \neq 1$

We outline how to extend the FESD method when an RK scheme with $c_{n_s} \neq 1$ is regarded. In contrast to the developments so far, with $c_{n_s} \neq 1$ the variables λ_{n,n_s}, μ_{n,n_s} do not correspond the boundary values $\lambda(t_{n+1})$ and $\mu(t_{n+1})$ anymore (since $t_n + c_{n_s} h_n < t_{n+1}$). We denote the boundary points in this case by λ_{n,n_s+1}, μ_{n,n_s+1}. They are computed by solving $\text{LP}(x_{n+1})$ for $n = 0, \dots N_{\text{FE}} - 2$:

$$0 = G_{\text{LP}}(x_{n+1}, \theta_{n,n_s+1}, \lambda_{n,n_s+1}, \mu_{n,n_s+1}). \tag{33}$$

We still exploit the continuity of $\lambda(\cdot)$ and $\mu(\cdot)$ (cf. Lemma 5), by replacing (27) with the following continuity conditions for their discrete-time counterparts for $n = 0, \dots, N_{\text{FE}} - 1$:

$$\lambda_{n,n_s+1} = \lambda_{n+1,0},\ \mu_{n,n_s+1} = \mu_{n+1,0}. \tag{34}$$

With slight abuse of notation, we add the new variables $\theta_{n,n_s+1}, \lambda_{n,n_s+1}$ and μ_{n,n_s+1} to the vectors $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and $\mathbf{M}$, respectively. The vector $\mathbf{Z}$ is redefined

accordingly. The cross complementarity conditions are now modified such that next to the stage points we include the boundary points with the index $n_\mathrm{s}+1$:

$$G_{\mathrm{cross}}(\boldsymbol{\Theta},\boldsymbol{\Lambda}) \coloneqq \begin{bmatrix} \sum_{m=1}^{n_\mathrm{s}}\sum_{m'=1,m'\neq m}^{n_\mathrm{s}+1} \theta_{0,m}^\top \lambda_{0,m'} + \sum_{m=1}^{n_\mathrm{s}}\sum_{m'=0,m'\neq m}^{n_\mathrm{s}+1} \theta_{1,m}^\top \lambda_{1,m'} \\ \vdots \\ \sum_{m=1}^{n_\mathrm{s}}\sum_{\substack{m'=0,\\ m'\neq m}}^{n_\mathrm{s}+1} \theta_{N_\mathrm{FE}-2,m}^\top \lambda_{N_\mathrm{FE}-2,m'} + \sum_{m=1}^{n_\mathrm{s}}\sum_{\substack{m'=0,\\ m'\neq m}}^{n_\mathrm{s}} \theta_{N_\mathrm{FE}-1,m}^\top \lambda_{N_\mathrm{FE}-1,m'} \end{bmatrix}.$$

For the whole control time we have in total $(N_\mathrm{FE}-1)(2n_f+1)$ new variables.

4 Convergence theory

In this section we present the main convergence result of the FESD method. First, we prove that even though the FESD system (32) is always over-determined it still has a locally isolated solution. Second, we show that the numerical solution approximation $\hat{x}_h(\cdot)$ generated by FESD converges to a solution $x(\cdot)$ in the sense of Definition 1, with the same order that the underlying RK method has for smooth ODE. Additionally, we prove that the numerical sensitivities converge to their correct values with high accuracy.

4.1 Main assumptions

We start by introducing some notation and stating some assumptions related to the FESD formulation (32), which are important for our theoretical study in this section.

Assumption 11 *(Runge-Kutta method) A Butcher tableau with the entries $a_{i,j}, b_i$ and c_i, $i,j \in \{1,\dots,n_\mathrm{s}\}$ related to an n_s-stage Runge-Kutta (RK) method is used in the FESD* (32)*. Moreover, we assume that:*

(a) If the same RK method is applied to the differential algebraic equation (14) *on an interval $[t_a,t_b]$, it has a global accuracy of $O(h^p)$ for the differential states.*
(b) The RK equations applied to (14) *have a locally isolated solution for a sufficiently small $h_n > 0$.*

This assumption aims to consider a broad class of RK methods, and both assumptions are standard assumptions [26].

Assumption 12 *(Solution existence) For given parameters s_0, q and T, there exists a solution to the FESD problem* (32)*, such that for all $n \in \{0,\dots,N_\mathrm{FE}-1\}$ it holds that $h_n \geq 0$.*

This assumption means that there exists a solution and that we can compute it. If the FESD method is used in direct optimal control, non-negativity of the step sizes can easily be achieved by adding box constraints on h_n. This is the strongest assumption we make in this paper. Ideally, one would prove the existence of solutions. Since the system is over-determined this cannot be done straightforwardly by applying standard existence results [20]. As we will show below, in practice numerical solvers have no trouble computing such solutions.

We state a technical assumption that ensures regularity of the FESD problem (32).

Assumption 13 *(Regularity) Given the complementarity pairs* $\Psi(\theta_{n,m}, \lambda_{n,m}) = 0$*, for all* $n = 0, \dots N_{\mathrm{FE}} - 1$ *there exists an* $m \in \{1, \dots, n_{\mathrm{s}}\}$ *and* $i \in \{1, \dots, n_f\}$*, such that the strict complementarity property holds, i.e.,* $\theta_{n,m,i} + \lambda_{n,m,i} > 0$*. Moreover, for the RK equations* (25) *it holds for all* $n = 0, \dots N_{\mathrm{FE}} - 1$*, that at least one entry of the vector* $\nabla_{h_n} G_{\mathrm{rk}}(x_{n+1}, Z_n, h_n, q)$ *is nonzero.*

Once all stage values are computed by solving (32), we can use some interpolation method to construct the solution approximation candidate in continuous time, cf. Assumption 11. For example, if we use a collocation-based IRK method continuous-time approximation $\hat{x}_n(t; h_n)$ on every finite element is easily obtained via Lagrange polynomials [26]. We append the approximation for every finite element and write

$$\hat{x}_h(t) = \hat{x}_n(t; h_n) \text{ if } t \in [t_n, t_{n+1}], \tag{35}$$

where $h = \max_{n \in \{0, \dots N_{\mathrm{FE}} - 1\}} h_n$. Similarly, continuous-time representations can be found for the algebraic variables, and we denote them compactly as $\hat{\lambda}_h(t)$, $\hat{\theta}_h(t)$ and $\hat{\mu}_h(t)$. Similar to the definitions in Section 2.2.3, the fixed active set in this case is denoted by $\mathcal{I}(\hat{x}_h(t)) = \hat{\mathcal{I}}_n,\ t \in (\hat{t}_{\mathrm{s},n}, \hat{t}_{\mathrm{s},n+1})$ and the active set at switching point $\hat{t}_{\mathrm{s},n}$ by $\mathcal{I}(\hat{x}_h(\hat{t}_{\mathrm{s},n})) = \hat{\mathcal{I}}^0_n$.

4.2 Solutions of the FESD problem are locally isolated

In this subsection, we analyze some properties of solutions of the FESD problem (32). For the convenience of the reader, we restate the problem but discard the trivial state transition map $s_1 = F_{\mathrm{fesd}}(\mathbf{Z}) = x_{N_{\mathrm{FE}}}$:

$$G_{\mathrm{fesd}}(\mathbf{Z}, \mathbf{h}, s_0, q, T) = \begin{bmatrix} G_{\mathrm{std}}(\mathbf{Z}, \mathbf{h}, s_0, q, T) \\ G_{\mathrm{cross}}(\mathbf{\Theta}, \mathbf{\Lambda}) \\ G_{\mathrm{eq}}(\mathbf{h}, \mathbf{\Theta}, \mathbf{\Lambda}) \\ \sum_{n=0}^{N_{\mathrm{FE}}-1} h_n - T \end{bmatrix} = 0. \tag{36}$$

Recall that $\mathbf{Z} = (\mathbf{x}, \mathbf{V}, \mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{M}) \in \mathbb{R}^{n_{\mathbf{z}}}$. Additionally, we have that $G_{\mathrm{std}} : \mathbb{R}^{n_{\mathbf{z}}} \times \mathbb{R}^{N_{\mathrm{FE}}} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R} \to \mathbb{R}^{n_{\mathbf{z}}}$, $G_{\mathrm{cross}} : \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{N_{\mathrm{FE}}-1}$ and $G_{\mathrm{eq}} : \mathbb{R}^{n_{N_{\mathrm{FE}}}} \times \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{N_{\mathrm{FE}}-1}$. Finally, we have that $G_{\mathrm{fesd}} : \mathbb{R}^{n_{\mathbf{z}}} \times \mathbb{R}^{N_{\mathrm{FE}}} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R} \to \mathbb{R}^{n_{\mathbf{z}}+2N_{\mathrm{FE}}-1}$. Again, for ease of exposition, we regard $c_{n_{\mathrm{s}}} = 1$ and give the extensions later with $n_\theta = N_{\mathrm{FE}} n_{\mathrm{s}} n_f$ and $n_\mu = N_{\mathrm{FE}} n_{\mathrm{s}}$.

The vectors $s_0 \in \mathbb{R}^{n_x}$, $q \in \mathbb{R}^{n_u}$ and $T \in \mathbb{R}$ are given parameters, hence we have $n_{\mathbf{Z}} + N_{\mathrm{FE}}$ unknowns and $n_{\mathbf{Z}} + 2N_{\mathrm{FE}} - 1$ equations. Consequently, for $N_{\mathrm{FE}} > 1$, which we always assume in FESD, the system (36) is overdetermined. However, we show in the next theorem that for a given active set $N_{\mathrm{FE}} - 1$ equations in (36) are implicitly satisfied, and we always end up with a square system. As a consequence, Eq. (36) has under reasonable assumptions a locally unique solutions. Nevertheless, since we do not know the active set a priori, we can also not know which equations are binding and which are implicitly satisfied.

Lemma 14 (Corollary 6.1 in [34]) *Let $A_1 \in \mathbb{R}^{k\times m}$ and $A_2 \in \mathbb{R}^{m\times q}$, then*

$$\operatorname{rank}(A_1) + \operatorname{rank}(A_2) - m \leq \operatorname{rank}(A_1 A_2) \leq \min(\operatorname{rank}(A_1), \operatorname{rank}(A_2)).$$

Theorem 15 *Suppose that Assumptions 11, 12 and 13 hold. Let s_0, q_0 and $T > 0$ be some fixed parameters such that $G_{\mathrm{fesd}}(\mathbf{Z}^*, \mathbf{h}^*, s_0, q, T) = 0$. Let $P^* \subseteq \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}$ be the set of all parameters $(\hat{s}_0, \hat{q}, \hat{T})$ such that $\mathbf{Z} \in \mathbb{R}^{n_{\mathbf{Z}}}$, which is the solution of $G_{\mathrm{fesd}}(\mathbf{Z}, \mathbf{h}, \hat{s}_0, \hat{q}, \hat{T}) = 0$, has the same active set as $\mathbf{Z}^*$. Additionally, suppose that $G_{\mathrm{fesd}}(\cdot)$ is continuously differentiable in s_0, q and T for all $(s_0, q, T) \in P^*$. Then there exists a neighborhood $P \subseteq P^*$ of (s_0, q_0, T) such that there exist continuously differentiable single valued functions $\mathbf{Z}^* : P \to \mathbb{R}^{n_{\mathbf{Z}}}$ and $\mathbf{h}^* : P \to \mathbb{R}^{N_{\mathrm{FE}}}$.*

Proof. We regard the active sets for every finite element $\hat{\mathcal{I}}_n$ for all $n \in \{0 \dots, N_{\mathrm{FE}} - 1\}$ that correspond to the solution $(\mathbf{Z}^*, \mathbf{h}^*)$. First, we look closer at the equations $G_{\mathrm{cross}}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$ and $G_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$. If two neighboring finite elements have the same active set, i.e., $\hat{\mathcal{I}}_n = \hat{\mathcal{I}}_{n+1}$, then the $(n+1)$-th entry of $G_{\mathrm{cross}}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*)$ is implicitly satisfied due to the point-wise complementarity conditions $\Psi(\Theta_n, \Lambda_n) = 0$ and $\Psi(\Theta_{n+1}, \Lambda_{n+1}) = 0$. Moreover, by construction we have $\eta_{n+1}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) > 0$ and the $(n+1)$-th entry of $G_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*, T) = 0$ is binding, i.e., it implies $h^*_{n+1} = h^*_n$. On the other hand, if $\hat{\mathcal{I}}_n \neq \hat{\mathcal{I}}_{n+1}$, we have by construction that $\eta_{n+1}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$ and then $(n+1)$-th entry of $G_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*, T) = 0$ vanishes, i.e., is satisfied for any h^*_n and h^*_{n+1}. However, the $(n+1)$-th entry of $G_{\mathrm{cross}}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$ is now binding, cf. Lemma 10.

We collect the binding n_1 cross complementarity conditions, with $0 \leq n_1 \leq N_{\mathrm{FE}} - 1$, in the equation $G^*_{\mathrm{cross}}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$, and the $N_{\mathrm{FE}} - 1 - n_1$ implicitly satisfied into $G^{\mathrm{res}}_{\mathrm{cross}}(\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$. Likewise, we collect the binding n_2 step equilibration conditions, with $1 \leq n_2 \leq N_{\mathrm{FE}} - 1$, in $G^*_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$. The remaining $N_{\mathrm{FE}} - 1 - n_2$ conditions are implicitly satisfied and are collected in $G^{\mathrm{res}}_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$. Note that $n_1 + n_2 = N_{\mathrm{FE}} - 1$. We highlight that $\sum_{n=0}^{N_{\mathrm{FE}}-1} h_n - T$ is always binding.

We can further simplify our system of equations by eliminating some degrees of freedom using $G^*_{\mathrm{eq}}(\mathbf{h}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*) = 0$. All components of this vector are of the form $\eta_n(h_n - h_{n+1})$ with $\eta_n > 0$. Therefore, we have n_2 equations of the form of $h_n = h_{n+1}$ and can remove n_2 degrees of freedom.

Furthermore, we can express any $h_j = T - \sum_{i=0, i\neq j}^{N_{\mathrm{FE}}-1} h_n$ and remove another degree of freedom. In total we removed $n_2 + 1$ degrees of freedom and can regard a reduced number of unknown step-sizes, which we denote by $\tilde{\mathbf{h}}^* \in \mathbb{R}^{n_1}, n_1 = N_{\mathrm{FE}} - n_2 - 1$. With a slight abuse of notation, we redefine the standard RK equations accordingly and obtain $G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T) = 0$ with $G_{\mathrm{std}} : \mathbb{R}^{n_\mathbf{Z}} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R} \to \mathbb{R}^{n_\mathbf{Z}}$.

To summarize, for a fixed active set we can rewrite (36) in a reduced form as

$$G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T) \coloneqq \begin{bmatrix} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T) \\ G^*_{\mathrm{cross}}(\mathbf{\Theta}^*, \mathbf{\Lambda}^*) \end{bmatrix} = 0, \tag{37}$$

with $G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \mathbf{h}^*, s_0, q, T) \in \mathbb{R}^{n_\mathbf{Z}+n_1}$. These conditions imply

$$G^{\mathrm{res}}_{\mathrm{fesd}}(\mathbf{h}^*, \mathbf{\Theta}^*, \mathbf{\Lambda}^*) \coloneqq \begin{bmatrix} G^{\mathrm{res}}_{\mathrm{cross}}(\mathbf{\Theta}^*, \mathbf{\Lambda}^*) \\ G^{\mathrm{res}}_{\mathrm{eq}}(\mathbf{h}^*, \mathbf{\Theta}^*, \mathbf{\Lambda}^*) \end{bmatrix} = 0, \tag{38}$$

with $G^{\mathrm{res}}_{\mathrm{fesd}}(\mathbf{h}^*, \mathbf{\Theta}^*, \mathbf{\Lambda}^*) \in \mathbb{R}^{N_{\mathrm{FE}}-1}$. Thus, for a given active set we can discard (38) and regard only the equivalent reduced problem (37), which is a square system of equations.

Next, we show that the Jacobian matrix $\nabla_{(\mathbf{Z},\tilde{\mathbf{h}})} G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top$ has full rank. This enables us to apply the implicit function theorem (cf. [19, Theorem 1B.1]) and establish the result of this theorem. We take a closer look at the matrix:

$$\begin{aligned} &\nabla_{(\mathbf{Z},\tilde{\mathbf{h}})} G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top \\ &= \begin{bmatrix} \nabla_\mathbf{Z} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top & \nabla_{\tilde{\mathbf{h}}} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top \\ \nabla_\mathbf{Z} G^*_{\mathrm{cross}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top & \nabla_{\tilde{\mathbf{h}}} G^*_{\mathrm{cross}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top \end{bmatrix}. \end{aligned}$$

Under Assumption 12, for a fixed active set and a fixed h_n^* the equation $G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T) = 0$ boils down to the RK equations for the differential algebraic equation (14). Due to Assumption 11 the RK system $G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T) = 0$ has a locally isolated solution. A necessary and sufficient condition for this property is the invertibility of the Jacobian $\nabla_\mathbf{Z} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top$ [19, Theorem 1B.8]. Thus, we have that $\operatorname{rank}(\nabla_\mathbf{Z} G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top) = n_\mathbf{Z}$. Second, due to the block diagonal structure of $\nabla_{\tilde{\mathbf{h}}} G^*_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)$ and Assumption 13 we can deduce that $\operatorname{rank}(\nabla_{\tilde{\mathbf{h}}} G^*_{\mathrm{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top) = n_1$. Third, due to the nonnegativity of $(\mathbf{\Theta}, \mathbf{\Lambda})$ and Assumption 13 by direct computation it can be verified that $\operatorname{rank}(\nabla_\mathbf{Z} G^*_{\mathrm{cross}}(\mathbf{\Theta}, \mathbf{\Lambda})^\top) = n_1$ and $\nabla_{\tilde{\mathbf{h}}} G^*_{\mathrm{cross}}(\mathbf{\Theta}, \mathbf{\Lambda})^\top = 0$.

We introduce more compact notation and summarize the results so far with:

- $M_1 = \nabla_\mathbf{Z} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top \in \mathbb{R}^{n_\mathbf{Z} \times n_\mathbf{Z}}$ with $\operatorname{rank}(M_1) = n_\mathbf{Z}$
- $M_2 = \nabla_{\tilde{\mathbf{h}}} G_{\mathrm{std}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top \in \mathbb{R}^{n_\mathbf{Z} \times n_1}$ with $\operatorname{rank}(M_2) = n_1$ and
- $M_3 = \nabla_\mathbf{Z} G_{\mathrm{cross}}(\mathbf{\Theta}, \mathbf{\Lambda})^\top \in \mathbb{R}^{n_1 \times n_\mathbf{Z}}$ with $\operatorname{rank}(M_3) = n_1$.

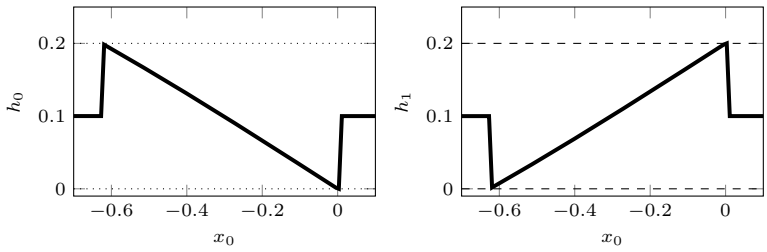

Fig. 5 Illustration of the discontinuity of the solution map of (36) for the PSS $\dot{x} \in 2 - \text{sign}(x) + x^2$ for $T = 0.2$ and $N_{\text{FE}} = 2$.

To show that $\nabla_{(\mathbf{Z},\tilde{\mathbf{h}})} G^*_{\text{fesd}}(\mathbf{Z}^*, \tilde{\mathbf{h}}^*, s_0, q, T)^\top$ has a rank of $n_{\mathbf{Z}} + n_1$, we show that the linear system

$$\begin{bmatrix} M_1 & M_2 \\ M_3 & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = 0,$$

with $v \in \mathbb{R}^{n_{\mathbf{Z}}}$ and $w \in \mathbb{R}^{n_1}$ has zero as the only solution.

From the first line in this linear system, we have that $v = -M_1^{-1} M_2 w$. Since $n_{\mathbf{Z}} > n_1$, from Lemma 14, we conclude that $\text{rank}(M_1^{-1} M_2) = n_1$. Next, from the second part of our linear system, we have that $-M_3 M_1^{-1} M_2 w = 0$. Again, using Lemma 14, we conclude that $\text{rank}(M_3 M_1^{-1} M_2) = n_1$. Hence, we have $w = 0$ and $v = 0$ to be the only solution of the regarded linear system. This completes the proof. □

Remark 16 *We note that one cannot apply more general forms of implicit function theorems for generalized and nonsmooth equations [19]. They usually require Lipschitz continuity of the solution map to reason about local uniqueness, but the solution map for FESD is not continuous in general, but only piecewise continuous.*

Example 5 *To illustrate the discontinuity of the solution map, we look at the example of* $\dot{x} \in 2 - \text{sign}(x) + x^2$, *with* $N_{\text{FE}} = 2$, $T = 0.2$ *and vary* $x_0 \in [-0.7, 0.1]$. *A solution approximation is obtained via FESD based on the Radau-IIA method of order 3. Consider an initial value* x_0 *such that no switch occurs and a perturbed initial value* $x_0 + \epsilon$ *where a single switch occurs on the time interval of interest. Clearly, in the first case, we have an equidistant grid with* $h_0 = h_1$, *and in the second case* h_0 *jumps to* $\hat{t}_{\text{s},1}$. *We conclude that* $h_0(x_0)$ *is not a Lipschitz function, see Figure 5 for an illustration.*

4.2.1 Extension for the case of $c_{n_\text{s}} \neq 1$

In the case of $c_{n_\text{s}} \neq 1$, we need to solve the additional LP (33) to obtain the boundary points. Note that if we have $g_i(x_{n+1}) = \min_{j \in \mathcal{J}} g_j(x_{n+1})$ for more than one i, the variables $\theta_{n,n_\text{s}+1}$ are not unique and the LP(x_{n+1}) has infinitely many solutions. However, these variables are neither used in the cross

complementarities nor step equilibration. Therefore, we can discard θ_{n,n_s+1} and simplify (33) to:

$$\begin{aligned}\lambda_{n,n_s+1} &= g(x_{n+1}) - \mu_{n,n_s+1}e,\\ \lambda_{n,n_s+1} &\geq 0,\end{aligned}$$

which has n_f+1 unknowns and n_f equalities and n_f inequalities for a given x_{n+1}. Now suppose that the first m_1 components of λ_{n,n_s+1} are zero (e.g., implied by cross complementarity) and the remaining m_2 are strictly positive, with $m_1+m_2=n_f$. We have that

$$g_i(x_{n+1}) = \mu_{n,n_s+1},\ i=1,\dots,m_1, \tag{39a}$$
$$\lambda_{n,n_s+1,i} = 0,\ i=1,\dots,m_1, \tag{39b}$$
$$\lambda_{n,n_s+1,j} = g_j(x_{n+1}) - \mu_{n,n_s+1},\ j=n_f-m_2+1,\dots,n_f. \tag{39c}$$

As the first m_1 relations all assign the same value to μ_{n,n_s+1}, we can discard m_1-1 of them and thus we end up with a system of $m_1+m_2+1=n_f+1$ equations and n_f+1 unknowns. This system has still the important property that $\mu_{n,n_s+1}=\min_i g_i(x_{n+1})$. With this simplification for an RK method with $c_{n_s}\neq 1$ we have $n_\theta = N_{\mathrm{FE}}n_s n_f$, $n_\lambda = n_\theta + (N_{\mathrm{FE}}-1)n_f$. The new variables are determined by the square linear system (39). Hence, it is straightforward to extend Theorem 15 for the case of $c_{n_s}\neq 1$.

4.3 Convergence and order of FESD

In this subsection we prove that under reasonable assumptions the sequence of approximations $\hat{x}_h(\cdot)$ generated by the FESD method converges with high order to a solution of (1) in the sense of Definition 1. Recall that $h=\max_{n\in\{0,\dots N_{\mathrm{FE}}-1\}} h_n$. The proof is inspired by the proof of Theorem 4.3 in [46]. We consider also $t_{s,0}=0$ as a switching point, since at this time point the active set for the first interval $(t_{s,0},t_{s,1})$ is determined.

Note that for generating solution approximations with FESD it is sufficient to consider only two finite elements at a time, i.e., $N_{\mathrm{FE}}=2$ in Eq. (36), and then to append the solutions in order to construct $\hat{x}_h(t),\ t\in[0,T]$ via Eq. (35). This requires of course to have only one switch in the regarded time interval, which can always be achieved with a sufficiently small h. We define the set of all discretization grid points as $\mathcal{G}=\{t_0,t_1,\dots,t_{N_{\mathrm{FE}}}\}$. We treat the cases of crossing a discontinuity or entering a sliding mode, i.e., the case of $\mathcal{I}_n\subset\mathcal{I}^0_{n+1}$ and $\mathcal{I}_{n+1}\subseteq\mathcal{I}^0_{n+1}$.

Theorem 17 *Suppose that $x(t)$ is a solution of (1) in the sense of Definition 1 for $t\in[0,T]$ with $x(0)=x_0$. Suppose the following is true:*

(a) the Assumptions 3 and 8 are satisfied and $\mathcal{I}_n\subset\mathcal{I}^0_{n+1}$ and $\mathcal{I}_{n+1}\subseteq\mathcal{I}^0_{n+1}$ holds for all $n=0,\dots,N_{\mathrm{sw}}$,
(b) the Assumption 11, 12 and 13 hold for the FESD problem (32).

Then $x(\cdot)$ is a limit point of the sequence of approximations $\hat{x}_h(\cdot)$, defined in Eq. (35) *as $h \downarrow 0$. Moreover, for sufficiently small $h > 0$, the solution of* (32) *generates a solution approximation $\hat{x}_h(t)$ on $[0, T]$ such that:*

$$|\hat{t}_{\mathrm{s},n} - t_{\mathrm{s},n}| = O(h^p) \text{ for every } n = 0, \dots, N_{\mathrm{sw}}, \tag{40a}$$

$$\|\hat{x}_h(t_n) - x(t_n)\| = O(h^p), \text{ for all } t_n \in \mathcal{G}. \tag{40b}$$

Proof. The proof will be carried out by induction, where we consider the switching events $n = 0, \dots, N_{\mathrm{sw}}$ and the corresponding time intervals $(t_{\mathrm{s},n}, t_{\mathrm{s},n+1})$, with a slight abuse of notation where $t_{\mathrm{s},N_{\mathrm{sw}}+1} = T$ is not necessarily a switching point. Regard $n = 0$, where we have trivially that $t_0 = 0$, thus

$$|\hat{t}_0 - t_0| = 0 = O(h^p), \ \|\hat{x}_h(0) - x(0)\| = 0 = O(h^p).$$

Moreover, $\mathcal{I}(x_0) = \mathcal{I}(\hat{x}_h(0)) = \mathcal{I}_0^0$.

Now we suppose (40) is true for n, i.e., at $t = t_{\mathrm{s},n}$. We show that the same statements are true for $n + 1$. By the induction hypothesis and due to continuity of $g_i(x), i \in \mathcal{J}$, we have that for sufficiently small h the equality $\mathcal{I}(\hat{x}(\hat{t}_{\mathrm{s},n})) = \mathcal{I}(x(t_{\mathrm{s},n})) = \mathcal{I}_n^0$ holds. Moreover, by Lipschitz continuity of $f_i(x)$ and $\nabla g_i(x), i \in \mathcal{J}$, it follows that (cf. Section 2.2.4)

$$M_{\mathcal{I}_n^0}(\hat{x}_h(\hat{t}_{\mathrm{s},n})) \to M_{\mathcal{I}_n^0}(x(t_{\mathrm{s},n})) \text{ as } h \downarrow 0.$$

According to Theorem 7 the solution of the LCP (18) corresponding to $M_{\mathcal{I}_n^0}(x(t_{\mathrm{s},n}))$ determines the new index set $\mathcal{I}_n = \{i \in \mathcal{I}_n^0 \mid \tilde{\theta}_i > 0\}$. Moreover, by Assumption 8 this LCP is strongly stable and due to Lemma 20 (cf. Appendix A), for sufficiently small $h > 0$ the solution of the LCP corresponding $M_{\mathcal{I}_n^0}(\hat{x}_h(\hat{t}_{\mathrm{s},n}))$ has a solution such that $\hat{\mathcal{I}}_n = \{i \mid \tilde{\theta}_i > 0, i \in \mathcal{I}_n^0\} = \{i \mid \theta_i > 0, i \in \mathcal{I}_n^0\} = \mathcal{I}_n$. Thus, we conclude that both the solution approximation and the solution *predicted* the same active set $\mathcal{I}_n$ in a neighborhood of $\hat{t}_{\mathrm{s},n}$ and $t_{\mathrm{s},n}$, respectively.

It is left to verify that such an active set $\mathcal{I}_n$ predicted by the solution approximation is indeed feasible for the FESD problem. Note that by the induction hypothesis and the reasoning above the solution approximation and $x(\cdot)$ have the same corresponding active set in a neighborhood of $t_{\mathrm{s},n}$. For a fixed active set, as a consequence of Proposition 4 the arising DAE (14) has a unique solution. Finally, under this setting with the given active sets in a neighborhood of $t_{\mathrm{s},n}$, according to Theorem 15 there is a locally unique solution to a FESD problem, thus we can construct an appropriate $\hat{x}_h(\cdot)$.

Note that one can make arbitrarily many integration steps with a fixed $\mathcal{I}_n$ before the next switch in time is reached. Again, due to Theorem 15 the corresponding FESD problem has a locally unique solution.

Now we provide an error estimate for the solution approximation until the next switching point. First, we define $\tilde{x}(t)$ to be the exact *extended* solution of the DAE (14) with the fixed active set $\mathcal{I}_n$ on the interval $t \in [t_{\mathrm{s},n}, \tilde{t}]$, with $\tilde{x}(t_{\mathrm{s},n}) = x(t_{\mathrm{s},n})$ and $\tilde{t} > t_{\mathrm{s},n+1}$. Obviously, it holds that $x(t) = \tilde{x}(t)$ for all $t \in [t_{\mathrm{s},n}, t_{\mathrm{s},n+1}]$. Second, from the discussions in Section 3.2.2 we know that

active-set changes can only happen at boundaries of the finite elements, thus it holds that $\hat{t}_{\mathrm{s},n} \in \mathcal{G}$ for all $n = 0, \ldots, \hat{N}_{\mathrm{sw}}$. Third, by the induction hypothesis we have $\|\hat{x}_h(\hat{t}_{\mathrm{s},n}) - x(\hat{t}_{\mathrm{s},n})\| = O(h^p)$. As noted above, for a fixed active set $\mathcal{I}_n$ and fixed h_n the FESD equations boil down to standard RK equations applied to (14). Thus, from Assumption 11 we have the estimate

$$\|\hat{x}_h(\hat{t}_{\mathrm{s},n+1}) - \tilde{x}(\hat{t}_{\mathrm{s},n+1})\| = O(h^p). \tag{41}$$

With the help of this estimate, in the next few steps we prove that $|\hat{t}_{\mathrm{s},n+1} - t_{\mathrm{s},n+1}| = O(h^p)$. It is assumed that we regard crossing a discontinuity or entering a sliding mode, i.e., the case of $\mathcal{I}_n \subset \mathcal{I}(x(t_{\mathrm{s},n+1}))$ and $\mathcal{I}_{n+1} \subseteq \mathcal{I}(x(t_{\mathrm{s},n+1}))$. We need to distinguish the two scenarios: I. $\hat{t}_{\mathrm{s},n+1} > t_{\mathrm{s},n+1}$ and II. $\hat{t}_{\mathrm{s},n+1} \leq t_{\mathrm{s},n+1}$.

Case I. Regard the following indices $j \in \mathcal{I}_n$ and $i \in \mathcal{I}(x(t_{\mathrm{s},n+1})) \setminus \mathcal{I}_n$. This means that $\min_k g_k(x((t_{\mathrm{s},n+1})) = g_i(x(t_{\mathrm{s},n+1})) = g_j(x(t_{\mathrm{s},n+1})) = \mu(t_{\mathrm{s},n+1})$ holds and one can locally regard the following *switching* function

$$\psi_{i,j}(x(t)) = g_i(x(t)) - g_j(x(t)) = \lambda_i(t) - \lambda_j(t).$$

Note that this function is Lipschitz continuous. It must by definition become zero when an active-set change happens.

Due to the strict complementarity assumed in Assumption 8 (see also part 9 of the proof of [46, Theorem 4.3], and the remarks after Assumption 8) we have at $t^-_{\mathrm{s},n+1}$ that $\dot{\lambda}_j(t^-_{\mathrm{s},n+1}) = 0$ and $\dot{\lambda}_i(t^-_{\mathrm{s},n+1}) < 0$. Therefore, it holds that:

$$\psi_{i,j}(x(t_{\mathrm{s},n+1})) = 0, \ \frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(t^-_{\mathrm{s},n+1})) < 0. \tag{42}$$

Obviously, the same assertion holds for $\tilde{x}(t)$. Moreover, due to the smoothness of $\tilde{x}(t)$, we have $\psi_{i,j}(\tilde{x}(t)) < 0$ for $t \in (t_{\mathrm{s},n+1}, t_{\mathrm{s},n+1} + \epsilon)$ for some $\epsilon > 0$.

Similarly, for the solution approximation we have $\psi_{i,j}(\hat{x}_h(t)) = g_i(\hat{x}_h(t)) - g_j(\hat{x}_h(t))$. Since $\hat{t}_{\mathrm{s},n+1} > t_{\mathrm{s},n+1}$, due to continuity of $\psi_{i,j}(\cdot)$ and $\hat{x}_h(\cdot)$ it follows that

$$\psi_{i,j}(\hat{x}_h(t_{\mathrm{s},n+1})) > 0 \text{ and } \frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(\hat{x}_h(t_{\mathrm{s},n+1})) < 0.$$

Now from Lipschitz continuity of $\psi_{i,j}(\cdot)$ and (41) we can establish that

$$\begin{aligned} |\underbrace{\psi_{i,j}(\hat{x}_h(\hat{t}_{\mathrm{s},n+1}))}_{=0} - \underbrace{\psi_{i,j}(\tilde{x}(\hat{t}_{\mathrm{s},n+1}))}_{<0}| &\leq L_\psi \|\hat{x}_h(\hat{t}_{\mathrm{s},n+1}) - \tilde{x}(\hat{t}_{\mathrm{s},n+1})\|, \\ |\psi_{i,j}(\tilde{x}(\hat{t}_{\mathrm{s},n+1}))| &= O(h^p). \end{aligned} \tag{43}$$

Note that in contrast to $\psi_{i,j}(x(t))$, the function $\psi_{i,j}(\tilde{x}(t))$ is smooth in a neighborhood of $t_{\mathrm{s},n+1}$. Thus, we look at the first-order Taylor approximation of $\psi_{i,j}(\cdot)$ at $\tilde{x}(t_{\mathrm{s},n+1})$.

$$\psi_{i,j}(\tilde{x}(t)) = \psi_{i,j}(\tilde{x}(t_{\mathrm{s},n+1})) + \frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(\tilde{x}(t_{\mathrm{s},n+1}))(t - t_{\mathrm{s},n+1}) + o(|t - t_{\mathrm{s},n+1}|).$$

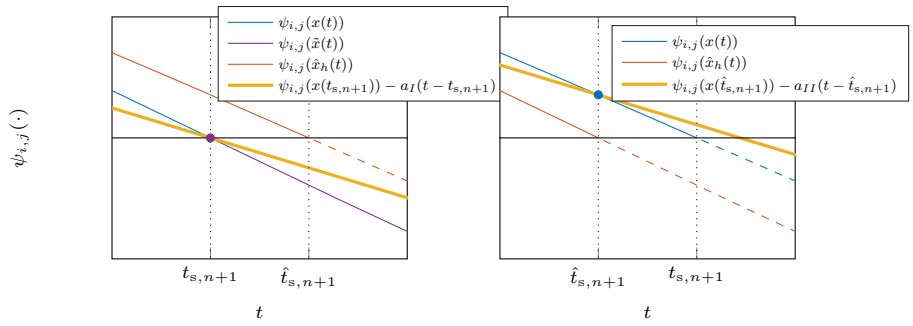

Fig. 6 The left plots shows an illustration of the argument of Case I: $\hat{t}_{s,n+1} > t_{s,n+1}$ and the right plot shows an illustration of the argument of Case II: $\hat{t}_{s,n+1} \leq t_{n+1}$, for establishing $|\hat{t}_{s,n+1} - t_{s,n+1}| = O(h^p)$.

Due to continuity, $\psi_{i,j}(\tilde{x}(t))$ is decreasing for $t \in [t_{s,n+1}, \hat{t}_{s,n+1}]$, there exists some positive constant a_I with $0 < a_I < |\frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(\tilde{x}(t_{s,n+1})|$ such that for sufficiently small h and $t \in [t_{s,n+1}, \hat{t}_{s,n+1}]$ it holds that

$$\psi_{i,j}(\tilde{x}(t)) \leq \psi_{i,j}(\tilde{x}(t_{s,n+1})) - a_I(t - t_{s,n+1}).$$

The arguments above are illustrated in the left plot of Figure 6. From the last inequality and (43) at $t = \hat{t}_{s,n+1}$ we have that $\psi_{i,j}(\tilde{x}(\hat{t}_{s,n+1})) < 0$, $\psi_{i,j}(\tilde{x}(t_{s,n+1})) = 0$, and conclude that

$$\hat{t}_{s,n+1} - t_{s,n+1} \leq O(h^p). \tag{44}$$

This completes the consideration of case I.

Case II. We apply similar arguments as for I. Under the assumption of $\hat{t}_{s,n+1} < t_{s,n+1}$, following similar lines as in the proof of [46, Theorem 4.3], we first prove $\hat{t}_{s,n+1} \to t_{s,n+1}$ and establish subsequently the convergence rate.

Regard the set $H = \{h > 0 \mid \hat{t}_{s,n+1} < t_{s,n+1}\}$. By the assumption of case II and the induction hypothesis, it holds that $\hat{t}_{s,n+1} \in [\hat{t}_{s,n}, t_{s,n+1}]$. Since this is a bounded set, there must be a subsequence $H' \subseteq H$ with $h \downarrow 0$ such that $\hat{t}_{s,n+1} \to \bar{t}$. We show now that $\bar{t} = \hat{t}_{s,n+1}$. We consider again the function $\psi_{i,j}(\cdot)$ for some $j \in \mathcal{I}_n$ and $i \in \mathcal{I}(x(t_{s,n+1})) \setminus \mathcal{I}_n$

$$\psi_{i,j}(x(t)) = g_i(x(t)) - g_j(x(t)) = \lambda_i - \lambda_j,$$

which becomes zero at an active-set change and is positive before. Similarly, active-set changes for the solution approximation happen when $\psi_{i,j}(\hat{x}_h(\hat{t}_{s,n+1})) = 0$. We remind the reader that earlier it was shown that $\hat{\mathcal{I}}_n = \mathcal{I}_n$.

By taking $h \downarrow 0$ and $h \in H'$ from (41) it follows that $\psi_{i,j}(x(\bar{t})) = 0$. By the definition of a switching point, there must be a $i \notin \mathcal{I}_n$, but $i \in \mathcal{I}(x(\bar{t}))$. However, this contradicts the assumption that $\mathcal{I}(x(t)) = \mathcal{I}_n$ for $t \in (t_{s,n}, t_{s,n+1})$ and we conclude that $\bar{t} \notin (t_{s,n}, t_{s,n+1})$.

On the other hand at $t_{\mathrm{s},n}$, due to strict complementarity in the active-set determining LCP (cf. Theorem7 and Assumption 8), if some $i \in \mathcal{I}(x(t_{\mathrm{s},n}))\setminus\mathcal{I}_n$ and $j \in \mathcal{I}_n$, we know that

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(t^+_{\mathrm{s,n}})) = g_i(x(t^+_{\mathrm{s,n}})) - g_j(x(t^+_{\mathrm{s,n}})) > 0.$$

Due to continuity of the functions $g_i(\cdot), i \in \mathcal{J}$, and the induction hypothesis, there exists some $\epsilon > 0$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}(g_i(\hat{x}_h(t)) - g_j(\hat{x}_h(t)) > 0 \text{ for } t \in [\hat{t}_{\mathrm{s},n}, \hat{t}_{\mathrm{s},n} + \epsilon].$$

However, when a switch happens the derivative in the last line must be negative at t (cf. Remark 6), thus $\hat{t}_{\mathrm{s},n+1} > \hat{t}_{\mathrm{s},n} + \epsilon$, i.e., with $h \downarrow 0, h \in H'$, $\bar{t} > t_{\mathrm{s},n} + \epsilon$. This means that $\bar{t} \neq t_{\mathrm{s},n}$ and the only option that is left is $\hat{t}_{\mathrm{s},n+1} \to \bar{t} = \hat{t}_{\mathrm{s},n+1}$ as $h \downarrow 0, h \in H'$.

Now we continue with establishing the convergence rate for $\hat{t}_{\mathrm{s},n+1} \to t_{\mathrm{s},n+1}$. From $\hat{t}_{\mathrm{s},n+1} \leq t_{\mathrm{s},n+1}$ we have from the definition of $\psi_{i,j}(\cdot)$ that $\psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1})) > 0$ and $\psi_{i,j}(\hat{x}_h(\hat{t}_{\mathrm{s},n+1})) = 0$. Using the fact that $\tilde{x}(\hat{t}_{\mathrm{s},n+1}) = x(\hat{t}_{\mathrm{s},n+1})$ and (41) we have

$$\begin{aligned}|\psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1}))| &= |\psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1})) - \psi_{i,j}(\hat{x}_h(\hat{t}_{\mathrm{s},n+1}))| \\ &\leq L_\psi \|\tilde{x}(\hat{t}_{\mathrm{s},n+1}) - \hat{x}_h(\hat{t}_{\mathrm{s},n+1})\| = O(h^p).\end{aligned}$$

We again use a first-order expansion:

$$\psi_{i,j}(x(t)) = \psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1})) + \frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1}))(t - \hat{t}_{\mathrm{s},n+1}) + o(|t - \hat{t}_{\mathrm{s},n+1}|).$$

Once again, due to assumption 8, we have that $\frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(t_{\mathrm{s},n+1})) < 0$. Note that $\frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(t^-_{\mathrm{s},n+1})) < 0$. From $\hat{t}_{\mathrm{s},n+1} \to t_{\mathrm{s},n+1}$ and continuity of $\psi_{i,j}(\cdot)$ it follows that for sufficiently small $h > 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_{i,j}(x(\hat{t}_{\mathrm{s},n+1})) < 0.$$

Using similar reasoning as in case I (see right plot of Figure 6 for an illustration of the argument), there exists some $a_{II} > 0$ such that from the last equation at $t = t_{\mathrm{s},n+1}$ it follows $0 \leq O(h^p) - a_{II}(t_{\mathrm{s},n+1} - \hat{t}_{\mathrm{s},n+1})$, i.e.,

$$t_{\mathrm{s},n+1} - \hat{t}_{\mathrm{s},n+1} \leq O(h^p). \tag{45}$$

Putting (44) and (45) together, we obtain the first part of the induction statement, i.e.,

$$|t_{\mathrm{s},n+1} - \hat{t}_{\mathrm{s},n+1}| = O(h^p). \tag{46}$$

To complete the induction step we must prove that (40b) holds for $t = \hat{t}_{\mathrm{s},n+1}$. For $\hat{t}_{\mathrm{s},n+1} \leq t_{\mathrm{s},n+1}$ we have $\tilde{x}(\hat{t}_{\mathrm{s},n+1}) = x(\hat{t}_{\mathrm{s},n+1})$ and the assertion

follows directly from (41). Note that for any other $t_n \in \mathcal{G}$ that is not a switching point (40b) follows immediately from Assumption 11.

It is left to investigate the case of $\hat{t}_{\mathrm{s},n+1} > t_{\mathrm{s},n+1}$. Using Lipschitz continuity of $x(\cdot)$, $\tilde{x}(\cdot)$, the fact that $\tilde{x}(t_{\mathrm{s},n+1}) = x(t_{\mathrm{s},n+1})$ and (46) one obtains

$$\begin{aligned}\|\hat{x}_h(\hat{t}_{\mathrm{s},n+1}) - x(\hat{t}_{\mathrm{s},n+1})\| &\leq \|\hat{x}_h(\hat{t}_{\mathrm{s},n+1}) - \tilde{x}(\hat{t}_{\mathrm{s},n+1})\| + \|x(t_{\mathrm{s},n+1}) - x(\hat{t}_{\mathrm{s},n+1})\| \\ + \|\tilde{x}(\hat{t}_{\mathrm{s},n+1}) - \tilde{x}(t_{\mathrm{s},n+1})\| &\leq O(h^p) + (L_x + L_{\tilde{x}})|t_{\mathrm{s},n+1} - \hat{t}_{\mathrm{s},n+1}| = O(h^p).\end{aligned}$$

Moreover, from Lipschitz continuity of $g_i(\cdot), i \in \mathcal{J}$ and the last inequality for sufficiently small $h > 0$ we have that $\mathcal{I}(x(t_{\mathrm{s},n+1})) = \mathcal{I}(\hat{x}(\hat{t}_{\mathrm{s},n+1}))$, which completes the induction step for $n+1$. With the use of an interpolation scheme, from (40) it follows that we can make a continuous-time approximation $\hat{x}_h(t)$ for $t \in [0, T]$ with the accuracy $O(h^{\bar{p}})$, $1 \leq \bar{p} \leq p$ for $t \notin \mathcal{G}$. Therefore, it follows that the sequence of approximations $\hat{x}_h(t)$ generated by the FESD method converges to a solution $x(t)$ in the sense of Definition 1 for $t \in [0, T]$. The proof is completed. □

In the next subsection, we illustrate the results of this theorem for several RK schemes. We compare the results obtained via FESD to the ones obtained with the standard RK discretization from Section 3.1.

4.4 Illustrating the integration order

Consider the nonsmooth IVP

$$\dot{x}(t) = \begin{cases} A_1 x, & c(x) < 0, \\ A_2 x, & c(x) > 0, \end{cases} \tag{47}$$

with $A_1 = \begin{bmatrix} 1 & \omega \\ -\omega & 1 \end{bmatrix}$, $A_2 = \begin{bmatrix} 1 & -\omega \\ \omega & 1 \end{bmatrix}$, $c(x) = x_1^2 + x_2^2 - 1$, $\omega = 2\pi$ and $x(0) = (e^{-1}, 0)$ for $t \in [0, T]$. The example trajectory is given in Figure 7. Following Section 2.2, we can write (47) in the form of the DCS (8), where:

$$F(x) = \begin{bmatrix} A_1 x & A_2 x \end{bmatrix}, \; g(x) = [1 \; -1]^\top c(x),$$

where $g(x)$ was obtained and via Eq. (13). It can be shown that the switch happens at $t_{\mathrm{s}} = 1$ and that $x(T) = (e^{(T-1)}\cos(2\pi(T-1)), -e^{(T-1)}\sin(2\pi(T-1)))$ for $T > t_{\mathrm{s}}$. Hence, we can determine the global integration error $E(T) = \|x(T) - \hat{x}_h(T)\|$. We regard solution approximations to this IVP obtained by standard explicit and implicit RK methods (26) and FESD (32) with $N_{\mathrm{FE}} = 2$ and different step sizes. Both integration methods are available in `NOSNOC` via the function `integrator_fesd()`. The regarded RK methods are listed in Table 1 together with their global integration error estimate. For explicit methods, we consider in this paper $n_{\mathrm{s}} \leq 4$. Several other IRK methods are available in `NOSNOC`, but we omit the full comparison for brevity. The nominal step h size is obtained by dividing T by the number of simulation steps N_{sim}. We take an irrational number for $T = \frac{\pi}{2}$ so that $t_{\mathrm{s}} = 1$ never coincides with

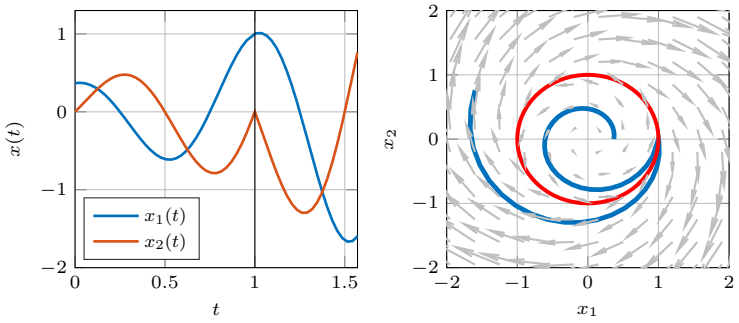

Fig. 7 Illustration of the solution to the nonsmooth IVP given by (47).

Method	Global error estimate
Radau-IIA	h^{2n_s-1}
Gauss-Legendre	h^{2n_s}
Lobatto-IIIA	h^{2n_s-2}
Lobatto-IIIC	h^{2n_s-2}
Explicit-RK	h^{n_s}

Table 1 List of analyzed RK methods and their accuracy order for ODE [26]. Note that for Explicit-RK methods the assertion in the table is true for $n_s \leq 4$, otherwise $p < n_s$.

a finite element boundary. Therefore, we avoid accidental switch detection via the discretization grid for the standard discretization.

The numerical error as a function of the step sizes for the RK and FESD methods from Table 1 are depicted in Figure 8. We can see that the standard discretization achieves in all cases only first-order accuracy (left plots). In contrast, the FESD method recovers in all cases the high accuracy order that the underlying RK method has for smooth ODE (right plots). This verifies the result of Theorem 17 in practice and demonstrates how FESD can be used as an event-based integrator without an external switch detection procedure. The *saturation* in the right plots of some high-accuracy methods is due to the round-off errors in floating point arithmetic which limit the possible accuracy on a computer.

4.5 Convergence of discrete-time sensitivities

One of the most important tools for computing stationary points and verifying their optimality in direct optimal control is the numerical sensitivities, i.e., the derivatives of the numerical solution approximation $\hat{x}_h(t; x_0)$ w.r.t. the initial value x_0 (and parameters) for some $t \in (0, T]$. Here we denote them by $\hat{X}_h(t; 0, x_0)$ or sometimes more compactly by $\hat{X}_h(t, x_0)$.

A fundamental limitation of standard discretization methods for nonsmooth systems (e.g., with the RK discretization from Subsection 3.1) is that $\hat{X}_h(t, x_0) \not\to X(t, x_0)$ as $h \downarrow 0$ [51]. In direct optimal control, this can cause

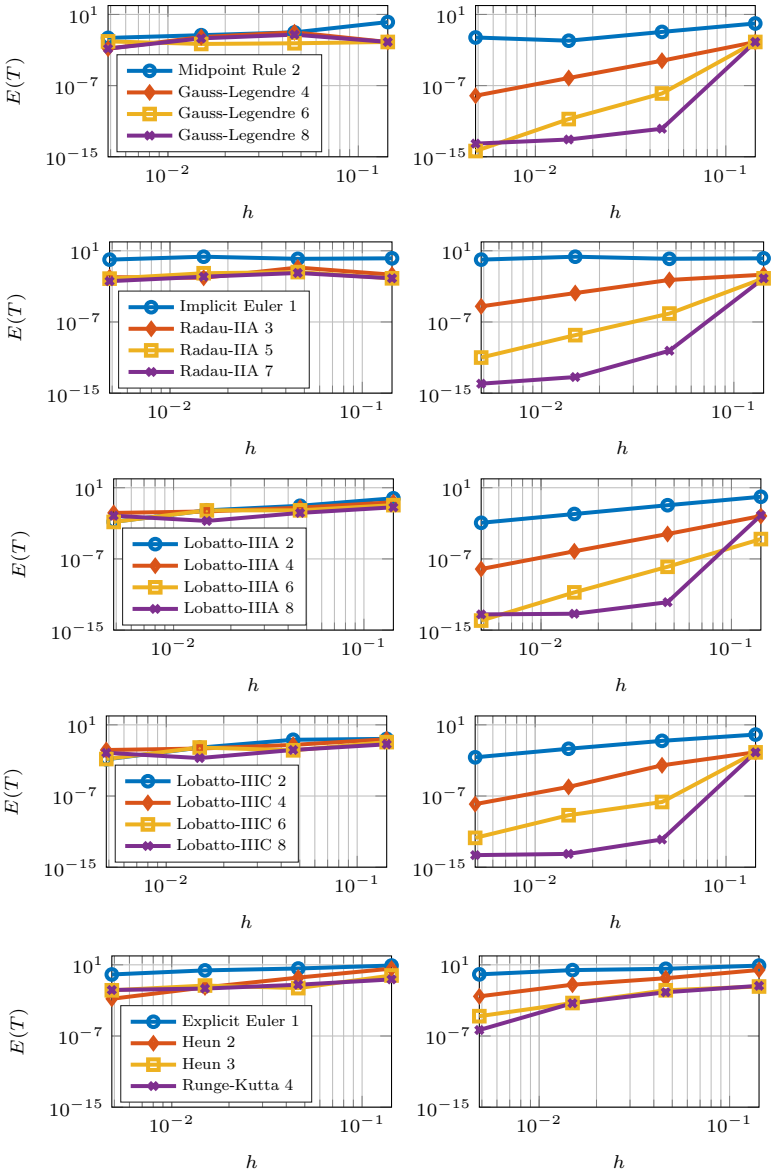

Fig. 8 Integration error $E(T) = \|x(T) - \hat{x}_h(T)\|$ vs. the step size h for different RK methods: standard (left plot) vs. FESD (right plot). The legend provides the name of the used RK method and its order of the global integration error.

convergence to artificial stationary points arbitrarily close to the initialization point [36,51]. Fortunately, the sensitivities of the solutions generated by the FESD method converge to their true values (cf. Subsection 2.4). This is shown in the next theorem, but before we state it, we make one more assumption on the time derivatives of the solution approximation.

Assumption 18 *(RK derivatives) Regard the RK methods from Assumption 11 applied to the differential algebraic equations* (14)*. Suppose that the derivatives of the numerical approximation for the same RK method converge with order* $1 \leq q \leq p$*, i.e.,* $\|\dot{\hat{x}}_h(t) - \dot{x}(t)\| = O(h^q),\ t \in \mathcal{G}$.

The aim of the assumption about the convergence of the derivatives of the numerical approximation is to cover a broad class of RK methods and the value of q depends on the specific choice of an RK method. For example, for collocation-based implicit RK methods for ODE in general it holds that $q = p - 1$ [25, Theorem 7.10]. More specifically, methods that contain the boundary point of a finite element denoted by $\hat{t}$, that is $c_{n_\mathrm{s}} = 1$ satisfy $p = q$. This assertion follows directly from Lipschitz continuity and the fact that the numerical approximations satisfy the ODE at every stage point:
$\|\dot{\hat{x}}_h(\bar{t}) - \dot{x}(\bar{t})\| = \|f_i(\hat{x}_h(\bar{t})) - f_i(x(\bar{t}))\| \leq L_{f_i}\|\hat{x}_h(\bar{t}) - x(\bar{t})\| = O(h^p)$.

Theorem 19 (Convergence to exact sensitivities) *Suppose the assumptions of Theorem 17 and Assumption 18 hold. Assume that a single active-set change happens at time* $t_{\mathrm{s},n}$*, i.e.,* $||\mathcal{I}_n| - |\mathcal{I}_{n+1}|| \leq 1, n = 0, \dots N_\mathrm{sw}$*. Then for* $h \downarrow 0$ *it holds that* $\hat{X}_h(t, x_0) \to X(t, x_0)$ *with the convergence rate*

$$\|\hat{X}_h(t, x_0) - X(t, x_0)\| = O(h^q),\ \textit{for all } t \in \mathcal{G}. \tag{48}$$

Proof. Regard partition of $[0, T]$: $0 < \tilde{t}_1 < \cdots < \tilde{t}_k < \cdots < \tilde{t}_{N_k} < T$ such that in the open interval between every two neighboring points there is a single switching point with $N_k > N_\mathrm{sw}$. Then by the chain rule we have

$$X(T; 0, x_0) = X(T; \tilde{t}_{N_k}, x(\tilde{t}_{N_k})) \cdots X(\tilde{t}_{k+1}; \tilde{t}_k, x(\tilde{t}_k)) \cdots X(\tilde{t}_1; 0, x_0).$$

W.l.o.g. assume that on $[0, \tilde{t}_1]$ a single switch occurs at $t_\mathrm{s} \in (0, \tilde{t}_1)$. We show convergence of the sensitivities on this interval. Convergence on $[0, T]$ follows by inductively applying the same arguments on every sub-interval $[\tilde{t}_k, \tilde{t}_{k+1}]$.

Regard the two smooth pieces of the approximation $\hat{x}_h(t)$: 1) $\hat{x}_{h,1}(t, x_0)$ for $t \leq \hat{t}_\mathrm{s}$ and, 2) $\hat{x}_{h,2}(t, y_0(x_0))$ where $y_0(x_0) = \hat{x}_{h,2}(0, y_0(x_0)) = \hat{x}_{h,1}(\hat{t}_\mathrm{s}, x_0)$. With this definition, we have for $t \geq \hat{t}_\mathrm{s}$

$$\hat{x}_{h,2}(t - \hat{t}_\mathrm{s}, y_0(x_0)) = \hat{x}_h(t, x_0). \tag{49}$$

From Theorem 17 we know that $|\hat{t}_\mathrm{s} - t_\mathrm{s}| = O(h^p)$. Obviously, the value of $\hat{t}_\mathrm{s}$ depends on x_0 and we know from Theorem 17 that we obtain implicitly at a switching point the condition

$$\psi_{i,j}(\hat{x}_{h,1}(\hat{t}_\mathrm{s}(x_0), x_0)) = g_i(\hat{x}_{h,1}(\hat{t}_\mathrm{s}(x_0), x_0)) - g_j(\hat{x}_{h,1}(\hat{t}_\mathrm{s}(x_0), x_0)) = 0, \tag{50}$$

where $i \notin \mathcal{I}_0$, $i \in \mathcal{I}_1$ and $j \in \mathcal{I}_0$.

For computing $\hat{X}_h(\cdot)$ on $[\hat{t}_\mathrm{s}, t_1]$ from Eq. (49) we have

$$\begin{aligned}\frac{\partial \hat{x}_h(t, x_0)}{\partial x_0} = \frac{\partial \hat{x}_{h,2}(t - \hat{t}_\mathrm{s}(x_0), y_0(x_0))}{\partial x_0} &= -\dot{\hat{x}}_{h,2}(t - \hat{t}_\mathrm{s}(x_0), y_0(x_0)) \nabla_{x_0} \hat{t}_\mathrm{s}(x_0)^\top \\ &+ \frac{\partial \hat{x}_{h,2}(t - \hat{t}_\mathrm{s}(x_0), y_0(x_0))}{\partial y_0} \nabla_{x_0} y_0(x_0).\end{aligned} \tag{51}$$

Next, we compute the expressions for the two unknowns $\nabla_{x_0}\hat{t}_{\mathrm{s}}(x_0)^\top$ and $\nabla_{x_0} y_0(x_0)$. Denote by $\hat{X}_{h,1}(t;0,x_0) = \frac{\partial \hat{x}_{h,1}(t,x_0)}{\partial x_0}$. Using the implicit function theorem for (50), we can compute

$$\nabla_{x_0}\hat{t}_{\mathrm{s}}(x_0)^\top = -\frac{\nabla\psi_{i,j}(\hat{x}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0))^\top \hat{X}_{h,1}(\hat{t}_{\mathrm{s}},x_0)}{\nabla\psi_{i,j}(\hat{x}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0))^\top \dot{\hat{x}}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0)}. \tag{52}$$

At $t = \hat{t}_{\mathrm{s}}$, we exploit the fact that $\hat{x}_{h,2}(0, y_0(x_0)) = y_0(x_0) = \hat{x}_{h,1}(\hat{t}_{\mathrm{s}}, x_0)$, thus

$$\nabla_{x_0} y_0(x_0) = \frac{\partial \hat{x}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0)}{\partial x_0} = \dot{\hat{x}}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0)\nabla_{x_0}\hat{t}_{\mathrm{s}}(x_0)^\top + \hat{X}_{h,1}(\hat{t}_{\mathrm{s}};x_0).$$

Combing the last line with (52) we obtain

$$\nabla_{x_0} y_0(x_0) = \left[I - \frac{\dot{\hat{x}}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0)\nabla\psi_{i,j}(\hat{x}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0))^\top}{\nabla\psi_{i,j}(\hat{x}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0))^\top \dot{\hat{x}}_{h,1}(\hat{t}_{\mathrm{s}}(x_0),x_0)}\right]\hat{X}_{h,1}(\hat{t}_{\mathrm{s}};x_0). \tag{53}$$

We are interested in $\frac{\partial x_h(t;x_0)}{\partial x_0}$ when $t \downarrow \hat{t}_{\mathrm{s}}$. Note that $\frac{\partial \hat{x}_{h,2}(t-\hat{t}_{\mathrm{s}}(x_0),y_0(x_0))}{\partial y_0} \to I$ as $t \downarrow \hat{t}_{\mathrm{s}}$. Thus from (51), (52) and (53) one obtains

$$\begin{aligned}
&\frac{\partial x_h(\hat{t}_{\mathrm{s}}^+;x_0)}{\partial x_0} = \hat{J}_h(\hat{x}_h(\hat{t}_{\mathrm{s}};x_0))\frac{\partial x_h(\hat{t}_{\mathrm{s}}^-;x_0)}{\partial x_0}, \text{ with}\\
&\hat{J}_h(\hat{x}_h(\hat{t}_{\mathrm{s}};x_0)) \coloneqq \left[I + \frac{(\dot{\hat{x}}_h(\hat{t}_{\mathrm{s}}^+,x_0) - \dot{\hat{x}}_h(\hat{t}_{\mathrm{s}}^-,x_0))\nabla\psi_{i,j}(\hat{x}_h(\hat{t}_{\mathrm{s}}^-,x_0))^\top}{\nabla\psi_{i,j}(\hat{x}_h(\hat{t}_{\mathrm{s}}^-,x_0))^\top \dot{\hat{x}}_h(\hat{t}_{\mathrm{s}}^-,x_0)}\right].
\end{aligned} \tag{54}$$

By the chain rule, we have that for $t > \hat{t}_{\mathrm{s}}$

$$\hat{X}_h(t;x_0) = \hat{X}_{h,2}(t;\hat{t}_{\mathrm{s}}^+,y_0)\hat{J}_h(\hat{x}_h(\hat{t}_{\mathrm{s}};x_0))\hat{X}_{h,1}(\hat{t}_{\mathrm{s}}^-;0,x_0). \tag{55}$$

First, note that for a fixed active set the FESD equations for $\hat{x}_{h,1}(t,x_0)$ and $\hat{x}_{h,2}(t,y_0)$ boil down to RK equations for the DAE (14) with fixed step size h_n, cf. Theorem 15. Differentiating the RK equations to obtain $\hat{X}_h(\cdot)$ results in the same RK method applied to the variational differential equations of the system at hand, thus the numerical sensitivities converge in this setting to the continuous-time sensitivities with the same accuracy $O(h^p)$ as for the solution of the system [6].

Second, as $h \downarrow 0$, due to assumption of this theorem, in $\hat{J}_h(\hat{x}_h(\hat{t}_{\mathrm{s}};x_0))$ the functions $\dot{x}_h(\cdot)$ converge to $f(x(\cdot))$ with order q. Due to Theorem 17 all other terms converge with order p. Thus, $\|\hat{J}_h(\hat{x}_h(\hat{t}_{\mathrm{s}};x_0)) - J(x(t_{\mathrm{s}};x_0))\| = O(h^q)$.

Summarizing the last two arguments and applying them inductively for every active-set change we conclude that $\hat{X}_h(t;x_0) \to X(t;x_0)$ as $h \downarrow 0$ with the order $q = \min(p,q)$. This completes the proof. □

The only restrictive assumption we make is that a single active-set change happens at a time. For multiple simultaneous switches the derivation becomes quite involved even in continuous-time case [23], hence we made this assumption for simplicity. The results of Theorem 19 are illustrated on a numerical example in the next subsection.

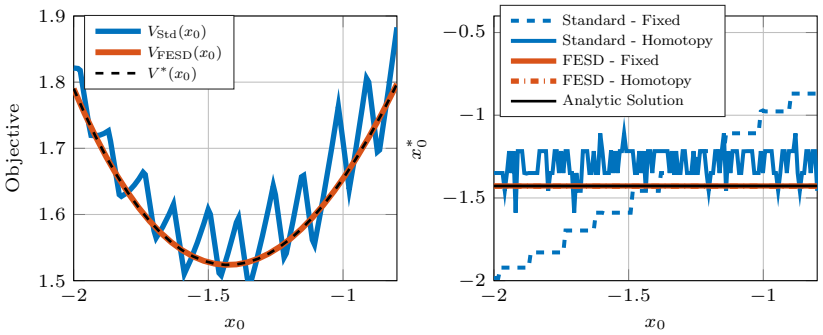

Fig. 9 Approximations of the objective $V(x_0)$ computed with FESD and a standard IRK Gauss-Legendre method compared to the true value are shown in the left plot. The right plot shows the value of the optimal solution x_0^* against different initial values x_0 for which a initial feasible solution guess was computed. The optimal solution is computed with FESD and a standard method with and without a homotopy approach for the underlying optimization problem.

4.6 Illustration of numerical sensitivity convergence

To demonstrate the improvements in FESD compared to standard methods we repeat the experiments from [36]. For this purpose, we look at the optimal control problem from [51]:

$$\min_{x_0\in\mathbb{R}, x(\cdot)\in\mathcal{C}^0} \quad \int_0^2 x(t)^2\,\mathrm{d}t + (x(2)-5/3)^2 \tag{56a}$$

$$\text{s.t.} \quad x(0) = x_0, \tag{56b}$$

$$\dot{x}(t) \in 2 - \operatorname{sign}(x(t)),\ t\in[0,2]. \tag{56c}$$

Note that the initial value x_0 is the only effective degree of freedom. Let $V_*(x_0)$ be the objective value for the unique feasible trajectory starting at $x(0)=x_0$. The equivalent reduced problem is given by

$$\min_{x_0\in\mathbb{R}} \quad V_*(x_0). \tag{57}$$

In the first experiment, we evaluate $V_*(x_0)$ by simulating the trajectory of (56c) for various x_0 with a standard RK method (26) and FESD (32). We pick the Gauss-Legendre method of fourth order with $n_\mathrm{s}=2$ and $N_\mathrm{FE}=25$. The results are depicted in the left plot in Figure 9. It can be seen that for standard RK methods, the trajectories would converge to the analytic solution, but the derivatives w.r.t. x_0 do not. In fact, many artificial minima arise [36,51]. On the other hand, the FESD solution matches the analytic solution.

In the second experiment, we solve the OCP (56) for different initial guesses x_0. The initial guess for the arising MPCC is obtained from a forward simulation of (56c) with the same method and same grid as in the discretized OCP. The arising MPCC are solved by a relaxation approach, where the complementarity constraints (60c) are replaced by $w_1^\top w_2 \le \sigma,\ w_1\ge 0,\ w_2\ge 0$.

In the first experiment, the MPCC are solved for a fixed $\sigma = 10^{-15}$. In the second case, we solve a sequence of NLP in a homotopy procedure where σ is decreased from $\sigma_0 = 1$ to $\sigma_N = 10^{-15}$, via the rule $\sigma_{k+1} = 0.1\sigma_k$. It was demonstrated in [36] that a homotopy approach with standard discretizations can lead to convergence to better local minima. The results are depicted in the right plot of Figure 9.

Remarkably, the FESD solution matches the analytic solution for all x_0 regardless of the MPCC strategy and initialization point (FESD - fixed and FESD - homotopy). This highlights that the numerical sensitivities converge to the true ones, cf. Theorem 19. On the other hand, as in [36], with the standard IRK method, the fixed parameter MPCC strategy leads to convergence to the artificial local minima close to the initialization. This results in a stair-like curve in right plot in Figure 9 (Standard - fixed). This is a consequence of the fact that the numerical derivatives are always wrong, i.e., the numerical approximation is "trapped" in one of the local minima of the standard approach, which are visible in the left plot (blue).

The homotopy strategy yields better local minima since in the earlier homotopy iterations the derivatives are still correct [40] (Standard - homotopy). However, the error of $O(h)$ to the analytic solution is still present since a standard method has at best accuracy of order one. This highlights the result of Theorems 17 and 19 and demonstrates the superiority of FESD to the naïve approach even on a very simple example.

5 FESD in direct optimal control

This section regards the use of FESD in direct optimal control, i.e., a first discretize then optimize approach. We consider the following continuous-time optimal control problem on a control horizon $[0, T_{\text{ctrl}}]$:

$$\min_{x(\cdot),u(\cdot),z(\cdot)} \quad \int_0^{T_{\text{ctrl}}} L_{\text{int}}(x(t),u(t))\mathrm{d}t + L_{\text{end}}(x(T_{\text{ctrl}})) \tag{58a}$$

$$\text{s.t.} \quad x_0 = \bar{x}_0, \tag{58b}$$

$$\dot{x}(t) = F(x(t),u(t))\theta(t), \qquad t \in [0, T_{\text{ctrl}}], \tag{58c}$$

$$0 = g(x(t)) - \lambda(t) - \mu(t)e, \qquad t \in [0, T_{\text{ctrl}}], \tag{58d}$$

$$1 = e^\top \theta(t), \qquad t \in [0, T_{\text{ctrl}}], \tag{58e}$$

$$0 \le \theta(t) \perp \lambda(t) \ge 0, \qquad t \in [0, T_{\text{ctrl}}], \tag{58f}$$

$$0 \le G_{\text{ineq}}(x(t),u(t)), \qquad t \in [0, T_{\text{ctrl}}], \tag{58g}$$

$$0 \le G_{\text{end}}(x(T_{\text{ctrl}})), \tag{58h}$$

where $\bar{x}_0$ is a given initial value, $u(t) \in \mathbb{R}^{n_u}$ is the control function, $z(t) = (\lambda(t), \theta(t), \mu(t)) \in \mathbb{R}^{2n_f+1}$ collects the algebraic variables. The function $L_{\text{int}} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ is the Lagrange objective term to be integrated and $L_{\text{end}} : \mathbb{R}^{n_x} \to \mathbb{R}$ is the terminal cost also called Mayer term. The path and terminal

constraints are collected in the functions $G_{\text{ineq}} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_{g1}}$ and $G_{\text{end}} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_{g2}}$, respectively.

5.1 A multiple shooting-type discretization

We discretize the OCP (58) with $N_{\text{ctrl}} \geq 1$ control intervals indexed by k. The control function approximation is taken to be piecewise constant on an equidistant grid. This is the usual choice in direct multiple shooting [13] and is required by many practical applications of feedback control, but could be easily generalized to any other local control parameterization. The constant controls are collected in $\mathbf{q} = (q_0, \ldots, q_{N_{\text{ctrl}}-1}) \in \mathbb{R}^{N_{\text{ctrl}} n_u}$. All internal variables of every control interval are additionally equipped with an index k. On every control interval k with fixed duration $T = \frac{T_{\text{ctrl}}}{N_{\text{ctrl}}}$, we apply a discretization (32) with N_{FE} internal finite elements. The state values at the control interval boundaries are collected in $\mathbf{s} = (s_0, \ldots, s_{N_{\text{ctrl}}}) \in \mathbb{R}^{(N_{\text{ctrl}}+1) n_x}$. The following vectors collect all internal variables of all discretization steps: $\mathcal{H} = (\mathbf{h}_0, \ldots, \mathbf{h}_{N_{\text{ctrl}}-1})$ and $\mathcal{Z} = (\mathbf{Z}_0, \ldots, \mathbf{Z}_{N_{\text{ctrl}}-1})$. The FESD discretization of the OCP (58) together with the aforementioned control discretization reads as:

$$\min_{\mathbf{s},\mathbf{q},\mathcal{H},\mathcal{Z}} \quad \sum_{k=1}^{N_{\text{ctrl}}-1} \hat{L}_{\text{int}}(s_k, \mathbf{x}_k, q_k) + \hat{L}_{\text{end}}(s_{N_{\text{ctrl}}}) \tag{59a}$$

$$\text{s.t.} \quad s_0 = \bar{x}_0, \tag{59b}$$

$$s_{k+1} = F_{\text{fesd}}(\mathbf{Z}_k), \qquad k = 0, \ldots, N_{\text{ctrl}} - 1, \tag{59c}$$

$$0 = G_{\text{fesd}}(\mathbf{Z}_k, \mathbf{h}_k, s_k, q_k, T_k), \qquad k = 0, \ldots, N_{\text{ctrl}} - 1, \tag{59d}$$

$$0 \leq G_{\text{ineq}}(\mathbf{x}_k, q_k), \qquad k = 0, \ldots, N_{\text{ctrl}} - 1, \tag{59e}$$

$$h_{\min} e \leq \mathbf{h}_k \leq h_{\max} e, \qquad k = 0, \ldots, N_{\text{ctrl}} - 1, \tag{59f}$$

$$0 \leq G_{\text{end}}(s_{N_{\text{ctrl}}}). \tag{59g}$$

where $\hat{L}_{\text{int}} : \mathbb{R}^{n_x} \times \mathbb{R}^{(N_{\text{FE}}+1) n_s n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ and $\hat{L}_{\text{end}} : \mathbb{R}^{n_x} \to \mathbb{R}$ are the discretized integral and terminal costs, respectively. The scalars $h_{\min}$ and $h_{\max}$ are the lower and upper bounds for the step sizes. The box constraint (59f), prohibits negative step sizes and bounds the variability of the step size and thus the local discretization errors. We want to recall that with this formulation at every control interval the constraint $\sum_{n=1}^{N_{\text{FE}}} h_{n,k} = T$ is imposed as part of (59d), cf. Eq (31).

5.2 Solving the discretized OCP

When solving the OCP (59) in practice we do not use C-functions (cf. Section 1 for a definition), but write the complementarity conditions explicitly. Therefore, the NLP (59) is a Mathematical Program with Complementarity Constraints (MPCC) and it can be compactly written as

$$\min_{w} \quad \varphi(w) \tag{60a}$$

$$\text{s.t.} \quad 0 \leq \zeta(w), \tag{60b}$$
$$0 \leq w_1 \perp w_2 \geq 0, \tag{60c}$$

where $w = (w_0, w_1, w_2) \in \mathbb{R}^{n_w}$ is a decomposition of the problem variables. MPCC are difficult nonsmooth optimization problems that violate the Mangasarian-Fromovitz constraint qualification at all feasible points [8], which makes standard NLP algorithms fail to converge. Fortunately, MPCC can often be solved efficiently via reformulations and homotopy approaches [8,27, 33,43]. In a homotopy procedure a sequence of more regular, smooth, and relaxed NLPs related to (60) are solved. Several relaxation, smoothing, and penalty homotopy approaches are implemented in MATLAB package `NOSNOC` [38]. They differ in how the complementarity constraints (60c) are treated. Under some regularity assumptions the complementarity constraints are satisfied exactly even after solving a finite sequence of NLP [8,43]. In `NOSNOC` the NLP is solved with `IPOPT` [54] via its `CasADi` [7] interface. In our numerical experiments, we do not provide an elaborate initial guess. For the states, we simply use the initial value for all stages and set the controls to zero. The homotopy approaches implemented in `NOSNOC` are quite robust and usually succeed in finding a good solution, as shown in the benchmark in [39].

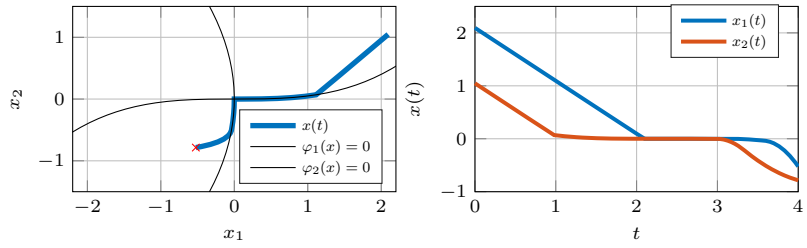

Fig. 10 A solution $x(t)$ to the OCP (61).

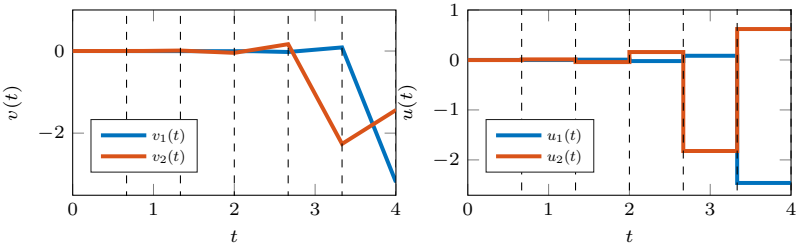

Fig. 11 The left plot shows the solution trajectories for $v(t)$. The right plot shows optimal controls obtained via the FESD discretization of the OCP (61). The vertical dashed lines highlight the control discretization grid.

5.3 A numerical optimal control example

The following optimal control problem demonstrates several features developed in this paper: FESD for Cartesian Products of Filippov systems (Sec. 2.3), handling multiple sliding modes, step equilibration (Sec. 3.2.3) and equidistant control discretization grids (Sec. 5.1). We also include a benchmark where we compare the accuracy and computational time of the standard and FESD methods. Regard the following OCP with $q \in \mathbb{R}^2, v \in \mathbb{R}^2, u \in \mathbb{R}^2$ and $x = (q, v)$:

$$\min_{x(\cdot),u(\cdot)} \quad \int_0^4 u(t)^\top u(t) + v(t)^\top v(t)\,\mathrm{d}t + \rho\,\|q(4) - q_{\text{final}}\|_1 \tag{61a}$$

$$\text{s.t.} \quad x(0) = (\frac{2\pi}{3}, \frac{\pi}{3}, 0, 0), \tag{61b}$$

$$\dot{x}(t) = \begin{bmatrix} -\text{sign}(c(x(t))) + v(t) \\ u(t) \end{bmatrix}, \quad t \in [0,4], \tag{61c}$$

$$-2e \leq v(t) \leq 2e, \quad t \in [0,4], \tag{61d}$$

$$-10e \leq u(t) \leq 10e, \quad t \in [0,4]. \tag{61e}$$

where $\rho = 10^3$, $\varphi_1(x) = q_1 + 0.15q_2^2$, $\varphi_2(x) = -0.05q_1^3 + q_2$ and the function $c(x) = (\varphi_1(x), \varphi_2(x))$ defines the region boundaries. It can be seen that for $v(t) = 0$ the vector fields of q point in all regions towards the origin in the (q_1, q_2) plane. Settings the control functions to zero, this results in trajectories going to the origin with sliding modes on the surfaces of discontinuity defined by $c(x) = 0$. On the other hand, by increasing the value of $v(t)$ via the control functions $u(t)$, the vector fields can change their direction and sliding modes can be left or not achieved at all.

The goal in the OCP is to reach $q_{\text{final}} = (-\frac{\pi}{6}, -\frac{\pi}{4})$ with a minimum control effort. The trajectory has to: (1) first reach $\varphi_1(x) = 0$; (2) slide towards $\mathcal{M} = \{q \in \mathbb{R}^2 \mid c(x) = 0\}$; (3) stay there for some time; (4) exit $\mathcal{M}$ and slide on $\varphi_2(x) = 0$; (5) and then leave the sliding mode as late as possible to reach q_{final}. The seemingly simple example comprises several difficult switching cases in its solution. The described solution is illustrated in Figure 10 and the optimal controls $u(t)$ and state $v(t)$ are depicted in Figure 11. The OCP is discretized with the FESD Radau-IIA method of order 3 with $n_{\text{s}} = 2$ and $N_{\text{ctrl}} = 6$ control intervals with $N_{\text{FE}} = 6$ finite elements on every control interval. This system is transformed as described in Section 2.3 with $n_{\text{sys}} = 2$, where $c_{i,1}(x) = \varphi_i(x), i = 1, 2$. The functions $g_{i,j}(x) \in \mathbb{R}^2, i, j = 1, 2$ are computed via Eq. (13).

This solution comprises four switches and different sliding modes on nonlinear manifolds including: twice sliding on co-dimension 1 manifolds and once on the co-dimension 2 manifold $\mathcal{M}$, and additionally leaving the sliding mode twice. One can see in Figure 11 that the control is discretized on an equidistant grid despite the variable length of the finite elements. This illustrates the multiple shooting-type discretizations described in Section 5.1.

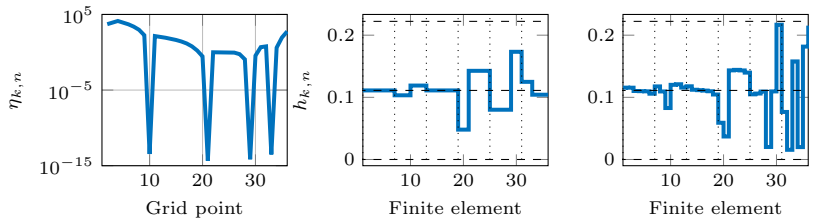

Fig. 12 The left plot depicts the switching indicator function $\eta(\cdot)$ at the solution of the OCP (61). The middle plot show the step size $h_{n,k}$ with step equilibration. The right plot shows the step sizes without step equilibration. The horizontal dashed lines correspond to the minimum, maximum, and nominal step size, the vertical dotted lines correspond to control interval boundaries.

In the first experiment, we demonstrate the effects of step equilibration for a solution of the OCP (61). The left plot in Figure 12 depicts the indicator function $\eta(\cdot)$. Clearly, the function is only zero if a switch occurs, cf. the right plot of Figure 10. The resulting step sizes $h_{k,n}$ with and without step equilibration are depicted in the middle and right plots, respectively. For the right plot we discard the step equilibration conditions $G_{\text{eq}}(\mathbf{h}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, T) = 0$. Obviously, without them, the optimizer varies the step size in a somewhat random way. On the other hand, with step equilibration, we obtain a piecewise equidistant grid, where the step size changes only when a switch occurs.

In the second experiment, we compare the accuracy of an OCP solution obtained with the standard and FESD method as a function of the CPU time. We take the optimal controls and perform a high-accuracy simulation of the system dynamics in (61), which we denote by $x_{\text{int}}(t)$. As a metric, we take the terminal constraint satisfaction of the high accuracy solution, i.e., $E(T) = \|x_{\text{int}}(T) - x_{\text{final}}\|$. We set $N_{\text{ctrl}} = 6$, and vary the number of finite elements per stage N_{FE} from 1 to 7 and the number of stage points n_{s} from 1 to 4. The experiment is performed for the following RK methods: Radau-IIA, Gauss-Legendre, Lobatto-IIIC, and Explicit-RK.

For Radau-IIA-FESD and Lobatto-IIIC-FESD we solved the arising MPCC with an elastic mode homotopy approach [8]. In the other scenarios, we were not able to solve all problems to convergence with the elastic mode approach. Therefore, the MPCC was solved with a relaxation homotopy approach, cf. [38] for implementation details. The second approach is slightly slower than elastic mode, but more robust, and all problems were solved successfully. The terminal error as a function of the total CPU time is given in Figure 13.

We can draw several conclusions from the experiments. Clearly, the FESD method outperforms the standard approach in all experiments. For example, for a CPU time of ≈ 1 second FESD achieves five orders of magnitude more accurate solutions than the standard time-stepping approach. A better solution than the most accurate one of the standard approaches can be achieved with FESD by an order of magnitude faster CPU time. The Radau-IIA and Lobatto-IIIC methods are the most efficient ones in this benchmark, whereas the Gauss-Legendre and Explicit-RK methods perform poorly. This is no sur-

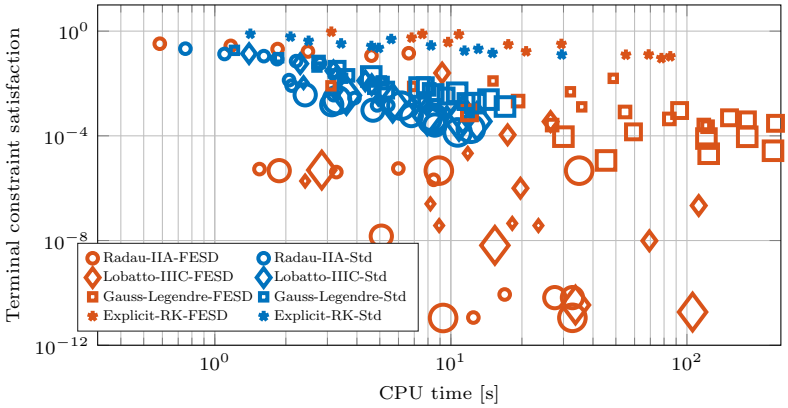

Fig. 13 Terminal constraint satisfaction vs. CPU time for the Standard and FESD method. The size of the marker indicates the number of stage points, the smallest corresponds to $n_s = 1$ and the largest to $n_s = 4$.

prise, since the solution trajectories contain sliding mode arcs which require solving nonlinear DAE of index 2. Radau-IIA and Lobatto-IIIC usually perform well and have good theoretical properties for higher index DAE, whereas Gauss-Legendre and Explicit RK even lose the high accuracy orders that they have for ODE [26]. The source code of all examples is available in the repository of the open source tool `NOSNOC` [1]. The same repository contains a few additional examples where FESD is used, including systems with state jumps that are transformed via time-freezing into PSS [37,38] and a nonsmooth mechanics simulation problem from [48], among others.

6 Summary

This paper presents a method that enables direct optimal control of a broad class of switched systems with high simulation accuracy, called Finite Elements with Switch Detection (FESD). We build a solid theoretical foundation and prove that FESD has the same accuracy as the underlying RK method has for smooth differential equations and that it delivers exact numerical sensitivities. An implementation of the FESD method described in this paper is available in the recently introduced open-source package `NOSNOC` [1]. Compared to standard discretizations, FESD achieves for a similar CPU time usually several orders of magnitude more accurate solutions.

A key advantage of the new approach for direct optimal control is that no guessing of the number or order of switches needs to be done. FESD can treat multiple or simultaneous switches and sliding modes. With time-freezing [40,37,35] many nonsmooth systems with state jumps can be recast into a piecewise smooth system. This allows us to treat many classes of nonsmooth systems in direct optimal control in a unified way.

In future work, one should relax some of the possibly restrictive assumptions in our theoretical analysis. We aim to extend FESD to other transformations of piecewise smooth systems into dynamic complementarity systems, e.g., the via the step function approach [5,17]. Some further open questions to be answered are: Are all limit points of the solution approximations candidates $\hat{x}_h(t)$ indeed solutions to the Filippov DI (2)? Do unique Filippov solutions to a given problem also imply a unique solution to the corresponding FESD problem (32)?

A Auxiliary results needed for Theorem 17

This lemma is about the active sets of perturbed LCP, where strict complementarity holds at a solution.

Lemma 20 ([46][Lemma A.2]) *Suppose that all entries of M are positive in* (19) *and all solutions of* $\mathrm{LCP}(M,q)$ *are strongly stable. If $\hat{M}_n \to M$, $\hat{q}_n \to q$, then* $\mathrm{SOL}(\hat{M}_n, \hat{q}_n) \to \mathrm{SOL}(M,q)$, *as $n \to \infty$, in the Hausdorff metric. Moreover, if $(w,\theta) \in \mathrm{SOL}(M,q)$, such that $w+\theta > 0$, then there is a $(\hat{w}_n, \hat{\theta}_n) \in \mathrm{SOL}(\hat{M}_n, \hat{q}_n)$ for sufficiently large n such that $\{i \mid \hat{\theta}_{n,i} > 0\} = \{i \mid \theta_i > 0\}$.*

References

1. NOSNOC. https://github.com/nurkanovic/nosnoc, 2022.
2. V. Acary, O. Bonnefon, and B. Brogliato. *Nonsmooth modeling and simulation for switched circuits*, volume 69. Springer Science & Business Media, 2010.
3. V. Acary and B. Brogliato. *Numerical methods for nonsmooth dynamical systems: applications in mechanics and electronics*. Springer Science & Business Media, 2008.
4. V. Acary and B. Brogliato. Implicit Euler numerical scheme and chattering-free implementation of sliding mode systems. *Systems & Control Letters*, 59(5):284–293, 2010.
5. V. Acary, H. De Jong, and B. Brogliato. Numerical simulation of piecewise-linear models of gene regulatory networks using complementarity systems. *Physica D: Nonlinear Phenomena*, 269:103–119, 2014.
6. J. Albersmeyer. *Adjoint-based algorithms and numerical methods for sensitivity generation and optimization of large scale dynamic systems*. PhD thesis, University of Heidelberg, 2010.
7. J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi – a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36, 2019.
8. M. Anitescu, P. Tseng, and S. J. Wright. Elastic-mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity properties. *Mathematical Programming*, 110(2):337–371, 2007.
9. X. J. Ban, J.-S. Pang, H. X. Liu, and R. Ma. Continuous-time point-queue models in dynamic network loading. *Transportation Research Part B: Methodological*, 46(3):360–380, 2012.
10. B. T. Baumrucker and L. T. Biegler. MPEC strategies for optimization of a class of hybrid dynamic systems. *Journal of Process Control*, 19(8):1248–1256, 2009.
11. A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427, 1999.
12. H. G. Bock, C. Kirches, A. Meyer, and A. Potschka. Numerical solution of optimal control problems with explicit and implicit switches. *Optimization Methods and Software*, 33(3):450–474, 2018.

13. H. G. Bock and K. J. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. In *Proceedings of the IFAC World Congress*, pages 242–247. Pergamon Press, 1984.
14. B. Brogliato and A. Tanwani. Dynamical systems coupled with monotone set-valued operators: Formalisms, applications, well-posedness, and stability. *SIAM Review*, 62(1):3–129, 2020.
15. M.-G. Cojocaru. Dynamic equilibria of group vaccination strategies in a heterogeneous population. *Journal of Global Optimization*, 40(1):51–63, 2008.
16. G. Colombo, B. S. Mordukhovich, and D. Nguyen. Optimization of a perturbed sweeping process by constrained discontinuous controls. *SIAM Journal on Control and Optimization*, 58(4):2678–2709, 2020.
17. L. Dieci and L. Lopez. Sliding motion on discontinuity surfaces of high co-dimension. a construction for selecting a filippov vector field. *Numerische Mathematik*, 117(4):779–811, 2011.
18. A. Dontchev and F. Lempio. Difference methods for differential inclusions: A survey. *SIAM review*, 34(2):263–294, 1992.
19. A. L. Dontchev and T. R. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, 2014.
20. F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume 1-2. Springer-Verlag, 2003.
21. A. F. Filippov. On certain questions in the theory of optimal control. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(1):76–84, 1962.
22. A. F. Filippov. Differential Equations with discontinuous right hand side. *AMS Transl.*, 42:199–231, 1964.
23. A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides: Control Systems*, volume 18. Springer Science & Business Media, 1988.
24. L. Guo and J. J. Ye. Necessary optimality conditions for optimal control problems with equilibrium constraints. *SIAM Journal on Control and Optimization*, 54(5):2710–2733, 2016.
25. E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I*. Springer Series in Computational Mathematics. Springer, Berlin, 2nd edition, 1993.
26. E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*. Springer, Berlin Heidelberg, 2nd edition, 1991.
27. J. Hall, A. Nurkanović, F. Messerer, and M. Diehl. A sequential convex programming approach to solving quadratic programs and optimal control problems with linear complementarity constraints. *IEEE Control Systems Letters*, 6:536–541, 2022.
28. A. Hauswirth, S. Bolognani, G. Hug, and F. Dörfler. Optimization Algorithms as Robust Feedback Controllers. *arXiv preprint arXiv:2103.11329*, 2021.
29. A. Kastner-Maresch. Implicit Runge-Kutta methods for differential inclusions. *Numerical functional analysis and optimization*, 11(9-10):937–958, 1990.
30. S. Katayama, M. Doi, and T. Ohtsuka. A moving switching sequence approach for nonlinear model predictive control of switched systems with state-dependent switches and state jumps. *International Journal of Robust and Nonlinear Control*, 30(2):719–740, 2020.
31. C. Kirches. A Numerical Method for Nonlinear Robust Optimal Control with Implicit Discontinuities and an Application to Powertrain Oscillations. Diploma thesis, University of Heidelberg, October 2006.
32. C. Kirches, J. Larson, S. Leyffer, and P. Manns. Sequential linearization method for bound-constrained mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 32(1):75–99, 2022.
33. S. Leyffer, G. López-Calva, and J. Nocedal. Interior methods for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 17(1):52–77, 2006.
34. G. Matsaglia and G. PH Styan. Equalities and inequalities for ranks of matrices. *Linear and multilinear Algebra*, 2(3):269–292, 1974.
35. A. Nurkanović, S. Albrecht, B. Brogliato, and M. Diehl. The time-freezing reformulation for numerical optimal control of complementarity lagrangian systems with state jumps. *Automatica*, 158:111295, 2023.

36. A. Nurkanović, S. Albrecht, and M. Diehl. Limits of MPCC Formulations in Direct Optimal Control with Nonsmooth Differential Equations. In *2020 European Control Conference (ECC)*, pages 2015–2020, 2020.
37. A. Nurkanović and M. Diehl. Continuous optimization for control of hybrid systems with hysteresis via time-freezing. *IEEE Control Systems Letters*, 6:3182–3187, 2022.
38. A. Nurkanović and M. Diehl. Nosnoc: A software package for numerical optimal control of nonsmooth systems. *IEEE Control Systems Letters*, 6:3110–3115, 2022.
39. A. Nurkanović, A. Pozharskiy, and M. Diehl. Solving mathematical programs with complementarity constraints arising in nonsmooth optimal control. *arXiv preprint arXiv:2312.11022*, 2023.
40. A. Nurkanović, T. Sartor, S. Albrecht, and M. Diehl. A Time-Freezing Approach for Numerical Optimal Control of Nonsmooth Differential Equations with State Jumps. *IEEE Control Systems Letters*, 5(2):439–444, 2021.
41. L. S. Pontryagin. *The mathematical theory of optimal processes*. Wiley, 1962.
42. R. Pytlak and D. Suski. Algorithms for optimal control of hybrid systems with sliding motion. *arXiv preprint arXiv:2101.04754*, 2021.
43. D. Ralph and S. J. Wright. Some properties of regularization and penalization schemes for mpecs. *Optimization Methods and Software*, 19(5):527–556, 2004.
44. J. B. Rawlings, D. Q. Mayne, and M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill, 2nd edition, 2017.
45. M. S. Shaikh and P. E. Caines. On the hybrid optimal control problem: Theory and algorithms. *IEEE Transactions on Automatic Control*, 52(9):1587–1603, 2007.
46. D. E. Stewart. A high accuracy method for solving ODEs with discontinuous right-hand side. *Numerische Mathematik*, 58(1):299–328, 1990.
47. D. E. Stewart. *High accuracy numerical methods for ordinary differential equations with discontinuous right-hand side*. PhD thesis, University of Queensland, St. Lucia, Queensland 4072, Australia, 1990.
48. D. E. Stewart. A numerical method for friction problems with multiple contacts. *The ANZIAM Journal*, 37(3):288–308, 1996.
49. D. E. Stewart. Rigid-body dynamics with friction and impact. *SIAM Review*, 42(1):3–39, 2000.
50. D. E. Stewart. *Dynamics with Inequalities: Impacts and Hard Constraints*, volume 59. SIAM, 2011.
51. D. E. Stewart and M. Anitescu. Optimal control of systems with discontinuous differential equations. *Numerische Mathematik*, 114(4):653–695, 2010.
52. K. Taubert. Converging multistep methods for initial value problems involving multivalued maps. *Computing*, 27(2):123–136, 1981.
53. A. Vieira, B. Brogliato, and C. Prieur. Quadratic optimal control of linear complementarity systems: First-order necessary conditions and numerical analysis. *IEEE Transactions on Automatic Control*, 65(6):2743–2750, 2020.
54. A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.