

Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index

Peter Kunkel · Volker Mehrmann

Received: 9 September 2006 / Accepted: 8 September 2008 / Published online: 9 October 2008
© Springer-Verlag London Limited 2008

Abstract We study optimal control problems for general unstructured nonlinear differential-algebraic equations of arbitrary index. In particular, we derive necessary conditions in the case of linear-quadratic control problems and extend them to the general nonlinear case. We also present a Pontryagin maximum principle for general unstructured nonlinear DAEs in the case of restricted controls. Moreover, we discuss the numerical solution of the resulting two-point boundary value problems and present a numerical example.

Keywords Nonlinear differential-algebraic equation · Optimal control · Maximum principle · Solvability · Necessary optimality condition · Behavior approach · Strangeness index · Regularization

This research was supported through the Research-in-Pairs Program at Mathematisches Forschungsinstitut Oberwolfach. V. Mehrmann's research was supported by *Deutsche Forschungsgemeinschaft*, through MATHEON, the DFG Research Center "Mathematics for Key Technologies" in Berlin.

P. Kunkel
Mathematisches Institut, Universität Leipzig,
Augustusplatz 10–11, 04109 Leipzig, Germany
e-mail: kunkel@mathematik.uni-leipzig.de

V. Mehrmann (✉)
Institut für Mathematik, MA 4-5, Technische Universität Berlin,
10623 Berlin, Germany
e-mail: mehrmann@math.tu-berlin.de

1 Introduction

We study optimal control problems

$$\mathcal{J}(x, u) = \mathcal{M}(x(\bar{t})) + \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x(t), u(t)) dt = \min! \quad (1)$$

subject to a constraint given by an initial value problem associated with a nonlinear system of differential-algebraic equations (*descriptor system*) consisting of

$$F(t, x, u, \dot{x}) = 0 \quad (2)$$

and

$$x(\underline{t}) = \underline{x}. \quad (3)$$

We assume that $F \in C^0(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_u \times \mathbb{D}_{\dot{x}}, \mathbb{R}^m)$ is sufficiently smooth, that $\mathbb{I} = [\underline{t}, \bar{t}] \subseteq \mathbb{R}$ is a (compact) interval, and that $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n, \mathbb{D}_u \subseteq \mathbb{R}^l$ are open sets.

Throughout the paper, we will frequently make use of the *behavior representation* (see [54]) of the control problem, i.e., we combine (x, u) into one *generalized state vector* z and then study the optimization problem

$$\mathcal{J}(z) = \mathcal{M}(z(\bar{t})) + \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, z(t)) dt = \min! \quad (4)$$

subject to the constraint

$$F(t, z, \dot{z}) = 0, \quad (5)$$

and the initial condition

$$[I_n \ 0]z(\underline{t}) = [I_n \ 0]\underline{z}. \quad (6)$$

Optimal control problems like (1)–(3) arise in the control of mechanical multibody systems [22, 24], electrical circuits [25, 26], chemical engineering [20, 21] or heterogeneous systems, where different models are coupled together [52].

The theory of optimal control problems for ordinary differential equations is well established since the middle of the twentieth century, see, e.g., [7, 27, 30–32, 59] and the references therein. For systems where the constraint is a differential-algebraic equation, the situation is much more difficult and the existing literature is more recent. First results, mainly for special cases such as linear constant coefficient systems or semi-explicit systems of index 1, were obtained in [6, 16, 48, 50, 51, 53, 58].

A major difficulty in deriving adjoint equations, optimality systems or even a maximum principle for general higher index DAEs is that for the potential candidates of adjoint equations and optimality systems, existence and uniqueness of solutions cannot be guaranteed, see [1, 3, 19, 40, 53, 57] for examples and discussion of the difficulties.

Due to these difficulties, the standard approach to deal with optimal control problems for DAEs is to first perform regularization and index reduction via feedback or differentiation. Conditions when such transformations exist have been studied in [9, 11, 12] and in their most general form in [37, 41]. Some of these results were reproduced and extended in a different setting in the recent work of [1, 40].

There also exist some papers that derive optimality conditions for specially structured higher index systems directly. For semi-explicit systems of index 1, a general maximum principle was proved in [53] and extended to systems up to differentiation index 3 in [57]. Further results for index 2 systems are presented in [24], for multibody systems in [10, 23] and for DAEs with properly stated leading term in [1, 2, 4, 5, 40].

In this paper, we take a more general approach and discuss general unstructured linear and nonlinear systems of arbitrary index. In particular, we generalize results of [53, 57]. We follow the strangeness index concept, see [38], and consider the system in a behavior setting as a general over- or underdetermined differential-algebraic system. For this behavior system a derivative array, see [15], is formed and from this array, a reduced control problem is determined that has the same solution set (in the behavior setting) but is essentially index 1. On the basis of this reduced system, the optimality conditions are then derived.

The paper is organized as follows. We first give a brief survey of the theoretical results on the strangeness index concept that will be needed in Sect. 2 and recall some general functional analytic results on optimization in Banach spaces. Then we derive necessary optimality conditions for optimal control problems subject to general linear and nonlinear unstructured DAEs in Sect. 3. Furthermore, a Pontryagin maximum principle for general DAEs will be presented. In Sect. 4, we describe another formulation of the optimality boundary value problem that can be used in the context of numerical methods and present a numerical example. Finally, we give some conclusions in Sect. 5.

2 Preliminaries

In this section, we will introduce some notation and recall some results on differential-algebraic equations and on optimization in Banach spaces. Throughout the paper, we assume that all functions are sufficiently smooth, i.e., sufficiently often continuously differentiable.

2.1 Notation

We will make frequent use of the *Moore-Penrose pseudoinverse* of a matrix valued function $A: \mathbb{I} \rightarrow \mathbb{R}^{m,n}$, which is the unique matrix function $A^+: \mathbb{I} \rightarrow \mathbb{R}^{n,m}$ that satisfies the four Penrose axioms

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)^T = AA^+, \quad (A^+A)^T = A^+A \quad (7)$$

pointwise, see, e.g., [18]. Note that if $A \in C^k(\mathbb{I}, \mathbb{R}^{m,n})$ and has constant rank on \mathbb{I} then $A^+ \in C^k(\mathbb{I}, \mathbb{R}^{n,m})$.

In the context of restricted control values, we must allow for bounded (with respect to the L_∞ -norm) and up to a finite number of points continuous control functions. We denote the set of all these functions on the interval \mathbb{I} by $L_\infty^c(\mathbb{I}, \mathbb{R}^l)$.

2.2 DAE theory

The theory of differential-algebraic equations has changed significantly in the past 20 years, see [8, 28, 38, 56]. We recall some necessary concepts and follow [38] in notation and style of presentation.

When studying control problems like (2), one can essentially distinguish two viewpoints. The first possibility is to take the behavior approach and to consider the optimization problem (4) subject to (5) and (6). For this underdetermined system one can study existence and uniqueness of solutions. In this setting, feedbacks are just standard equivalence transformations. If one carries out a transformation to canonical or condensed form, then index reduction and regularization via feedback follow directly, see [38]. If it is clear that the variables u really describe input variables and the variables x states, as is often the case in practice, then the second possibility is to keep the states and controls separate. In this case, one has to distinguish whether solutions exist for all controls in a given input set \mathbb{U} or whether there exist controls at all for which the system is solvable.

To discuss these questions, we consider the following solution concept.

Definition 1 Consider system (2) with a given fixed input function u that is sufficiently smooth. A function $x : \mathbb{I} \rightarrow \mathbb{R}^n$ is called a *solution* of (2) if $x \in C^1(\mathbb{I}, \mathbb{R}^n)$ and x satisfies (2) pointwise. It is called a *solution of the initial value problem (2)–(3)* if x is a solution of (2) and satisfies (3). An initial condition (3) is called *consistent* if the corresponding initial value problem has at least one solution.

It is possible to weaken this solution concept [33, 45, 49, 55]. In particular, it will turn out that it is actually necessary to slightly weaken this solution concept to define underlying Banach space operators with appropriate properties. But to do so, we first must introduce some additional theory. Note, however, that under the assumption of sufficient smoothness we will always be in the case of Definition 1.

Definition 2 A control problem of the form (2) with a given set of controls \mathbb{U} is called *consistent* (with \mathbb{U}) if there exists an input function $u \in \mathbb{U}$ for which there exists a solution x in the sense of Definition 1.

It is called *regular* [locally with respect to a given solution (\hat{x}, \hat{u}) of (2)] if it has a unique solution for every sufficiently smooth input function u in a neighborhood of \hat{u} and every initial value in a neighborhood of $\hat{x}(\underline{t})$ that is consistent for the system with input function u .

To analyze the properties of the system, in [36] for the square nonlinear case, in [41] for the rectangular linear case, and in [37] for the general over- and underdetermined case, hypotheses have been formulated which lead to an index concept, the so-called strangeness index. See [38] for a detailed derivation and analysis of this concept.

To summarize the strangeness index concept, we consider the constraint system in the form (5). As in [36], we introduce a nonlinear derivative array, see also [14, 17], of the form

$$F_\ell(t, z, \dot{z}, \dots, z^{(\ell+1)}) = 0, \tag{8}$$

which stacks the original equation and all its derivatives up to level ℓ in one large system, i.e.,

$$F_\ell(t, z, \dot{z}, \dots, z^{(\ell+1)}) = \begin{bmatrix} F(t, z, \dot{z}) \\ \frac{d}{dt} F(t, z, \dot{z}) \\ \vdots \\ \frac{d^\ell}{dt^\ell} F(t, z, \dot{z}) \end{bmatrix}. \tag{9}$$

Partial derivatives of F_ℓ with respect to selected variables p from $(t, z, \dot{z}, \dots, z^{(\ell+1)})$ are denoted by $F_{\ell;p}$, e.g.,

$$F_{\ell;z} = \frac{\partial}{\partial z} F_\ell, \quad F_{\ell;\dot{z}, \dots, z^{(\ell+1)}} = \left[\frac{\partial}{\partial \dot{z}} F_\ell \quad \dots \quad \frac{\partial}{\partial z^{(\ell+1)}} F_\ell \right].$$

A corresponding notation is also used for partial derivatives of other functions.

To analyze existence and uniqueness of solutions, we need to introduce the solution set of the nonlinear algebraic equation associated derivative array F_μ for some integer μ . We denote this solution set by

$$\mathbb{L}_\mu = \{z_\mu \in \mathbb{I} \times \mathbb{R}^n \times \mathbb{R}^n \times \dots \times \mathbb{R}^n \mid F_\mu(z_\mu) = 0\}. \tag{10}$$

Then we make the following hypothesis, see [38].

Hypothesis 3 Consider the general system of nonlinear differential-algebraic equations (5). There exist integers $\mu, r, a, d,$ and v such that \mathbb{L}_μ is not empty and such that for every $z_\mu^0 = (t_0, z_0, \dot{z}_0, \dots, z_0^{(\mu+1)}) \in \mathbb{L}_\mu$ there exists a (sufficiently small) neighborhood in which the following properties hold:

1. The set $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)n+1}$ forms a manifold of dimension $(\mu + 2)n + 1 - r$.
2. We have $\text{rank } F_{\mu;z, \dot{z}, \dots, z^{(\mu+1)}} = r$ on \mathbb{L}_μ .
3. We have $\text{corank } F_{\mu;z, \dot{z}, \dots, z^{(\mu+1)}} - \text{corank } F_{\mu-1;z, \dot{z}, \dots, z^{(\mu)}} = v$ on \mathbb{L}_μ , where the corank is the dimension of the corange and the convention is used that $\text{corank } F_{-1;z} = 0$.

4. We have $\text{rank } F_{\mu; \dot{z}, \dots, z^{(\mu+1)}} = r - a$ on \mathbb{L}_μ such that there exist smooth full rank matrix functions Z_2 and T_2 of size $(\mu + 1)m \times a$ and $n \times (n - a)$, respectively, satisfying

$$Z_2^T F_{\mu; \dot{z}, \dots, z^{(\mu+1)}} = 0, \quad \text{rank } Z_2^T F_{\mu; z} = a, \quad Z_2^T F_{\mu; z} T_2 = 0 \tag{11}$$

on \mathbb{L}_μ .

5. We have $\text{rank } F_{\dot{z}} T_2 = d = m - a - v$ on \mathbb{L}_μ such that there exists a smooth full rank matrix function Z_1 of size $n \times d$ satisfying $\text{rank } Z_1^T F_{\dot{z}} T_2 = d$.

Note that, since the Gram–Schmidt orthonormalization is a continuous process, we can assume without loss of generality that the matrix function Z_2 , T_2 , and Z_1 have pointwise orthonormal columns.

As in [36, 38], we call the smallest possible μ for which Hypothesis 3 is valid the *strangeness index* of (5). Systems with vanishing strangeness index are called *strangeness free*. The strangeness index is closely related to the differentiation index, see [8], but it should be observed that it allows over- and underdetermined systems and the counting is different, since in the strangeness index concept ordinary differential equations and purely algebraic equations both have $\mu = 0$. See [38] for a detailed analysis of the relationship between different index concepts.

It has been shown in [37] that Hypothesis 3 implies locally (via the implicit function theorem) the existence of a *reduced system* given by

$$\begin{aligned} \text{(a)} \quad & \hat{F}_1(t, z_1, z_2, z_3, \dot{z}_1, \dot{z}_2, \dot{z}_3) = 0, \\ \text{(b)} \quad & \hat{F}_2(t, z_1, z_2, z_3) = 0, \end{aligned} \tag{12}$$

with $\hat{F}_1 = Z_1^T F$, where $(z_1, z_2, z_3) \in \mathbb{R}^d \times \mathbb{R}^{n-a-d} \times \mathbb{R}^a$ is a suitable splitting of the unknown z . Part 4 of Hypothesis 3 guarantees that (12b) can be solved for z_3 according to $z_3 = \mathcal{R}(t, z_1, z_2)$. Eliminating z_3 and \dot{z}_3 in (12a) with the help of this relation and its derivative then leads to

$$\begin{aligned} & \hat{F}_1(t, z_1, z_2, \mathcal{R}(t, z_1, z_2), \dot{z}_1, \dot{z}_2, \mathcal{R}_t(t, z_1, z_2) + \mathcal{R}_{z_1}(t, z_1, z_2)\dot{z}_1 \\ & + \mathcal{R}_{z_2}(t, z_1, z_2)\dot{z}_2) = 0. \end{aligned}$$

By part 5 of Hypothesis 3, we may assume without loss of generality that this system can (locally) be solved for \dot{z}_1 leading to the system

$$\begin{aligned} \dot{z}_1 &= \mathcal{L}(t, z_1, z_2, \dot{z}_2), \\ z_3 &= \mathcal{R}(t, z_1, z_2). \end{aligned} \tag{13}$$

Obviously, in this system, interpreted as a DAE, $z_2 \in C^1(\mathbb{I}, \mathbb{R}^{n-a-d})$ can be chosen arbitrarily (at least when staying in the domain of definition of \mathcal{R} and \mathcal{L}), while the resulting system has locally a unique solution for z_1 and z_3 , provided that a consistent initial condition is given. This means that z_2 can be interpreted as a control, see also the

discussion in Sect. 3.2. We summarize these observations in the following theorem, see [37,38].

Theorem 4 *Let F in (2) be sufficiently smooth and satisfy Hypothesis 3 with μ , a , d , v . Then every sufficiently smooth solution of (5) also solves the reduced problems (12) and (13) consisting of d differential and a algebraic equations.*

Under some further assumptions, the converse of Theorem 4 holds as well, see again [37,38].

Theorem 5 *Let F in (2) be sufficiently smooth and satisfy Hypothesis 3 with μ , a , d , v and $\mu+1$ (replacing μ), a , d , v . Let $z_{\mu+1}^0 \in \mathbb{L}_{\mu+1}$ be given and let the parameterization of $\mathbb{L}_{\mu+1}$ include \dot{z}_2 . Then, for every function $z_2 \in C^1(\mathbb{I}, \mathbb{R}^{n-a-d})$ with $z_2(t_0) = z_{2,0}$, $\dot{z}_2(t_0) = \dot{z}_{2,0}$, the reduced DAEs (12) and (13) have unique solutions z_1 and z_3 satisfying $z_1(t_0) = z_{1,0}$. Moreover, the so-obtained function $z = (z_1, z_2, z_3)$ locally solves the original problem.*

The quantity v in Theorem 4, which has not been addressed yet, measures the number of equations in the original system that give rise to trivial equations $0 = 0$, i.e., it counts the number of redundancies in the system. Together with a and d it gives a complete classification of the m equations into d differential equations, a algebraic equations and v trivial equations. Of course, trivial equations can be simply removed without altering the solution set.

If the variable z is a combined vector of states and controls, then, since (12) consists of original variables, these can again be split into parts stemming from x and u . It has been shown in [37,41]; see also [38], how this system can then be treated in the control context concerning solvability, regularizability, and model consistency.

Theorem 6 *Suppose that the control problem (2) in the form (5) satisfies Hypothesis 3 with μ , a , d , v and assume that $d + a = n$. Then there (locally) exists a state feedback $u = K(t, x)$ satisfying the initial condition*

$$u(\underline{t}) = \underline{u} = K(\underline{t}, \underline{x}), \quad \dot{u}(\underline{t}) = \dot{\underline{u}} = K_t(\underline{t}, \underline{x}) + K_x(\underline{t}, \underline{x})\dot{\underline{x}} \quad (14)$$

such that the resulting closed loop reduced problem is regular and strangeness free.

Corollary 7 *Suppose that the control problem (2) in the form (5) satisfies Hypothesis 3 with μ , a , d , v and with $\mu + 1$ (replacing μ), a , d , v and assume that $d + a = n$. Furthermore, let u be a control in the sense that u and \dot{u} can be chosen as part of the parametrization of $\mathbb{L}_{\mu+1}$ at $z_{\mu+1}^0 \in \mathbb{L}_{\mu+1}$. Let $u = K(t, x)$ be a state feedback which satisfies the initial conditions (14) and yields a regular and strangeness-free closed loop reduced system. Then, the closed loop reduced problem has a unique solution satisfying the initial values given by $z_{\mu+1}^0$. Moreover, this solution locally solves the closed loop problem*

$$F(t, x, K(t, x), \dot{x}) = 0.$$

Similar results are given in [37,38] for output control problems, but in this paper we restrict our attention to optimal control problems without output equation.

Note that due to the application of the implicit function theorem, the above results are only valid locally. For linear problems, the local results automatically hold globally. In the general nonlinear case, however, when global constructions are needed as in the present case, a more detailed analysis is required. We will present corresponding results in Sect. 3.2.

2.3 Optimization in Banach spaces

We recall some results from general optimization theory, see, e.g., [60]. For this consider the optimization problem

$$\mathcal{J}(z) = \min! \tag{15}$$

subject to the constraint

$$\mathcal{F}(z) = 0, \tag{16}$$

where

$$\mathcal{J} : \mathbb{D} \rightarrow \mathbb{R}, \quad \mathcal{F} : \mathbb{D} \rightarrow \mathbb{Y}, \quad \mathbb{D} \subseteq \mathbb{Z} \text{ open},$$

with real Banach spaces \mathbb{Z}, \mathbb{Y} . Let, furthermore,

$$z^* \in \mathbb{M} = \{z \in \mathbb{D} \mid \mathcal{F}(z) = 0\}.$$

Then we have the following theorem which is due to [47].

Theorem 8 *Let \mathcal{J} be Fréchet differentiable in z^* and let \mathcal{F} be a submersion in z^* , i.e., let \mathcal{F} be Fréchet differentiable in a neighborhood of z^* with Fréchet derivative $D\mathcal{F}(z^*) : \mathbb{Z} \rightarrow \mathbb{Y}$ surjective and kernel $D\mathcal{F}(z^*)$ continuously projectable.*

If z^ is a local minimum of (15), then there exists a unique Λ in the dual space \mathbb{Y}^* of \mathbb{Y} with*

$$D\mathcal{J}(z^*)\Delta z + \Lambda(D\mathcal{F}(z^*)\Delta z) = 0 \quad \text{for all } \Delta z \in \mathbb{Z}. \tag{17}$$

The functional Λ in Theorem 8 is called the *Lagrange multiplier*, associated with the constraint (16).

In general, we are interested in function representations of the Lagrange multiplier functional Λ . Such representations are obtained by the following theorem.

Lemma 9 *Let $\mathbb{Y} = C^0(\mathbb{I}, \mathbb{R}^m) \times \mathbb{V}$ with a vector space $\mathbb{V} \subseteq \mathbb{R}^m$ and let $(\lambda, \gamma) \in \mathbb{Y}$. Then*

$$\Lambda(g, r) = \int_t^{\bar{t}} \lambda(t)^T g(t) dt + \gamma^T r$$

defines a linear form $\Lambda \in \mathbb{Y}^*$, which conversely uniquely determines $(\lambda, \gamma) \in \mathbb{Y}$.

A sufficient condition that guarantees that the minimum is also unique, is given by the following theorem, which, e.g., covers linear-quadratic control problems with positive definite reduced Hessian.

Theorem 10 *Suppose that $\mathcal{F} : \mathbb{Z} \rightarrow \mathbb{Y}$ is affine linear and that $\mathcal{J} : \mathbb{Z} \rightarrow \mathbb{R}$ is strictly convex on \mathbb{M} , i.e.,*

$$\begin{aligned} \mathcal{J}(\alpha z_1 + (1 - \alpha)z_2) &< \alpha \mathcal{J}(z_1) + (1 - \alpha)\mathcal{J}(z_2) \\ &\text{for all } z_1, z_2 \in \mathbb{M} \text{ with } z_1 \neq z_2 \text{ and for all } \alpha \in (0, 1), \end{aligned}$$

then the optimization problem (15) subject to (16) has a unique minimum.

For the analysis and solution of optimal control problems subject to constraints given by differential-algebraic equations, we will have to carry out changes of variables and linear or nonlinear feedbacks. To see how these effect the minimization problem, consider a local diffeomorphism $\phi : \mathbb{Z} \rightarrow \mathbb{Z}$ in a neighborhood of \tilde{z}^* with $z^* = \phi(\tilde{z}^*)$. If we transform the optimization problem (15) and the constraint (16) to the new variable \tilde{z} via $z = \phi(\tilde{z})$, then we obtain the transformed optimization problem

$$\tilde{\mathcal{J}}(\tilde{z}) = \min!$$

subject to the constraint

$$\tilde{\mathcal{F}}(\tilde{z}) = 0,$$

where

$$\tilde{\mathcal{J}}(\tilde{z}) = \mathcal{J}(\phi(\tilde{z})), \quad \tilde{\mathcal{F}}(\tilde{z}) = \mathcal{F}(\phi(\tilde{z})).$$

If \tilde{z}^* satisfies the necessary condition (17) in the form

$$D\tilde{\mathcal{J}}(\tilde{z}^*)\Delta\tilde{z} + \Lambda(D\tilde{\mathcal{F}}(\tilde{z}^*)\Delta\tilde{z}) = 0 \quad \text{for all } \Delta\tilde{z} \in \mathbb{Z},$$

then

$$D\mathcal{J}(\phi(\tilde{z}^*))D\phi(\tilde{z}^*)\Delta\tilde{z} + \Lambda(D\mathcal{F}(\phi(\tilde{z}^*))D\phi(\tilde{z}^*)\Delta\tilde{z}) = 0 \quad \text{for all } \Delta\tilde{z} \in \mathbb{Z}.$$

With $\Delta z = D\phi(\tilde{z}^*)\Delta\tilde{z}$ we then have

$$D\mathcal{J}(z^*)\Delta z + \Lambda(D\mathcal{F}(z^*)\Delta z) = 0 \quad \text{for all } \Delta z \in \mathbb{Z},$$

and thus, z satisfies the necessary condition (17) for the optimization problem (15) subject to (16).

If the controls are restricted by $u(t) \in \mathcal{U}$, then we must admit bang-bang controls. In this case, the optimal solution is obtained via versions of the Pontryagin maximum principle.

For the *Bolza problem* to determine $(\bar{t}, x, u) \in \mathbb{R} \times C^0(\mathbb{I}, \mathbb{R}^n) \times L^\infty_c(\mathbb{I}, \mathbb{R}^l)$ as a solution of

$$\mathcal{J}(x, u) = \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x(t), u(t)) \, dt = \min! \tag{18}$$

subject to

$$\begin{aligned} x(t) &= \underline{x} + \int_{\underline{t}}^t f(s, x(s), u(s)) \, ds \\ 0 &= h(\bar{t}, x(\bar{t})), \quad h \in C(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n), \\ u(t) &\in \mathcal{U} \subset \mathbb{R}^l \quad \text{for all } t \in \mathbb{I}, \end{aligned} \tag{19}$$

one has the following theorem.

Theorem 11 *If (\bar{t}^*, x^*, u^*) is a local solution of the Bolza problem (18) subject to (19) with unknowns (\bar{t}, x, u) , then there exist scalars $\alpha_0, \alpha_1, \dots, \alpha_n$, which all do not vanish simultaneously, $\alpha_0 \geq 0$, and a multiplier $\lambda \in C^0(\mathbb{I}, \mathbb{R}^n)$ such that, with $H(t, x, u, \lambda, \alpha_0) = \lambda^T f(t, x, u) - \alpha_0 \mathcal{K}(t, x, u)$ and $\alpha = (\alpha_1, \dots, \alpha_n)^T$,*

$$\begin{aligned} H(t, x^*(t), u^*(t), \lambda(t), \alpha_0) &= \max_{u \in \mathcal{U}} H(t, x^*(t), u, \lambda(t), \alpha_0), \\ \dot{\lambda}(t) &= -\nabla_x H(t, x^*(t), u^*(t), \lambda(t), \alpha_0), \\ \dot{x}^*(t) &= \nabla_\lambda H(t, x^*(t), u^*(t), \lambda(t), \alpha_0), \\ \lambda(\bar{t}^*) &= -\alpha^T \nabla_{x(\bar{t})} h(\bar{t}^*, x^*(\bar{t}^*)), \end{aligned} \tag{20}$$

for all $t \in \mathbb{I}$, where u^* is continuous.

3 Necessary conditions

In this section, we will derive necessary optimality conditions for the minimization of (1) subject to (2) and (3). We first start with the special case of a linear-quadratic optimal control problem and then extend the results to the general case.

3.1 Linear-quadratic optimal control problems

The linear-quadratic optimal control problem for differential-algebraic equations has been well studied for constant coefficient systems, see [50] and the references therein, and variable coefficient problems, which possess a specific canonical form, in

[35]. These results are based on the idea to first use index reduction and feedback regularization to transform the problem into a regular, strangeness-free problem and then to use the analysis for this case. Recently, in [1, 3, 40] this problem was studied again in a different setting for some restricted classes of linear DAEs with small index.

Here, we study the general case of unstructured linear-quadratic optimal problems, i.e., we study the cost functional

$$\mathcal{J}(x, u) = \frac{1}{2}x(\bar{t})^T Mx(\bar{t}) + \frac{1}{2} \int_t^{\bar{t}} (x^T Wx + 2x^T Su + u^T Ru) dt, \tag{21}$$

with $W \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $S \in C^0(\mathbb{I}, \mathbb{R}^{n,l})$, $R \in C^0(\mathbb{I}, \mathbb{R}^{l,l})$, and we assume, furthermore, that W and R are pointwise symmetric and also that $M \in \mathbb{R}^{n,n}$ is symmetric. As constraint we consider the initial value problem for a general linear differential-algebraic equations with variable coefficients of the form

$$E\dot{x} = Ax + Bu + f, \quad x(t) = \underline{x}, \tag{22}$$

with $E \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $B \in C^0(\mathbb{I}, \mathbb{R}^{n,l})$, $f \in C^0(\mathbb{I}, \mathbb{R}^n)$, and $\underline{x} \in \mathbb{R}^n$. Our aim is to determine optimal controls $u \in \mathbb{U} = C^0(\mathbb{I}, \mathbb{R}^l)$.

We could discuss a more general situation and allow the coefficient functions E , A to be non-square but, as it has been shown in [37], this case can always be transformed to the regular square case. To avoid unnecessary technicalities, we therefore restrict ourselves here to the regular square case.

For a better readability of the more complicated formulas, we omit here and in the following the argument t of the involved coefficient functions.

In the case of linear ordinary differential equations, corresponding to the case $E(t) = I$ in (22), the initial value problem has a unique solution $x \in C^1(\mathbb{I}, \mathbb{R}^n)$ for every $u \in \mathbb{U}$, $f \in C^0(\mathbb{I}, \mathbb{R}^n)$, and initial value $\underline{x} \in \mathbb{R}^n$. In contrast to this, in the case of differential-algebraic equations, where $E(t)$ may be singular, the equation is not necessarily (uniquely) solvable for any $u \in \mathbb{U}$ and also the initial conditions may be restricted, see [38]. Furthermore, it will be necessary to consider solutions $x \in \mathbb{X}$, where \mathbb{X} usually is a larger space than $C^1(\mathbb{I}, \mathbb{R}^n)$.

For our analysis, we consider the system in behavior form (5)

$$\mathcal{E}\dot{z} = \mathcal{A}z + f, \tag{23}$$

with

$$\mathcal{E} = [E \ 0], \quad \mathcal{A} = [A \ B].$$

Its associated derivative array is given by

$$M_\ell(t)\dot{z}_\ell = N_\ell(t)z_\ell + g_\ell(t), \tag{24}$$

where

$$\begin{aligned}
 (M_\ell)_{i,j} &= \binom{i}{j} \mathcal{E}^{(i-j)} - \binom{i}{j+1} \mathcal{A}^{(i-j-1)}, \quad i, j = 0, \dots, \ell, \\
 (N_\ell)_{i,j} &= \begin{cases} \mathcal{A}^{(i)} & \text{for } i = 0, \dots, \ell, \quad j = 0, \\ 0 & \text{otherwise,} \end{cases} \\
 (z_\ell)_j &= z^{(j)}, \quad j = 0, \dots, \ell, \\
 (g_\ell)_i &= f^{(i)}, \quad i = 0, \dots, \ell.
 \end{aligned}$$

We assume that this system has a well-defined strangeness index μ according to Hypothesis 3 and, furthermore, as we have already stated before, we assume that there is no consistency condition for the inhomogeneities, i.e., $v = 0$. Under the assumptions of Theorem 5, the initial value problem (22) is equivalent (in the sense that it has the same set of solutions) to the *reduced system*

$$\hat{E}\dot{x} = \hat{A}x + \hat{B}u + \hat{f}, \quad x(\underline{t}) = \underline{x}, \tag{25}$$

where

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix}, \quad \hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix} \tag{26}$$

with

$$\begin{aligned}
 \hat{E}_1 &= Z_1^T E, \quad [\hat{A}_1 \quad \hat{B}_1] = Z_1^T [A \quad B], \quad \hat{f}_1 = Z_1^T f, \\
 [\hat{A}_2 \quad \hat{B}_2] &= Z_2^T N_\mu [I_{n+l} \quad 0 \quad \dots \quad 0]^T, \quad \hat{f}_2 = Z_2^T g_\mu.
 \end{aligned}$$

By construction, the reduced system (25) is strangeness free. In particular, the matrix function \hat{E}_1 has full row rank d and $[\hat{A}_2 T_2' \quad \hat{B}_2]$ has full row rank a with a matrix function T_2' satisfying $\hat{E}_1 T_2' = 0$ and $T_2'^T T_2' = I_a$. Due to the fact that the solution set has not changed, one can consider the minimization of (21) subject to (25) instead of (22). Unfortunately, (25) still may not be solvable for all $u \in \mathbb{U}$. But, since $[\hat{A}_2 T_2' \quad \hat{B}_2]$ has full row rank, it follows from Theorem 6, see also [41], that there exists a linear feedback

$$u = Kx + w, \tag{27}$$

with $K \in C^0(\mathbb{I}, \mathbb{R}^{l,n})$ such that in the closed loop system

$$\hat{E}\dot{x} = (\hat{A} + \hat{B}K)x + \hat{B}w + \hat{f}, \quad x(\underline{t}) = \underline{x}, \tag{28}$$

the matrix function $(\hat{A}_2 + \hat{B}_2 K)T_2'$ is pointwise nonsingular, implying that the DAE in (28) is regular and strangeness free for every given $w \in \mathbb{U}$.

If we insert the feedback (27) in (25), then we obtain an optimization problem for the variables x, w instead of x, u , and according to the analysis in Sect. 2.3, these

problems and the solutions are directly transferable to each other. For this reason we may in the following assume w.l.o.g. that the differential-algebraic system (22) is regular and strangeness free as a free system without control, i.e., when $u = 0$.

Under these assumptions it is then known, see, e.g., [38], that there exist $P \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ pointwise orthogonal such that

$$\begin{aligned} \tilde{E} &= PEQ = \begin{bmatrix} E_{1,1} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{A} = PAQ - PE\dot{Q} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \\ \tilde{B} &= PB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \tilde{f} = Pf = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad x = Q\tilde{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ \underline{x} &= Q\tilde{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \end{aligned} \tag{29}$$

with $E_{1,1} \in C(\mathbb{I}, \mathbb{R}^{d,d})$ and $A_{2,2} \in C(\mathbb{I}, \mathbb{R}^{a,a})$ pointwise nonsingular. To get solvability of (22) for arbitrary $u \in \mathbb{U}$ and $f \in C(\mathbb{I}, \mathbb{R}^n)$, in view of

$$E\dot{x} = EE^+E\dot{x} = E \frac{d}{dt}(E^+Ex) - E \frac{d}{dt}(E^+E)x,$$

we have to interpret (22) as

$$E \frac{d}{dt}(E^+Ex) = (A + E \frac{d}{dt}(E^+E))x + Bu + f, \quad (E^+Ex)(\underline{t}) = \underline{x}, \tag{30}$$

which allows the larger solution space, see [34],

$$\mathbb{X} = C^1_{E^+E}(\mathbb{I}, \mathbb{R}^n) = \left\{ x \in C^0(\mathbb{I}, \mathbb{R}^n) \mid E^+Ex \in C^1(\mathbb{I}, \mathbb{R}^n) \right\} \tag{31}$$

equipped with the norm

$$\|x\|_{\mathbb{X}} = \|x\|_{C^0} + \left\| \frac{d}{dt}(E^+Ex) \right\|_{C^0}. \tag{32}$$

One should note that the choice of the initial value \underline{x} is restricted by the requirement in (30).

Following [34], we can use in (16) the constraint function

$$\mathcal{F} : \mathbb{X} \rightarrow \mathbb{Y} = C^0(\mathbb{I}, \mathbb{R}^n) \times \text{range } E^+(\underline{t})E(\underline{t})$$

given by

$$\mathcal{F}(x) = \left(E \frac{d}{dt}(E^+Ex) - (A + E \frac{d}{dt}(E^+E))x - Bu - f, (E^+Ex)(\underline{t}) - \underline{x} \right).$$

Our task now is to show that this \mathcal{F} satisfies the assumptions of Theorem 8.

From (30) we obtain

$$PEQQ^T \frac{d}{dt} (QQ^T E + P^T PEQQ^T x) = (PAQ + PEQQ^T \frac{d}{dt} (QQ^T E + P^T PEQQ^T) Q) Q^T x + PBu + Pf,$$

or equivalently

$$\tilde{E} Q^T \frac{d}{dt} (Q\tilde{E}^+ \tilde{E}\tilde{x}) = (\tilde{A} + P P^T \tilde{E} Q^T \dot{Q} + \tilde{E} Q^T \frac{d}{dt} (Q\tilde{E}^+ \tilde{E} Q^T) Q) \tilde{x} + \tilde{B}u + \tilde{f}.$$

Using the product rule and cancelling equal terms on both sides we obtain

$$\tilde{E} Q^T Q \frac{d}{dt} (\tilde{E}^+ \tilde{E}\tilde{x}) = (\tilde{A} + \tilde{E} Q^T \dot{Q} + \tilde{E} \frac{d}{dt} (\tilde{E}^+ \tilde{E}) + \tilde{E} \tilde{E}^+ \tilde{E} \dot{Q}^T Q) \tilde{x} + \tilde{B}u + \tilde{f}.$$

Since by definition $\tilde{E} \tilde{E}^+ \tilde{E} = \tilde{E}$ and $\dot{Q}^T Q + Q^T \dot{Q} = 0$, we then obtain

$$\tilde{E} \frac{d}{dt} (\tilde{E}^+ \tilde{E}\tilde{x}) = (\tilde{A} + \tilde{E} \frac{d}{dt} (\tilde{E}^+ \tilde{E})) \tilde{x} + \tilde{B}u + \tilde{f}, \quad (\tilde{E}^+ \tilde{E}\tilde{x})(\underline{t}) = \tilde{x}, \tag{33}$$

i.e., (30) transforms covariantly with pointwise orthogonal P and Q . If we partition P and Q conformably to (29) as

$$P = \begin{bmatrix} Z'^T \\ Z^T \end{bmatrix}, \quad Q = [T' \ T],$$

then $Z^T E = 0, ET = 0$, and we can write (33) as

$$\begin{bmatrix} E_{1,1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad \begin{bmatrix} x_1(\underline{t}) \\ 0 \end{bmatrix} = \begin{bmatrix} \underline{x}_1 \\ 0 \end{bmatrix}.$$

Since $A_{2,2}$ is pointwise nonsingular, this system is uniquely solvable for arbitrary continuous functions u, f_1 , and f_2 , and for any \underline{x}_1 , with solution components satisfying

$$x_1 \in C^1(\mathbb{I}, \mathbb{R}^d), \quad x_2 \in C^0(\mathbb{I}, \mathbb{R}^a)$$

such that

$$x = Q\tilde{x} = [T' \ T] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{X}.$$

In particular, this construction defines a solution operator of the form

$$S : \mathbb{U} \times \mathbb{Y} \rightarrow \mathbb{X}, \quad (u, f, \underline{x}) \mapsto x. \tag{34}$$

The Fréchet derivative $D\mathcal{F}(z)$ of \mathcal{F} at $z \in \mathbb{Z} = \mathbb{X} \times \mathbb{U}$ is given by

$$D\mathcal{F}(z)\Delta z = \left(E \frac{d}{dt}(E^+ E \Delta x) - (A + E \frac{d}{dt}(E^+ E))\Delta x - B \Delta u, (E^+ E \Delta x)(\underline{t}) \right).$$

For $(g, r) \in \mathbb{Y}$, the equation $D\mathcal{F}(z) = (g, r)$ then takes the form

$$E \frac{d}{dt}(E^+ E \Delta x) - (A + E \frac{d}{dt}(E^+ E))\Delta x - B \Delta u = g, \quad (E^+ E \Delta x)(\underline{t}) = r.$$

A possible solution is given by $u = 0$ and $\Delta x = \mathcal{S}(0, g, r)$; hence, $D\mathcal{F}(z)$ is surjective. Moreover, the kernel is given by

$$\begin{aligned} \text{kernel}(D\mathcal{F}(z)) &= \{(\Delta x, \Delta u) \mid E \frac{d}{dt}(E^+ E \Delta x) - (A + E \frac{d}{dt}(E^+ E))\Delta x - B \Delta u = 0, \\ &\quad (E^+ E \Delta x)(\underline{t}) = 0\} \\ &= \{(\Delta x, \Delta u) \mid \Delta x = \mathcal{S}(\Delta u, 0, 0), \Delta u \in \mathbb{U}\} \subseteq \mathbb{X} \times \mathbb{U}. \end{aligned}$$

Observe that $\text{kernel}(D\mathcal{F}(z))$ is parameterized with respect to Δu and that

$$\mathcal{P}(z) = \mathcal{P}(x, u) = (\mathcal{S}(u, 0, 0), u)$$

defines a projection $\mathcal{P} : \mathbb{Z} \rightarrow \mathbb{Z}$ onto $\text{kernel}(D\mathcal{F}(z))$. Here,

$$\|(\mathcal{S}(u, 0, 0), u)\|_{\mathbb{Z}} = \|\mathcal{S}(u, 0, 0)\|_{\mathbb{X}} + \|u\|_{\mathbb{U}}, \quad \text{and} \quad \|\mathcal{S}(u, 0, 0)\|_{\mathbb{X}} = \|x\|_{\mathbb{X}},$$

where x is the solution of the homogeneous problem

$$E \frac{d}{dt}(E^+ E x) - (A + E \frac{d}{dt}(E^+ E))x - B u = 0, \quad (E^+ E x)(\underline{t}) = 0. \tag{35}$$

Replacing again $x = Q\tilde{x}$ as in (29), we can write (35) as

$$\begin{bmatrix} E_{1,1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u, \quad x_1(\underline{t}) = 0,$$

or equivalently

$$E_{1,1}\dot{x}_1 = (A_{1,1} - A_{1,2}A_{2,2}^{-1}A_{2,1})x_1 + (B_1 - A_{1,2}A_{2,2}^{-1}B_2)u, \quad x_1(\underline{t}) = 0, \tag{36}$$

$$x_2 = -A_{2,2}^{-1}(A_{2,1}x_1 + B_2u). \tag{37}$$

The variation of the constant formula for the ODE in (36) yields the estimate $\|x_1\|_{C^0} + \|\dot{x}_1\|_{C^0} \leq c_1\|u\|_{\mathbb{U}}$, with a constant c_1 , and thus $\|x_2\|_{C^0} \leq c_2\|u\|_{\mathbb{U}}$ with a constant c_2 . Altogether, using (32) we then get the estimate

$$\begin{aligned} \|x\|_{\mathbb{X}} &= \|x\|_{C^0} + \left\| \frac{d}{dt}(E^+ E x) \right\|_{C^0} = \|Q\tilde{x}\|_{C^0} + \left\| \frac{d}{dt}(E^+ E T' x_1) \right\|_{C^0} \\ &= \|Q\tilde{x}\|_{C^0} + \left\| \frac{d}{dt}(E^+ E T')x_1 + (E^+ E T')\dot{x}_1 \right\|_{C^0} \leq c_3\|u\|_{\mathbb{U}}, \end{aligned}$$

with a constant c_3 . With this we have shown that \mathcal{P} is continuous and thus kernel $(D\mathcal{F}(z))$ is continuously projectable. Hence, we can apply Theorem 8 and obtain the existence of a unique Lagrange multiplier $\Lambda \in \mathbb{Y}^*$. To determine Λ , we make the ansatz

$$\Lambda(g, r) = \int_{\underline{t}}^{\bar{t}} \lambda^T g \, dt + \gamma^T r. \tag{38}$$

Using the cost function (21) we have

$$D\mathcal{J}(z)\Delta z = x(\bar{t})^T M \Delta x(\bar{t}) + \int_{\underline{t}}^{\bar{t}} (x^T W \Delta x + x^T S \Delta u + u^T S^T \Delta x + u^T R \Delta u) \, dt,$$

and in a local minimum $z = (x, u)$ we obtain that for all $(\Delta x, \Delta u) \in \mathbb{X} \times \mathbb{U}$ the relationship

$$\begin{aligned} 0 &= x(\bar{t})^T M \Delta x(\bar{t}) + \int_{\underline{t}}^{\bar{t}} (x^T W \Delta x + x^T S \Delta u + u^T S^T \Delta x + u^T R \Delta u) \, dt \\ &\quad + \int_{\underline{t}}^{\bar{t}} \lambda^T \left(E \frac{d}{dt} (E^+ E \Delta x) - (A + E \frac{d}{dt} (E^+ E)) \Delta x - B \Delta u \right) \, dt \\ &\quad + \gamma^T (E^+ E \Delta x)(\underline{t}) \end{aligned} \tag{39}$$

has to hold. If $\lambda \in C^1_{E^+E}(\mathbb{I}, \mathbb{R}^n)$, then, using the fact that $E = EE^+E = (EE^+)^T E$, we have by partial integration

$$\begin{aligned} &\int_{\underline{t}}^{\bar{t}} \lambda^T E \frac{d}{dt} (E^+ E \Delta x) \, dt \\ &= \int_{\underline{t}}^{\bar{t}} \lambda^T (EE^+)^T E \frac{d}{dt} (E^+ E \Delta x) \, dt \\ &= \int_{\underline{t}}^{\bar{t}} (EE^+\lambda)^T E \frac{d}{dt} (E^+ E \Delta x) \, dt \\ &= \lambda^T EE^+E \Delta x \Big|_{\underline{t}}^{\bar{t}} - \int_{\underline{t}}^{\bar{t}} \frac{d}{dt} \left[(EE^+\lambda)^T E \right] (E^+ E \Delta x) \, dt \end{aligned}$$

$$\begin{aligned}
 &= \lambda^T E \Delta x \Big|_{\underline{t}}^{\bar{t}} - \int_{\underline{t}}^{\bar{t}} \left[\frac{d}{dt} (EE^+\lambda)^T E + (EE^+\lambda)^T \dot{E} \right] (E^+ E \Delta x) dt \\
 &= \lambda^T E \Delta x \Big|_{\underline{t}}^{\bar{t}} - \int_{\underline{t}}^{\bar{t}} \left[\frac{d}{dt} (EE^+\lambda)^T E \Delta x + (EE^+\lambda)^T \dot{E} E^+ E \Delta x \right] dt.
 \end{aligned}$$

Therefore, we can rewrite (39) as

$$\begin{aligned}
 0 &= \int_{\underline{t}}^{\bar{t}} \left(x^T W + u^T S^T - \frac{d}{dt} (EE^+\lambda)^T E - (EE^+\lambda)^T \dot{E} E^+ E - \lambda^T A - \lambda^T E \frac{d}{dt} (E^+ E) \right) \\
 &\quad \times \Delta x dt + \int_{\underline{t}}^{\bar{t}} (x^T S + u^T R - \lambda^T B) \Delta u dt + x(\bar{t})^T M \Delta x(\bar{t}) + \lambda^T(\bar{t}) E(\bar{t}) \Delta x(\bar{t}) \\
 &\quad - \lambda^T(\underline{t}) E(\underline{t}) \Delta x(\underline{t}) + \gamma^T (E^+ E \Delta x)(\underline{t}).
 \end{aligned}$$

If we first choose $\Delta x = 0$ and vary over all $\Delta u \in \mathbb{U}$, then we obtain the necessary *optimality condition*

$$S^T x + Ru - B^T \lambda = 0. \tag{40}$$

Varying then over all $\Delta x \in \mathbb{X}$ with $\Delta x(\underline{t}) = \Delta x(\bar{t}) = 0$, we obtain the *adjoint equation*

$$Wx + Su - E^T \frac{d}{dt} (EE^+\lambda) - E^+ E \dot{E}^T E E^+ \lambda - A^T \lambda - \frac{d}{dt} (E^+ E) E^T \lambda = 0. \tag{41}$$

Varying finally over $\Delta x(\underline{t}) \in \mathbb{R}^n$ and $\Delta x(\bar{t}) \in \mathbb{R}^n$, respectively, yields the initial condition

$$(E^+(\underline{t})E(\underline{t}))^T \gamma = E^T(\underline{t})\lambda(\underline{t}), \text{ i.e., } \gamma = E(\underline{t})^T \lambda(\underline{t}) \tag{42}$$

and the end condition

$$Mx(\bar{t}) + E(\bar{t})^T \lambda(\bar{t}) = 0, \tag{43}$$

respectively.

Observe that the condition (43) can only be satisfied when $Mx(\bar{t}) \in \text{cokernel } E(\bar{t})$. This extra requirement for the cost term involving the final state was observed already for constant coefficient systems in [50] and in a different setting in [40]. If this condition on M holds, then from (43) we obtain $\lambda(\bar{t}) = -E^+(\bar{t})^T Mx(\bar{t})$.

Using the identity

$$\begin{aligned}
 EE^+ \dot{E}E^+ E + E \frac{d}{dt}(E^+ E) &= EE^+ (\dot{E}E^+ E + E \frac{d}{dt}(E^+ E)) \\
 &= EE^+ \frac{d}{dt}(EE^+ E) = EE^+ \dot{E},
 \end{aligned}$$

we obtain the initial value problem for the adjoint equation in the form

$$E^T \frac{d}{dt}(EE^+ \lambda) = Wx + Su - (A + EE^+ \dot{E})^T \lambda, \quad (EE^+ \lambda)(\bar{t}) = -E^+(\bar{t})^T Mx(\bar{t}). \tag{44}$$

As we had to interpret (22) in the form (30) for the correct choice of the spaces, (44) is the correct interpretation of the problem

$$\frac{d}{dt}(E^T \lambda) = Wx + Su - A^T \lambda, \quad \lambda(\bar{t}) = -E^+(\bar{t})^T Mx(\bar{t}). \tag{45}$$

Note again that these re-interpretations are not crucial when the coefficient functions are sufficiently smooth. The formulation (45) now suggests the following definition.

Definition 12 Let (E, A) be a pair of matrix functions with $E \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ and $A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$. The pair $(-E^T, (A + \dot{E})^T)$ is called the *adjoint pair* of (E, A) .

The notion ‘‘adjoint pair’’ is not only justified by the above construction but also by the following property.

Theorem 13 Let (E, A) have the adjoint pair $(-E^T, (A + \dot{E})^T)$. Then $(-E^T, (A + \dot{E})^T)$ has an adjoint pair which is given by (E, A) .

Proof Obviously, we have $-E^T \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ and $(A + \dot{E})^T \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$. Hence, the pair $(-E^T, (A + \dot{E})^T)$ has an adjoint pair given by $(-(-E^T)^T, [(A + \dot{E})^T - \dot{E}^T]^T) = (E, A)$. □

It is possible to show that if a pair of matrix functions has a well-defined differentiation index ν then its adjoint pair also has a well-defined differentiation index ν . Since we do not need this result in the course of this paper we omit a proof of this observation. A more important property of the adjoint pair especially for the treatment of concrete problems is its behavior under equivalence transformations. For this, let $P, Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ be pointwise nonsingular and let $\tilde{E} = PEQ$ and $\tilde{A} = PAQ - PE\dot{Q}$. Assuming that (E, A) possesses an adjoint pair, we see that (\tilde{E}, \tilde{A}) possesses an adjoint pair as well which is given by

$$\begin{aligned}
 &(-\tilde{E}^T, (\tilde{A} + \dot{\tilde{E}})^T) \\
 &= (-Q^T E^T P^T, Q^T A^T P^T - \dot{Q}^T E^T P^T + \dot{Q}^T E^T P^T + Q^T \dot{E}^T P^T + Q^T E^T \dot{P}^T) \\
 &= (-Q^T E^T P^T, Q^T (A + \dot{E})^T P^T + Q^T E^T \dot{P}^T).
 \end{aligned}$$

The latter representation then states that the adjoint pair of the transformed pair is equivalent to the adjoint pair of the original pair. Hence, we are always allowed to transform a given pair into a suitable form before we build the adjoint pair.

Returning to the adjoint equation and the optimality condition, we will study the action of the special equivalence transformation of (29) on these equations. Using that $(EE^+)^T = EE^+$, we obtain for (44) the transformed system

$$QE^T P^T P \frac{d}{dt} (P^T PEQQ^T E^+ P^T P\lambda) = Q^T WQQ^T x + Q^T Su - (Q^T A^T P^T + Q^T \dot{E} P^T PEQQ^T E^+ P^T) P\lambda.$$

Setting

$$\tilde{W} = Q^T WQ, \quad \tilde{S} = Q^T S, \quad \tilde{\lambda} = P\lambda, \quad \tilde{M} = Q(\bar{t})^T M Q(\bar{t}),$$

we obtain

$$\tilde{E} P \frac{d}{dt} (P^T \tilde{E} \tilde{E}^+ \tilde{\lambda}) = \tilde{W} \tilde{x} + \tilde{S} u - \left(\tilde{A}^T + \dot{Q}^T Q \tilde{E}^T + Q^T (Q \tilde{E}^T \dot{P} + Q \dot{E}^T P + \dot{Q} \tilde{E}^T P) P^T \tilde{E} \tilde{E}^+ \right) \tilde{\lambda}$$

or equivalently

$$\tilde{E} P \dot{P}^T \tilde{E} \tilde{E}^+ \tilde{\lambda} + \tilde{E} \frac{d}{dt} (\tilde{E} \tilde{E}^+ \tilde{\lambda}) = \tilde{W} \tilde{x} + \tilde{S} u - \left(\tilde{A}^T + \dot{Q}^T Q \tilde{E}^T + \tilde{E}^T \dot{P} P^T \tilde{E} \tilde{E}^+ + \dot{\tilde{E}}^T \tilde{E} \tilde{E}^+ + Q^T \dot{Q} \tilde{E}^T \tilde{E} \tilde{E}^+ \right) \tilde{\lambda}.$$

Using the orthogonality of P, Q , which implies that $\dot{Q}^T Q + Q^T \dot{Q} = 0$ and $\dot{P}^T P + P^T \dot{P} = 0$, we obtain

$$\tilde{E} \frac{d}{dt} (\tilde{E} \tilde{E}^+ \tilde{\lambda}) = \tilde{W} \tilde{x} + \tilde{S} u - (\tilde{A} + \tilde{E} \tilde{E}^+ \dot{\tilde{E}}) \tilde{\lambda}.$$

For the initial condition, we obtain accordingly

$$(\tilde{E} \tilde{E}^+ \tilde{\lambda})(\bar{t}) = (PEQQ^T E^+ P^T P\lambda)(\bar{t}) = (PEE^+ \lambda)(\bar{t}) = -P(\bar{t}) E^+(\bar{t})^T Q(\bar{t}) Q(\bar{t})^T M Q(\bar{t}) Q(\bar{t})^T x(\bar{t}) = -\tilde{E}^+(\bar{t})^T \tilde{M} \tilde{x}(\bar{t}).$$

Thus, we have shown that (44) transforms covariantly and that we may consider (44) in the condensed form associated with (29). Setting (with conformable partitioning)

$$\tilde{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{2,1} & W_{2,2} \end{bmatrix}, \quad \tilde{S} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \quad \tilde{M} = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix}, \quad (46)$$

we obtain the system

$$\begin{aligned} E_{1,1}^T \dot{\lambda}_1 &= W_{1,1} x_1 + W_{1,2} x_2 + S_1 u - (A_{1,1} + \dot{E}_{1,1})^T \lambda_1 - A_{2,1}^T \lambda_2, \\ \lambda_1(\bar{t}) &= -E_{1,1}^{-T}(\bar{t}) (M_{1,1} x_1(\bar{t}) + M_{1,2} x_2(\bar{t})), \\ 0 &= W_{2,1} x_1 + W_{2,2} x_2 + S_2 u - A_{1,2}^T \lambda_1 - A_{2,2}^T \lambda_2. \end{aligned}$$

We immediately see that as a differential-algebraic equation in λ this system is again regular and strangeness free. In particular, since $A_{2,2}$ is pointwise nonsingular, this system yields a unique solution $\lambda \in C^1_{EE^+}(\mathbb{I}, \mathbb{R}^n)$ for every $(x, u) \in \mathbb{Z}$.

If $(x, u) \in \mathbb{Z}$ is a local minimum, then from (43) and (44) we can determine Lagrange multipliers $\lambda \in C^1_{EE^+}(\mathbb{I}, \mathbb{R}^n)$ and $\gamma \in \text{cokernel } E(\bar{t})$. It is, however, not clear, whether this λ also satisfies the optimality condition (40).

Suppose that this is not the case, i.e., for the given (x, u, λ) we have

$$S^T x + Ru - B^T \lambda \neq 0. \tag{47}$$

Then there exists $\Delta u \in \mathbb{U}$ with

$$\int_{\underline{t}}^{\bar{t}} (x^T S + u^T R - \lambda^T B) \Delta u \, dt \neq 0.$$

Using this Δu , we have a unique $\Delta x \in \mathbb{X}$ satisfying

$$E \frac{d}{dt} (E^+ E \Delta x) = (A + E \frac{d}{dt} (E^+ E)) \Delta x + B \Delta u, \quad (E^+ E \Delta x)(\underline{t}) = 0,$$

which implies that for $z + \varepsilon \Delta z = (x, u) + \varepsilon(\Delta x, \Delta u)$, we have $\mathcal{F}(z + \varepsilon \Delta z) = 0$ and

$$\begin{aligned} & \mathcal{J}(z + \varepsilon \Delta z) - \mathcal{J}(z) \\ &= \varepsilon \left[x(\bar{t})^T M \Delta x(\bar{t}) + \int_{\underline{t}}^{\bar{t}} (x^T W \Delta x + x^T S \Delta u + u^T S^T \Delta x + u^T R \Delta u) \, dt \right] + \mathcal{O}(\varepsilon^2) \\ &= \varepsilon \left[x(\bar{t})^T M \Delta x(\bar{t}) + \int_{\underline{t}}^{\bar{t}} ((x^T W + u^T S^T) \Delta x + (x^T S + u^T R) \Delta u) \, dt \right] + \mathcal{O}(\varepsilon^2) \\ &= \varepsilon \left[x(\bar{t})^T M \Delta x(\bar{t}) + \int_{\underline{t}}^{\bar{t}} \left(\frac{d}{dt} (EE^+ \lambda)^T E + \lambda^T (A + EE^+ \dot{E}) \right) \Delta x \, dt \right. \\ & \quad \left. + \int_{\underline{t}}^{\bar{t}} (x^T S + u^T R) \Delta u \, dt \right] + \mathcal{O}(\varepsilon^2) = \varepsilon \left[x(\bar{t})^T M \Delta x(\bar{t}) + (\lambda^T E \Delta x) \Big|_{\underline{t}}^{\bar{t}} \right. \\ & \quad \left. - \int_{\underline{t}}^{\bar{t}} (EE^+ \lambda)^T \dot{E} E^+ E \Delta x \, dt - \int_{\underline{t}}^{\bar{t}} \lambda^T E \frac{d}{dt} (E^+ E \Delta x) \, dt + \int_{\underline{t}}^{\bar{t}} \lambda^T (A + EE^+ \dot{E}) \Delta x \, dt \right] \end{aligned}$$

$$\begin{aligned}
 & \left. + \int_{\underline{t}}^{\bar{t}} (x^T S + u^T R) \Delta u \, dt \right] + \mathcal{O}(\varepsilon^2) = \varepsilon \left[\int_{\underline{t}}^{\bar{t}} \lambda^T \left((A + EE^+ \dot{E}) \Delta x \right. \right. \\
 & \left. \left. - E \frac{d}{dt} (E^+ E \Delta x) - EE^+ \dot{E} E^+ E \Delta x \right) dt + \int_{\underline{t}}^{\bar{t}} (x^T S + u^T R) \Delta u \, dt \right] + \mathcal{O}(\varepsilon^2) \\
 & = \varepsilon \left[\int_{\underline{t}}^{\bar{t}} \lambda^T \left((A + EE^+ \dot{E}) - (A + E \frac{d}{dt} (E^+ E)) - EE^+ \dot{E} E^+ E \right) \Delta x \, dt \right. \\
 & \left. + \int_{\underline{t}}^{\bar{t}} (x^T S + u^T R - \lambda^T B) \Delta u \, dt \right] + \mathcal{O}(\varepsilon^2) \\
 & = \varepsilon \left[\int_{\underline{t}}^{\bar{t}} (x^T S + u^T R - \lambda^T B) \Delta u \, dt \right] + \mathcal{O}(\varepsilon^2).
 \end{aligned}$$

Since ε can take any positive and negative values, it follows that z was not a local minimum. Hence, the function λ defined by (44) must satisfy (40).

It thus follows that the functional that is defined via (38), (44) and $\gamma = E(\underline{t})^T \lambda(\underline{t})$ as in (42) has the property (17) and is, therefore, the desired Lagrange multiplier. Furthermore, it is then clear that $(z, \lambda) = (x, u, \lambda)$ is a local minimum of the unconstrained optimization problem

$$\begin{aligned}
 \hat{\mathcal{J}}(z, \lambda) &= \mathcal{J}(z) + \Lambda(\mathcal{F}(z)) \\
 &= \frac{1}{2} x(\bar{t})^T M x(\bar{t}) + \frac{1}{2} \int_{\underline{t}}^{\bar{t}} (x^T W x + 2x^T S u + u^T R u) \, dt \\
 &\quad + \int_{\underline{t}}^{\bar{t}} \lambda^T \left(E \left(\frac{d}{dt} (E^+ E x) - (A + E \frac{d}{dt} (E^+ E)) x - B u - f \right) dt \right. \\
 &\quad \left. + \gamma^T \left((E^+ E x)(\underline{t}) - \underline{x} \right) \right) = \min! \tag{48}
 \end{aligned}$$

In summary, we have proved the following theorem.

Theorem 14 Consider the optimal control problem (21) subject to (22) with a consistent initial condition. Suppose that (22) is strangeness free as a behavior system and that $\text{range } M \subseteq \text{cokernel } E(\bar{t})$.

If $(x, u) \in \mathbb{X} \times \mathbb{U}$ is a solution to this optimal control problem, then there exists a Lagrange multiplier function $\lambda \in C^1_{E^+E}(\mathbb{I}, \mathbb{R}^n)$, such that (x, λ, u) satisfy the

optimality boundary value problem

$$\begin{aligned}
 \text{(a)} \quad & E \frac{d}{dt}(E^+Ex) = (A + E \frac{d}{dt}(E^+E))x + Bu + f, \quad (E^+Ex)(\underline{t}) = \underline{x}, \\
 \text{(b)} \quad & E^T \frac{d}{dt}(EE^+\lambda) = Wx + Su - (A + EE^+\dot{E})^T \lambda, \\
 & (EE^+\lambda)(\bar{t}) = -E^+(\bar{t})^T Mx(\bar{t}), \\
 \text{(c)} \quad & 0 = S^T x + Ru - B^T \lambda.
 \end{aligned} \tag{49}$$

It should be noted again that the assumption in Theorem 14 that the system (22) is regular and strangeness-free in the behavior formulation is not a restriction, since we can always assume that we have already obtained the reduced system (25) which has this property. The same is true for the requirement of consistent initial conditions, which are easily obtained from the reduced system. The third assumption can be easily guaranteed as well, since usually the weight on the final state is something that is chosen independently of the model.

In principle, it is possible to reformulate Theorem 14 in terms of the original data. But this involves the whole index reduction procedure and, hence, would be very technical without giving further analytical insight. Compare with Sect. 4.1 where a numerical approach is discussed.

An important question for the numerical computation of optimal controls is when the optimality system (49) is regular and strangeness free and whether the strangeness index of (49) is related to the strangeness index of the original system. For other index concepts like the tractability index this question has been discussed in [2,4,5,40].

Theorem 15 *The DAE in (49) is regular and strangeness free if and only if*

$$\hat{R} = \begin{bmatrix} 0 & A_{2,2} & B_2 \\ A_{2,2}^T & W_{2,2} & S_2 \\ B_2^T & S_2^T & R \end{bmatrix} \tag{50}$$

is pointwise nonsingular, where we used the notation of (29).

Proof Consider the reduced system (25) associated with the DAE (22) and derive the boundary value problem (49) from this reduced system. If we carry out the change of basis with orthogonal transformations leading to the normal form (29), then we obtain the transformed boundary value problem

$$\begin{aligned}
 \text{(a)} \quad & E_{1,1}\dot{x}_1 = A_{1,1}x_1 + A_{1,2}x_2 + B_1u + f_1, \quad x_1(\underline{t}) = \underline{x}_1, \\
 \text{(b)} \quad & 0 = A_{2,1}x_1 + A_{2,2}x_2 + B_2u + f_2, \\
 \text{(c)} \quad & E_{1,1}^T \dot{\lambda}_1 = W_{1,1}x_1 + W_{1,2}x_2 + S_1u - (A_{1,1} + \dot{E}_{1,1})^T \lambda_1 - A_{2,1}^T \lambda_2, \\
 & \lambda_1(\bar{t}) = -E_{1,1}(\bar{t})^{-T} M_{1,1}x_1(\bar{t}), \\
 \text{(d)} \quad & 0 = W_{2,1}x_1 + W_{2,2}x_2 + S_2u - A_{1,2}^T \lambda_1 - A_{2,2}^T \lambda_2, \\
 \text{(e)} \quad & 0 = S_1^T x_1 + S_2^T x_2 + Ru - B_1^T \lambda_1 - B_2^T \lambda_2.
 \end{aligned} \tag{51}$$

We can rewrite (51) in a symmetrized way as

$$\begin{aligned}
 & \left[\begin{array}{cc|ccc} 0 & E_{1,1} & 0 & 0 & 0 \\ -E_{1,1}^T & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \begin{bmatrix} -\dot{\lambda}_1 \\ \dot{x}_1 \\ -\dot{\lambda}_2 \\ \dot{x}_2 \\ \dot{u} \end{bmatrix} \\
 & = \left[\begin{array}{cc|ccccc} 0 & A_{1,1} & 0 & A_{1,2} & B_1 \\ (A_{1,1} + \dot{E}_{1,1})^T & W_{1,1} & A_{2,1}^T & W_{2,1}^T & S_1 \\ \hline 0 & A_{2,1} & 0 & A_{2,2} & B_2 \\ A_{1,2}^T & W_{2,1} & A_{2,2}^T & W_{2,2} & S_2 \\ B_1^T & S_1^T & B_2^T & S_2^T & R \end{array} \right] \begin{bmatrix} -\lambda_1 \\ x_1 \\ -\lambda_2 \\ x_2 \\ u \end{bmatrix} + \begin{bmatrix} f_1 \\ 0 \\ f_2 \\ 0 \\ 0 \end{bmatrix}. \quad (52)
 \end{aligned}$$

Obviously, this DAE is regular and strangeness free if and only if the symmetric matrix function \hat{R} is pointwise nonsingular. □

If (22) itself is regular and strangeness-free as a free system with $u = 0$, then $A_{2,2}$ is pointwise nonsingular. In our analysis, we have shown that this property can always be achieved, but note that we do not need that $A_{2,2}$ is pointwise nonsingular to obtain a regular and strangeness-free optimality system (49).

On the other hand for \hat{R} to be pointwise nonsingular, it is clearly necessary that $[A_{2,2} \ B_2]$ has pointwise full row rank. This condition is equivalent to the condition that the behavior system (23) belonging to the reduced problem satisfies Hypothesis 3 with $\mu = 0$ and $v = 0$, see [41] for a detailed discussion of this issue and also for an extension of these results to the case of control systems with output equations.

Example 16 An example of a control problem of the form (22) that is not directly strangeness free in the behavior setting is discussed in [1, p. 50]. This linear-quadratic control problem has the coefficients

$$\begin{aligned}
 E &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad f = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \\
 M &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad W = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad S = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad R = 1,
 \end{aligned}$$

and the initial condition $x_1(0) = \alpha$, $x_2(0) = 0$. A possible reduced system (25) is given by

$$\hat{E} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \hat{A} = A, \quad \hat{B} = B.$$

Observe that the corresponding free system of this reduced problem (i.e., with $u = 0$) itself is regular and strangeness free. It follows that the adjoint equation and the optimality condition are given by

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \\ \dot{\lambda}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix},$$

$$0 = - \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + u,$$

respectively, with the end condition $\lambda_1(\bar{t}) = -x_1(\bar{t})$.

We obtain that the matrix function \hat{R} in (50) given by

$$\hat{R} = \left[\begin{array}{cc|cc|c} 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 & 1 \end{array} \right]$$

is pointwise nonsingular, and hence the boundary value problem (49) is regular and strangeness free. Moreover, it has a unique solution which is given by

$$x_1 = \alpha \left(1 - \frac{t}{2 + \bar{t}} \right), \quad x_2 = \lambda_3 = 0, \quad x_3 = u = -\lambda_2 = -\frac{\alpha}{2 + \bar{t}}, \quad \lambda_1 = -\frac{2\alpha}{2 + \bar{t}}.$$

Example 17 In [40], the optimal control problem to minimize

$$\mathcal{J}(x, u) = \int_0^{\bar{t}} (x_1(t)^2 + u(t)^2) dt$$

subject to

$$\frac{d}{dt} \left(\begin{bmatrix} 0 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u, \quad x_2(0) = x_{2,0}$$

is discussed. Obviously, x_1 does not enter the DAE and therefore rather plays the role of a control than of a state. Consequently, the corresponding free system is not regular. Rewriting the system as

$$\begin{bmatrix} 0 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} u, \quad x_2(0) = x_{2,0},$$

and analyzing this system in our described framework, we first of all observe that this system possesses a strangeness index and that it is even regular and strangeness free

as a behavior system. A possible reduced system (25) is given by

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad x_2(0) = x_{2,0}.$$

Also here the corresponding free system is not regular although it is strangeness free. Moreover, we can read off

$$\hat{R} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

which is obviously pointwise nonsingular. Hence, the boundary value problem (49) is regular and strangeness free.

In view of Definition 12 together with (45) one may be tempted to drop the assumptions of Theorem 14 and to consider directly the *formal optimality boundary value problem* given by

- (a) $E\dot{x} = Ax + Bu + f, \quad x(\underline{t}) = \underline{x},$
- (b) $\frac{d}{dt}(E^T\lambda) = Wx + Su - A^T\lambda, \quad (E^T\lambda)(\bar{t}) = -Mx(\bar{t}),$ (53)
- (c) $0 = S^T x + Ru - B^T\lambda.$

But it was already observed in [1,40,50] that it is in general not correct to just consider this system. First of all, as we have shown, the cost matrix M for the final state has to be in the correct cokernel, since otherwise the initial value problem may not be solvable due to a wrong number of conditions. An example for this is given in [1,40]. A further difficulty arises from the fact that the formal adjoint equation (53b) may be a high index equation in the variable λ and thus extra differentiability conditions may arise which may not be satisfied, see [39] for an example. In particular, they may not be solvable due to additional initial conditions or due to lack of smoothness. If, however, the cost functional is positive semidefinite, then one can show that any solution of the formal optimality system yields a minimum and thus constitutes a sufficient condition. This was, e.g., shown for ODE optimal control in [13], for linear constant coefficient DAEs in [50], and in a specific setting for linear DAEs with variable coefficients in [1]. The general result is given in Theorem 18 of [39], where also further issues like differential-algebraic Riccati equations and alternative ways to formulate the cost function are discussed.

We can summarize the results of this section as follows. The necessary optimality condition for the optimal control problem (21) subject to (22) is given by (49) and not by the formal optimality system (53). If, however, (53) has a solution, then it corresponds to a minimum of the optimal control problem. If no index reduction is performed, then a necessary condition for the DAE in (21) to be regular and strangeness free is that the DAE (22) itself is regular and strangeness free as a behavior system.

3.2 General nonlinear problems

In this section, we discuss the general nonlinear optimal control problem to minimize (1) subject to (2) and (3). We assume that all describing functions are sufficiently smooth and that the system described in the behavior setting (5) satisfies Hypothesis 3 with $v = 0$.

Let $z \in C^0(\mathbb{I}, \mathbb{R}^{n+l})$ be a potential candidate for a minimum of (4) subject to (5) and (6). In particular, let z be part of a (continuous) path

$$(t, z(t), \mathcal{P}(t)) \in \mathbb{L}_{\mu+1} \quad \text{for all } t \in \mathbb{I}, \tag{54}$$

cp. Theorem 4.34 of [38]. Due to Hypothesis 3 there exist

$$Z_2 \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)n,a}), \quad T_2 \in C^0(\mathbb{I}, \mathbb{R}^{n+l,n+l-a}), \quad Z_1 \in C^0(\mathbb{I}, \mathbb{R}^{n,d}),$$

with the described properties. Let

$$Z'_2 \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}), \quad T'_2 \in C^0(\mathbb{I}, \mathbb{R}^{n+l,a}), \quad Z'_1 \in C^0(\mathbb{I}, \mathbb{R}^{n,n-d}),$$

be such that

$$[Z'_2 \ Z_2], \quad [T'_2 \ T_2], \quad [Z'_1 \ Z_1]$$

are pointwise orthogonal. Furthermore, there exist

$$T_1 \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)(n+l),(\mu+1)l+a}), \quad T'_1 \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)(n+l),(\mu+1)n-a})$$

such that

$$[T'_1 \ T_1]$$

is pointwise orthogonal and

$$Z'_2(t)^T F_{\mu;\dot{z},\dots,z^{(\mu+1)}}(t, z(t), \mathcal{P}(t))T_1(t) = 0 \quad \text{for all } t \in \mathbb{I}.$$

If we define a function \mathcal{H} via

$$\mathcal{H}(t, z, p, \phi) = \begin{bmatrix} F_{\mu}(t, z, p) + Z_2(t)\phi \\ T_1(t)^T(p - \mathcal{P}(t)) \end{bmatrix}, \tag{55}$$

then

$$\begin{aligned} \text{(a)} \quad & \mathcal{H}(t, z(t), \mathcal{P}(t), 0) = 0, \\ \text{(b)} \quad & \mathcal{H}_{p,\phi}(t, z(t), \mathcal{P}(t), 0) = \begin{bmatrix} F_{\mu;\dot{z},\dots,z^{(\mu+1)}}(t, z(t), \mathcal{P}(t)) & Z_2(t) \\ T_1(t)^T & 0 \end{bmatrix}. \end{aligned} \tag{56}$$

By construction $\mathcal{H}_{p,\phi}(t, z(t), \mathcal{P}(t), 0)$ is nonsingular for all $t \in \mathbb{I}$ and thus we can locally solve for p and ϕ as

$$\phi = \hat{F}_2(t, z), \quad p = \hat{\mathcal{P}}(t, z).$$

We have, in particular, that

$$\hat{F}_2(t, z(t)) = 0, \quad \hat{\mathcal{P}}(t, z(t)) = \mathcal{P}(t)$$

and

$$F_\mu(t, z(t), \hat{\mathcal{P}}(t, z)) + Z_2(t)\hat{F}_2(t, z) = 0 \quad \text{for all } (t, z),$$

and hence

$$F_{\mu;z} + F_{\mu;\dot{z}, \dots, z^{(\mu+1)}}\hat{\mathcal{P}}_z + Z_2\hat{F}_{2;z} = 0,$$

which implies

$$\hat{F}_{2;z}(t, z(t)) = -Z_2(t)^T F_{\mu;z}(t, z(t), \mathcal{P}(t)),$$

i.e., $\hat{F}_{2;z}$ has full row rank along $(t, z(t))$. The equation

$$\hat{F}_2(t, z) = 0 \tag{57}$$

thus is just the requirement that z satisfies at time t all constraints that are contained in (2).

With the change of variables

$$z = T_2 z_1 + T_2' z_2, \quad z_1 = T_2^T z, \quad z_2 = T_2'^T z$$

(57) turns into

$$\hat{F}_2(t, T_2(t)z_1 + T_2'(t)z_2) = 0. \tag{58}$$

If we set $z_1(t) = T_2^T(t)z(t)$, $z_2(t) = T_2'(t)^T z(t)$ then it follows that for all $t \in \mathbb{I}$

- (a) $\hat{F}_2(t, T_2(t)z_1(t) + T_2'(t)z_2(t)) = 0$,
- (b) $\hat{F}_{2;z}(t, z(t))T_2'(t)$ is nonsingular.

Thus, we can solve (58) for z_2 as

$$z_2 = \mathcal{R}(t, z_1) \tag{59}$$

and we have

$$z_2(t) = \mathcal{R}(t, z_1(t)) \quad \text{for all } t \in \mathbb{I}. \tag{60}$$

Since Hypothesis 3 also holds for the transformed system

$$\tilde{F}(t, z_1, z_2, \dot{z}_1, \dot{z}_2) = F(t, T_2 z_1 + T_2' z_2, \frac{d}{dt}(T_2 z_1 + T_2' z_2)), \tag{61}$$

we obtain from (54) a path

$$(t, z_1(t), z_2(t), \tilde{\mathcal{P}}(t)) \in \tilde{\mathbb{L}}_{\mu+1} \quad \text{for all } t \in \mathbb{I}, \tag{62}$$

where $\tilde{\mathbb{L}}_{\mu+1}$ is the solution set associated with (61). Besides (60) we have

$$p_2(t) = \mathcal{R}_t(t, z_1(t)) + \mathcal{R}_{z_1}(t, z_1(t))p_1(t), \tag{63}$$

where we use the partition

$$[I_{n+l} \ 0 \ \dots \ 0] \tilde{\mathcal{P}} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix},$$

compare the proof of Theorem 4.34 in [38]. From (61) and (62) we then obtain

$$\begin{aligned} \tilde{F}(t, z_1(t), z_2(t), p_1(t), p_2(t)) &= F(t, T_2(t)z_1(t) + T_2'(t)\dot{z}_2(t), \dot{T}_2(t)z_1(t) \\ &+ T_2(t)p_1(t) + \dot{T}_2'(t)z_2(t) + T_2'(t)p_2(t)) = 0, \quad \text{for all } t \in \mathbb{I}, \end{aligned} \tag{64}$$

in which we can eliminate z_2, p_2 via (59) and (63), respectively. If we define

$$\begin{aligned} \tilde{F}_1(t, z_1, p_1) &= Z_1(t)^T F(t, T_2(t)z_1 + T_2'(t)\mathcal{R}(t, z_1), \\ &\dot{T}_2(t)z_1 + T_2(t)p_1(t) + \dot{T}_2'(t)\mathcal{R}(t, z_1) + T_2'(t)(\mathcal{R}_t(t, z_1) + \mathcal{R}_{z_1}(t, z_1)p_1)), \end{aligned}$$

then $(t, z_1(t), p_1(t))$ solves $\tilde{F}_1(t, z_1, p_1) = 0$.

Furthermore,

$$\tilde{F}_{1;p_1}(t, z_1(t), p_1(t)) = Z_1(t)^T F_z(t, z(t), p(t))(T_2(t) + T_2'(t)\mathcal{R}_{z_1}(t, z_1(t))), \tag{65}$$

where $[I_{n+l} \ 0 \ \dots \ 0] \mathcal{P} = p$.

To determine $\mathcal{R}_{z_1}(t, z_1(t))$, one observes that from

$$\hat{F}_2(t, T_2(t)z_1(t) + T_2'(t)\mathcal{R}_{z_1}(t, z_1(t))) = 0 \quad \text{for all } t \in \mathbb{I},$$

it follows that

$$\hat{F}_{2;z}(t, z(t))(T_2(t) + T_2'(t)\mathcal{R}_{z_1}(t, z_1(t))) = 0 \quad \text{for all } t \in \mathbb{I}$$

and hence, using (57) we obtain

$$Z_2(t)^T F_{\mu; z}(t, z(t), \mathcal{P}(t))(T_2(t) + T_2'(t)\mathcal{R}_{z_1}(t, z_1(t))) = 0 \quad \text{for all } t \in \mathbb{I}.$$

By the construction of Z_2 , T_2 , and T_2' , we immediately obtain that

$$\mathcal{R}_{z_1}(t, z_1(t)) = 0 \quad \text{for all } t \in \mathbb{I}$$

and that $\tilde{F}_{1; p_1}(t, z_1(t), p_1(t))$ has full row rank for all $t \in \mathbb{I}$. Thus, there exists a pointwise orthogonal matrix function $[V' \ V] \in C^0(\mathbb{I}, \mathbb{R}^{d+l, d+l})$, with

$$Z_1(t)^T F_{\tilde{z}}(t, z(t), p(t))T_2(t)[V'(t) \ V(t)] = [\Sigma(t) \ 0], \tag{66}$$

with pointwise nonsingular Σ . Making a change of variables

$$z_1 = V'z_3 + Vz_4, \quad z_3 = V'^T z_1, \quad z_4 = V^T z_1, \tag{67}$$

and introducing

$$p_3 = \dot{V}'^T z_1 + V'^T p_1, \quad p_4 = \dot{V}^T z_1 + V^T p_1,$$

gives

$$p_1 = \dot{V}'z_3 + V'p_3 + \dot{V}z_4 + Vp_4,$$

and we obtain

$$\tilde{F}_1(t, V'(t)z_3(t) + V(t)z_4(t), \dot{V}'(t)z_3(t) + V'(t)p_3(t) + \dot{V}(t)z_4(t) + V(t)p_4(t)) = 0$$

for all $t \in \mathbb{I}$.

If we define

$$\mathcal{H}(t, z_3, z_4, p_3, p_4) = \tilde{F}_1(t, V'(t)z_3 + V(t)z_4, \dot{V}'(t)z_3 + V'(t)p_3 + \dot{V}(t)z_4 + V(t)p_4),$$

then it follows that

- (a) $\mathcal{H}(t, z_3(t), z_4(t), p_3(t), p_4(t)) = 0$,
- (b) $\mathcal{H}_{p_3}(t, z_3(t), z_4(t), p_3(t), p_4(t)) = Z_1^T(t)F_{\tilde{z}}(t, z(t), p(t))T_2(t)V'(t)$,

and we can solve for p_3 according to

$$p_3 = \mathcal{L}(t, z_3, z_4, p_4). \tag{68}$$

If we insert (67) into (59) we, moreover, obtain

$$z_2 = \mathcal{R}(t, V'(t)z_3 + V(t)z_4). \tag{69}$$

Note that by construction p_3 and p_4 represent the derivatives of z_3 and z_4 , respectively. If we require that z_3 and z_4 are continuously differentiable and that \mathcal{P} satisfies $p_3(t) = \dot{z}_3(t)$ and $p_4(t) = \dot{z}_4(t)$ for all $t \in \mathbb{I}$, then we notice that $z_4 \in C^1(\mathbb{I}, \mathbb{R}^l)$ plays the role of a control in the sense that one can choose it freely in $C^1(\mathbb{I}, \mathbb{R}^l)$ and with an appropriate initial condition $z_3(\underline{t})$ we obtain a unique solution of the ODE $\dot{z}_3 = \mathcal{L}(t, z_3, z_4, \dot{z}_4)$ corresponding to (68). Setting then $(x_1, x_2, u) = (z_3, z_2, z_4)$ we can rewrite (68), (69) as

$$\begin{aligned} \text{(a)} \quad & \dot{x}_1 = \mathcal{L}(t, x_1, u, \dot{u}), \\ \text{(b)} \quad & x_2 = \mathcal{R}(t, x_1, u), \end{aligned} \tag{70}$$

where we have used the same notation as in (69) for the function \mathcal{R} in the renamed variables.

The appearance of \dot{u} in (70) and the implied higher smoothness requirement for u cannot be avoided in the general case. However, the structure of the problem often implies that actually \dot{u} is not present in (70), see, e.g., [38, Remark 4.36]. We therefore use

$$\begin{aligned} \text{(a)} \quad & \dot{x}_1 = \mathcal{L}(t, x_1, u), \\ \text{(b)} \quad & x_2 = \mathcal{R}(t, x_1, u) \end{aligned} \tag{71}$$

instead of (70) and allow u to be only continuous.

We can also get rid of the term \dot{u} by defining \dot{u} as a new control w and replacing (70) by

$$\begin{aligned} \text{(a)} \quad & \dot{x}_1 = \mathcal{L}(t, x_1, x_3, w), \\ \text{(b)} \quad & \dot{x}_3 = w, \\ \text{(c)} \quad & x_2 = \mathcal{R}(t, x_1, x_3). \end{aligned} \tag{72}$$

Note that this reformulation changes the original input function and can be performed beforehand. Simultaneously, it requires a change in the corresponding solution spaces.

If we transform the cost function correspondingly, then the optimal control problem (1) changes to

$$\mathcal{J}(x_1, x_2, u) = \mathcal{M}(x_1(\bar{t}), x_2(\bar{t})) + \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x_1, x_2, u) \, dt = \min! \tag{73}$$

subject to (71) with initial condition $x_1(\underline{t}) = \underline{x}_1$.

Let $z = (x_1, x_2, u)$ be a (local) solution of this optimal control problem, where

$$x_1 \in C^1(\mathbb{I}, \mathbb{R}^d), \quad x_2 \in C^0(\mathbb{I}, \mathbb{R}^a), \quad u \in C^0(\mathbb{I}, \mathbb{R}^l).$$

Then (according to the standard theory for control problems with algebraic and differential constraints), there exist Lagrange multipliers

$$\lambda_1 \in C^1(\mathbb{I}, \mathbb{R}^d), \quad \lambda_2 \in C^0(\mathbb{I}, \mathbb{R}^a), \quad \gamma \in \mathbb{R}^d$$

such that $(x_1, x_2, u, \lambda_1, \lambda_2, \gamma)$ solves the unconstrained problem

$$\begin{aligned} \mathcal{M}(x_1(\bar{t}), x_2(\bar{t})) + \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x_1, x_2, u) \, dt + \gamma^T(x_1(\underline{t}) - \underline{x}_1) \\ + \int_{\underline{t}}^{\bar{t}} \lambda_1^T(\dot{x}_1 - \mathcal{L}(t, x_1, u)) \, dt + \int_{\underline{t}}^{\bar{t}} \lambda_2^T(x_2 - \mathcal{R}(t, x_1, u)) \, dt = \min! \end{aligned} \tag{74}$$

in $\mathbb{W} = \mathbb{Z} \times \mathbb{Y}$ with

$$\mathbb{Z} = C^1(\mathbb{I}, \mathbb{R}^d) \times C^0(\mathbb{I}, \mathbb{R}^a) \times C^0(\mathbb{I}, \mathbb{R}^l), \quad \mathbb{Y} = C^1(\mathbb{I}, \mathbb{R}^d) \times C^0(\mathbb{I}, \mathbb{R}^a) \times \mathbb{R}^d.$$

From (74), we obtain the necessary condition

$$\begin{aligned} \mathcal{M}_{x_1}(x_1(\bar{t}), x_2(\bar{t}))\Delta x_1(\bar{t}) + \mathcal{M}_{x_2}(x_1(\bar{t}), x_2(\bar{t}))\Delta x_2(\bar{t}) \\ + \int_{\underline{t}}^{\bar{t}} (\mathcal{K}_{x_1}(t, x_1, x_2, u)\Delta x_1 + \mathcal{K}_{x_2}(t, x_1, x_2, u)\Delta x_2 + \mathcal{K}_u(t, x_1, x_2, u)\Delta u) \, dt \\ + \int_{\underline{t}}^{\bar{t}} \lambda_1^T(\Delta \dot{x}_1 - \mathcal{L}_{x_1}(t, x_1, u)\Delta x_1 - \mathcal{L}_u(t, x_1, u)\Delta u) \, dt \\ + \int_{\underline{t}}^{\bar{t}} \Delta \lambda_1^T(\dot{x}_1 - \mathcal{L}_{x_1}(t, x_1, u)) \, dt \\ + \int_{\underline{t}}^{\bar{t}} \lambda_2^T(\Delta x_2 - \mathcal{R}_{x_1}(t, x_1, u)\Delta x_1 - \mathcal{R}_u(t, x_1, u)\Delta u) \, dt \\ + \gamma^T \Delta x_1(\underline{t}) + \Delta \gamma^T(x_1(\underline{t}) - \underline{x}_1) = 0 \end{aligned}$$

for all $(\Delta x_1, \Delta x_2, \Delta u, \Delta \lambda_1, \Delta \lambda_2, \Delta \gamma) \in \mathbb{W}$. Variation over $\Delta \lambda_1, \Delta \lambda_2,$ and $\Delta \gamma_1$ as usual reproduces the constraints. Moreover, when comparing with the linear case we

only have to perform the replacements

$$\begin{aligned}
 & \text{(a) } x_1(\bar{t})^T M \leftarrow [\mathcal{M}_{x_1}(x_1(\bar{t}), x_2(\bar{t})) \ \mathcal{M}_{x_2}(x_1(\bar{t}), x_2(\bar{t}))], \\
 & \text{(b) } x^T W + u^T S \leftarrow [\mathcal{K}_{x_1}(t, x_1, x_2, u) \ \mathcal{K}_{x_2}(t, x_1, x_2, u)], \\
 & \text{(c) } x^T S^T + u^T R \leftarrow \mathcal{K}_u(t, x_1, x_2, u), \\
 & \text{(d) } E \leftarrow \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \ A \leftarrow \begin{bmatrix} \mathcal{L}_{x_1}(t, x_1, u) & 0 \\ \mathcal{R}_{x_1}(t, x_1, u) & -I \end{bmatrix}, \ B \leftarrow \begin{bmatrix} \mathcal{L}_u(t, x_1, u) \\ \mathcal{R}_u(t, x_1, u) \end{bmatrix},
 \end{aligned}
 \tag{75}$$

in (39).

In this way, we obtain the boundary value problem of necessary optimality conditions

$$\begin{aligned}
 & \text{(a) } \dot{x}_1 = \mathcal{L}(t, x_1, u), \ x_1(\underline{t}) = \underline{x}_1, \\
 & \text{(b) } x_2 = \mathcal{R}(t, x_1, u), \\
 & \text{(c) } \dot{\lambda}_1 = \mathcal{K}_{x_1}(t, x_1, x_2, u)^T - \mathcal{L}_{x_1}(t, x_1, x_2, u)^T \lambda_1 - \mathcal{R}_{x_1}(t, x_1, u)^T \lambda_1, \\
 & \quad \lambda_1(\bar{t}) = -\mathcal{M}_{x_1}(x_1(\bar{t}), x_2(\bar{t}))^T, \\
 & \text{(d) } 0 = \mathcal{K}_{x_2}(t, x_1, x_2, u)^T + \lambda_2, \\
 & \text{(e) } 0 = \mathcal{K}_u(t, x_1, x_2, u)^T - \mathcal{L}_u(t, x_1, u)^T \lambda_1 - \mathcal{R}_u(t, x_1, u)^T \lambda_2, \\
 & \text{(f) } \gamma = \lambda_1(\underline{t})
 \end{aligned}
 \tag{76}$$

proving the following result.

Theorem 18 *Let z be a local solution of (4) subject to (5) and (6) in the sense that the transformed $(x_1, x_2, u) \in \mathbb{Z}$ is a local solution of (73) subject to (71) and $x_1(\underline{t}) = \underline{x}_1$. Then there exist unique Lagrange multipliers $(\lambda_1, \lambda_2, \gamma) \in \mathbb{Z}$ such that $(x_1, x_2, u, \lambda_1, \lambda_2, \gamma)$ solves the boundary value problem (76).*

Note that Theorem 18 implicitly contains the assumption that Hypothesis 3 holds, giving the implicitly defined functions \mathcal{L} and \mathcal{R} and the splitting of x and λ . A reformulation of (76) in terms of the original data is in general not possible. Even when dealing with (1)–(3) numerically, we are forced to evaluate implicitly defined functions by some nonlinear system solver, see also Sect. 4.2.

Remark 19 The preceding result can be generalized to constraints that additionally contain end conditions, i.e., conditions on parts of $x(\bar{t})$. The observation is the same as for ODEs. In particular, for every additional scalar end condition we lose one scalar condition on values $\lambda(\bar{t})$.

3.3 A maximum principle for general DAEs

If we, furthermore, allow the input functions to be constrained, then according to (18), (19) we have the problem

$$\mathcal{J}(x, u) = \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x(t), u(t)) \, dt = \min! \tag{77}$$

subject to

$$\begin{aligned} 0 &= F(t, x, u, \dot{x}), \quad x(\underline{t}) = \underline{x}, \\ 0 &= h(\bar{t}, x(\bar{t})), \quad h \in C(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n), \\ u(t) &\in \mathcal{U} \subset \mathbb{R}^l \quad \text{for all } t \in \mathbb{I}. \end{aligned} \tag{78}$$

Since the control u is restricted we must (at least) allow for an optimal control u that has a finite number of jumps, i.e., $u \in L^\infty_c(\mathbb{I}, \mathbb{R}^l)$. In view of (71) we must then also allow that the algebraic variables x_2 in a reduced formulation possess jumps at the same locations. Moreover, we must say in what sense the differential-algebraic equation in the constraint is to be satisfied when we allow for jumps in the input.

Starting with $z \in L^\infty_c(\mathbb{I}, \mathbb{R}^{n+l})$ and thus with a (piecewise continuous) path $(t, z(t), \mathcal{P}(t))$ as a potential candidate for a minimum, Hypothesis 3 will no longer guarantee that we can perform the construction in the beginning of Sect. 3.2. In particular, we need additional assumptions that yield the necessary smooth transformations and the necessary implications of the implicit function theorem. These assumptions must guarantee local existence of solutions in the required spaces and (even in the presence of jumps) the possibility of smooth variations. In this respect the assumptions made during the following construction are indispensable. Note that in many applications these additional assumptions hold due to the structure of the given problem.

Following the beginning of Sect. 3.2, we first assume that in spite of the lack of smoothness the projector functions $Z'_1, Z_1, Z'_2, Z_2, T'_1, T_1, T'_2, T_2$ are still at least continuous. Instead of (55) we define

$$\tilde{\mathcal{H}}(t, z, p, \tilde{p}, \phi) = \begin{bmatrix} F_\mu(t, z, p) + Z_2(t)\phi \\ T_1(t)^T(p - \tilde{p}) \end{bmatrix}, \tag{79}$$

which can locally be solved according to

$$\phi = \tilde{F}_2(t, z, \tilde{p}), \quad p = \tilde{\mathcal{P}}(t, z, \tilde{p}).$$

Here we must assume that at a point where the solution path has a jump, the whole jump lies in the domain of the implicitly defined functions. In particular, we then have that

$$\hat{F}_2(t, z(t)) = 0 \quad \text{for all } t \in \mathbb{I},$$

where $\hat{F}_2(t, z) = \tilde{F}_2(t, z, \mathcal{P}(t))$. Again, $\hat{F}_{2;z}$ has full row rank along $(t, z(t))$ and $\hat{F}_2(t, z) = 0$ represents all the algebraic constraints that are contained in the DAE.

Proceeding in the same way with the following constructions and corresponding assumptions, we arrive at the reduced formulation consisting of (68) and (69) with no p_4 present in (68). Again $z_4 \in L^\infty_c(\mathbb{I}, \mathbb{R}^l)$ plays the role of a control. Given an

initial value $z_3(\underline{t})$ we require that

$$z_3(t) = z_3(\underline{t}) + \int_{\underline{t}}^t p_3(s) \, ds$$

holds for the given path. Renaming the variables as before, we get the reduced problem

$$\begin{aligned} \text{(a)} \quad & x_1(t) = x_1(\underline{t}) + \int_{\underline{t}}^t \mathcal{L}(s, x_1(s), u(s)) \, ds, \\ \text{(b)} \quad & x_2 = \mathcal{R}(t, x_1, u), \end{aligned} \tag{80}$$

which reflects that x_1 does not need to be continuously differentiable on the whole interval \mathbb{I} .

Note that a state feedback as in (27) makes the constraint $u(t) \in \mathcal{U}$ state dependent. But, if feedback controls are necessary to regularize the DAE, then the original formulation of the control problem should be seen to be incorrectly posed and thus constraints should be prescribed for w and not for u .

Transforming \mathcal{J} and h to the new variables, eliminating the variable x_2 with the help of the algebraic constraint, and assuming that the so obtained h does not depend on u (since the quantity $u(\bar{t})$ does not make sense), we get as interpretation of (77) and (78) a problem of the form

$$\hat{\mathcal{J}}(x_1, u) = \int_{\underline{t}}^{\bar{t}} \hat{\mathcal{K}}(t, x_1(t), u(t)) \, dt = \min! \tag{81}$$

subject to

$$\begin{aligned} x_1(t) &= x_1(\underline{t}) + \int_{\underline{t}}^t \mathcal{L}(s, x_1(s), u(s)) \, ds, \\ 0 &= \hat{h}(\bar{t}, x_1(\bar{t})), \quad h \in C(\mathbb{R} \times \mathbb{R}^d, \mathbb{R}^d), \\ u(t) &\in \mathcal{U} \subset \mathbb{R}^l \quad \text{for all } t \in \mathbb{I} \end{aligned} \tag{82}$$

for the determination of an optimal $(\bar{t}, x_1, u) \in \mathbb{R} \times C^0(\mathbb{I}, \mathbb{R}^d) \times L^\infty_c(\mathbb{I}, \mathbb{R}^d)$. Of course, the missing part x_2 is then given by (80b).

Theorem 20 *If (\bar{t}^*, x_1^*, u^*) is a local solution of the Bolza problem (81) subject to (82) with unknowns (\bar{t}, x_1, u) , then there exist scalars $\alpha_0, \alpha_1, \dots, \alpha_d$, which do not all vanish simultaneously, $\alpha_0 \geq 0$, and a multiplier $\lambda \in C^0(\mathbb{I}, \mathbb{R}^d)$ such that, with $H(t, x_1, u, \lambda, \alpha) = \lambda^T \mathcal{L}(t, x_1, u) - \alpha_0 \hat{\mathcal{K}}(t, x_1, u)$ and $\alpha = (\alpha_1, \dots, \alpha_n)^T$,*

$$\begin{aligned} H(t, x_1^*(t), u^*(t), \lambda(t), \alpha) &= \max_{u \in \mathcal{U}} H(t, x_1^*(t), u, \lambda(t), \alpha), \\ \dot{\lambda}(t) &= -\nabla_{x_1} H(t, x_1^*(t), u^*(t), \lambda(t), \alpha), \\ \dot{x}_1^*(t) &= \nabla_\lambda H(t, x_1^*(t), u^*(t), \lambda(t), \alpha), \\ \lambda(\bar{t}^*) &= -\alpha^T \nabla_{x_1(\bar{t})} \hat{h}(\bar{t}^*, x_1^*(\bar{t}^*)), \end{aligned} \tag{83}$$

for all $t \in \mathbb{I}$, where u^* is continuous.

As already stated in the case of Theorem 18, it is virtually impossible in the general case to formulate Theorem 20 in terms of the original problem (77), (78).

Theorem 20 covers Theorem 18 in the case of a continuous control u when we fix \bar{t} omitting the condition on it described by h in Theorem 20 and when we omit the costs on the final state described by \mathcal{M} in Theorem 18. Of course, we could have formulated generalized versions of both theorems such that this would directly be the case, but we chose the restricted versions because here we concentrate merely on the DAE aspect of the results and not on the most general possible formulations.

As a simple application of Theorem 20 we consider a semi-explicit differential-algebraic equation of index 1 in the constraints. In particular, we consider the problem

$$\mathcal{J}(x_1, x_2, u) = \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t, x_1(t), x_2(t), u(t)) dt = \min!$$

subject to

$$\begin{aligned} \dot{x}_1 &= f(t, x_1, x_2, u), \quad x_1(\underline{t}) = \underline{x}_1 \\ 0 &= g(t, x_1, x_2, u), \\ u(t) &\in \mathcal{U} \subset \mathbb{R}^l \quad \text{for all } t \in \mathbb{I}, \end{aligned}$$

with the assumption that g_{x_2} is everywhere nonsingular. Given a (local) solution (x_1, x_2, u) the implicit function theorem yields that x_2 is determined in terms of (t, x_1, u) according to

$$x_2 = \mathcal{R}(t, x_1, u),$$

while the differential part reads

$$\dot{x}_1 = \mathcal{L}(t, x_1, u) = f(t, x_1, \mathcal{R}(t, x_1, u), u).$$

Accordingly, we get

$$\hat{\mathcal{K}}(t, x_1, u) = \mathcal{K}(t, x_1, \mathcal{R}(t, x_1, u), u).$$

Thus, the structure of the differential-algebraic equation immediately gives a suitable reformulation fitting to Theorem 20. With

$$H(t, x_1, u, \lambda, \alpha_0) = \lambda^T f(t, x_1, \mathcal{R}(t, x_1, u), u) - \alpha_0 \mathcal{K}(t, x_1, \mathcal{R}(t, x_1, u), u)$$

the essential part of (83) reads (omitting arguments)

$$\begin{aligned} \dot{\lambda}(t) &= -((f_{x_1}^T + \mathcal{R}_{x_1}^T f_{x_2}^T)\lambda + \alpha_0(\mathcal{K}_{x_1}^T + \mathcal{R}_{x_1}^T \mathcal{K}_{x_2}^T)), \\ \dot{x}_1 &= f(t, x_1, \mathcal{R}(t, x_1, u), u). \end{aligned}$$

This corresponds (up to several technical differences) to the results given in [57]. The same applies to the other results of [57] which deal with differential-algebraic equations in Hessenberg form of index 2 and differential-algebraic equations of index 3 that arise in the modeling of multibody systems. The latter case, however, is treated by an index reduction which increases the number of differential components. Hence, one must pay attention to the correct choice of the boundary conditions.

4 Numerical methods for optimal control problems

In this section, we discuss the numerical solution of the optimality boundary value problems (49) and (76), respectively. In contrast to the analytical treatment, for the numerical solution we may neither just assume that the free system (with $u = 0$) is strangeness free, since the regularizing feedbacks can only be computed during the integration, nor can we work with implicitly defined functions as contained in (76). Instead, we have to work with the original data and possibly their derivatives.

We again first study the case of linear systems with variable coefficients.

4.1 Numerical methods for linear-quadratic optimal control problems

To incorporate (if necessary) an index reduction, we use the functions as in (25) that can be determined in every time step from the given data (including derivatives of the coefficient functions) but we have to note that the projection functions Z_1^T and Z_2^T are not realized as smooth functions in numerical methods such as the code GENDA [42], although such smooth realizations exist. It would be just too expensive to carry the computation of smooth realizations along. Since most numerical integration methods such as Runge–Kutta methods or BDF methods, see [8, 29], are invariant under transformations from the left, the non-smooth realizations yield the same results.

Taking into account that the coefficient functions in (49) are only available through index reduction and assuming sufficient smoothness of the data, we write (49) in terms of (26) as

$$\begin{aligned}
 \text{(a)} \quad & \hat{E}_1 \dot{x} = \hat{A}_1 x + \hat{B}_1 u + \hat{f}_1, \quad (\hat{E}_1^+ \hat{E}_1 x)(\underline{t}) = \underline{x}, \\
 \text{(b)} \quad & 0 = \hat{A}_2 x + \hat{B}_2 u + \hat{f}_2, \\
 \text{(c)} \quad & \frac{d}{dt}(\hat{E}_1^T \lambda_1) = Wx + Su - \hat{A}_1^T \lambda_1 - \hat{A}_2^T \lambda_2, \\
 & \lambda_1(\bar{t}) = -[\hat{E}_1^+(\bar{t})^T \quad 0] Mx(\bar{t}), \\
 \text{(d)} \quad & 0 = S^T x + Ru - \hat{B}_1^T \lambda_1 - \hat{B}_2^T \lambda_2.
 \end{aligned} \tag{84}$$

The missing smoothness of Z_1, Z_2 is not a problem in (84a) and (84b), but as constructed, the unknowns λ_1, λ_2 are in general not smooth if Z_1, Z_2 are non-smooth. If, however, we choose Z_1 and Z_2 to have orthonormal columns, then at least $Z_1 Z_1^T$ and $Z_2 Z_2^T$ are smooth, since they represent orthogonal projections onto the corresponding

image spaces. Thus, with

$$\hat{E}_1^T \lambda_1 = E^T Z_1 \lambda_1 = E^T Z_1 Z_1^T Z_1 \lambda_1 = E^T Z_1 Z_1^T \hat{\lambda}_1$$

and $\hat{\lambda}_1 = Z_1 \lambda_1$, we can obtain smooth coefficients for the unknown $\hat{\lambda}_1$. However, we have to add the condition that $\hat{\lambda}_1 \in \text{range } Z_1$ to the system. If we complete Z_1 via Z'_1 to a pointwise orthogonal matrix function, then we can express this as

$$Z'_1{}^T \hat{\lambda}_1 = 0. \tag{85}$$

Note that here again it plays no role whether Z' is constructed as a smooth function.

Due to

$$\begin{aligned} [\hat{A}_2 \hat{B}_2]^T \lambda_2 &= [I_{n+l} \ 0 \ \dots \ 0] N_\mu^T Z_2 \lambda_2 \\ &= [I_{n+l} \ 0 \ \dots \ 0] N_\mu^T Z_2 Z_2^T Z_2 \lambda_2 \\ &= [I_{n+l} \ 0 \ \dots \ 0] N_\mu^T Z_2 Z_2^T \hat{\lambda}_2, \end{aligned}$$

with $\hat{\lambda}_2 = Z_2 \lambda_2$, we can proceed analogously for λ_2 . In particular, we complete Z_2 via Z'_2 to a pointwise orthogonal matrix function and require

$$Z'_2{}^T \hat{\lambda}_2 = 0. \tag{86}$$

Adding (85) and (86) to boundary value problem (84) yields the new boundary value problem

- (a) $\hat{E}_1 \dot{x} = \hat{A}_1 x + \hat{B}_1 u + \hat{f}_1, (\hat{E}_1^+ \hat{E}_1 x)(\underline{t}) = \underline{x},$
- (b) $0 = \hat{A}_2 x + \hat{B}_2 u + \hat{f}_2,$
- (c) $\frac{d}{dt}(E^T Z_1 Z_1^T \hat{\lambda}_1) = Wx + Su - A^T \hat{\lambda}_1 - [I_n \ 0 \ | \ 0 \ 0 \ | \ \dots \ | \ 0 \ 0] N_\mu^T \hat{\lambda}_2,$
 $(Z_1^T \hat{\lambda}_1)(\bar{t}) = -[\hat{E}_1^+(\bar{t})^T \ 0] Mx(\bar{t}), \tag{87}$
- (d) $0 = S^T x + Ru - B^T \hat{\lambda}_1 - [0 \ I_l \ | \ 0 \ 0 \ | \ \dots \ | \ 0 \ 0] N_\mu^T \hat{\lambda}_2,$
- (e) $0 = Z_1^T \hat{\lambda}_1,$
- (f) $0 = Z_2^T \hat{\lambda}_2.$

As constructed, this boundary value problem now allows for a smooth solution independent of non-smooth realizations of the coefficient functions. The parts which are not explicitly represented in the original data and their derivatives can be obtained from them by the standard index reduction, see [38]. Compare also with the following discussion of the nonlinear case.

Since the coefficient $E^T Z_1 Z_1^T$ is smooth, we can discretize (87) for example with BDF methods or using the boundary value methods introduced in [43,44,46]. This is justified by the following observation.

Lemma 21 *The boundary value problem (87) is regular and strangeness free iff the boundary value problem (49) is regular and strangeness free.*

Proof Consider the coefficients of the boundary value problem given by

$$\mathcal{E} = \begin{bmatrix} \hat{E}_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & E^T Z_1 Z_1^T & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} \hat{A}_1 & \hat{B}_1 & 0 & 0 \\ \hat{A}_2 & \hat{B}_2 & 0 & 0 \\ W & S & -A^T - \frac{d}{dt}(E^T Z_1 Z_1^T) & -\tilde{A}_{3,4} \\ S^T & R & -B^T & -\tilde{A}_{4,4} \\ 0 & 0 & Z_1'^T & 0 \\ 0 & 0 & 0 & Z_2'^T \end{bmatrix}, \tag{88}$$

with

$$\tilde{A}_{3,4} = [I_n \ 0 \ 0 \ 0 \ | \ \dots \ | \ 0 \ 0] N_\mu^T, \quad \tilde{A}_{4,4} = [0 \ I_l \ 0 \ 0 \ | \ \dots \ | \ 0 \ 0] N_\mu^T,$$

where obviously $\text{rank } \mathcal{E} = 2d$. With

$$\mathcal{Z} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ I_1 & 0 & 0 & 0 & 0 \\ 0 & T_2' & 0 & 0 & 0 \\ 0 & I_l & 0 & 0 & 0 \\ 0 & 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}, \quad \mathcal{T} = \begin{bmatrix} T_2' & 0 & 0 & 0 \\ 0 & I_l & 0 & 0 \\ 0 & 0 & Z_1' & 0 \\ 0 & 0 & 0 & I \end{bmatrix},$$

describing corange and kernel of \mathcal{E} , we have $\mathcal{Z}^T \mathcal{E} = 0, \mathcal{E} \mathcal{T} = 0$, and thus, the DAE associated with (88) is regular and strangeness free if and only if

$$\mathcal{Z}^T \mathcal{A} \mathcal{T} = \begin{bmatrix} \hat{A}_2 T_2' & \hat{B}_2 & 0 & 0 \\ T_2'^T W T_2' & T_2'^T S & -T_2'^T (A^T + \frac{d}{dt}(E^T Z_1 Z_1^T)) Z_2'^T & -T_2'^T \tilde{A}_{3,4} \\ S^T T_2' & R & -B^T Z_1' & -\tilde{A}_{4,4} \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & Z_2'^T \end{bmatrix} \tag{89}$$

is nonsingular. Omitting the block row and block column containing I_a and multiplying the last block column with $[Z_2' \ Z_2]$, we see that $\mathcal{Z}^T \mathcal{A} \mathcal{T}$ is nonsingular if and only if

$$\begin{bmatrix} \hat{A}_2 T_2' & \hat{B}_2 & 0 & 0 \\ T_2'^T W T_2' & T_2'^T S & -T_2'^T \tilde{A}_{3,4} Z_2' & -T_2'^T \tilde{A}_{3,4} Z_2 \\ S^T T_2' & R & -\tilde{A}_{4,4} Z_2' & -\tilde{A}_{4,4} Z_2 \\ 0 & 0 & I_a & 0 \end{bmatrix},$$

is nonsingular. Since

$$\begin{aligned} T_2'^T \tilde{A}_{3,4} Z_2 &= T_2'^T [I_n \ 0 \mid 0 \ 0 \mid \cdots \mid 0 \ 0] N_\mu^T Z_2 = T_2'^T \hat{A}_2^T, \\ \tilde{A}_{4,4} Z_2 &= [0 \ I_l \mid 0 \ 0 \mid \cdots \mid 0 \ 0] N_\mu^T Z_2 = \hat{B}_2^T, \end{aligned}$$

and observing that

$$A_{2,2} = \hat{A}_2 T_2', \quad B_2 = \hat{B}_2, \quad W_{2,2} = T_2'^T W T_2', \quad S_2^T = S^T T_2',$$

this is the case if and only \hat{R} as in (50) is nonsingular. □

4.2 Numerical methods for nonlinear optimal control problems

The numerical solution of the boundary value problem (76) is approached in a similar way as for the linear case. To represent this boundary value problem in the original data, we proceed as for the integration of a differential-algebraic equation, see [38]. The differential equations (76a) are represented by the equations $Z_1^T F(t, x, \dot{x}, u) = 0$ with Z_1 defined by Hypothesis 3 and the algebraic equations (76b) are implied by the derivative array $F_\mu(t, x, u, p) = 0$. The remaining equations are defined via linearization and correspond to (87c–d) of the linear case such that in (87c–d) the replacements

$$E \leftarrow F_{\dot{x}}, \quad A \leftarrow -F_x, \quad B \leftarrow -F_u, \quad Mx(\bar{t}) \leftarrow \mathcal{M}_x(x(\bar{t}))^T$$

and

$$N_\mu [I_n \ 0 \mid 0 \ 0 \mid \cdots \mid 0 \ 0]^T \leftarrow -F_{\mu;x}, \quad N_\mu [0 \ I_l \mid 0 \ 0 \mid \cdots \mid 0 \ 0]^T \leftarrow -F_{\mu;u}$$

apply. Introducing $\hat{\lambda}_1, \hat{\lambda}_2$ and adding (85) and (86) to the nonlinear boundary value problem, we end up with the boundary value problem (omitting arguments)

- (a) $Z_1^T F = 0, (\hat{E}_1^+ \hat{E}_1 x)(\underline{t}) = \underline{x},$
- (b) $F_\mu = 0,$
- (c) $\frac{d}{dt}(F_x^T Z_1 Z_1^T \hat{\lambda}_1) = \mathcal{K}_x^T + F_x^T \hat{\lambda}_1 + F_{\mu;x} \hat{\lambda}_2,$
 $(Z_1^T \hat{\lambda}_1)(\bar{t}) = -[\hat{E}_1^+(\bar{t})^T \ 0] \mathcal{M}_x(x(\bar{t}))^T, \tag{90}$
- (d) $0 = \mathcal{K}_u^T + F_u^T \hat{\lambda}_1 + F_{\mu;u}^T \hat{\lambda}_2,$
- (e) $0 = Z_1'^T \hat{\lambda}_1,$
- (f) $0 = Z_2'^T \hat{\lambda}_2.$

Note that we have presented the boundary conditions in terms of the linearization. Of course, it is possible to state them in terms of the original data and the involved

projections as the other relations. However, the resulting formulas are relatively complicated, since they are directly related to the problem of the consistent initialization of differential-algebraic equations, and we refrain here from presenting these.

4.3 A numerical example

We have discretized the boundary value problem (90) by means of the midpoint rule for the differential equations in the state variables and the trapezoidal rule for the differential equations in the adjoint variables together with the algebraic constraints at all grid points and simple divided differences for the term $\frac{d}{dt}(F_x^T Z_1 Z_1^T \hat{\lambda}_1)$.

To use numerical differentiation to generate the necessary Jacobians, the relations starting with $Z_1^T, Z_1'^T$, and $Z_2'^T$, respectively, were used with smooth projectors $Z_1 Z_1^T, Z_1' Z_1'^T$, and $Z_2' Z_2'^T$ instead. Although this introduces a rank deficiency in the Jacobians with respect to the rows, we can expect the Gauß–Newton method to converge at least superlinearly due to the consistency of the solution with the overdetermined relations. Note that this would not be necessary if we generated the Jacobians utilizing the structure of the equations.

The preceding approach was implemented in FORTRAN double precision. As one of the test problems we selected a nonlinear optimal control problem for a multibody system taken from [10].

A model problem for a motor controlled pendulum to be driven into its equilibrium with minimal costs is given by

$$\begin{aligned}
 J(x, u) &= \int_0^3 u(t)^2 dt = \min! \\
 \text{s.t.} \quad \dot{x}_1 &= x_3, & x_1(0) &= \frac{1}{2}\sqrt{2}, & g &= 9.81, \\
 \dot{x}_2 &= x_4, & x_2(0) &= -\frac{1}{2}\sqrt{2}, \\
 \dot{x}_3 &= -2x_1x_5 + x_2u, & x_3(0) &= 0, \\
 \dot{x}_4 &= -g - 2x_2x_5 - x_1u, & x_4(0) &= 0, \\
 0 &= x_1^2 + x_2^2 - 1, & x_5(0) &= -\frac{1}{2}gx_2(0).
 \end{aligned}$$

It is known that the differential-algebraic equation in the constraint satisfies Hypothesis 3 with $\mu = 2, a = 3, d = 2$, and $v = 0$. Hence, only two scalar initial values are sufficient to describe the initial state. We chose them to be $x_2(0) = -\frac{1}{2}\sqrt{2}$ and $x_3(0) = 0$. Similarly, $x_1(3) = 0$ and $x_3(3) = 0$ are sufficient to describe the equilibrium at the end point. According to Remark 19 no end conditions for the Lagrange multipliers occur. As initial trajectory we took

$$\begin{aligned}
 x_1(t) &= \frac{1}{2}\sqrt{2} - \frac{1}{6}\sqrt{2}t, & x_3(t) &= 0, \\
 x_2(t) &= -\sqrt{1 - x_1(t)^2}, & x_4(t) &= 0, & x_5(t) &= -\frac{1}{2}gx_2(t),
 \end{aligned}$$

Table 1 Course of the Gauß–Newton iteration

k	$\ \Delta w_k\ _2$
1	0.140D+03
2	0.223D+03
\vdots	\vdots
16	0.561D+01
17	0.103D+01
18	0.610D-02
19	0.318D-06
20	0.966D-11

with all other unknowns set to zero on an equidistant grid of 60 intervals. The required tolerance for the Gauß–Newton method was 10^{-7} . See Table 1 for the course of the iteration, where k counts the iterations and $\|\Delta w_k\|_2$ denotes the Euclidian norm of the corresponding Gauß–Newton correction.

In the Gauss–Newton iteration, there is an initial phase with a bad convergence behavior due to a bad initial guess which could be remedied by damping. But in the final phase one easily recognizes quadratic convergence. The obtained final value of the cost function was $J_{\text{opt}} = 3.82$ which is, up to discretization errors, in coincidence with the value given in [10].

Note that the implementation used here is by no means efficient. This would require to incorporate the structure of the problem when setting up the Jacobian and solving the linear subproblems. See [43,44] for techniques that may be applied.

5 Conclusions

We have presented the optimal control theory for general unstructured linear and nonlinear systems of differential-algebraic equations. In particular, we have generalized results of [53,57] to arbitrary index. We have derived necessary conditions and a maximum principle and have shown how these can be numerically solved. The results are illustrated by a numerical example.

Acknowledgments We thank two anonymous referees for several comments and suggestions that helped to improve the readability of the paper.

References

1. Backes A (2006) Optimale Steuerung der linearen DAE im Fall Index 2. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Berlin
2. Balla K, Kurina G, März R (2006) Index criteria for differential algebraic equations arising from linear-quadratic optimal control problems. *J Dyn Control Syst* 12:289–311
3. Balla K, Linh VH (2005) Adjoint pairs of differential-algebraic equations and Hamiltonian systems. *Appl Numer Math* 53:131–148
4. Balla K, März R (2002) A unified approach to linear differential algebraic equations and their adjoints. *Z Anal Anwendungen* 21:783–802

5. Balla K, März R (2004) Linear boundary value problems for differential algebraic equations. *Math Notes* 5:3–17
6. Bender D, Laub A (1987) The linear quadratic optimal regulator problem for descriptor systems. *IEEE Trans Automat Control* 32:672–688
7. Boltyanskii V, Gamkrelidze R, Mishenko E, Pontryagin LS (1962) The mathematical theory of optimal processes. Interscience, New York
8. Brenan KE, Campbell SL, Petzold LR (1996) Numerical solution of initial-value problems in differential algebraic equations, 2nd edn. SIAM, Philadelphia
9. Bunse-Gerstner A, Mehrmann V, Nichols NK (1992) Regularization of descriptor systems by derivative and proportional state feedback. *SIAM J Matr Anal Appl* 13:46–67
10. Büskens C, Gerds M (2000) Numerical solution of optimal problems with DAEs of higher index. In: *Proceedings of the Workshop: Optimalsteuerungsprobleme in der Luft und Raumfahrt*, pp 27–38. Sonderforschungsbereich 255: Transatmosphärische Flugsysteme, Hieronymus, München
11. Byers R, Geerts T, Mehrmann V (1997) Descriptor systems without controllability at infinity. *SIAM J Cont* 35:462–479
12. Byers R, Kunkel P, Mehrmann V (1997) Regularization of linear descriptor systems with variable coefficients. *SIAM J Cont* 35:117–133
13. Campbell SL (1980) Singular systems of differential equations I. Pitman, San Francisco
14. Campbell SL (1987) Comment on controlling generalized state-space (descriptor) systems. *Int J Control* 46:2229–2230
15. Campbell SL (1987) A general form for solvable linear time varying singular systems of differential equations. *SIAM J Math Anal* 18:1101–1115
16. Cobb JD (1983) A further interpretation of inconsistent initial conditions in descriptor-variable systems. *IEEE Trans Automat Control* AC 28:920–922
17. Campbell SL, Gear CW (1995) The index of general nonlinear DAEs. *Numer Math* 72:173–196
18. Campbell SL, Meyer CD (1979) Generalized inverses of linear transformations. Pitman, San Francisco
19. Devdariani EN, Ledyaeu YuS (1999) Maximum principle for implicit control systems. *Appl Math Optim* 40:79–103
20. Diehl M, Leineweber DB, Schäfer A, Bock HG, Schlöder JP (2002) Optimization of multiple-fraction batch distillation with recycled waste cuts. *AIChE J* 48(12):2869–2874
21. Diehl M, Uslu I, Findeisen R, Schwarzkopf S, Allgöwer F, Bock HG, Bürner T, Gilles ED, Kienle A, Schlöder JP, Stein E (2001) Real-time optimization for large scale processes: nonlinear model predictive control of a high purity distillation column. In: Grötschel M, Krumke SO, Rambau J (eds), *Online optimization of large scale systems: state of the art*, pp 363–384. Springer, Berlin
22. Eich-Soellner E, Führer C (1998) Numerical methods in multibody systems. Teubner, Stuttgart
23. Gerds M (2003) Optimal control and real-time optimization of mechanical multi-body systems. *Z Angew Math Mech* 83:705–719
24. Gerds M (2006) Local minimum principle for optimal control problems subject to index two differential algebraic equations systems. *J Optim Theory Appl* 130:441–460
25. Günther M, Feldmann U (1999) CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations. *Surv Math Ind* 8:97–129
26. Günther M, Feldmann U (1999) CAD-based electric-circuit modeling in industry II. Impact of circuit configurations and parameters. *Surv Math Ind* 8:131–157
27. Gabasov R, Kirillova F (1976) The qualitative theory of optimal processes. Marcel Dekker, New York
28. Griepentrog E, März R (1986) Differential-algebraic equations and their numerical treatment. Teubner Verlag, Leipzig
29. Hairer E, Wanner G (1996) Solving ordinary differential equations II: stiff and differential-algebraic problems, 2nd edn. Springer, Berlin
30. Hesteness MR (1966) Calculus of variations and optimal control theory. Wiley, New York
31. Ioffe AD, Tichomirov VM (1979) Theorie der Extremalaufgaben. VEB Deutscher Verlag der Wissenschaften, Berlin
32. Kirsch A, Warth W, Werner J (1978) Notwendige Optimalitätsbedingungen und ihre Anwendung. Springer, Berlin
33. Kunkel P, Mehrmann V (1996) Generalized inverses of differential-algebraic operators. *SIAM J Matr Anal Appl* 17:426–442
34. Kunkel P, Mehrmann V (1996) A new class of discretization methods for the solution of linear differential algebraic equations with variable coefficients. *SIAM J Numer Anal* 33:1941–1961

35. Kunkel P, Mehrmann V (1997) The linear quadratic control problem for linear descriptor systems with variable coefficients. *Math Control Signals Sys* 10:247–264
36. Kunkel P, Mehrmann V (1998) Regular solutions of nonlinear differential-algebraic equations and their numerical determination. *Numer Math* 79:581–600
37. Kunkel P, Mehrmann V (2001) Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Math Control Signals Sys* 14:233–256
38. Kunkel P, Mehrmann V (2006) Differential-algebraic equations. Analysis and numerical solution. EMS Publishing House, Zürich
39. Kunkel P, Mehrmann V (2006) Necessary and sufficient conditions in the optimal control for general nonlinear differential-algebraic equations. Preprint 355, DFG Research Center Matheon, TU Berlin, Berlin, Germany
40. Kurina GA, März R (2004) On linear-quadratic optimal control problems for time-varying descriptor systems. *SIAM J Cont Optim* 42:2062–2077
41. Kunkel P, Mehrmann V, Rath W (2001) Analysis and numerical solution of control problems in descriptor form. *Math Control Signals Sys* 14:29–61
42. Kunkel P, Mehrmann V, Rath W, Weickert J (1997) A new software package for linear differential-algebraic equations. *SIAM J Sci Comput* 18:115–138
43. Kunkel P, Mehrmann V, Stöver R (2004) Multiple shooting for unstructured nonlinear differential-algebraic equations of arbitrary index. *SIAM J Numer Anal* 42:2277–2297
44. Kunkel P, Mehrmann V, Stöver R (2004) Symmetric collocation for unstructured nonlinear differential-algebraic equations of arbitrary index. *Numer Math* 98:277–304
45. Kunkel P, Mehrmann V, Schmidt M, Seufer I, Steinbrecher A (2006) Weak formulations of linear differential-algebraic systems. Technical Report 16, Institut für Mathematik, TU Berlin, Berlin, Germany. <http://www.math.tu-berlin.de/preprints/>
46. Kunkel P, Stöver R (2002) Symmetric collocation methods for linear differential-algebraic boundary value problems. *Numer Math* 91:475–501
47. Ljusternik L (1934) On constrained extrema of functionals. *Math Sb* 41:390–401 (in Russian)
48. Lin J-Y, Yang Z-H (1988) Optimal control for singular systems. *Int J Control* 47:1915–1924
49. März R (2004) Solvability of linear differential algebraic equations with properly stated leading terms. *Res Math* 45:88–105
50. Mehrmann V (1991) The autonomous linear quadratic control problem. Springer, Berlin
51. Müller PC (1996) Stability and optimal control of nonlinear descriptor systems. In: Proceedings of 3rd International Symposium Methods and Models in Automation and Robotics (MMAR 96), vol 1, pp 17–26, Szczecin, Poland. University of Szczecin
52. Otter M, Elmqvist H, Mattson SE (2006) Multi-domain modeling with modelica. In: Fishwick P (ed) CRC Handbook of Dynamic System Modeling, pp 36.1–36.27. Chapman and Hall/CRC Press, London
53. de Pinho MDR, Vinter RB (1997) Necessary conditions for optimal control problems involving nonlinear differential algebraic equations. *J Math Anal Appl* 212:493–516
54. Polderman JW, Willems JC (1998) Introduction to mathematical systems theory: a behavioural approach. Springer, New York
55. Rabier PJ, Rheinboldt WC (1996) Classical and generalized solutions of time-dependent linear differential-algebraic equations. *Lin Alg Appl* 245:259–293
56. Rabier PJ, Rheinboldt WC (2002) Theoretical and numerical analysis of differential-algebraic equations, volume VIII of Handbook of Numerical Analysis. Elsevier, Amsterdam
57. Roubicek T, Valasek M (2002) Optimal control of causal differential-algebraic systems. *J Math Anal Appl* 269:616–641
58. Schultz VH (1996) Reduced SQP methods for large-scale optimal control problems in DAE with application to path planning problems for satellite mounted robots. Dissertation, Universität Heidelberg, Interdisz. Zentrum für wissenschaftliches Rechnen
59. Vinter R, Vinter RB (2000) Optimal control. Birkhäuser, Boston
60. Zeidler E (1985) Nonlinear functional analysis and its applications III. Variational methods and optimization. Springer, New York