# Multilevel Optimization Problems with Linear Differential-Algebraic Equations

vorgelegt von

**M. Sc. Daniel Steffen Bankmann**
0000-0001-6381-1466

an der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

**Promotionsausschuss**

| | |
|---|---|
| Vorsitzender: | Prof. Dr. Martin Skutella |
| Gutachter: | Prof. Dr. Volker Mehrmann |
| Gutachter: | Prof. Dr. Karl Worthmann |
| Gutachter: | Prof. Dr. Paul Van Dooren |

Tag der wissenschaftlichen Aussprache: 16. Dezember 2020

Berlin 2021

# Acknowledgments

First of all, I would like to thank the referees for agreeing to examine this thesis. I thank my advisor, Volker Mehrmann, for introducing me to the topic, being open for discussion and proofreading the preliminary manuscripts. I would also like to thank Paul Van Dooren for the insightful discussions about the analytic center.

Moreover, I would like to thank my colleagues and friends at TU Berlin, in particular Andreas, Benni, Christoph, Marine, Matze, Micha, Paul, Punit, Philipp, Riccardo, Robert, for all the inspiring discussions, lunch breaks, and board game nights. Also, thanks to Benni for proofreading introduction and conclusion.

Last but not least, I would like to thank my friends, family, and my partner for the continuous support.

This work was partially supported by the German Research Foundation (DFG) in the projects B03 in the TRR-CRC 154 Mathematical Modeling, Simulation and Optimization using the Example of Gas Networks and as part of the project 'Distributed Dynamic Security Control in Next-Generation Electrical Power Systems' with the project identification number 361092219 of the priority program 'DFG SPP 1984 – Hybrid and multimodal energy systems: System theory methods for the transformation and operation of complex networks'.

# Zusammenfassung

Diese Arbeit befasst sich mit verschiedenen mehrstufigen Optimierungsproblemen im Kontext allgemeiner parameterabhängiger linearer differentiell-algebraischer Gleichungen (DAEs).

Im ersten Teil analysieren wir Probleme, bei denen die unterste Stufe aus einem Optimalsteuerungsproblem für lineare DAEs besteht. Das Optimalsteuerungsproblem hängt von Variablen höherer Stufen ab, die die Rolle eines Parameters übernehmen. Die Lösung des Optimalsteuerungsproblems wird anschließend in das Optimierungsproblem der nächsten Stufe eingesetzt. Wir berechnen die Lösung des Optimalsteuerungsproblems mit Hilfe der notwendigen Optimalitätsbedingungen, die ein differentiell-algebraisches Randwertproblem darstellen.

Ein wesentlicher Bestandteil zur Lösung der übergeordneten Optimierungsprobleme sind Sensitivitäten des Optimalsteuerungsproblems. Wir analysieren verschiedene Möglichkeiten zur Berechnung der Sensitivitäten und diskutieren kurz die Folgen für DAEs mit höherem Index. Außerdem stellen wir mögliche numerische Methoden für die Berechnung der Sensitivitäten vor. Wir wenden die Ergebnisse auf ein mehrstufiges Optimalsteuerungsproblem mit einer oberen Stufe bestehend aus einem nichtlinearen quadratischen Ausgleichsproblem an, und stellen ein Konvergenzergebnis und numerische Beispiele zur Verfügung.

Weiterhin leiten wir Formeln für das analytische Zentrum der Lösungsmenge von linearen Matrixungleichungen (LMIs), die passive Übertragungsfunktionen definieren, her. Für die Berechnung des analytischen Zentrums werden numerische Methoden entwickelt. Es wird auch gezeigt, dass das analytische Zentrum hilfreiche Robustheitseigenschaften hat, wenn es zur Darstellung passiver Systeme verwendet wird. Die Ergebnisse werden durch numerische Beispiele veranschaulicht.

Darüber hinaus analysieren wir bestimmte Robustheitsmaße, die bei der robusten Stabilisierung linearer zeitinvarianter Systeme in Bezug auf eine bestimmte Konditionszahl auftreten, die bei der Berechnung von invarianten Unterräu-

men eines Matrizenbüschels bei linear-quadratischen Optimalsteuerungsproblemen auftreten. Wir zeigen, dass diese Konditionszahl bei der Berechnung der Robustheitsmaße einfließen kann und veranschaulichen mögliche Konsequenzen anhand eines Beispiel.

# Abstract

This thesis is about various multilevel optimization problems in the context of general parameter-dependent linear differential-algebraic equations (DAEs).

In the first part we analyze problems where the lowest level consists of an optimal control problem for linear DAEs. The optimal control problem depends on higher level variables that take the role of a parameter. A solution of the optimal control problem is fed into the next level optimization problem. We compute the solution of the optimal control with the help of the necessary optimality conditions which constitute a differential-algebraic boundary value problem.

An essential ingredient for solving the higher level optimization problems is sensitivity information of the optimal control problem. We analyze different possibilities for the computation of the sensitivities and briefly discuss implications for higher index systems. We also present possible numerical methods for the computation of the sensitivities. We apply the results to a multilevel optimal control problem with a nonlinear least-squares upper level and provide a convergence result and numerical examples.

We also derive formulas for the analytic center of the solution set of linear matrix inequalities (LMIs) defining passive transfer functions. Numerical methods are developed for the computation of the analytic center. It is also shown that the analytic center has nice robustness properties when it is used to represent passive systems. The results are illustrated by numerical examples.

Also, we analyze certain robustness measures appearing in the robust stabilization of linear time-invariant systems with respect to a certain condition number appearing in the computation of deflating subspaces of a matrix pencil associated with the linear quadratic optimal control problem. We show, that this condition number can be incorporated into the computation of the robustness measures and illustrate possible consequences with an example.

# Declaration of Contributions

Chapter 6 and Appendices A and B and parts of the abstract and the introduction are essentially copies of the article

[BMNV20b] D. Bankmann, V. Mehrmann, Y. Nesterov, and P. Van Dooren. "Computation of the Analytic Center of the Solution Set of the Linear Matrix Inequality Arising in Continuous- and Discrete-Time Passivity Analysis". *Vietnam J. Math.*: (2020). DOI: 10.1007/s10013-020-00427-x. arXiv: 1904.08202 (cit. on pp. 113, 137, 138).

and are presented here in their post-print version with minor modifications.

The co-authors of this article are

- Prof. Dr. Volker Mehrmann, Technische Universität Berlin,
- Prof. Dr. Yurii Nesterov, Université catholique de Louvain,
- Prof. Dr. Paul Van Dooren, Université catholique de Louvain.

The research in Part I was closely supervised by Volker Mehrmann and Paul Van Dooren. Contributions of the author, including proofs and proof ideas, are mainly given in Sections 6.2 to 6.4 and Appendices A and B and the development of the research software package [BMNV20a]. Proofs in Section 6.4 without contribution of the author have been removed.

The research in Part I has been performed by the author, whereas Volker Mehrmann provided some of the research topics and the supervision.

The research in Chapter 7 was supervised by Volker Mehrmann. Kyle McKee, University of Alberta, assisted by performing numerical experiments and with discussions about the research.

The following software packages have emerged from the research and are publicly available.

[Ban20] D. Bankmann. *Code and examples for the computation of solutions to multilevel optimal control problems with differential algebraic equations*. Zenodo, 2020. DOI: 10.5281/zenodo.3971868 (cit. on pp. 163, 174).

[BMNV20a] D. Bankmann, V. Mehrmann, Y. Nesterov, and P. Van Dooren. *Code and examples for the paper 'Computation of the analytic center of the solution set of the linear matrix inequality arising in continuous- and discrete-time passivity analysis'*. Zenodo, 2020. DOI: 10.5281/zenodo.3997097 (cit. on pp. 134, 175, 176).

The software packages are written by the author. An overview and a short introduction are available in Appendix C.

# Contents

# Nomenclature

| | |
|---|---|
| $\mathring{\mathbb{A}}$ | interior of a set $\mathbb{A}$ |
| $\overline{\mathbb{A}}$ | closure of a set $\mathbb{A}$ |
| $\partial\mathbb{A}$ | boundary of a set $\mathbb{A}$ |
| $\mathbb{N}$ | $:= \{1, 2, \ldots\}$; set of natural numbers |
| $\mathbb{N}_0$ | $:= \mathbb{N} \cup \{0\}$ |
| $\mathbb{R}$ | field of real numbers |
| $\mathbb{R}^+$ | set of positive real numbers |
| $\mathbb{R}_0^+$ | set of non-negative real numbers |
| $\mathbb{C}$ | field of complex numbers |
| $\Im(z)$ | imaginary part of a complex number $z \in \mathbb{C}$ |
| $\Re(z)$ | real part of a complex number $z \in \mathbb{C}$ |
| $\mathbb{I}$ | $\subseteq \mathbb{R}$; interval |
| $\mathbb{K}$ | $\in \{\mathbb{C}, \mathbb{R}\}$ |
| $\Theta \subseteq \mathbb{R}^p$ | parameter space |
| $e_i^{(n)}$ | $:= e_i$; $i$-th unit vector in $\mathbb{R}^n$ |
| $I_n$ | identity matrix of size $n \times n$ |
| $\det A$ | determinant of a matrix $A \in \mathbb{K}^{n \times n}$ |
| $\mathbb{H}_n$ | set of Hermitian matrices $A \in \mathbb{C}^{n \times n}$ |

| | |
|---|---|
| $A^+$ | Moore-Penrose pseudo inverse of a matrix $A \in \mathbb{K}^{m\times n}$ |
| $A^g$ | a generalized inverse of a matrix $A \in \mathbb{K}^{m\times n}$ such that $AA^gA = A$ |
| $A^{\mathsf{H}}$ | conjugate transpose of a matrix $A \in \mathbb{K}^{m\times n}$ |
| $A^T$ | transpose of a matrix $A \in \mathbb{K}^{m\times n}$ |
| $A^{+\mathsf{H}}$ | $:= (A^+)^{\mathsf{H}} = (A^{\mathsf{H}})^+$ |
| $\mathrm{vec}A$ | vectorization operator which maps a matrix $A \in \mathbb{K}^{m\times n}$ to a vector $\mathrm{vec}A \in \mathbb{K}^{mn}$ by vertically stacking the columns of $A$ |
| $A \otimes B$ | Kronecker product of two matrices $A \in \mathbb{K}^{m\times n}$, $B \in \mathbb{K}^{k\times l}$ |
| $\|\cdot\|_2$ | 2-norm of a matrix or vector |
| $\|\cdot\|_F$ | Frobenius norm of a matrix |
| $\mathcal{C}^k(V_1, V_2)$ | $k \in \mathbb{N}_0$; space of $k$ times continuously differentiable functions from $V_1$ to $V_2$ |
| $\mathcal{C}^k_{\mathrm{pw}}(V_1, V_2, \{v_1, \dots, v_q\})$ | $k \in \mathbb{N}_0$; space of piecewise $k$ times continuously differentiable functions from $V_1$ to $V_2$, with discontinuous points $v_1, ..., v_l \in V_1$ |
| $\mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n)$ | $:= \{x \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n) \mid E^+Ex \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)\}$ |
| $\mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^n)$ | $:= \{\lambda \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n) \mid EE^+\lambda \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)\}$ |
| $\mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^n, \{v_1, \dots, v_q\})$ | space of functions $\lambda$, for which $EE^+\lambda$ is a piecewise smooth function with $EE^+\lambda \in \mathcal{C}^1_{\mathrm{pw}}(\mathbb{I}, \mathbb{K}^n, \{v_1, \dots, v_q\})$ and $\lambda \in \mathcal{C}^0_{\mathrm{pw}}(\mathbb{I}, \mathbb{K}^n, \{v_1, \dots, v_q\})$ |
| $\Sigma_{m,n}(\mathbb{K})$ | space of data $(E, A, B, f)$ with $E, A \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times n}), B \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times m}), f \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n)$ |

# 1. Introduction

This thesis is devoted to the analysis and the development of numerical algorithms for optimization problems arising in the context of parameter-dependent optimal control problems and parameter-dependent passivity problems. While on a more abstract level the problems being analyzed are all similar, they vary in a lot of properties, in particular when it comes down to the actual underlying structure. Optimal control problems and passivity problems are defined for certain system classes. Throughout this thesis we focus on system classes where the dynamics are described by differential-algebraic equations (DAEs) or the special case of ordinary differential equations (ODEs).

This work is split up into two parts. In the first part, we investigate multilevel optimal control problems in the linear-quadratic setting with linear DAEs. These are multilevel optimization problems, where the lowest level is an optimal control problem depending on higher level variables. These higher level variables take the role of a parameter for the optimal control task. In the next level, another optimization problem is solved, which depends on the solution of the lowest level for the current value of higher level variables.

We focus on analyzing bilevel optimal control problems – multilevel optimal control problems with exactly two levels, a *lower* and an *upper* level. Bilevel optimization and bilevel optimal control problems have recently attained a lot of attention. For plain algebraic bilevel optimization problems an overview can be found in [CMS07; Dem02; DKPK15]. Bilevel optimal control problems have been considered with ordinary, differential-algebraic, or partial differential equations as lower level dynamics in various settings, see e. g., [DG19; DF19; HW19; Meh16; Meh17a; Meh17b; MW16; PG16; PG17; Ye97], and the references therein. For the sake of simplicity we assume in this work that the lower level optimal control problem is uniquely solvable, and its solution is differentiable with respect to the set of parameters.

One commonly analyzed bilevel optimal control problem is parameter estimation, where the lower level is given by an optimal control problem and the

upper level is given by a nonlinear least squares problem. Another application is given by co-design problems, see, e. g., [JWBJ15; WWB16], where both levels optimize the same objective function subject to changes in system parameters. Parameter estimation problems appear regularly in engineering problems such as humanoid locomotion [MTL09]. In the literature, proposed methods for the computation of numerical solutions mainly include so-called direct methods, where the underlying optimal control problem is discretized before optimization. Indirect methods using the necessary conditions for optimality are also available, see [Boc87; BP84; HSB12; KB06; LBBS03; LSBS03] and the references therein for a comparison. It is well-known, that for optimal control problems under certain circumstances both approaches, direct and indirect, lead to correct results, see [Hag00]. In the DAE setting, however, the situation is more complicated. Besides well-known numerical issues with higher index DAEs, see [KM06; KM08], even in the case of strangeness-free DAEs one may suffer from a drop in the order of convergence, in particular if the leading coefficient in front of the derivative term has a time-depending kernel [CK13].

This thesis aims at laying the theoretical foundation for analyzing such multilevel optimal control problems based on the necessary conditions for general linear time-varying and parameter-dependent DAEs. These multilevel optimal control problems essentially need sensitivity information of the optimal control problem with respect to the higher level parameters. Based on the strangeness concept for DAEs [KM06] and a related flow definition [Bau14; Bau17; KM06] we develop a forward and an adjoint equation for general parameter-dependent boundary value problems, which then can be applied to the boundary value problem of necessary conditions of the parameter depending optimal control problem. We use these systems to derive formulas for the computation of the sensitivities. While for the case of ordinary equations this problem is solved in [SP02], the situation for DAEs is more complicated as additional regularity requirements have to be fulfilled. For DAEs only some of the subproblems as the development of the adjoint equation for self-adjoint boundary value problems have been solved, see [BL05; BM00; KMS14; KM08; KM11a; KM04] for more details. Also, some of the cited articles base their results on the tractability index for DAEs, which has its own advantages and disadvantages, see [Meh15] for a comparison. Here, we recover some of those results for the case of strangeness-free DAEs and generalize the results in [SP02] for the computation of the sensitivities.

For the numerical solution we discuss a multiple shooting approach and develop a differential Riccati equation approach. Also, we provide test examples for verification. To conclude this part, we analyze the different numerical errors in solving the multilevel optimal control problem, where the upper level is given by a nonlinear least-squares problem. We study the numerical errors for solving the necessary optimality conditions for the optimal control problem and the corresponding sensitivity system. Our analysis results in an error-estimator that links these errors to the global error of the Gauss-Newton method [DS96] used to solve the higher level problem.

In the second part of the thesis we consider optimization problems that also have more than one level of optimization. However, in contrast to the first part, these optimizations do not rely on the solution of the optimal control problem. Instead, other quantities are optimized in the context of the computation of solutions of optimal control problems and estimates for the passivity radius.

Formulas are derived for the analytic center of the solution set of linear matrix inequalities (LMIs) defining passive transfer functions. The algebraic Riccati equations that are usually associated with such systems are related to boundary points of the convex set defined by the solution set of the LMI. It is shown that the analytic center is described by closely related matrix equations, and their properties are analyzed for continuous- and discrete-time systems. Numerical methods are derived to solve these equations via steepest descent and Newton methods. It is also shown that the analytic center has nice robustness properties when it is used to represent passive systems. The results are illustrated by numerical examples. The analytic center proves to be a good choice regarding computation of the passivity radius, see [BMV19], which can be considered as a bilevel optimization problem. Also, the analytic center for the passivity LMI turns out to be an useful object for the determination of a good port-Hamiltonian formulation. Here, 'good' means finding the right coordinate system that is robust in the sense of numerical errors [BMV19].

We also consider the computation of the condition number of the transformation matrix that is used in the computation of deflating subspaces associated with the solution of linear time-invariant control systems. In [MX00] several methods for solving algebraic Riccati equations are evaluated in terms of solving the robust stabilization problem. Robust stabilization is achieved, if by choosing certain weights in an optimal control setting, the real values of the eigenvalues of the closed-loop matrix lie below a predefined threshold. We analyze how

optimizing this certain condition affects the overall performance of the robust stabilization problem.

# 2. Multilevel Optimizations in Optimal Control

In this chapter we introduce multilevel optimization problems in the context of optimal control problems. The following definitions cover the optimization problems we are facing in this work, c. f., Pages xiii to xiv for basic notation.

**Definition 2.1** (Multilevel optimization problem)**.** Let $\mathbb{I} \subseteq \mathbb{R}$ be an interval. Further, let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and sets $\mathbb{X}_1, \dots, \mathbb{X}_l$, $\mathbb{U}_1, \dots, \mathbb{U}_l$, $l \in \mathbb{N}$ be given, where $\mathbb{X}_i$ are either subsets of $\mathcal{C}^k(\mathbb{I}, \mathbb{K}^{n_l})$, $k \in \mathbb{N}$, $n_l \in \mathbb{N}$, or empty sets and $\mathbb{U}_i$ are Banach spaces. Moreover, set $\mathbb{Z}_i := \mathbb{X}_i \times \mathbb{U}_i$, $i = 1, \dots, l$ and let $S_i$ be mappings such that $S_i z_i = u_i$ for $z_i = (x_i, u_i) \in \mathbb{Z}_i$. Also, if $\mathbb{X}_i \subseteq \mathcal{C}^k(\mathbb{I}, \mathbb{K}^{n_l})$ denote by $\mathbb{X}_i'$ a subset of $\mathcal{C}^{k-1}(\mathbb{I}, \mathbb{K}^{n_l})$; otherwise $\mathbb{X}_i' = \{\}$. For $i = 1, \dots, l-1$ define objective functions

$$\mathcal{K}_i : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_{i+1} \to \mathbb{R},$$

constraint functions

$$\mathcal{G}_i : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_{i-1} \times \mathbb{I} \times \mathbb{X}_i' \times \mathbb{Z}_i \times \mathbb{Z}_{i+1} \to \mathbb{R}^{c_i},$$

$c_i \in \mathbb{N}_0$, and solution functions

$$\mathcal{L}_{i+1} : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_i \to \mathbb{Z}_{i+1}.$$

Furthermore, let $\mathcal{K}_l : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_l \to \mathbb{R}$ and $\mathcal{G}_l : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_{l-1} \times \mathbb{I} \times \mathbb{X}_l' \times \mathbb{Z}_l \to \mathbb{R}^{c_l}$, $c_l \in \mathbb{N}_0$. Also, for brevity, let

$$\mathcal{L}_{l+1} : \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_l \to \{\}$$

denote the 'empty function', i. e., $\mathcal{K}_l(z_1, \dots, z_l, \mathcal{L}_{l+1}(z_1, \dots, z_l)) := \mathcal{K}_l(z_1, \dots, z_l)$. Then a *multilevel optimization problem* consists of $l$ recursively defined optimizations of the form

$$\begin{aligned} &\min_{u_i = S_i z_i} \mathcal{K}_i(z_1, \dots, z_i, \mathcal{L}_{i+1}(z_1, \dots, z_i)) \\ &\quad \text{s.t.} \quad \mathcal{G}_i(z_1, \dots, z_{i-1}, t, \dot{x}_i, z_i, \mathcal{L}_{i+1}(z_1, \dots, z_i)) = 0, \quad i = 1, \dots, l, \end{aligned} \tag{2.1}$$

where $z_1, \dots, z_{i-1} \in \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_{i-1}$ are given and $z_{i+1}^* := \mathcal{L}_{i+1}(z_1, \dots, z_i)$ denotes the unique solution of the optimization problem at level $i+1, i < l$. ▷

**Definition 2.2** (Multilevel optimal control problem)**.** Let all the assumptions of Definition 2.1 hold. If in addition the lowermost level fulfills

$$\mathbb{Z}_l \subseteq \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n_l}) \times \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{m_l}), \quad m_l \in \mathbb{N}, \tag{2.2}$$

then the Optimization Problem 2.1 is called *multilevel optimal control problem*. Optimal control problems are introduced in more detail in Section 3.4. ▷

*Remark* 2.1.

1. The given definition of a multilevel optimization or multilevel optimal control problem suggests to solve the optimizations recursively. The first optimization to be solved is for $i = l$, the last for $i = 1$. The number $i$ is thus often referred to as the *level* of the optimization problem, where $i = l$ is the *lowermost* level and $i = 1$ the *uppermost* level.

   In the particular case of $l = 2$ we have a *bilevel* optimization problem and we denote the two levels by *lower* level with subscript l and *upper* level with subscript u, respectively.

   Analogously, the variables $z_1, \dots, z_{i-1}$ are called *higher level variables* and $z_i, \dots, z_l$ are called *lower level variables*.

2. The assumption (2.2) and the conditions $\mathbb{X}_l \in \mathcal{C}^k(\mathbb{I}, \mathbb{K}^{n_l})$, $\mathbb{X}_l' \in \mathcal{C}^{k-1}(\mathbb{I}, \mathbb{K}^{n_l})$ may need to be weakened in certain cases. This is discussed in Section 4.3.

3. We also may need to include inequality constraints in Definition 2.1 and the optimizations in (2.1). See Chapter 7 for an example.

4. It is important to note that we explicitly require that there is a solution operator $\mathcal{L}_i$ with a unique solution on each level. The problem gets much more involved, if one allows non-unique solutions, see, e. g., [Dem02] for an introduction in the case of plain algebraic bilevel optimization problems and the work on bilevel optimal control problems [Meh17a]. ▷

The thesis is divided into two parts. In Part I we focus on multilevel optimal control problems as in Definition 2.2. In Part II we consider more general problems, that do not fit Definition 2.2, but Definition 2.1. These optimizations do not

directly depend on the solutions of the dynamical system involved. Rather, other quantities are optimized. While we focus on bilevel optimization problems, the developed results and techniques can also be applied to multilevel optimization problems.

# Part I.

# Solution Based Multilevel Optimal Control Problems

In this part we study multilevel optimal control problems that fit Definition 2.2. Preliminary material that ranges from matrix perturbation theory to parameter-dependent differential-algebraic systems and their optimal control is introduced in Chapter 3. We discuss sensitivities, i. e., parameter derivatives of solutions of optimal control problems of linear parameter-dependent differential-algebraic equations (DAEs) in Chapter 4. Finally, we turn to a concrete bilevel optimal control problem in Chapter 5, explain how it fits Definition 2.2, and link the numerical errors of both levels.

Starting from a general parameter-dependent differential-algebraic control system

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

with *time* $t \in \mathbb{I}$, time-differentiable *state* $x$, parameters $\theta \in \Theta$ and continuous *input* $u$, we perform a linearization step such that we obtain a linearized DAE of the form

$$E(t,\theta)\dot{x} = A(t,\theta)x + B(t,\theta)u + f. \tag{I.1}$$

We discuss conditions under which we can separate the time-parameter space $\mathbb{I} \times \theta$ into disjoint subsets, where the *characteristic values* of the linear DAE (I.1), c. f., Hypothesis 3.3, stay constant. Starting from one such system, e. g., with time-parameter space $\mathbb{I}_1 \times \Theta_1$, we set up the corresponding linear-quadratic optimal control problem. Given weight functions $Q(t,\theta), S(t,\theta)$, and $R(t,\theta)$ on $\mathbb{I}_1 \times \Theta_1$ and a weight $K(\theta)$ on the final state we want to minimize the objective function

$$\mathcal{J}(x,u) \coloneqq x(t_\mathrm{f})^\mathsf{H} K x(t_\mathrm{f}) + \int_{t_0}^{t_\mathrm{f}} \begin{pmatrix} x \\ u \end{pmatrix}^\mathsf{H} \begin{bmatrix} Q & S \\ S^\mathsf{H} & R \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \mathrm{d}t \tag{I.2}$$

over all input functions $u(t)$ such that (I.1) is fulfilled with an initial condition $x(t_0) = x_0$.

We compute the minimizer of (I.2) with the help of the necessary conditions for optimality, which constitute a boundary value problem of the form

$$E\dot{x} = (A + \dot{E})x + Bu + f, \tag{I.3a}$$
$$-E^{\mathsf{H}}\dot{\lambda} = Qx + Su + (A + \dot{E})^{\mathsf{H}}\lambda, \tag{I.3b}$$
$$0 = S^{\mathsf{H}}x + Ru - B^{\mathsf{H}}\lambda, \tag{I.3c}$$

with corresponding boundary conditions

$$(E^{+}Ex)(t_0) = x_0, \qquad (EE^{+}\lambda)(t_{\mathrm{f}}) = (E^{+\mathsf{H}}Kx)(t_{\mathrm{f}}),$$

where $\lambda$ is a Lagrange multiplier function and $E^{+}$ is the Moore-Penrose inverse of $E$.

There are at least two ways for determining the necessary conditions (I.3). We can proceed by first deriving the *strangeness-free* formulation

$$\hat{E}(t,\theta)\dot{x} = \hat{A}(t,\theta)x + \hat{B}(t,\theta)u + \hat{f}$$

of the original dynamical equation (I.1), see the forthcoming discussion after Hypothesis 3.3, and then writing down the corresponding system of necessary conditions. Another approach is to directly write down a formal system of necessary conditions. The differences between those approaches are discussed in Subsection 3.4.5.

A summary of the general procedure for the computation of a solution of the bilevel optimal control problem is depicted in Figure I.1. The possible procedures for determining the sensitivities of the system of necessary conditions is pictured in more detail in Figure I.2. We discuss them briefly in Chapter 4, before we focus on the boldly marked paths in Figure I.2.

First, similar to the derivation of the necessary conditions (I.3), which themselves constitute a DAE with possibly higher index, we need to decide, whether we want to proceed with a strangeness-free formulation or by using some kind of formal approach. The latter case is discussed in Section 4.6. In the former case, we need to decide, which kind of index reduction is appropriate, which is investigated in Subsection 3.4.6.

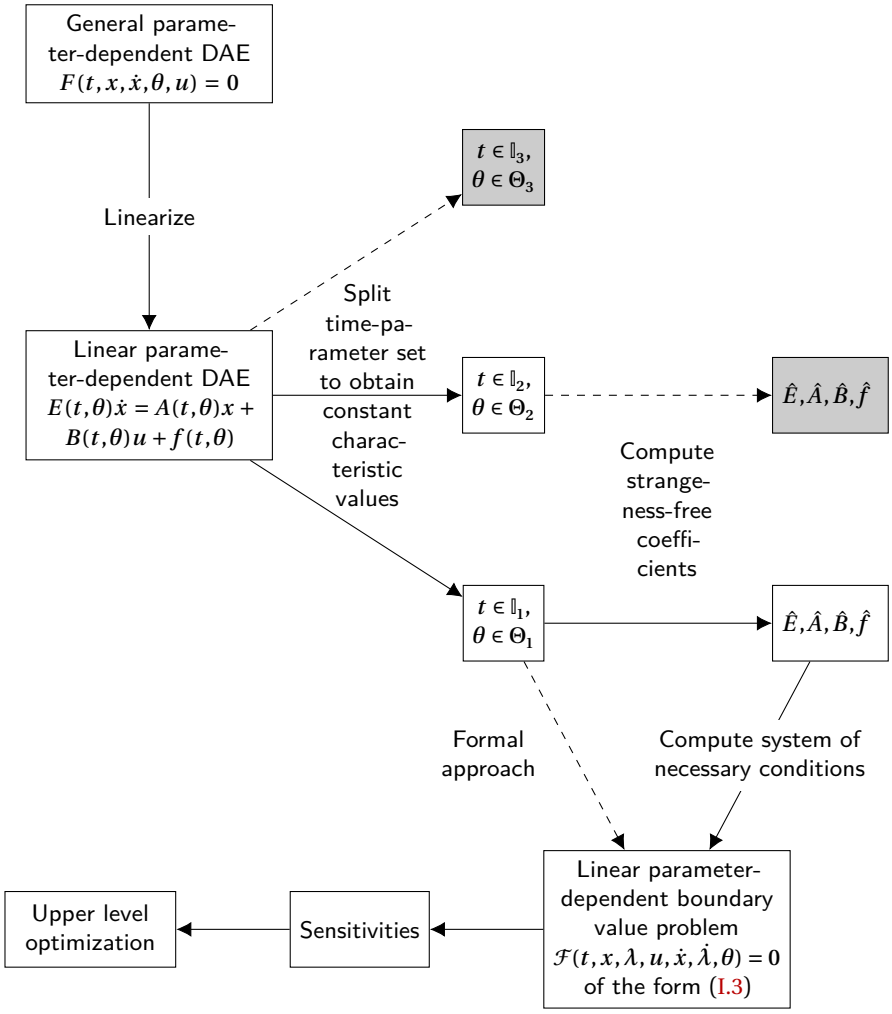



Figure I.1.: Full picture of the computation of sensitivities in a bilevel optimal control problem.

Once, we have fixed a strangeness-free formulation of the system of necessary conditions (I.3), we distinguish between two approaches, the forward system approach and the adjoint system approach. The forward system approach is discussed in Section 4.1, where we differentiate the system of necessary conditions. The adjoint approach is discussed in Section 4.2, where we develop a corresponding adjoint boundary value problem of the necessary conditions (I.3). Based on a solution of that adjoint system, we derive formulas for the computation of the sensitivities.

Using the sensitivity information, a concrete bilinear optimal control problem, also known as parameter estimation, is studied in Chapter 5, where the higher level optimization is given by a nonlinear least squares problem. We assume that the nonlinear least squares problem is solved by the Gauss-Newton method [DS96]. The numerical solution of the lower level optimal control problem is only correct up to the discretization and roundoff errors. We analyze the relation of these discretization errors with the convergence behavior of the Gauss-Newton method by using perturbation theory established in [GLN07]. We also check validity of the result with a numerical example.

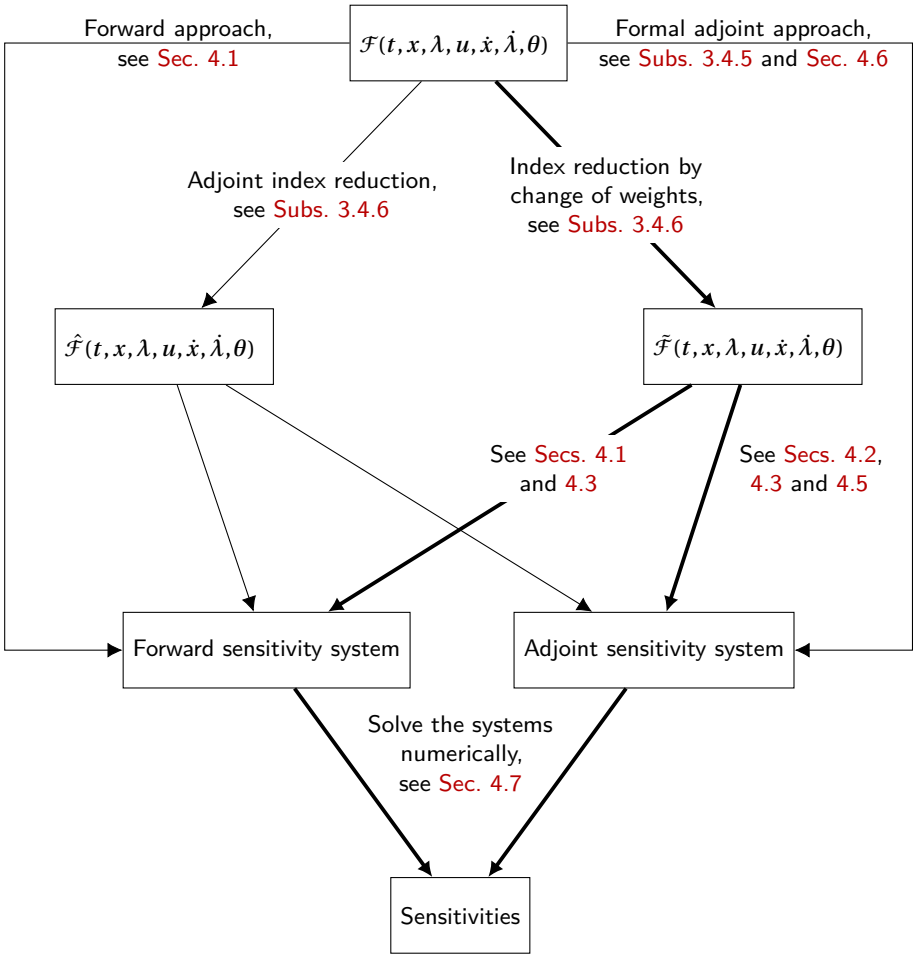



Figure I.2.: Different ways of computing the sensitivities from a given system of necessary conditions $\mathcal{F}$ of the form (I.3). All systems are linear with respect to the variables $\lambda$, $x$, and $u$ and possibly their time derivatives. Paths marked in bold are analyzed in more detail in this thesis in Subsection 3.4.6 and Chapter 4.

# 3. Preliminaries

In this chapter, we introduce definitions and results that are used throughout the forthcoming chapters. For basic notation see Pages xiii to xiv.

## 3.1. Generalized Inverses

First, we look at generalized inverses, see, e. g., [CM79].

**Definition 3.1.** (Pseudo inverse) Let $A \in \mathbb{K}^{n\times m}$ be given. Then the *Moore-Penrose pseudo inverse* or short *pseudo inverse* is uniquely defined by the four properties

1. $A^+AA^+ = A^+$;
2. $AA^+A = A$;
3. $(A^+A)^{\mathsf{H}} = A^+A$;
4. $(AA^+)^{\mathsf{H}} = AA^+$. ▷

There is also a weaker notion of a generalized inverse. We call $A^{\mathrm{g}}$ a *generalized inverse* of $A$ if it only fulfills the first condition in Definition 3.1. Note, that $A^g$ is not necessarily unique anymore.

**Proposition 3.1.** *Let $A \in \mathbb{K}^{n\times m}$, $B \in \mathbb{K}^{m\times r}$, $C \in \mathbb{K}^{s\times m}$ and isometric matrices $U \in \mathbb{K}^{l\times n}$, $V \in \mathbb{K}^{k\times m}$ be given. Then, the following holds.*

1. *$P_A \coloneqq A^+A$ and $\check{P}_A \coloneqq AA^+$ are orthogonal projectors onto $\operatorname{im}A^{\mathsf{H}}$ and $\operatorname{im}A$, respectively, i. e., both quantities are Hermitian and we have $P_A^2 = P_A$ and $\check{P}_A^2 = \check{P}_A$;*
2. *$(UA)^+ = A^+U^{\mathsf{H}}$ and $(AV^{\mathsf{H}})^+ = VA^+$;*
3. *$(P_AB)^+ = (P_AB)^+P_A$;*

4. $(CP_A)^+ = P_A(CP_A)^+$. ▷

*Proof.* See, e. g., [GV96]. □

Next, we discuss differentiability of parameter-dependent matrix functions. The following differentiation formulas are well-known.

**Lemma 3.2** ([GP73])**.** *Let an open set $\mathbb{O} \subseteq \mathbb{R}^p$ and $A \in \mathcal{C}^1(\mathbb{O}, \mathbb{K}^{n\times m})$ be given. Then the derivative of the pseudo-inverse $A^+$ with respect to the arguments can be expressed on $\mathbb{O}$ as*

$$\frac{\partial A^+}{\partial \theta} = -A^+ \frac{\partial A}{\partial \theta} A^+ A^+ (A^+)^{\mathsf{H}} \frac{\partial A^{\mathsf{H}}}{\partial \theta} (I_n - AA^+) + (I_m - A^+A) \frac{\partial A^{\mathsf{H}}}{\partial \theta} (A^+)^{\mathsf{H}} A^+. \tag{3.1}$$

*In particular for the projectors $P_A$ and $\check{P}_A$ we have*

$$\frac{\partial P_A}{\partial \theta} = A^+ \frac{\partial A}{\partial \theta} P_A^\perp + \left( A^+ \frac{\partial A}{\partial \theta} P_A^\perp \right)^{\mathsf{H}}$$

*and*

$$\frac{\partial \check{P}_A}{\partial \theta} = \check{P}_A^\perp \frac{\partial A}{\partial \theta} A^+ + \left( \check{P}_A^\perp \frac{\partial A}{\partial \theta} A^+ \right)^{\mathsf{H}},$$

*where the conjugate transpose in the second term of both formulas is applied for each term $A^+ \frac{\partial A}{\partial \theta_i} P_A^\perp$ or $\check{P}_A^\perp \frac{\partial A}{\partial \theta_i} A^+$ individually, i. e.,*

$$\left( A^+ \frac{\partial A}{\partial \theta} P_A^\perp \right)^{\mathsf{H}} := \left[ (A^+ \frac{\partial A}{\partial \theta_i} P_A^\perp)^{\mathsf{H}} \right]_{i=1,\dots,p}$$

*and*

$$\left( \check{P}_A^\perp \frac{\partial A}{\partial \theta} A^+ \right)^{\mathsf{H}} := \left[ \left( \check{P}_A^\perp \frac{\partial A}{\partial \theta_i} A^+ \right)^{\mathsf{H}} \right]_{i=1,\dots,p}.$$ ▷

*Remark* 3.1. Note, that the resulting quantity in (3.1) is a function, whose images are *tensors of order* 3, i. e., $\frac{\partial A(\theta)}{\partial \theta} \in \mathbb{K}^{n\times m\times p}$. For our purposes, we can view this object as a stack of matrices with the third index being the *stacking dimension*. In particular, products of those objects can be viewed as matrix products for every index in the stacking dimension.

Later we differentiate products of the form $A(\theta)x(\theta)$ with respect to parameters $\theta \in \mathbb{R}^p$. The product rule still holds in that case, i. e., we get

$$(Ax)_\theta = A_\theta x + Ax_\theta,$$

where $A_\theta x$ is a product of a rank 3 tensor with a vector and can be rewritten in matrix terms by using the convention

$$A_\theta x \coloneqq \begin{bmatrix} A_{\theta_1} & \dots & A_{\theta_p} \end{bmatrix} (I_p \otimes x),$$

where $\otimes$ denotes the Kronecker product, see, e. g., [KM06, p. 248]. ▷

## 3.2. Matrix Perturbations

Let us look at a few basic matrix perturbation results. The next result establishes an estimate of the difference between the pseudo inverse of a matrix and the pseudo inverse of a perturbed matrix in terms of the perturbed pseudo inverse and the error.

**Lemma 3.3.** *Let matrices $\tilde{J}, J \in \mathbb{K}^{m\times n}$ with $\operatorname{rk}\tilde{J} = \operatorname{rk} J$ be given and set $J_\mathfrak{E} \coloneqq \tilde{J} - J$. Then,*

$$\|\tilde{J}^+ - J^+\|_2 \leq \|\tilde{J}^+\|_2 \|J^+\|_2 \|J_\mathfrak{E}\|_2 \leq \frac{\|\tilde{J}^+\|_2^2 \|J_\mathfrak{E}\|_2}{1 - \|\tilde{J}^+\|_2 \|J_\mathfrak{E}\|_2}. \qquad ▷$$

*Proof.* The first inequality is a standard result which can be found in [MZ10]. The second part is an immediate consequence of Weyl's inequality for singular values [Wey12]. □

The bound is sharp as can be seen from the following example.

**Example 3.2.** Let

$$J = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad J_\mathfrak{E} = \begin{bmatrix} \frac{1}{100} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{J} = \begin{bmatrix} \frac{101}{100} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then

$$\|\tilde{J}^+ - J^+\|_2 = \frac{1}{101} = \frac{\frac{100}{101}\frac{1}{101}}{1 - \frac{1}{101}} = \frac{\|\tilde{J}^+\|_2 (\|\tilde{J}^+\|_2 \|J_\mathfrak{E}\|_2)}{1 - \|\tilde{J}^+\|_2 \|J_\mathfrak{E}\|_2}. \qquad ▷$$

We continue with deriving an estimate of the difference between the product of two matrices and its perturbed version.

**Lemma 3.4.** *Let matrices $A_\mathfrak{E}, \tilde{A}, A \in \mathbb{K}^{m\times n}$, $B_\mathfrak{E}, \tilde{B}, B \in \mathbb{K}^{n\times p}$, and $C \in \mathbb{K}^{n\times n}$ be given such that $\tilde{A} = A + A_\mathfrak{E}$ and $\tilde{B} = B + B_\mathfrak{E}$. Then*

$$\|\tilde{A}C\tilde{B} - ACB\|_2 \leq \left(\|\tilde{A}\|_2\|B_\mathfrak{E}\|_2 + \|\tilde{B}\|_2\|A_\mathfrak{E}\|_2 + \|A_\mathfrak{E}\|_2\|B_\mathfrak{E}\|_2\right)\|C\|_2. \qquad \triangleright$$

*Proof.* The assertion follows immediately by inserting the relations for $A$ and $B$ and an application of the triangular inequality. □

We can finally bound the error of the product of a matrix with the pseudo inverse of a second matrix.

**Lemma 3.5.** *Let matrices $A_\mathfrak{E}, \tilde{A}, A \in \mathbb{K}^{m\times n}$ and $B_\mathfrak{E}, \tilde{B}, B \in \mathbb{K}^{n\times p}$ be given such that $\tilde{A} = A + A_\mathfrak{E}$ and $\tilde{B} = B + B_\mathfrak{E}$. Also, assume that* $\mathrm{rk}A = \mathrm{rk}\tilde{A}$. *Then*

$$\|\tilde{A}^+\tilde{B} - A^+B\|_2 \leq \frac{\|\tilde{A}^+\|_2^2\|\tilde{B}\|_2}{1 - \|\tilde{A}^+\|_2\|A_\mathfrak{E}\|_2}\|A_\mathfrak{E}\|_2 + \frac{\|\tilde{A}^+\|_2}{1 - \|\tilde{A}^+\|_2\|A_\mathfrak{E}\|_2}\|B_\mathfrak{E}\|_2. \qquad \triangleright$$

*Proof.* First we apply Lemma 3.4 and obtain that

$$\|\tilde{A}^+\tilde{B} - A^+B\|_2 \leq \|\tilde{A}^+\|_2\|B_\mathfrak{E}\|_2 + \|\tilde{B}\|_2\|(A^+)_\mathfrak{E}\|_2 + \|(A^+)_\mathfrak{E}\|_2\|B_\mathfrak{E}\|_2,$$

where $(A^+)_\mathfrak{E} = \tilde{A}^+ - A^+$. By Lemma 3.3 we further deduce that

$$\|(A^+)_\mathfrak{E}\|_2 \leq \frac{\|\tilde{A}^+\|_2^2\|A_\mathfrak{E}\|_2}{1 - \|\tilde{A}^+\|_2\|A_\mathfrak{E}\|_2},$$
$$\|\tilde{A}^+\|_2 + \|(A^+)_\mathfrak{E}\|_2 \leq \frac{\|\tilde{A}^+\|_2}{1 - \|\tilde{A}^+\|_2\|A_\mathfrak{E}\|_2}.$$

Thus the assertion follows. □

## 3.3. Nonlinear Least Squares Problems

Let $\mathbb{I} \subseteq [0,\infty)$ be an interval and $\Theta \subseteq \mathbb{R}^p$. Given $(t_i, \xi_i) \in \mathbb{I} \times \mathbb{K}^n$, $i = 1, \ldots, q$, and a function $f : \mathbb{I} \times \Theta \to \mathbb{K}^n$, a nonlinear least squares problem consists of the unconstrained minimization problem

$$\min_{\theta\in\Theta} \quad \sum_{k=1}^{q} \|f(t_k, \theta) - \xi_k\|^2. \tag{3.2}$$

Define the residuals $r_k(\theta) \coloneqq f(t_k, \theta) - \xi_k$ and the residual vector

$$r(\theta) \coloneqq \begin{pmatrix} r_1(\theta) \\ \vdots \\ r_q(\theta) \end{pmatrix}. \tag{3.3}$$

Optimization Problem 3.2 can be solved with a simplified Newton approach, where we can use, that most of the information of the Hessian is already given by the Jacobian [DS96]. The following approach is called *Gauss-Newton method* and, given a starting point $\theta_k \in \Theta$, computes the next iterate by

$$J(\theta_k)^{\mathsf{H}} J(\theta_k) \Delta_k = -J^{\mathsf{H}}(\theta_k) r(\theta_k), \tag{3.4a}$$

$$\theta_{k+1} = \theta_k + \Delta_k, \tag{3.4b}$$

where $J$ denotes the Jacobian of $r$ with respect to $\theta$ and is assumed to have point-wise full column rank.

In [DS96] the following approximation result is derived.

**Theorem 3.6** ([DS96])**.** *Let $\theta^* \in \Theta$ be a minimizer of Optimization Problem 3.2 and let $\lambda$ be the smallest eigenvalue of $J(\theta^*)^{\mathsf{H}} J(\theta^*)$. Assume that there is an open set $\mathbb{O} \subseteq \Theta$ with $\theta^* \in \mathbb{O}$ such that*

1. *$J(\theta)$ is Lipschitz continuous in $\mathbb{O}$ with a Lipschitz constant equal to $\gamma_J$;*
2. *The Jacobian $J(\theta)$ is bounded, i. e., $\|J(\theta)\|_2 \leq \alpha$ for some $\alpha \in \mathbb{R}_+$ and all $\theta \in \mathbb{O}$.*
3. *There exists $0 < \sigma < \lambda$ such that*

$$\|(J(\theta^*) - J(\theta)) r(\theta^*)\|_2 \leq \sigma \|\theta^* - \theta\|_2$$

*for all $\theta \in \mathbb{O}$ and let $c \in \mathbb{R}$ such that $1 < c < \lambda/\sigma$.*

*Further, let $\theta_0$ be given. Then, there exists $\varepsilon > 0$ such that, if $\|\theta^* - \theta_0\| < \varepsilon$, then the iterates $\{\theta_k\}_k$ generated by the Gauss-Newton method (3.4) converge to $\theta^*$.* ▷

In the following we summarize perturbation results from [GLN07]. If we allow perturbations $e_k$ in the solution of the normal equation (3.4a), i. e., the perturbed Gauss-Newton scheme reads

$$J(\theta_k)^{\mathsf{H}} J(\theta_k) \Delta_k = -J(\theta_k)^{\mathsf{H}} r_k(\theta_k) + e_k, \tag{3.5a}$$

$$\theta_{k+1} = \theta_k + \Delta_k, \tag{3.5b}$$

then we obtain the following convergence result.

**Theorem 3.7** ([GLN07]). *Let all the assumptions of Theorem 3.6 hold and let $c \in \mathbb{R}$ be given such that $1 < c < \lambda/\sigma$. Choose $\beta_k$ such that*

$$0 \leq \beta_k \leq \hat{\beta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)},$$
$$\left\| e_k \right\|_2 \leq \beta_k \left\| J(\theta_k)^{\mathsf{H}} r(\theta_k) \right\|_2.$$

*Then there exists $\varepsilon > 0$ such that, if $\|\theta^* - \theta_0\| < \varepsilon$, the iterates $\{\theta_k\}_k$ generated by the perturbed Gauss-Newton method (3.5) converge to $\theta^*$.* ▷

If, on the other hand, we only have an approximation $\tilde{J}(\theta_k)$ of the Jacobian $J(\theta_k)$, i. e.,

$$\tilde{J}(\theta_k)^{\mathsf{H}} \tilde{J}(\theta_k) \Delta_k = -\tilde{J}(\theta_k)^{\mathsf{H}} r_k(\theta_k), \tag{3.6a}$$
$$\theta_{k+1} = \theta_k + \Delta_k, \tag{3.6b}$$

then we have the following converge result.

**Theorem 3.8** ([GLN07]). *Let all the assumptions of Theorem 3.6 hold and let $c \in \mathbb{R}$ be given such that $1 < c < \lambda/\sigma$. Choose $\eta_k$ such that*

$$0 \leq \eta_k \leq \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)},$$
$$\frac{\left\| J(\theta_k)^{\mathsf{H}} J(\theta_k) (J^+(\theta_k) - \tilde{J}^+(\theta_k)) r(\theta_k) \right\|_2}{\left\| J(\theta_k) r(\theta_k) \right\|_2} \leq \eta_k < \hat{\eta}.$$

*Then there exists $\varepsilon > 0$ such that, if $\|\theta^* - \theta_0\| < \varepsilon$, the iterates $\{\theta_k\}_k$ generated by the perturbed Gauss-Newton method (3.6) converge to $\theta^*$.* ▷

If we combine both perturbations, i. e.,

$$\tilde{J}(\theta_k)^{\mathsf{H}} \tilde{J}(\theta_k) \Delta_k = -\tilde{J}(\theta_k)^{\mathsf{H}} r_k(\theta_k) + e_k, \tag{3.7a}$$
$$\theta_{k+1} = \theta_k + \Delta_k, \tag{3.7b}$$

then we have another convergence result.

**Theorem 3.9** ([GLN07]). *Let all the assumptions of Theorem 3.6 hold and let $c \in \mathbb{R}$ be given such that $1 < c < \lambda/\sigma$. Choose $\eta_k$ and $\beta_k$ such that*

$$0 \leq \eta_k \leq \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)},$$
$$\left\| e_k \right\|_2 \leq \beta_k \left\| \tilde{J}(\theta_k)^{\mathsf{H}} r(\theta_k) \right\|_2,$$
$$0 \leq \beta_k \leq \frac{\eta_k \left\| \tilde{J}(\theta_k)^{\mathsf{H}} r(\theta_k) \right\|_2 - \left\| J(\theta_k)^{\mathsf{H}} J(\theta_k) (J^+(\theta_k) - \tilde{J}^+(\theta_k)) r(\theta_k) \right\|_2}{\left\| \tilde{J}(\theta_k)^{\mathsf{H}} r(\theta_k) \right\|_2 \left\| J(\theta_k)^{\mathsf{H}} J(\theta_k) (\tilde{J}(\theta_k)^{\mathsf{H}} \tilde{J}(\theta_k))^{-1} \right\|_2}.$$

*Then there exists $\varepsilon > 0$ such that, if $\|\theta^* - \theta_0\| < \varepsilon$, the iterates $\{\theta_k\}_k$ generated by the perturbed Gauss-Newton method (3.7) converge to $\theta^*$.* ▷

## 3.4. Linear Differential-Algebraic Equations and Control Systems

Consider control systems of the form

$$E\dot{x} = Ax + Bu + f \tag{3.8}$$

with given matrix functions $E \in \mathcal{C}^2(\mathbb{I}, \mathbb{K}^{n\times n})$, $A \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times n})$, $B \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times m})$, and inhomogeneity $f \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n)$, where $\mathbb{I} \subseteq \mathbb{R}$ is a (possibly unbounded) interval. In the control world, the time-dependent functions $x$ and $u$ are called *state* and *input*, respectively. We call $(x, u)$ with $x \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)$ and $u \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^m)$ a solution of the control problem (3.8), if $x(t)$ and $u(t)$ fulfill

$$E(t)\dot{x}(t) = Ax(t) + Bu(t) + f(t) \tag{3.9}$$

for all $t \in \mathbb{I}$. Then $(x, u)$ lies in the *behavior* $\mathfrak{B}_{(E,A,B,f)}$ which is the set of all such solutions for given data $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$, where $\Sigma_{m,n}(\mathbb{K})$ denotes the set of all admissible data. The control system can be casted to a pure differential-algebraic formulation, where formally there are no inputs and every variable is a state, by defining a new system

$$\mathcal{E}\dot{z} = \mathcal{A}z + f \tag{3.10}$$

where

$$\mathcal{E} := \begin{bmatrix} E & 0 \end{bmatrix}, \quad \mathcal{A} := \begin{bmatrix} A & B \end{bmatrix}, \quad z := \begin{bmatrix} x \\ u \end{bmatrix}.$$

This formulation is called the *behavioral* formulation [IM05] and originally goes back to Willems for ordinary differential equations (ODEs), see e. g., [PW98]. Imposing the additional condition

$$x(t_0) = x_0 \tag{3.11}$$

for some $x_0 \in \mathbb{K}^n$ we arrive at an initial value problem. An initial value problem (3.8), (3.11) is called solvable, if the corresponding initial value (3.11) is *consistent*, i. e., there exists $(x, u) \in \mathfrak{B}_{(E,A,B,f)}$ that fulfills (3.11). In contrast to ODEs, the set of consistent initial values for differential-algebraic equations (DAEs) is in general not be the whole space $\mathbb{K}^n$ and depends on the initial time $t_0$.

The control system is called *solvable* for some consistent initial value $x_0$ and sufficiently smooth $u$ and $f$, if there exists $x$ such that $(x, u) \in \mathfrak{B}_{(E,A,B,f)}$ and (3.11) is fulfilled. If that is the case for all consistent $x_0$ and sufficiently smooth $u$ and $f$, then the system is called *solvable*. If in addition, such $x$ is uniquely determined, the control system is called *uniquely solvable* or *regular*.

Since for DAEs the required regularity on $u$ and $f$ depends on the system structure, we present a more precise definition of what 'sufficiently smooth' means in Subsection 3.4.1.

In the following, we do not anymore distinguish between the formal notation in Equation (3.8) and the pointwise notation in Equation (3.9), whenever the correct interpretation is clear from the context.

*Remark* 3.2. Sometimes, we are dealing with pure DAEs without inputs that are of the form

$$E\dot{x} = Ax + f, \tag{3.12}$$

i. e., $m = 0$ in (3.8). The results in this subsection can be applied accordingly by setting $m = 0$. ▷

Solvability and uniqueness of solutions of the control system can be analyzed using the strangeness concept [KM06]. It builds on the so-called derivative array [Cam87] that we apply to the system in behavior form (3.10). Assuming sufficient smoothness of the data, we obtain an *inflated* system

$$\mathcal{M}_l \dot{z}_l = ((e_1^{(l)})^T \otimes \mathcal{N}_l) z_l + g_l \tag{3.13}$$

with $\mathcal{M}_l \in \mathbb{K}^{nl\times nl}$ and $\mathcal{N}_l \in \mathbb{K}^{nl\times n}$, where each subblock is defined by

$$
\begin{aligned}
(\mathcal{M}_l)_{ij} &= \binom{i}{j}\mathcal{E}^{(i-j)} - \binom{i}{j+1}\mathcal{A}^{(i-j-1)},\ i,j = 0,\dots,l,\\
(\mathcal{N}_l)_i &= \mathcal{A}^{(i)},\ i = 0,\dots,l,\\
(z_l)_i &= z^{(i)},\ i = 0,\dots,l,\\
(g_l)_i &= f^{(i)},\ i = 0,\dots,l.
\end{aligned}
$$

Note, that we slightly changed the definition of $\mathcal{N}_l$ compared to [KM06] to simplify some of the quantities involved in forthcoming terms such that it is more consistent with the nonlinear case.

Throughout this work we assume that there are no vanishing equations, i. e., consistency conditions on the right-hand side $f$. For more details, see [CKM12; KM06]. Then we can formulate the following regularity assumptions.

**Hypothesis 3.3** ([KM06])**.** There exist integers $n_\mu$, $n_a$, and $n_d$ such that the inflated pair $(\mathcal{M}_{n_\mu},\mathcal{N}_{n_\mu})$ associated with the given pair of matrix functions $(\mathcal{E},\mathcal{A})$ has the following properties:

1. For all $t \in \mathbb{I}$ we have $\operatorname{rk}\mathcal{M}_{n_\mu}(t) = (n_\mu+1)n - n_a$ such that there exists a smooth matrix function $Z_2$ of size $(n_\mu+1)n \times n_a$ and pointwise maximal rank satisfying $Z_2^{\mathsf{H}}\mathcal{M}_{n_\mu} = 0$.

2. For all $t \in \mathbb{I}$ we have $\operatorname{rk}\hat{\mathcal{A}}_2(t) = n_a$, where $\hat{\mathcal{A}}_2 = Z_2^{\mathsf{H}}\mathcal{N}_{n_\mu}$ such that there exists a smooth matrix function $T_2$ of size $n \times n_d$, $n_d = n - n_a$, and pointwise maximal rank satisfying $\hat{\mathcal{A}}_2 T_2 = 0$.

3. For all $t \in \mathbb{I}$ we have $\operatorname{rk}\mathcal{E}(t)T_2(t) = n_d$ such that there exists a smooth matrix function $Z_1$ of size $n \times n_d$ and pointwise maximal rank satisfying $\operatorname{rk}\hat{E}_1 T_2 = n_d$ with $\hat{E}_1 = Z_1^{\mathsf{H}}E$.

The integers $n_u$, $n_a$ $n_d$ are called *characteristic values* of $(\mathcal{E},\mathcal{A})$. ▷

The quantities in Hypothesis 3.3 are independent under *global equivalence transformations*.

**Definition 3.4** ([KM06])**.** Let $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ and invertible matrix functions $T_\mathrm{l} \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times n})$ an $T_\mathrm{r} \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times n})$ be given. Then the system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{f}) \in \Sigma_{m,n}(\mathbb{K})$ given by

$$\tilde{E} = T_\mathrm{l} E T_\mathrm{r}, \qquad \tilde{A} = T_\mathrm{l} A T_\mathrm{r} - T_\mathrm{l} E \dot{T}_\mathrm{r}, \qquad \tilde{B} = T_\mathrm{l} B, \qquad \tilde{f} = T_\mathrm{l} f$$

is said to be *globally equivalent* to $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{f})$.

If, additionally, $U = T_\mathrm{l}$ and $V = T_\mathrm{r}$ are pointwise unitary matrices, then the system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{f})$ is called *unitarily equivalent* to $(E, A, B, f)$. ▷

Under the assumptions of Hypothesis 3.3 and using the notation of (3.13) we can formulate the *strangeness-free* formulation given by

$$\hat{E}\dot{x} = \hat{A}x + \hat{B}u + \hat{f}, \tag{3.14}$$

where

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} Z_1^\mathsf{H} E \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} \coloneqq \begin{bmatrix} \hat{A}_1 & \hat{B}_1 \\ \hat{A}_2 & \hat{B}_2 \end{bmatrix} \coloneqq \begin{bmatrix} Z_1^\mathsf{H} & 0 \\ 0 & Z_2^\mathsf{H} \end{bmatrix} \begin{bmatrix} \mathcal{A} \\ \mathcal{N}_{n_\mu} \end{bmatrix},$$
$$\hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix} = \begin{bmatrix} Z_1^\mathsf{H} & 0 \\ 0 & Z_2^\mathsf{H} \end{bmatrix} g_l.$$

A system in the form (3.14) is strangeness-free, if it fulfills Hypothesis 3.3 with $n_\mu = 0$, i. e., if and only if the matrix

$$\begin{bmatrix} \hat{E}_1 & 0 \\ \hat{A}_2 & \hat{B}_2 \end{bmatrix} \tag{3.15}$$

has full row rank. It immediately follows that the strangeness-free formulation is strangeness-free. Note, that all selection matrices $Z_1$, $Z_2$, $T_2$ can be chosen smooth and with orthonormal columns, by, e. g., using the Gram-Schmidt orthonormalization process, which itself is a smooth process. Computing a smooth singular value decomposition is another tool for guaranteeing existence of smooth and orthonormal $Z_1$, $Z_2$, and $T_2$ in Hypothesis 3.3.

**Lemma 3.10** ([KM06, Theorem 3.9])**.** *Let $E \in C^l(\mathbb{I}, \mathbb{K}^{n\times n})$, $l \in \mathbb{N}_0 \cup \{\infty\}$ be given with $\operatorname{rk} E(t) = r$ for all $t \in \mathbb{I}$. Then there exist pointwise unitary (and therefore nonsingular) functions $U \in C^l(\mathbb{I}, \mathbb{K}^{n\times n})$ and $V \in C^l(\mathbb{I}, \mathbb{K}^{n\times n})$, such that*

$$U^\mathsf{H} E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

*with pointwise nonsingular* $\Sigma \in C^l(\mathbb{I}, \mathbb{K}^{r\times r})$. ▷

Another consequence is, that these matrices can be orthogonally complemented to unitary matrices by smooth and isometric matrices

$$Z_1' \in \mathbb{K}^{n\times n_a}, Z_2' \in \mathbb{K}^{(n_\mu+1)n\times(n_\mu+1)n-n_a}, T_2' \in \mathbb{K}^{n\times n_a}. \tag{3.16}$$

In the following, we merely follow [KM06; KM08].

We need to distinguish between two different systems. The *controlled* system $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ with inputs allowed to be arbitrary, free variables and the *uncontrolled* system $(E, A, f) \in \Sigma_{0,n}(\mathbb{K})$, where we formally set $u \equiv 0$ or $m = 0$. We immediately obtain from (3.15) that if the uncontrolled system is strangeness-free, then also the controlled system is strangeness-free. The opposite direction does not hold in general. In that case, we can apply a feedback of the form $u = Fx + v$ with $F \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{m\times n})$ and $v \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^m)$ being the new input [KM06]. The feedback can be chosen in such a way that

$$\begin{bmatrix} \hat{E}_1 \\ \hat{A}_2 + \hat{B}_2 F \end{bmatrix}$$

is pointwise invertible. Hence, after renaming of variables and suitable regularity assumptions we can safely assume, that the system under consideration is also strangeness-free as an uncontrolled system. Note, that for time-varying systems the feedback can in general only be computed during the integration and thus it may be pointless to include this procedure into a numerical algorithm as a preprocessing step.

Sometimes, for theoretical analysis, it turns out to be useful to transform the system to another representation. Using a unitarily equivalent system

$$(\tilde{E}, \tilde{A}) = (UEV, UAV - PE\dot{V})$$

according to Definition 3.4 does not change the solution behavior in the sense that $x(t)$ is a solution of the original system $(E, A)$ with inhomogeneity $f$ if and only if $\tilde{x}(t) := V^{\mathsf{H}} x(t)$ solves the new system $(\tilde{E}, \tilde{A})$ with inhomogeneity $\tilde{f} = Uf$. For numerical analysis, we want to avoid variable transformations, as they require differentiation of transformation matrices $V$ and possibly mix variables from different physical domains and orders of magnitude. Most numerical

integrators, however, do not apply such a transformation and directly operate on the original data.

A particular globally equivalent system for a system in strangeness-free from (3.14) is the *semi-explicit* form. It is given by

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{3.17}$$

with invertible $E_{11} \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n_d \times n_d})$ and $A$ partitioned accordingly with invertible $A_{22} \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n_a \times n_a})$.

### 3.4.1. Regularity Requirements

To adequately reflect the smoothness properties of the inhomogeneity $f$, in particular, when it is continuous only, we need to weaken the assumptions on differentiability of $x$. Since

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

we can rewrite (3.8) as

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x + Bu + f$$

or as

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x + f \tag{3.18}$$

if $m = 0$. For $f \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n)$ and $u \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^m)$ we therefore find solutions

$$x \in \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n) := \{x \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n) \mid E^+Ex \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)\}. \tag{3.19}$$

Similarly, for a system with $E^{\mathsf{H}}$ as leading matrix, we define the solution space

$$\mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^n) := \{\lambda \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n) \mid EE^+\lambda \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)\}. \tag{3.20}$$

This solution space becomes important for analyzing adjoint equations, see Subsection 3.4.5 and Sections 4.3 and 4.5 for more details.

In the case of semi-explicit systems (3.17) we have

$$E^+E = \begin{bmatrix} I_{n_d} & 0 \\ 0 & 0 \end{bmatrix} = EE^+$$

and thus (3.17) can be written as

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ 0 \end{pmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{3.21}$$

### 3.4.2. Flow Formulation

Analysis of solvability and uniqueness of solutions can be done via so-called flow formulations, which capture the whole solution behavior in a single flow operator. For DAEs in strangeness-free form (3.14) we obtain an explicit flow formulation [Bau14; Bau17] by first defining an equivalent projected system. Let

$$\dot{x}_\mathrm{d} = D_\mathrm{d}(t)x_\mathrm{d} + f_\mathrm{d}(t), \tag{3.22a}$$
$$x_\mathrm{a} = -D_\mathrm{a}(t)x_\mathrm{d} - f_\mathrm{a}(t), \tag{3.22b}$$

where

$$\begin{aligned} P_z = E^+E, \qquad \mathcal{P}_z = (I_n - D_\mathrm{a})P_z, \qquad P_z^\perp = I_n - P_z, \\ D_\mathrm{d} = (\hat{E}^+\hat{A} + \dot{P}_z)\mathcal{P}_z, \qquad f_\mathrm{d} = \hat{E}^+\hat{f} - (\hat{E}^+\hat{A} + \dot{P}_z)f_\mathrm{a}, \\ D_\mathrm{a} = P_z^\perp(\hat{A}_2P_z^\perp)^+\hat{A}_2, \qquad f_\mathrm{a} = (\hat{A}_2P_z^\perp)^+\hat{f}_2. \end{aligned} \tag{3.23}$$

Note, that both $x_\mathrm{d}$ and $x_\mathrm{a}$ are functions in the full space $\mathbb{K}^n$ in contrast to $x_1$ and $x_2$ in (3.21). Then, the flow is given by the affine linear operator

$$\Phi^{t_0}_{(E,A,f)}(t, x_0) = \Phi^{t_0}_{(E,A)}(t)x_0 + \int_{t_0}^{t} \Phi^{s}_{(E,A)}(t)f_\mathrm{d}(s)\mathrm{d}s - f_\mathrm{a}(t), \tag{3.24}$$

where the *homogeneous flow* is given by the operator $\Phi^{t_0}_{(E,A)}(t) = (\mathcal{P}_z\Phi^{t_0}_{D_\mathrm{d}}P_z)(t)$ and $\Phi^{t_0}_{D_\mathrm{d}}$ corresponds to the usual flow definition for ODEs, i. e., $\Phi^{t_0}_{D_\mathrm{d}}$ solves

$$\dot{\Phi}^{t_0}_{D_\mathrm{d}} = D_\mathrm{d}\Phi^{t_0}_{D_\mathrm{d}}, \qquad \Phi^{t_0}_{D_\mathrm{d}}(t_0) = I_n.$$

The homogeneous flow matrix $\Phi^{t_0}_{(E,A)}$ admits certain properties which are a generalization of the properties well-known for the ODE case.

**Proposition 3.11** ([Bau17]). *Let $s, t \in \mathbb{I}$ and the corresponding homogeneous flows $\Phi^s_{(E,A)}$ and $\Phi^{t_0}_{(E,A)}$ be given. Then we have*

1. $\Phi^s_{(E,A)}(t)\Phi^{t_0}_{(E,A)}(s) = \mathcal{P}_z(s)$;
2. $\Phi^t_{(E,A)}(s)\Phi^s_{(E,A)}(t) = \mathcal{P}_z(t)$;
3. $\Phi^s_{(E,A)}(t)$ *is a generalized inverse of* $\Phi^{t_0}_{(E,A)}(s)$, *i. e.,* $(\Phi^s_{(E,A)}(t))^{\mathrm{g}} = \Phi^{t_0}_{(E,A)}(s)$.

*In the case of an invertible $E$ we immediately obtain that $\Phi^s_{(E,A)}(t) = (\Phi^{t_0}_{(E,A)}(s))^{-1}$.* ▷

Since we are dealing with strangeness-free DAEs it follows immediately that the selector matrices $T_2$ and $Z_1$ can be directly read off the smooth singular value decomposition of $E$, i. e., $E = Z_1 E_{11} T_2^{\mathsf{H}}$ for invertible $E_{11}$ with $\operatorname{rk} E_{11} = \operatorname{rk} E$. Also, the complementary matrices $Z_1'$ and $T_2'$ of (3.16) represent the kernel of $E^{\mathsf{H}}$ and $E$, respectively. Consequently, $P_z = E^+E = T_2 T_2^{\mathsf{H}}$ and $P_z^\perp = T_2'(T_2')^{\mathsf{H}}$.

Let us also introduce the projectors

$$\check{P}_z = EE^+, \qquad \check{P}_z^\perp = I_n - \check{P}_z,$$

which can be computed by the products $Z_1 Z_1^{\mathsf{H}}$ and $Z_1'(Z_1')^{\mathsf{H}}$. Then we have the following result for unitary equivalence.

**Proposition 3.12.** *Let $(E, A, f) \in \Sigma_{0,n}(\mathbb{K})$ be given and let $(\tilde{E}, \tilde{A}, \tilde{f}) \in \Sigma_{0,n}(\mathbb{K})$ be unitarily equivalent via $U$ and $V$. Then, we have the following.*

1. *The quantities* (3.23) *for the transformed system are given by*

$$\tilde{P}_z = V^{\mathsf{H}} P_z V, \quad \tilde{P}_z^\perp = V^{\mathsf{H}} P_z^\perp V, \quad \tilde{\mathcal{P}}_z = V^{\mathsf{H}} \mathcal{P}_z V, \tag{3.25a}$$

$$\tilde{D}_{\mathrm{a}} = V^{\mathsf{H}} D_{\mathrm{a}} V, \quad \tilde{f}_{\mathrm{a}} = V^{\mathsf{H}} f_{\mathrm{a}}, \tag{3.25b}$$

$$\tilde{D}_{\mathrm{d}} = V^{\mathsf{H}} D_{\mathrm{d}} V + \dot{V}^{\mathsf{H}} P_z V, \quad \tilde{f}_{\mathrm{d}} = V^{\mathsf{H}} f_{\mathrm{d}} V + \dot{V}^{\mathsf{H}} P_z V. \tag{3.25c}$$

2. *The quantities* (3.23) *are independent of unitary transformations from the left and the transformation $Z_1$ to strangeness-free form* (3.14).
3. *Further, the quantities* (3.23) *can be represented in terms of the original data without explicitly forming $Z$ or $Z_1$ or $Z_1'$ or parts of them, i. e., we have*

$$\begin{aligned} D_{\mathrm{d}} &= (E^+A + \dot{P}_z)\mathcal{P}_z, \qquad f_{\mathrm{d}} = E^+ f - (E^+A + \dot{P}_z) f_{\mathrm{a}}, \\ D_{\mathrm{a}} &= (\check{P}_z^\perp A P_z^\perp)^+ A, \qquad f_{\mathrm{a}} = (\check{P}_z^\perp A P_z^\perp)^+ f. \end{aligned} \tag{3.26}$$

▷

*Proof.* We consider the equivalent system $\tilde{E} = UEV$, $\tilde{A} = UAV - UE\dot{V}$, and $\tilde{f} = Uf$. Let the transformation matrices $Z_1, Z_1'$ be given according to Hypothesis 3.3 and set

$$\begin{bmatrix} \hat{\tilde{A}}_2 & \hat{\tilde{f}}_2 \end{bmatrix} := (\tilde{Z}_1')^{\mathsf{H}} \begin{bmatrix} \tilde{A} & \tilde{f} \end{bmatrix}.$$

Then, the transformation matrix $\tilde{Z}_1'$ fulfills

$$\tilde{Z}_1' = UZ_1'\tilde{P}$$

for some pointwise orthogonal matrix function $\tilde{P}$. Thus,

$$\tilde{P}_z := \tilde{E}^+\tilde{E} = V^{\mathsf{H}}E^+U^{\mathsf{H}}UEV = V^{\mathsf{H}}P_zV$$

and

$$\tilde{\check{P}}_z^{\perp} = \tilde{Z}_1'(\tilde{Z}_1')^{\mathsf{H}} = UZ_1'(Z_1')^{\mathsf{H}}U^{\mathsf{H}} = U\check{P}_z^{\perp}U^{\mathsf{H}}. \tag{3.27}$$

Note, that by Proposition 3.1

$$P_z^{\perp}(\hat{A}_2P_z^{\perp})^+ = ((Z_1')^{\mathsf{H}}AP_z^{\perp})^+ = (\check{P}_zAP_z^{\perp})^+Z_1'.$$

Hence,

$$D_{\mathrm{a}} = P_z^{\perp}(\hat{A}_2P_z^{\perp})^+\hat{A}_2 = (\check{P}_z^{\perp}AP_z^{\perp})^+\check{P}_z^{\perp}A = (\check{P}_z^{\perp}AP_z^{\perp})^+A$$

does not need explicit availability of $Z_1'$. By (3.27), Proposition 3.1 and noting that $\tilde{\check{P}}_z^{\perp}\tilde{A} = \tilde{\check{P}}_z^{\perp}UAV$, we conclude that

$$\tilde{D}_{\mathrm{a}} = (\tilde{\check{P}}_z^{\perp}\tilde{A}\tilde{P}_z^{\perp})^+\tilde{\check{P}}_z^{\perp}\tilde{A} = (U\check{P}_z^{\perp}AP_z^{\perp}V)^+U\check{P}_z^{\perp}AV = V^{\mathsf{H}}D_{\mathrm{a}}V.$$

The remaining quantities in Equations (3.25a) and (3.25b) follow analogously. For the quantities in Equation (3.25c) we first note, that $E^+A = \hat{E}^+\hat{A}$. Then, we obtain for $\dot{\tilde{P}}_z$ that

$$\dot{\tilde{P}}_z = \dot{V}^{\mathsf{H}}P_zV + V^{\mathsf{H}}\dot{P}_zV + V^{\mathsf{H}}P_z\dot{V}$$

and thus

$$\begin{aligned}\tilde{D}_{\mathrm{d}} = V^{\mathsf{H}}E^+(A\mathcal{P}_zV - E\dot{V}V^{\mathsf{H}}\mathcal{P}_zV) + V^{\mathsf{H}}\dot{P}_z\mathcal{P}_zV + \dot{V}^{\mathsf{H}}P_z\mathcal{P}_zV + V^{\mathsf{H}}P_z\dot{V}V^{\mathsf{H}}\mathcal{P}_zV \\ = V^{\mathsf{H}}D_{\mathrm{d}}V + \dot{V}^{\mathsf{H}}P_zV.\end{aligned}$$

The remaining quantity $\tilde{f}_{\mathrm{d}}$ follows analogously. □

The next result shows, that for the flow formulation we can use either the original representation of the DAE (3.12) or the reformulated version (3.18) with weaker solution requirements.

**Proposition 3.13.** *The system* (3.12) *and the reformulated version* (3.18) *are represented by the same flow representation* (3.23)*. In particular, solutions of the flow equation* (3.24) *with* $x_\mathrm{d} \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)$ *and* $x_\mathrm{a} \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n)$ *define solutions* $x = x_\mathrm{d} + x_\mathrm{a} \in \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n)$. ▷

*Proof.* Follows immediately by inspecting the proof of the flow formulation in [Bau14] and by noting that $x_\mathrm{d} = E^+Ex$ and $\check{P}_z^\perp(A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E)) = \check{P}_z^\perp A$. □

We conclude this subsection with the following result.

**Lemma 3.14.** *The differential part of the homogeneous flow matrix* $\Phi^{t_0}_{(E,A)}$ *is a flow for the matrix* $T_2^\mathsf{H} D_\mathrm{d} T_2 + \dot{T}_2^\mathsf{H} T_2$*, i. e., the matrix*

$$T_2^\mathsf{H}(t)\Phi^{t_0}_{(E,A)}(t)T_2(t_0)$$

*is invertible for all admissible* $t_0, t \in \mathbb{I}$. ▷

*Proof.* The result is a direct consequence of the findings in [KM06, Section 7.2]. For a direct proof let

$$T = \begin{bmatrix} T_2 & T_2' \end{bmatrix},$$

which by the above convention (3.16) is invertible. Then by global equivalence and Proposition 3.12 we have

$$T^\mathsf{H}(t)\Phi^{t_0}_{D_\mathrm{d}}(t)T(t_0) = \Phi^{t_0}_{\check{D}_\mathrm{d}}(t) = \Phi^{t_0}_{T^\mathsf{H} D_\mathrm{d} T + \dot{T}^\mathsf{H} P_z T}(t)$$

for all $t \in \mathbb{I}$. By simple calculations it follows that

$$(T_2'(t))^\mathsf{H} D_\mathrm{d} T_2(t) + (\dot{T}_2'(t))^\mathsf{H} P_z(t) T_2(t) = (T_2')^\mathsf{H}(t)\dot{T}_2(t) + (\dot{T}_2'(t))^\mathsf{H} T_2(t) = 0$$

and thus

$$\Phi^{t_0}_{\check{D}_\mathrm{d}}(t) = \begin{bmatrix} \Phi^{t_0}_{11}(t) & \Phi^{t_0}_{12}(t) \\ 0 & \Phi^{t_0}_{22}(t) \end{bmatrix}$$

with invertible $\Phi^{t_0}_{11}(t)$. Hence,

$$T_2^\mathsf{H}(t)\Phi^{t_0}_{(E,A)}(t)T_2(t_0) = T_2^\mathsf{H}(t)\Phi^{t_0}_{D_\mathrm{d}}(t)T_2(t_0) = \Phi^{t_0}_{11}(t)$$

is also invertible. □

### 3.4.3. Boundary Value Problems

We consider boundary value problems for strangeness-free DAEs of the form

$$E\dot{x} = Ax + f(t), \tag{3.28a}$$
$$\Gamma_0 x(t_0) + \Gamma_f x(t_f) = \gamma, \tag{3.28b}$$

where $\Gamma_0, \Gamma_f \in \mathbb{K}^{n\times n}$, and $\gamma \in \mathbb{K}^n$. Note, that Equation (3.28a) corresponds to the strangeness-free DAE (3.8) with no inputs and (3.28b) defines boundary conditions. Setting $\gamma = x_0$, $\Gamma_0 = I_n$, and $\Gamma_f = 0$ we are in the initial value problem case (3.11). Here, we seek for a solution $x \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)$ that is smooth on the whole interval $\mathbb{I}$. The following theorem provides a characterization for unique solvability in the case of strangeness-free DAEs.

**Theorem 3.15** ([KMS05]). *Let $\Gamma_0, \Gamma_f \in \mathbb{K}^{n\times n}$ be given such that for an isometric $Z_\Gamma \in \mathbb{K}^{n\times n_d}$*

$$\operatorname{rk} Z_\Gamma^H \begin{bmatrix} \Gamma_0 & \Gamma_f \end{bmatrix} = n_d = \operatorname{rk} \begin{bmatrix} \Gamma_0 & \Gamma_f \end{bmatrix}.$$

*Also, let $\Phi^{t_0}_{(E,A)}$ denote the flow and $T_2$ be the selection matrix from Hypothesis 3.3.*

*Then the boundary value problem (3.28) is uniquely solvable for every $\gamma \in \operatorname{im} Z_\Gamma$, if and only if*

$$Z_\Gamma^H (\Gamma_0 + \Gamma_f \Phi^{t_0}_{(E,A)}(t_f)) T_2(t_0) \tag{3.29}$$

*is invertible.* ▷

*Proof.* This is an immediate consequence of the findings in [KMS05] for the arbitrary index case, however, for a slightly different notation. For a direct proof see the proof of Theorem 3.16. □

We can also add inner value conditions by replacing (3.28b) with

$$\sum_{i=0}^{q} \Gamma_i x(t_i) = \gamma, \tag{3.30}$$

where $t_0 < t_1 < \ldots < t_q := t_f$, $\Gamma_i \in \mathbb{K}^{n\times n}$, $i = 0, \ldots, q$, and $\Gamma_q := \Gamma_f$. Then we have the following generalization.

**Theorem 3.16.** *Let $\Gamma_0, \ldots, \Gamma_f \in \mathbb{K}^{n\times n}$ given such that for an isometric $Z_\Gamma \in \mathbb{K}^{n\times n_d}$*

$$\operatorname{rk} Z_\Gamma^H \begin{bmatrix} \Gamma_0 & \ldots & \Gamma_f \end{bmatrix} = n_d = \operatorname{rk} \begin{bmatrix} \Gamma_0 & \ldots & \Gamma_f \end{bmatrix}. \tag{3.31}$$

*Also, let $\Phi_{(E,A)}^{t_0}$ denote the flow and $T_2$ be the selection matrix from Hypothesis 3.3.*

*Then the boundary value problem (3.28a) and (3.30) is uniquely solvable for every $\gamma \in \operatorname{im} Z_\Gamma$ with $x(t_0) \in P_z(t_0)$, if and only if*

$$Z_\Gamma^{\mathsf{H}}\big(\Gamma_0 + \sum_{i=1}^{q} \Gamma_i \Phi_{(E,A)}^{t_0}(t_i)\big) T_2(t_0) \tag{3.32}$$

*is invertible.* ▷

*Proof.* Let us look at a parametrization of all solutions of the DAE (3.28a), which is given by (3.24) in terms of $x_{0,\mathrm{d}} \in \operatorname{im} P_z(t_0)$. This is a bijective mapping with $\operatorname{rk} P_z(t_0) = n_d$, where the remaining algebraic variables $x_{0,\mathrm{a}}$ of $x_0$ are recovered from (3.22b).

Replacing $x(t_i)$ in (3.30) by $\Phi_{(E,A,f)}^{t_0}(t_i, x_{0,\mathrm{d}})$, we obtain that

$$Z_\Gamma^{\mathsf{H}}\big(\Gamma_0 + \sum_{i=1}^{q} \Gamma_i \Phi_{(E,A)}^{t_0}(t_i)\big) T_2(t_0) T_2^{\mathsf{H}}(t_0) x_{0,\mathrm{d}} = Z_\Gamma^{\mathsf{H}} \tilde{\gamma} \tag{3.33}$$

for some $\tilde{\gamma}$ independent of $x_{0,\mathrm{d}}$. Note, that by (3.31) $\tilde{\gamma} \in \operatorname{im} Z_\Gamma$ if and only if $\gamma \in \operatorname{im} Z_\Gamma$. Thus, (3.33) can be uniquely solved for $x_{0,\mathrm{d}}$ for every $\gamma \in Z_\Gamma$, if and only if (3.32) is invertible. □

In Chapter 4 we encounter situations where we can only expect solutions $x$ that are smooth inside the subintervals $(t_i, t_{i+1})$, which makes $x(t)$ a piecewise smooth solution. More precisely, given $t_0 < \ldots < t_q = t_\mathrm{f}$, we look for solutions $x : \mathbb{I} \to \mathbb{K}^n$, which fulfill $x|_{(t_i,t_{i+1})} \in \mathcal{C}^1((t_i, t_{i+1}), \mathbb{K}^n)$, $i = 0, \ldots, q$. We denote the set of all such *piecewise smooth solutions* by

$$\mathcal{C}^1_{\mathrm{pw}}(\mathbb{I}, \mathbb{K}^n, \{t_0, \ldots, t_q\}).$$

In addition, we may have to allow jumps at intermediate values of $x$, which corresponds to *distributional inhomogeneities*.

**Definition 3.5** ($\delta$-distribution)**.** Let $\tau \in \mathring{\mathbb{I}} = (t_0, t_\mathrm{f})$ and a function $g \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^n)$ be given. Then the *$\delta$-distribution $\delta_\tau$* is defined by the formal notation

$$\int_{t_0}^{t_\mathrm{f}} \delta_\tau(t) g(t) \mathrm{d}t := g(\tau). \tag{3.34}$$

▷

*Remark* 3.3. In Equation (3.34) we represented the $\delta$-distribution formally as a time-dependent function. This is motivated by the fact, that the $\delta$-distribution can be approximated by

$$\lim_{\varepsilon\to 0}\int_{t_0}^{t_\mathrm{f}} \delta_\tau^\varepsilon(t)g(t)\mathrm{d}t = g(\tau),$$

while $\lim_{\varepsilon\to 0}\delta_\tau^\varepsilon(t) = 0$ for $t \neq \tau$ and some $\delta_\tau^\varepsilon \in \mathcal{C}^0(\mathbb{I},\mathbb{K})$.

Further, the integral in (3.34) is not a Riemann or Lebesgue integral. However, we assume that conventional linearity properties hold in combination with standard integrals. ▷

**Definition 3.6.** Let $\mathbb{1}_{[t_0,\tau)}(t)$ denote the jump function on $\mathbb{I}$ which is 1, whenever $t < \tau$ and 0 otherwise. Then we define the derivative

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{1}_{[t_0,\tau)}(t) \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{1}_{[t_0,\tau]}(t) \coloneqq \delta_\tau(t).$$

*Remark* 3.4. Since $\mathbb{1}_{[t_0,\tau)}(t) = \mathbb{1}_{(t,t_\mathrm{f}]}(\tau) = 1 - \mathbb{1}_{[t_0,t]}(\tau)$ for all $t,\tau \in \mathring{\mathbb{I}}$, we also have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{1}_{[t_0,\tau)}(t) = -\delta_\tau(t).$$

**Definition 3.7** (Distributional inhomogeneities)**.** Let $\tau \in \mathring{\mathbb{I}} = (t_0,t_\mathrm{f})$ and a function $g \in \mathcal{C}^0(\mathbb{I},\mathbb{K}^n)$ with $g(\tau) \in \check{P}_z(\tau)$ be given. We can replace $f$ in Equation (3.28a) by the formal term $\delta_\tau(t)g(t)$ to obtain the formal representation

$$E\dot{x} = Ax + \delta_\tau(t)g(t).$$

For $x \in \mathcal{C}^1_{\mathrm{pw}}(\mathbb{I},\mathbb{K}^n,\{\tau\})$ this has to be understood as the evaluation of

$$0 = \int_{t_0}^{t_\mathrm{f}} \lambda^\mathsf{H}(E\dot{x} - Ax - \delta_\tau(t)g(t))\mathrm{d}t \tag{3.35}$$

for any $\lambda \in \mathcal{C}^1(\mathbb{I},\mathbb{K}^n)$. ▷

*Remark* 3.5. The integral in (3.35) can be reformulated by partial integration as

$$0 = \lambda^\mathsf{H}Ex|_{t_0}^{t_\mathrm{f}} - \int_{t_0}^{t_\mathrm{f}} \Big(\frac{\mathrm{d}}{\mathrm{d}t}(E^\mathsf{H}\lambda)^\mathsf{H} + \lambda^\mathsf{H}A\Big)x + \lambda^\mathsf{H}\delta_\tau(t)g(t))\mathrm{d}t$$

for any $\lambda \in \mathcal{C}^1(\mathbb{I},\mathbb{K}^n)$. Thus, for $t_0 \to \tau-$, $t_f \to \tau+$ we obtain that

$$E(\tau)(x(\tau+)-x(\tau-)) = g(\tau),$$

since $\lambda$ can be chosen arbitrarily. This also makes clear, why we assume, that $g(\tau) \in \operatorname{im} \check{P}_z(\tau)$ in Definition 3.7.

Note, that since $x$ is piecewise smooth with possible discontinuity at $\tau$ we could naturally set $x(\tau)$ either to $x(\tau+)$ or to $x(\tau-)$. As the evaluation of $x$ at such an intermediate point $\tau$ does not matter in this thesis, we consider solutions, that differ at such points as equivalent in the sense of an equivalence relation. ▷

There are a couple of small inconsistencies and implicit assumptions that come with the use of this formal notation. Also, here we restrict to a special case, where solutions are not distributional and at most have jumps at prescribed points. For a comprehensive and detailed study of those problems we refer to [Jan71; KM06; RR88; RR89; Tre09] and the references therein.

If we have a distributional and/or discontinuous inhomogeneity we can use the following trick.

**Lemma 3.17.** *Let us consider the boundary value problem* (3.28) *with some* $\gamma \in \mathbb{K}^n, \Gamma_0, \Gamma_f$, *and* $Z_\Gamma$. *Further, let* $\tilde{f}$ *be given by* $\tilde{f} = \alpha \mathbb{1}_{[t_0,\tau)} + \beta\delta_\tau$ *for some* $\alpha \in \mathcal{C}^0(\mathbb{I},\mathbb{K}^n)$, $\tau \in (t_0,t_f)$, *and* $\beta \in \mathcal{C}^0(\mathbb{I},\mathbb{K}^n)$ *with* $\beta(\tau) \in \check{P}_z(\tau)$. *Then, the following are equivalent:*

1. *The boundary value problem* (3.28) *with sufficiently smooth* $f$ *is uniquely solvable in the sense of Theorem 3.15 and the unique solution* $x$ *lies in* $\mathcal{C}^1(\mathbb{I},\mathbb{K}^n)$;

2. *The boundary value problem* (3.28) *with* $f = 0$ *is uniquely solvable in the sense of Theorem 3.15 and the unique solution* $x$ *lies in* $\mathcal{C}^1(\mathbb{I},\mathbb{K}^n)$;

3. *The boundary value problem* (3.28) *with* $f = \tilde{f} = \alpha \mathbb{1}_{[t_0,\tau)} + \beta\delta_\tau$ *has a unique piecewise smooth solution* $x$ *in* $\mathcal{C}^1_{\mathrm{pw}}(\mathbb{I},\mathbb{K}^n,\{\tau\})$ *for any* $\gamma \in \operatorname{im} Z_\Gamma$;

4. *The augmented system*

$$E\dot{x}_1 = Ax_1 + \alpha, \tag{3.36a}$$
$$E\dot{x}_2 = Ax_2, \tag{3.36b}$$

$$\tilde{Z}_\Gamma^{\mathsf{H}}\left(\begin{bmatrix}\Gamma_0 & 0\\ 0 & 0\end{bmatrix}\begin{pmatrix}x_1(t_0)\\ x_2(t_0)\end{pmatrix}+\begin{bmatrix}0 & 0\\ E(\tau) & -E(\tau)\end{bmatrix}\begin{pmatrix}x_1(\tau)\\ x_2(\tau)\end{pmatrix}+\begin{bmatrix}0 & \Gamma_\mathrm{f}\\ 0 & 0\end{bmatrix}\begin{pmatrix}x_1(t_\mathrm{f})\\ x_2(t_\mathrm{f})\end{pmatrix}\right) = \tilde{Z}_\Gamma^{\mathsf{H}}\begin{pmatrix}\gamma\\ \beta(\tau)\end{pmatrix}, \quad (3.36c)$$

*with $\tilde{Z}_\Gamma$ given by*

$$\tilde{Z}_\Gamma = \begin{bmatrix}Z_\Gamma & 0\\ 0 & Z_1(\tau)\end{bmatrix}$$

*has a unique smooth solution on $\mathbb{I}$ for any $\gamma \in \operatorname{im} Z_\Gamma$ with $x_i(t_0) \in P_z(t_0)$, $i = 1,2$.*

*In that case the solution of the boundary value problem (3.28) with $f = \tilde{f} = g + \beta\delta_\tau$ can be written as $x(t) = \mathbb{1}_{[t_0,\tau)}x_1(t) + \mathbb{1}_{[\tau,t_\mathrm{f})}x_2(t)$.* ▷

*Proof.* First, we show that Statement 3 is equivalent to Statement 4.

Let the boundary value problem (3.28) with $f = \alpha\mathbb{1}_{[t_0,\tau)} + \beta\delta_\tau$ have a unique and piecewise smooth solution $x$. First note, that since the $\delta$-function is localized at time point $\tau$, we know that $x$ solves the DAE $E\dot{x} = Ax + \alpha$ on $[t_0,\tau)$ and $E\dot{x} = Ax$ on $(\tau, t_\mathrm{f}]$. Thus, we can define $x_1 : \mathbb{I} \to \mathbb{K}^n$ and $x_2 : \mathbb{I} \to \mathbb{K}^n$ as the smooth extension of $x|_{[t_0,\tau)}$ and $x|_{(\tau,t_\mathrm{f}]}$ on the whole interval $\mathbb{I}$, respectively. Then, with the help of Remark 3.5 we easily confirm that (3.36c) is fulfilled.

Vice versa, we can define $x(t) = \mathbb{1}_{[t_0,\tau)}x_1(t) + \mathbb{1}_{[\tau,t_\mathrm{f})}x_2(t)$, which solves the boundary value problem (3.28) with $f = \tilde{f} = \alpha\mathbb{1}_{[t_0,\tau)} + \beta\delta_\tau$ by using Remark 3.5 and relation (3.36c).

Uniqueness follows from the fact, that there is a one-to-one correspondence between parts of the solutions, which is a unique mapping except for $t = \tau$.

The equivalence of Statement 1 and Statement 2 follows directly by Theorem 3.15 as it is independent of the concrete inhomogeneity $f$.

It remains to show the equivalence of Statement 2 and Statement 4. First, we note that due to the coupling structure the flow of Equations (3.36a) and (3.36b) is given by $\Phi^{t_0}_{(I_2\otimes E, I_2\otimes A)} = I_2 \otimes \Phi^{t_0}_{(E,A)}$. Thus, by Theorem 3.16 we have that Statement 4

is equivalent to the regularity of the matrix

$$\begin{aligned}\tilde{Z}_\Gamma^{\mathsf{H}}\Bigg(\begin{bmatrix}\Gamma_0 & 0\\ 0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0\\ E(\tau) & -E(\tau)\end{bmatrix}\begin{bmatrix}\Phi_{(E,A)}^{t_0}(\tau) & 0\\ 0 & \Phi_{(E,A)}^{t_0}(\tau)\end{bmatrix}\\ + \begin{bmatrix}0 & \Gamma_\mathrm{f}\\ 0 & 0\end{bmatrix}\begin{bmatrix}\Phi_{(E,A)}^{t_0}(t_\mathrm{f}) & 0\\ 0 & \Phi_{(E,A)}^{t_0}(t_\mathrm{f})\end{bmatrix}\Bigg)\begin{bmatrix}T_2(t_0) & 0\\ 0 & T_2(t_0)\end{bmatrix}\\ = \tilde{Z}_\Gamma^{\mathsf{H}}\begin{bmatrix}\Gamma_0 & \Gamma_\mathrm{f}\Phi_{(E,A)}^{t_0}(t_\mathrm{f})\\ \big(E\Phi_{(E,A)}^{t_0}\big)(\tau) & -\big(E\Phi_{(E,A)}^{t_0}\big)(\tau)\end{bmatrix}\begin{bmatrix}T_2(t_0) & 0\\ 0 & T_2(t_0)\end{bmatrix}.\end{aligned} \tag{3.37}$$

Since

$$(Z_1^{\mathsf{H}} E\Phi_{(E,A)}^{t_0})(\tau)T_2(t_0) = (E_{11}T_2^{\mathsf{H}}\Phi_{(E,A)}^{t_0})(\tau)T_2(t_0)$$

is invertible by Lemma 3.14, we conclude that the matrix (3.37) is invertible if and only if (3.29) is invertible, which, by Theorem 3.15 is equivalent to Statement 2. □

### 3.4.4. Optimal Control

In this section we introduce optimal control problems for DAEs, where we mainly follow [KM08; KM11b]. For additional resources, see, e. g., [Ger12; Lib12].

We again consider the control system (3.8), where the input in Equation (3.8) is used as a free variable. In optimal control, one wants to restrict $u$ in such a way, that together with the resulting state $x$ a certain objective function is minimized. A typical objective function for linear control systems is a *quadratic* function of the form

$$\mathcal{J}(x,u) \coloneqq x(t_\mathrm{f})^{\mathsf{H}} K x(t_\mathrm{f}) + \int_{t_0}^{t_\mathrm{f}} \begin{pmatrix}x(t)\\ u(t)\end{pmatrix}^{\mathsf{H}}\begin{bmatrix}Q & S\\ S^{\mathsf{H}} & R\end{bmatrix}\begin{pmatrix}x(t)\\ u(t)\end{pmatrix}\mathrm{d}t, \tag{3.38}$$

where we require, that

$$\begin{bmatrix}Q & S\\ S^{\mathsf{H}} & R\end{bmatrix} \geq 0, \tag{3.39}$$

$K \in \mathbb{K}^{n\times n}$ is Hermitian and positive semi-definite and $R \in \mathbb{K}^{m\times m}$ is Hermitian and positive definite. The first summand is a weight on the final state, where the second summand weighs the trajectory on the whole time interval.

Using Lagrange's theorem, see e. g., [Zei85], and using the notation from the definition of the spaces in (3.19) and (3.20) we can write down the necessary conditions for an optimal solution.

**Theorem 3.18** ([KM08])**.** *Let the control system* (3.10) *be strangeness-free as a controlled system. Further, let* $\operatorname{im} K \subseteq \operatorname{im} E^{\mathsf{H}}(t_{\mathrm{f}})$ *and* $x_0$ *be a consistent initial condition. Then, for every optimal* $(x,u)$ *with* $x \in \mathcal{C}^1_{E^+E}(\mathbb{I},\mathbb{K}^m)$ *and* $u \in \mathcal{C}^0(\mathbb{I},\mathbb{K}^m)$ *there exists a function* $\lambda \in \mathcal{C}^1_{EE^+}(\mathbb{I},\mathbb{K}^n)$ *such that* $(\lambda,x,u)$ *solves*

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x + Bu + f, \tag{3.40a}$$

$$-E^{\mathsf{H}}\frac{\mathrm{d}}{\mathrm{d}t}(EE^+\lambda) = -Qx - Su + (A + EE^+\dot{E})^{\mathsf{H}}\lambda, \tag{3.40b}$$

$$0 = -S^{\mathsf{H}}x - Ru + B^{\mathsf{H}}\lambda, \tag{3.40c}$$

*with boundary conditions*

$$(E^+Ex)(t_0) = x_0, \qquad (EE^+\lambda)(t_{\mathrm{f}}) = -(E^{+\mathsf{H}}Kx)(t_{\mathrm{f}}). \tag{3.40d}$$

▷

Using the condition (3.39) we also have, that any solution $(\lambda,x,u)$ of (3.40) defines an optimal solution $(x,u)$ of the optimal control problem with objective function (3.38). The boundary conditions (3.40d) can be written in the general framework as in Subsection 3.4.3 by setting

$$\Gamma_0 = \begin{bmatrix} 0 & (E^+E)(t_0) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \Gamma_{\mathrm{f}} = \begin{bmatrix} 0 & 0 & 0 \\ (EE^+)(t_{\mathrm{f}}) & E^{+\mathsf{H}}(t_{\mathrm{f}})K & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad z = \begin{pmatrix} \lambda \\ x \\ u \end{pmatrix}.$$

For a theoretical analysis, let us assume that the uncontrolled system $(E,A)$ is strangeness-free and regular. Let us choose $V \in \mathcal{C}^0(\mathbb{I},\mathbb{K}^{n\times n})$ and $W \in \mathcal{C}^1(\mathbb{I},\mathbb{K}^{n\times n})$ pointwise unitary such that

$$\begin{aligned} \tilde{E} = VEW = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \qquad \tilde{A} = VAW - VA\dot{W} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \\ \tilde{B} = VB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \qquad \tilde{f} = Vf = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \qquad \tilde{x} = W^{-1}x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ \tilde{x}_0 = W^{-1}x_0 = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}, \end{aligned} \tag{3.41}$$

where $E_{11} \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n_d \times n_d})$ is pointwise invertible and

$$\begin{bmatrix} A_{22} & B_2 \end{bmatrix}$$

has pointwise full rank. The weights of the objective function (3.38) transform appropriately, i. e.,

$$\begin{aligned} \tilde{Q} = VQV^{\mathsf{H}} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^{\mathsf{H}} & Q_{22} \end{bmatrix}, \qquad \tilde{S} = VS = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \qquad \tilde{R} = R, \\ \tilde{K} = VKV^{\mathsf{H}} = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^{\mathsf{H}} & K_{22} \end{bmatrix}. \end{aligned} \tag{3.42}$$

In these coordinates the system of necessary conditions simplifies even more.

**Corollary 3.19** ([KM08])**.** *Let the control system* (3.10) *be strangeness-free as an uncontrolled system. Further, consider the transformed quantities* (3.41), (3.42)*, let* $\operatorname{im} \tilde{K} \subseteq \operatorname{im} \tilde{E}^{\mathsf{H}}(t_{\mathrm{f}})$*, and let* $\tilde{x}_0$ *be a consistent initial condition.*

*Then, for every optimal* $(\tilde{x}, u)$ *there exists a function* $\tilde{\lambda} \in \mathcal{C}^1_{\tilde{E}\tilde{E}^+}(\mathbb{I}, \mathbb{K}^n)$ *such that* $(\tilde{\lambda}, \tilde{x}, u)$ *solves* $\mathcal{E}\dot{\tilde{z}} = \mathcal{A}\tilde{z} + \hat{\tilde{f}}(t)$*, where*

$$\mathcal{E} = \begin{bmatrix} 0 & E_{11} & 0 & 0 & 0 \\ -E_{11}^{\mathsf{H}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} 0 & A_{11} & 0 & A_{12} & B_1 \\ (A_{11} + \dot{E}_{11})^{\mathsf{H}} & -Q_{11} & A_{21}^{\mathsf{H}} & -Q_{21}^{\mathsf{H}} & -S_1 \\ 0 & A_{21} & 0 & A_{22} & B_2 \\ A_{12}^{\mathsf{H}} & -Q_{21}^{\mathsf{H}} & A_{22}^{\mathsf{H}} & -Q_{22} & -S_2 \\ B_1^{\mathsf{H}} & -S_1^{\mathsf{H}} & B_2^{\mathsf{H}} & -S_2^{\mathsf{H}} & -R \end{bmatrix}, \tag{3.43}$$

$$\tilde{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}, \qquad \tilde{z} = \begin{pmatrix} \lambda_1 \\ x_1 \\ \lambda_2 \\ x_2 \\ u \end{pmatrix}, \qquad \hat{\tilde{f}} = \begin{pmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

*with boundary conditions*

$$x_1(t_0) = x_{1,0}, \qquad \lambda_1(t_{\mathrm{f}}) = -(E_{11}^{-\mathsf{H}} K_{11} x_1)(t_{\mathrm{f}}). \tag{3.44}$$

▷

Note, that in this formulation, we formally differentiate the input $u$. The resulting derivative $\dot{u}$, however, does not appear in the resulting equations. Thus, we can use this system description formally for algebraic manipulations, even if the input is not differentiable.

Further note, that the requirement that $\operatorname{im} \tilde{K} \subseteq \operatorname{im} \tilde{E}^{\mathsf{H}}(t_{\mathrm{f}})$ implies that

$$\tilde{K} = \begin{bmatrix} K_{11} & 0 \\ 0 & 0 \end{bmatrix}.$$

Even if the original control system is strangeness-free as a controlled system, this does not necessarily mean, that the system of necessary conditions as in Equations (3.43) and (3.44) is again strangeness-free.

**Theorem 3.20** ([KM08])**.** *The system of necessary conditions* (3.43) *is strangeness-free and regular if and only if the matrix*

$$\hat{R} \coloneqq \begin{bmatrix} 0 & A_{22} & B_2 \\ A_{22}^{\mathsf{H}} & -Q_{22} & -S_2 \\ B_2^{\mathsf{H}} & -S_2^{\mathsf{H}} & -R \end{bmatrix} \tag{3.45}$$

*is invertible for all* $t \in \mathbb{I}$. ▷

*Proof.* The proof follows as in [KM08] by noting that we multiplied the second, fourth and fifth block row of (3.43) by $-1$. □

It can be immediately seen that condition (3.45) is fulfilled only if

$$\begin{bmatrix} A_{22} & B_2 \end{bmatrix} \tag{3.46}$$

has full rank, which is fulfilled in our case of strangeness-free controlled systems. Hence, in particular also for systems that are strangeness-free as an uncontrolled system.

**Lemma 3.21.** *Let* $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ *be strangeness-free as a controlled system and let* $\hat{R}$ *be given by* (3.45)*. If* (3.46) *has full row rank and if the submatrix*

$$\begin{bmatrix} Q_{22} & S_2 \\ S_2^{\mathsf{H}} & R \end{bmatrix} \tag{3.47}$$

*in* (3.45) *is positive definite, then* $\hat{R}$ *is invertible.* ▷

*Proof.* First note, that $R$ is invertible. Taking the Schur complement of $\hat{R}$ with respect to the 0 block, we obtain that $\hat{R}$ is invertible, if

$$\begin{bmatrix} A_{22} & B_2 \end{bmatrix} \begin{bmatrix} Q_{22} & S_2 \\ S_2^H & R \end{bmatrix} \begin{bmatrix} A_{22}^H \\ B_2^H \end{bmatrix}$$

is invertible. This is guaranteed by the fact that (3.46) has full row rank and the matrix (3.47) is positive definite and thus possesses an invertible Cholesky factorization. □

### 3.4.5. Adjoint Equations

In this subsection we mostly follow [KM11a]. The homogeneous part of equation (3.40b) in Theorem 3.18 can be rewritten as

$$-E^H\dot{\lambda} = (A + \dot{E})^H\lambda \tag{3.48}$$

and is usually referred to as the *adjoint equation* of the DAE (3.12). If $(E, A, B, f)$ is not strangeness-free, we have to replace $E$ and $A$ in (3.47) with their strangeness-free coefficients $\hat{E}$ and $\hat{A}$ of (3.14).

It is tempting to use the adjoint equations with respect to the original data $E, A$ also when the original DAE is not strangeness-free, i. e., the assumptions of Theorem 3.18 are not fulfilled. These what we call *formal* adjoints in contrast to the *true* adjoint (3.48) still show a lot of the underlying structure. The formal adjoint is defined by

$$(\check{E}, \check{A}) := (-E^H, (A + \dot{E})^H). \tag{3.49}$$

In the strangeness-free case Equation (3.40b) is stated with respect to the original data. Hence, in this case the formal adjoint adjoint coincides with the true adjoint equation.

**Definition 3.8.** Let $(E, A)$ be given. Then the system $(E, A)$ is called *self-adjoint* if the formal adjoint $(\check{E}, \check{A})$ equals the original coefficients $(E, A)$. ▷

*Remark* 3.6. The systems of necessary conditions (3.40) and (3.43) constitute self-adjoint DAEs. For (3.43) this immediately follows from the definition. The necessary conditions (3.40) are stated in the weak setting (3.18). The corresponding adjoint equation then reads

$$-E^H\frac{\mathrm{d}}{\mathrm{d}t}(EE^+\lambda) = (A + EE^+\dot{E})^H\lambda.$$

It is straightforward to check, that self-adjointness of the operator associated with $(E,A)$ is equivalent to the conditions $E = -E^{\mathsf{H}}$ and

$$(A + EE^{+}\dot{E})^{\mathsf{H}} = A + E\frac{\mathrm{d}}{\mathrm{d}t}E^{+}E.$$

Thus, the characterization of self-adjointness is equivalent for systems in the form (3.12) and (3.18). ▷

We have the following lemma.

**Lemma 3.22** ([KM11a]). *Let $(\tilde{E},\tilde{A})$ be globally equivalent to $(E,A)$. Then the adjoint pair of $(\tilde{E},\tilde{A})$ equals the transformed adjoint pair of $(E,A)$.* ▷

One can even show, that the index of the original DAE (3.12) is invariant under taking the adjoint. Similarly, one can also define *formal* necessary conditions.

**Proposition 3.23** ([KM11a]). *Let the original DAE Equation (3.12) have a well-defined differentiation-index $n_\nu \geq 1$ with characteristic value $n_d$, i. e., its strangeness index is also well-defined. Then the formal adjoint (3.49) also has a well-defined differentiation index, which equals $n_\nu$, and has the same characteristic value $n_d$.* ▷

This directly implies, that the adjoint equation of a strangeness-free DAE for which the differentiation-index is defined is also strangeness-free. In particular, the strangeness-free form of the adjoint system is given by the coefficients

$$\hat{\hat{E}} = \begin{bmatrix} \hat{\hat{E}}_1 \\ 0 \end{bmatrix}, \qquad \hat{\hat{A}} = \begin{bmatrix} \hat{\hat{A}}_1 \\ \hat{\hat{A}}_2 \end{bmatrix}.$$

Similarly, one can also define *formal* necessary conditions in the setting of non-vanishing strangeness index. Based on the *true* necessary conditions (3.40), we set

$$E\dot{x} = (A + \dot{E})x + Bu + f, \tag{3.50a}$$
$$-E^{\mathsf{H}}\dot{\lambda} = -Qx - Su + (A + \dot{E})^{\mathsf{H}}\lambda, \tag{3.50b}$$
$$0 = -S^{\mathsf{H}}x - Ru + B^{\mathsf{H}}\lambda \tag{3.50c}$$

with corresponding boundary conditions

$$x(t_0) = x_0, \qquad (E^{\mathsf{H}}\lambda)(t_\mathrm{f}) = -Kx(t_\mathrm{f}). \tag{3.50d}$$

Again, the boundary conditions (3.50d) can be reformulated in the form (3.44).

We have the following characterization for solvability of the formal necessary conditions.

**Theorem 3.24** ([KM11a])**.** *Let all data of the system* (3.50) *be sufficiently smooth. Further, let the formal system necessary conditions* (3.50) *have a solution* $(x, u, \lambda)$*. Then there exists a function* $\check{\lambda}$ *which replaces* $\lambda$ *such that* $(x, u, \check{\lambda})$ *solves the true system of necessary conditions* (3.40). ▷

However, this result does not imply the validity of the opposite direction. Let us suppose we do not want to solve the true necessary conditions, because we have to go through the process of index reduction, which involves solving the larger inflated system (3.13). Then, if the formal necessary conditions have a solution, we can directly solve the higher index system. Of course, if one needs $\lambda$, then this approach cannot be used. See [KM08] for a more detailed discussion.

### 3.4.6. Obtaining Numerical Solutions of the Necessary Conditions

For computing a numerical solution of the necessary conditions (3.40) there are multiple possibilities.

If we are allowed to assume that the system of necessary conditions is regular and strangeness-free, following [KM08], we can proceed as follows. For a numerical algorithm to work, we need to compute all the data defining the necessary conditions. The system defined in Equations (3.43) and (3.44) is not suitable for the numerical integration, since it involves variable transformations. Variable transformations can be very ill-conditioned when combining variables that have different physical meaning and thus have completely different behavior and orders of magnitude. Also, variable transformations can lead to non-smooth behavior. See e. g., [KM06] for more details.

Furthermore, the transformations $Z_1$ and $Z_2$ in Equation (3.14) are in general time-dependent smooth functions, which are very expensive to store. Thus, usually, numerical integration methods such as `GENDA`, see [KMS02], or `GELDA`, see [KMRW97], use non-smooth realizations and the fact that transformations from the left do not alter solutions of commonly used integration schemes such as Runge-Kutta or BDF methods. Instead, the derivation of the reduced system (3.14) only uses transformations from the left. Unfortunately, the adjoint equation in (3.40) needs transformations from the right and the data become

functions that are not transformable to smooth functions from the left anymore. To overcome this problem, the following approach is introduced in [KM08]. We define new variables $\hat{\lambda}_1 = Z_1\lambda_1$ and $\hat{\lambda}_2 = Z_2\lambda_2$. Note, that $Z_1 Z_1^{\mathsf{H}}$ and $Z_2 Z_2^{\mathsf{H}}$ are again smooth functions as long as the characteristic values stay constant on $\mathbb{I}$. An equivalent boundary value problem also suitable for numerical algorithms is then given by

$$\hat{E}_1 \frac{\mathrm{d}}{\mathrm{d}t}(\hat{E}_1^+ \hat{E}_1 x) = (\hat{A}_1 + \frac{\mathrm{d}}{\mathrm{d}t}(\hat{E}_1^+ \hat{E}_1))x + \hat{B}_1 u + \hat{f}_1, \tag{3.51a}$$

$$0 = \hat{A}_2 x + \hat{B}_2 u + \hat{f}_2, \tag{3.51b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(E^{\mathsf{H}} Z_1 Z_1^{\mathsf{H}} \hat{\lambda}_1) = Qx + Su - A^{\mathsf{H}}\hat{\lambda}_1 - \mathcal{N}_{n_\mu}^{\mathsf{H}}\hat{\lambda}_2, \tag{3.51c}$$

$$0 = S^{\mathsf{H}} x + Ru - B^{\mathsf{H}}\hat{\lambda}_1 - \mathcal{N}_{n_\mu}^{\mathsf{H}}\hat{\lambda}_2, \tag{3.51d}$$

$$0 = (Z_1')^{\mathsf{H}}\hat{\lambda}_1, \tag{3.51e}$$

$$0 = (Z_2')^{\mathsf{H}}\hat{\lambda}_2 \tag{3.51f}$$

with boundary conditions

$$(\hat{E}_1^+ \hat{E}_1 x)(t_0) = x_0, \qquad (Z_1^{\mathsf{H}}\hat{\lambda}_1)(t_\mathrm{f}) = -\begin{bmatrix}(\hat{E}_1^+)^{\mathsf{H}} & 0\end{bmatrix} K x(t_\mathrm{f}).$$

If the system of necessary conditions (3.40) is not strangeness-free, we cannot use this approach anymore. There are a few possibilities, to obtain a strangeness-free system of necessary conditions, where each of them have their own advantages and disadvantages. We describe them briefly in the following subsections. This corresponds to the first reduction step in Figure I.2.

#### Two-step index reduction

The most straight-forward way to obtain a strangeness-free formulation for the necessary conditions is a two-step index reduction. First, we write down the true necessary conditions (3.40) for the strangeness-free formulation (3.14). As pointed out in Subsection 3.4.4, this system is not necessarily strangeness-free anymore. Thus, applying a second index-reduction gives us a strangeness-free formulation of the necessary conditions. According to [KMS14] it is also possible to perform this index reduction while keeping the self-adjoint structure of the equation.

However, this approach is not feasible for computing a numerical solution of the necessary conditions, if either the original control system (3.8) or the necessary conditions are not already strangeness-free. Otherwise, this approach involves building the inflated system for the necessary conditions based on the computed data in (3.51). This may be very ill-conditioned, since differentiating computed quantities can lead to arbitrary errors as roundoff errors are non-smooth. On the other hand, this approach is still feasible for the more costly smooth implementations of the selection matrices $Z_1$ and $Z_2$, see [KM06] for a possible procedure for the computation of those selection matrices.

*Remark* 3.7. In the case of constant coefficient systems, the selection matrices are constant and thus automatically smooth. ▷

#### Reducing the formal necessary conditions

Another approach is writing down the formal necessary conditions (3.50) for the original data and then performing the self-adjoint index-reduction as it was pointed out in [KM11a]. The advantage of that approach is, that quantities and derivatives are available in terms of the original data. The big disadvantage, though, is that the inhomogeneity may need additional smoothness requirements and that the boundary conditions may be inconsistent, i. e., they may lead to contradictions, see the discussion in [KM11a, Section 4] and [Bac06] for more details and examples.

#### Changing the weights

According to Theorem 3.20 the system of necessary conditions is regular and strangeness-free if the matrix $\hat{R}$ is invertible, which by Lemma 3.21 is the case if the weight matrix on the algebraic variables

$$\mathcal{W}_{\mathrm{a}} = \begin{bmatrix} Q_{22} & S_2 \\ S_2^{\mathsf{H}} & R \end{bmatrix}$$

is positive definite and

$$\begin{bmatrix} A_{22} & B_2 \end{bmatrix}$$

has full rank. The latter holds by the assumption that the system $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ is strangeness-free in the behavior setting, while the former can always

be achieved by changing the original weights. This is feasible in the sense, that the weights are parameters chosen by application, so they are subject to change. Of course, if we change the weights, we alter the problem and possibly also its solution.

### 3.4.7. Linear Parameter-Dependent Differential-Algebraic Equations

Linear DAEs may also depend on parameters $\theta$ and as such are of the form

$$E(t,\theta)\dot{x} = A(t,\theta)x + f(t,\theta), \tag{3.52}$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$. Analysis of existence and uniqueness of solutions is similar to DAEs depending on time only, as long as the characteristic values of the DAE do not change with the parameters. If the set of parameters is bounded we can decompose the definition set into countably many disjoint parameter sets, whose closure span the closure of the original set.

**Theorem 3.25.** *Let $\mathbb{I} \times \Theta \subseteq \mathbb{R} \times \mathbb{R}^p$ be a compact subset with $\overline{\mathring{\mathbb{I}} \times \mathring{\Theta}} = \mathbb{I} \times \Theta$ on which $E(t,\theta), A(t,\theta)$, and $f(t,\theta)$ are sufficiently smooth. Then there are at most countably many pairwise disjoint domains $\mathbb{A}_i \subseteq \mathbb{I} \times \Theta$ with*

$$\overline{\mathring{\mathbb{A}}_i} = \overline{\mathbb{A}_i}$$

*such that all the rank conditions in Hypothesis 3.3 are constant on each set and*

$$\overline{\bigcup_i \mathbb{A}_i} = \mathbb{I} \times \Theta. \tag{3.53}$$

▷

*Proof.* It follows along the lines of [CM79, Theorem 10.5.2], that the set

$$\mathbb{B} := \{(t,\theta) \in \mathbb{I} \times \Theta \mid \operatorname{rk} E(t,\theta) \text{ is not continuous}\}$$

is closed and has no interior. Thus, $(\mathbb{I} \times \Theta) \setminus \mathbb{B}$ contains at most countably many pairwise disjoint domains $\mathbb{A}_i$ on which $\operatorname{rk} E(t,\theta)$ is constant with $\overline{\mathring{\mathbb{A}}_i} = \overline{\mathbb{A}_i}$ such that

$$\overline{\bigcup_i \mathbb{A}_i} \subseteq \mathbb{I} \times \Theta.$$

Suppose now, that

$$\mathbb{I} \times \Theta \setminus \overline{\bigcup_i \mathbb{A}_i} \neq \{\},$$

which is a relatively open set. Then, again, by the fact that $\mathbb{B}$ is closed and has no interior, $\mathbb{I} \times \Theta \setminus \overline{\bigcup_i \mathbb{A}_i}$ must contain additional sets of the form of $\mathbb{A}_i$ or a subset of the boundary $\partial(\mathbb{I} \times \Theta)$. The former contradicts the fact, that we collected all of those $\mathbb{A}_i$ and the latter violates the assumption that $\overline{\mathring{\mathbb{I}} \times \mathring{\Theta}} = \mathbb{I} \times \Theta$. Hence, (3.53) must hold.

The argument can be repeated finitely many times for the remaining matrix functions in Hypothesis 3.3 on each $\mathbb{A}_i$ to iteratively refine the family of subsets. □

Note, that Theorem 3.25 only guarantees existence of *some* subsets $\mathbb{A}_i \subseteq \mathbb{I} \times \Theta$. They do not need to be decomposable as $\mathbb{A}_i = \mathbb{I} \times \Theta_i \subseteq \mathbb{I} \times \Theta$. There is also a straight-forward generalization of the local form of the smooth singular value decomposition [KM06, Theorem 3.9] to parameter-dependent matrix functions.

**Lemma 3.26** (Local smooth full rank decomposition)**.** *Let $E \in \mathcal{C}^l(\mathbb{I} \times \Theta, \mathbb{K}^{n \times n})$, $l \in \mathbb{N}_0$ be a smooth function in $t$ and $\theta$ with constant rank $r$. Then, for any $(t_0, \theta_0) \in \mathbb{I} \times \Theta$ there exists a relatively open subset $\mathbb{A}_0 \subseteq \mathbb{I} \times \Theta$ with $(t_0, \theta_0) \in \mathbb{A}_0$ on which smooth transformation matrices $V \in \mathcal{C}^l(\mathbb{A}_0, \mathbb{K}^{n \times n})$ and $W \in \mathcal{C}^l(\mathbb{A}_0, \mathbb{K}^{n \times n})$ exist such that*

$$VEW = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

*where $\Sigma_{11} \in \mathcal{C}^l(\mathbb{A}_0, \mathbb{K}^{r \times r})$ is a smooth and invertible function.* ▷

*Proof.* The proof follows along the lines of the first part of the proof of [KM06, Theorem 3.9]. □

*Remark* 3.8. Lemma 3.26 is a local version of [KM06, Theorem 3.9] for matrix functions depending on more than one parameter. To the best knowledge of the author, a global version is not available in general. ▷

*Remark* 3.9 (Order of differentiation). We denote time differentiation of variable $x$ by $\dot{x}$. Differentiation with respect to other variables is written in the subscript, i. e.,

$$x_\theta \coloneqq \frac{\partial}{\partial \theta} x.$$

When mixing both notations, then the differentiation in the subscript takes precedence over time differentiation, i. e., we have

$$\dot{x}_\theta \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}(x_\theta).$$

We can safely ignore this convention whenever time differentiation and differentiation with respect to $\theta$ commute. ▷

It is a well-known fact, that for ODEs depending on parameters, i. e., set, $E = I_n$ in Equation (3.52), smoothness of the coefficients carries over to smoothness of the solution. For the following theorem let us recall Remark 3.1.

**Theorem 3.27.** *Let $E = I_n$ in Equation (3.52), $\mathbb{I} \subseteq \mathbb{R}$ be a compact interval, and $\Theta \subseteq \mathbb{R}^n$ be open. Further, let $A \in \mathcal{C}^0(\mathbb{I} \times \Theta, \mathbb{K}^{n\times n})$, $f \in \mathcal{C}^0(\mathbb{I} \times \Theta, \mathbb{K}^n)$, and an initial condition $x_0 \in \mathcal{C}^0(\Theta, \mathbb{K}^n)$ be given.*

1. *Equation (3.52) has a unique solution for any fixed $\theta_0 \in \Theta$ and initial condition $x(t_0) = x_0(\theta_0)$.*

2. *If the data is sufficiently smooth, that is, we have $A(t,\cdot) \in \mathcal{C}^1(\Theta, \mathbb{K}^{n\times n})$, $\frac{\partial}{\partial\theta}A(\cdot,\theta) \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times n\times p})$, $f(t,\cdot) \in \mathcal{C}^1(\Theta, \mathbb{K}^n)$, $\frac{\partial}{\partial\theta}f(\cdot,\theta) \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times p})$, and $x_0 \in \mathcal{C}^1(\Theta, \mathbb{K}^n)$, it follows that every solution $x$ of (3.52) fulfills $x(t,\cdot) \in \mathcal{C}^1(\Theta, \mathbb{K}^n)$ and $\dot{x}(t,\cdot) \in \mathcal{C}^1(\Theta, \mathbb{K}^n)$ for all $t \in \mathbb{I}$. In particular,*

$$\frac{\partial^2}{\partial\theta\partial t}x = \frac{\partial^2}{\partial t\partial\theta}x,$$

*and $x_\theta$ solves*

$$\dot{x}_\theta = Ax_\theta + A_\theta x + f_\theta.$$ ▷

*Proof.* Statement 1 is essentially a corollary of the Picard-Lindelöf uniqueness result. See, e. g., [GJ16] for a proof of Statement 1.

For Statement 2 note, that from [Wal98, Chapter 13] it follows, that $x \in \mathcal{C}^1(\mathbb{I} \times \Theta, \mathbb{K}^n)$ and $\frac{\mathrm{d}}{\mathrm{d}t}x_\theta$ exists. Thus, we can differentiate Equation (3.52) with respect to $\theta$ at some fixed $\theta_0$ to obtain the equation

$$\frac{\partial^2}{\partial\theta\partial t}x = Ax_\theta + A_\theta x + f_\theta.$$

Further, note, that the differential equation

$$\dot{y} = Ay + A_\theta x + f_\theta$$

has a unique solution for given $x$ and initial condition $y(t_0, \theta_0) = \frac{\partial}{\partial\theta} x_0(\theta_0)$. One can show, that indeed $y = x_\theta$, see again [Wal98, Chapter 13], and thus time and parameter differentiation commute. □

*Remark* 3.10. Statement 2 of Theorem 3.27 can be extended to higher order derivatives of $x$ with respect to the parameters $\theta$. This requires more smoothness on the data with respect to the parameters. See [Wal98, Chapter 13] for more details. ▷

When the characteristic values stay constant, the result of Theorem 3.27 can be easily extended to the DAE case, since we have a smooth transformation to an ODE by using the smooth decomposition of Lemma 3.26. See Chapter 4 for more details. Unfortunately, when the characteristic values are allowed to change with respect to the parameters, this result does not hold anymore as the following example shows.

**Example 3.9.** Let

$$E(t,\theta) = \begin{bmatrix}\theta\end{bmatrix}, \quad A(t,\theta) = \begin{bmatrix}1\end{bmatrix}, \quad f(t,\theta) = \alpha, \quad \alpha \in \mathbb{R},$$

be given, where $\Theta = [0,1]$, $\mathbb{I} = [0,1]$, and $x_0(\theta) = \theta$. Then the solution of the DAE is given by

$$x(t,\theta) = \begin{cases} e^{\frac{t}{\theta}}\theta + (e^{\frac{t}{\theta}} - 1)\alpha, & \text{if } \theta > 0, \\ -\alpha, & \text{if } \theta = 0. \end{cases}$$

Letting $\theta$ go to 0, we see that the solution blows up and goes to infinity, if $\alpha \neq -1$. Thus, the solution becomes non-smooth at $\theta = 0$, even though the coefficients are smooth. In this particular case, the system $(E,A)$ is unstable for all positive $\theta$. Setting $\Theta = [-1,0]$ instead with $x_0(\theta) = \theta + 1$, we see that

$$\lim_{\theta\to 0} x(t,\theta) = \begin{cases} 0, & t > 0, \\ 1, & t = 0. \end{cases}$$

▷

This result is not very promising as we see that we not only loose smoothness, but also may have solutions that blow up. This behavior is similar to what was observed in several occasions for time-varying DAEs with non-constant characteristic values or for so-called hybrid DAEs, see, e. g., [BL02; KM18; LT12; LPS99; MW09; Wun08] and the references therein. Hence, in the following we restrict us to the case of constant characteristic values, this corresponds to the first branch in Figure I.1.

# 4. Sensitivities

In order to compute solutions of multilevel optimal control problems, we are interested in computing the change of solutions with respect to higher level variables. Sensitivity analysis for ordinary differential equations (ODEs) and also differential-algebraic equations (DAEs) has been addressed by many authors. In [SP02], sensitivity analysis is done for implicit ODEs with boundary conditions. In [CLPS03] the case of general index 1 DAEs and DAEs of index 2 in Hessenberg form with given initial values have been treated. Adjoint equations for the tractability index have been analyzed in [BL05; BM00]. For a comparison of the different index concepts we refer to, e. g., [Meh15].

In the following, we want to combine both approaches to define a forward and an adjoint system for the computation of sensitivities of DAEs with given boundary conditions. For the adjoint system we consider strangeness-free DAEs.

Let us consider the parameter-dependent version of the differential-algebraic boundary value problem (3.28) given by

$$E(t,\theta)\dot{x} = A(t,\theta)x + f(t,\theta), \tag{4.1a}$$

$$\Gamma_0(\theta)x(t_0,\theta) + \Gamma_f(\theta)x(t_f,\theta) = \gamma(\theta) \tag{4.1b}$$

with sufficiently smooth $E, A, f$ on $\mathbb{I} \times \Theta \subseteq \mathbb{R} \times \mathbb{R}^p$ and $\Gamma_0, \Gamma_f, \gamma$ on $\Theta$. We also add an output equation

$$y(t,\theta) = C(t,\theta)x + g(t,\theta), \tag{4.2}$$

which selects $r$ variables of interest with the help of a sufficiently smooth matrix function $C$ with $C(t,\theta) \in \mathbb{K}^{r\times n}$. We are interested in computing the partial derivatives $y_\theta(\tau,\hat{\theta})$ at given timepoints $\tau \in [t_0, t_f]$ and parameter values $\hat{\theta} \in \Theta$. In the following we most of the times drop the explicit dependence on $t$ and $\theta$. Further, we assume, that the rank conditions in Hypothesis 3.3 are constant on $\mathbb{I} \times \Theta$.

In this chapter we first introduce the so-called forward-system in Section 4.1 for the computation of the sensitivities. In Section 4.2 we introduce an alternative approach using adjoint equations. In Section 4.3 we discuss how these

approaches can be extended to situations with fewer regularity assumptions on time differentiability of the state, which is used for the application to the necessary conditions in optimal control as introduced in Subsection 3.4.4. These sections are concluded by Section 4.4, which provides a comparison of the approaches. In Section 4.5 we apply these results to the necessary conditions in optimal control and discuss implications for higher index cases in Section 4.6. We finally discuss approaches for obtaining the sensitivities numerically in Section 4.7. Also, compare with Figures I.1 and I.2.

## 4.1. The Forward System

One straightforward approach for computing the sensitivities is to differentiate, both, boundary value problem (4.1) and output equation (4.2) with respect to $\theta$. Assuming enough smoothness in the solutions we can interchange differentiation with respect to the parameters and time similar to Theorem 3.27 and obtain the new system

$$E\dot{x}_\theta = Ax_\theta + \tilde{F}, \tag{4.3a}$$
$$y_\theta = Cx_\theta + \tilde{G}, \tag{4.3b}$$
$$\tilde{\Gamma} = \Gamma_0 x_\theta(t_0) + \Gamma_f x_\theta(t_f), \tag{4.3c}$$

where $\tilde{F} \coloneqq A_\theta x - E_\theta \dot{x} + f_\theta$, $\tilde{G} \coloneqq g_\theta + C_\theta x$, and $\tilde{\Gamma} \coloneqq \gamma_\theta - (\Gamma_0)_\theta x(t_0) - (\Gamma_f)_\theta x(t_f)$. For the products of differentiated matrix functions with vectors let us recall Remark 3.1. Note, that $\tilde{F}, \tilde{G}$, and $\tilde{\Gamma}$ depend on a solution $x(\theta)$ of the original system (4.1) for a given parameter $\theta$.

**Lemma 4.1.** *Consider the parameter-dependent boundary value problem (4.1), where all involved matrix functions in (4.1) and (4.2) are assumed to be sufficiently smooth such that all rank conditions in Hypothesis 3.3 are constant on $\mathbb{I} \times \Theta$. Moreover, assume that the coefficients of the corresponding reduced system given pointwise in $\theta$ by (3.14) are sufficiently smooth in time and parameters.*

*Then, if for a fixed $\hat{\theta} \in \Theta$ there is a solution $x$ of (4.1), then $x$ is differentiable with respect to $\theta$ and $x_\theta$ solves (4.3).* ▷

*Proof.* Let $x$ solve the boundary value problem (4.1). Then $x$ also solves the reformulated system (3.14) for this fixed $\hat{\theta}$.

Now, using the flow representation (3.24), we obtain that there is a solution $(x_d, x_a)$ of (3.22) with $x = x_d + x_a$. According to Theorem 3.27 we can differentiate $x_d$ and $\dot{x}_d$ with respect to $\theta$ such that $\dot{x}_{d,\theta} = \frac{\partial}{\partial\theta}\dot{x}_d = \frac{\mathrm{d}}{\mathrm{d}t}x_{d,\theta}$. Consequently, using Equation (3.22b), time and parameter derivatives of $x_a$ also exist and commute. Hence, time and parameter derivatives also exist and commute for $x = x_d + x_a$. Thus, all quantities in Equations (4.1) and (4.2) are sufficiently smooth with respect to $\theta$ and differentiation of Equations (4.1) and (4.2) immediately proves that $x_\theta$ solves Equation (4.3). □

*Remark* 4.1.

1. When we solve system (4.3), we need to solve a system in the matrix-valued function $x_\theta(t) \in \mathbb{K}^{n\times p}$ and thus the effort for solving such a system grows with the number of parameters.

2. Note, that we assumed that the whole state $x$ is differentiable with respect to time. In view of Subsection 3.4.4 this is quite restrictive. This becomes more evident in Sections 4.2 and 4.3.

3. Both systems (4.1) and (4.3) share the same matrix pair $(E, A)$ for the homogeneous dynamics. Thus, from the discussion in Section 3.4 it is clear, that if system (4.1) is regular, then system (4.3) is regular as well.

4. However, this does not mean, that there is a one-to-one correspondence between the coefficients of (4.1) and (4.3). For example, take any coefficients which are constant in $\theta$ such that (4.1) is regular. Then, for any admissible value of $\gamma$ we obtain different solutions $x$ while the coefficients of (4.3) stay untouched and (4.3) has the unique solution $x_\theta = 0$.

5. Still, if the system (4.1) is strangeness-free and fulfills the assumptions of Theorem 3.15, then also the forward system (4.3) fulfills the assumptions of Theorem 3.15, i. e., uniqueness in the sense of Theorem 3.15 of the original system translates to uniqueness of the forward system.

6. For the quantities of the flow formulation, we need to use representation (3.25) instead of the representation (3.23) as the transformation matrices $Z_1$ and $T_2$ may not be available on the whole set $\mathbb{I} \times \Theta$ as smooth matrix functions, see also Lemma 3.26. ▷

## 4.2. Adjoint Sensitivities

Instead of directly solving the forward system (4.3a) there is an alternative approach for computing the sensitivities $y_\theta$. This can be done by considering adjoint equations similar as in Subsection 3.4.4. Let us first analyze the properties of adjoint equations coming from the boundary value problem (4.1). Throughout this section we assume the following.

**Assumption 4.2.** *For all $\theta \in \Theta$ it holds that*

1. *the DAE (4.1a) is regular and strangeness-free;*
2. *the boundary coefficients (4.1b) satisfy*

$$\Gamma_0 E^+ E = \Gamma_0$$

*and*

$$\Gamma_\mathrm{f} E^+ E = \Gamma_\mathrm{f};$$

3. *there exists isometric $Z_\Gamma \in \mathbb{K}^{n\times n_d}$ such that*

$$\operatorname{rk} Z_\Gamma^\mathsf{H} \begin{bmatrix} \Gamma_0 & \Gamma_\mathrm{f} \end{bmatrix} = n_d = \operatorname{rk} \begin{bmatrix} \Gamma_0 & \Gamma_\mathrm{f} \end{bmatrix}$$

*and the boundary inhomogeneity $\gamma$ lies in $\operatorname{im} Z_\Gamma$.* ▷

The following lemma provides a characterization of boundary conditions for an adjoint boundary equation. Also, let us define the *adjoint system coefficients*

$$\check{E} \coloneqq -E^\mathsf{H}, \qquad \check{A} \coloneqq (A + \dot{E})^\mathsf{H} \tag{4.4}$$

and recall that $P_z = E^+E$ and $\check{P}_z = EE^+$.

**Lemma 4.3.** *Let Assumption 4.2 hold and a fixed $\theta_0 \in \Theta$ be given. Further, let $\check{\Gamma}_{0,11} \in \mathbb{K}^{n_d\times n_d}$ and $\check{\Gamma}_{\mathrm{f},11} \in \mathbb{K}^{n_d\times n_d}$ be given such that*

$$\operatorname{im} \begin{bmatrix} \check{\Gamma}_{0,11}^\mathsf{H} \\ \check{\Gamma}_{\mathrm{f},11}^\mathsf{H} \end{bmatrix} = \ker Z_\Gamma^\mathsf{H} \begin{bmatrix} -\Gamma_0 T_2(t_0) & \Gamma_\mathrm{f} T_2(t_\mathrm{f}) \end{bmatrix} = \ker \begin{bmatrix} -\Gamma_{0,11} & \Gamma_{\mathrm{f},11} \end{bmatrix} \tag{4.5}$$

*and set*

$$\check{\Gamma}_0 \coloneqq Z_\Gamma \check{\Gamma}_{0,11} T_2^\mathsf{H}(t_0), \qquad \check{\Gamma}_\mathrm{f} \coloneqq Z_\Gamma \check{\Gamma}_{\mathrm{f},11} T_2^\mathsf{H}(t_\mathrm{f}). \tag{4.6}$$

*Then, the boundary value problem*

$$\check{E}\dot{\mu} = \check{A}\mu, \tag{4.7a}$$
$$\check{\gamma} = \check{\Gamma}_0(\check{E}\mu)(t_0) + \check{\Gamma}_f(\check{E}\mu)(t_f) \tag{4.7b}$$

*is uniquely solvable for every* $\check{\gamma} \in \operatorname{im} Z_\Gamma$*, if and only if the original boundary value problem* (4.1) *for the same fixed* $\theta_0$ *is uniquely solvable for every* $\gamma \in \operatorname{im} Z_\Gamma$. ▷

Before we prove Lemma 4.3, we need the following lemma.

**Lemma 4.4.** *Let* $\Phi^{t_0}_{(E,A)}$ *and* $\Phi^{t_f}_{(\check{E},\check{A})}$ *be the flows of the DAEs* (4.1a) *and* (4.7a)*, respectively, for a fixed* $\hat{\theta} \in \Theta$*. Then, it holds that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left((\Phi^{t_f}_{(\check{E},\check{A})})^{\mathsf{H}} E \Phi^{t_0}_{(E,A)}\right) = 0. \tag{4.8}$$

▷

*Proof.* This identity has been shown in a different index setting and different flow definition [BM02]. However, with the flow definition in Subsection 3.4.2 we obtain that

$$\begin{aligned}(\Phi^{t_f}_{(\check{E},\check{A})})^{\mathsf{H}} E \Phi^{t_0}_{(E,A)} &= \left(EE^+\Phi^{t_f}_{(\check{E},\check{A})}\right)^{\mathsf{H}} E(E^+E\Phi^{t_0}_{(E,A)}) \\ &= (EE^+)(t_f)\left(\Phi^{t_f}_{\check{D}_d}\right)^{\mathsf{H}} E \Phi^{t_0}_{D_d}(E^+E)(t_0),\end{aligned} \tag{4.9}$$

where $\check{D}_d$ corresponds to $D_d$ in (3.23) for the adjoint system (4.7a).

We first show that Equation (4.8) is invariant under coordinate transformations. To this end, let unitary matrix functions $U \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times n})$ and $V \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times n})$ be given such that $(\tilde{E}, \tilde{A}) = (U^{\mathsf{H}}EV, U^{\mathsf{H}}AV - U^{\mathsf{H}}E\dot{V})$, which again is a strangeness-free system. To this end, let us recall the transformation result of Proposition 3.12. Then, using Equation (3.25c), the projected flow fulfills $\Phi^{t_0}_{\tilde{D}_d} = V^{\mathsf{H}}\Phi^{t_0}_{D_d}V(t_0)$. Similarly, for the adjoint system we obtain that $\Phi^{t_f}_{\check{\tilde{D}}_d} = U^{\mathsf{H}}\Phi^{t_f}_{\check{D}_d}U(t_f)$. Thus, with the help of (4.9) we conclude that (4.8) is equivalent to

$$(\tilde{E}\tilde{E}^+)(t_f)\left(\Phi^{t_f}_{\check{\tilde{D}}_d}\right)^{\mathsf{H}} \tilde{E}\Phi^{t_0}_{\tilde{D}_d}(\tilde{E}^+\tilde{E})(t_0) = 0.$$

Hence, we are allowed to assume that the system in Equation (4.8) is already in semi-explicit strangeness-free form. Thus,

$$P_z = \begin{bmatrix} I_{n_d} & 0 \\ 0 & 0 \end{bmatrix}, \quad D_\mathrm{a} = \begin{bmatrix} 0 & 0 \\ A_{22}^{-1}A_{21} & 0 \end{bmatrix},$$
$$\mathcal{P}_z = \begin{bmatrix} I_{n_d} & 0 \\ -A_{22}^{-1}A_{21} & 0 \end{bmatrix}, \quad D_\mathrm{d} = \begin{bmatrix} E_{11}^{-1}(A_{11} - A_{12}A_{22}^{-1}A_{21}) & 0 \\ 0 & 0 \end{bmatrix},$$

and for the adjoint equation

$$\check{P}_z = \begin{bmatrix} -I_{n_d} & 0 \\ 0 & 0 \end{bmatrix}, \quad \check{D}_\mathrm{a} = \begin{bmatrix} 0 & 0 \\ A_{22}^{-\mathsf{H}}A_{12}^{\mathsf{H}} & 0 \end{bmatrix},$$
$$\check{\mathcal{P}}_z = \begin{bmatrix} I_{n_d} & 0 \\ -A_{22}^{-\mathsf{H}}A_{12}^{\mathsf{H}} & 0 \end{bmatrix}, \quad \check{D}_\mathrm{d} = \begin{bmatrix} -E_{11}^{-\mathsf{H}}(A_{11}^{\mathsf{H}} + \dot{E}_{11}^{\mathsf{H}} - A_{21}^{\mathsf{H}}A_{22}^{-\mathsf{H}}A_{12}^{\mathsf{H}}) & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, what remains to show is the statement for the implicit ODE

$$E_{11}\dot{x}_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})x_1,$$

however, this immediately follows from the same statement for (implicit) ODEs, see [Boy01; SP02]. □

*Proof of Lemma 4.3.* Let the original boundary value problem (4.1) for the fixed $\theta_0$ be given in semi-explicit form (3.17). Then, also the adjoint boundary value problem (4.7a) is in semi-explicit form given by the coefficients

$$\check{E} = \begin{bmatrix} -E_{11}^{\mathsf{H}} & 0 \\ 0 & 0 \end{bmatrix}, \qquad \check{A} = \begin{bmatrix} A_{11}^{\mathsf{H}} + \dot{E}_{11}^{\mathsf{H}} & A_{21}^{\mathsf{H}} \\ A_{12}^{\mathsf{H}} & A_{22}^{\mathsf{H}} \end{bmatrix}.$$

It is feasible to assume the semi-explicit form since globally equivalent systems have a one-to-one correspondence between solutions. Also, recall Lemma 3.22. We now have to show, how the boundary conditions are transformed. To this end, let unitary matrix functions $Z \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{n\times n})$ and $T \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times n})$ be given such that

$$(\tilde{E}, \tilde{A}) = (Z^{\mathsf{H}}ET, Z^{\mathsf{H}}AT - Z^{\mathsf{H}}E\dot{T})$$

is in semi-explicit form (3.17). Then, boundary condition (4.1b) is transformed to

$$\gamma = \Gamma_0 T_2(t_0)x_1(t_0) + \Gamma_\mathrm{f} T_2(t_\mathrm{f})x_1(t_\mathrm{f})$$

and thus after premultiplication with $Z_\Gamma^{\mathsf{H}}$ we obtain

$$Z_\Gamma^{\mathsf{H}}\gamma = \Gamma_{0,11}x_1(t_0) + \Gamma_{\mathrm{f},11}x_1(t_\mathrm{f}).$$

Analogously, for the adjoint boundary conditions (4.7b) we obtain after premultiplication with $Z_\Gamma^{\mathsf{H}}$ that

$$Z_\Gamma^{\mathsf{H}}\check{\gamma} = \check{\Gamma}_{0,11}E_{11}^{\mathsf{H}}(t_0)\mu_1(t_0) + \check{\Gamma}_{\mathrm{f},11}E_{11}^{\mathsf{H}}(t_\mathrm{f})\mu_1(t_\mathrm{f}).$$

Let $\Phi_{(E,A)}^{t_0}$ and $\Phi_{(\check{E},\check{A})}^{t_\mathrm{f}}$ be the homogeneous flows of the DAEs (4.1a) and (4.7a), respectively, with

$$\Phi_{(E,A)}^{t_0}(t_0) = \begin{bmatrix} I_{n_d} & 0 \\ -\left(A_{22}^{-1}A_{21}\right)(t_0) & 0 \end{bmatrix}, \qquad \Phi_{(\check{E},\check{A})}^{t_\mathrm{f}}(t_\mathrm{f}) = \begin{bmatrix} I_{n_d} & 0 \\ -\left(A_{22}^{-\mathsf{H}}A_{12}^{\mathsf{H}}\right)(t_\mathrm{f}) & 0 \end{bmatrix}.$$

Then, by Theorem 3.15 the original boundary value problem (4.1) is uniquely solvable for every $\gamma \in \operatorname{im} Z_\Gamma$, if and only if

$$Z_\Gamma^{\mathsf{H}}\left(\Gamma_0\Phi_{(E,A)}^{t_0}(t_0)T_2(t_0) + \Gamma_\mathrm{f}\Phi_{(E,A)}^{t_0}(t_\mathrm{f})T_2(t_\mathrm{f})\right) = \Gamma_{0,11} + \Gamma_{\mathrm{f},11}\Phi_{D_{d,11}}^{t_0}$$

is invertible, where $D_{\mathrm{d},11} = E_{11}^{-1}(A_{11} - A_{12}A_{22}^{-1}A_{21})$ denotes the first diagonal block of $D_\mathrm{d}$. Analogously, the adjoint boundary value problem (4.7a) is uniquely solvable for every $\check{\gamma} \in \operatorname{im} Z_\Gamma$ with $\mu(t_\mathrm{f}) \in \check{P}_z(t_\mathrm{f})$, if and only if

$$Z_\Gamma^{\mathsf{H}}\left(\check{\Gamma}_0\left(E^{\mathsf{H}}\Phi_{(\check{E},\check{A})}^{t_\mathrm{f}}Z_1\right)(t_0) + \check{\Gamma}_\mathrm{f}\left(E^{\mathsf{H}}\Phi_{(\check{E},\check{A})}^{t_\mathrm{f}}Z_1\right)(t_\mathrm{f})\right) = \check{\Gamma}_{0,11}\left(E_{11}^{\mathsf{H}}\Phi_{\check{D}_{d,11}}^{t_\mathrm{f}}\right)(t_\mathrm{f}) + \check{\Gamma}_{\mathrm{f},11}E_{11}^{\mathsf{H}}(t_\mathrm{f})$$

is invertible, where $\check{D}_{\mathrm{d},11} = -E_{11}^{-\mathsf{H}}(A_{11}^{\mathsf{H}} + \dot{E}_{11}^{\mathsf{H}} - A_{21}^{\mathsf{H}}A_{22}^{-\mathsf{H}}A_{12}^{\mathsf{H}})$ denotes the first diagonal block of $\check{D}_\mathrm{d}$. Hence, we have reduced the assertion to the problem stated for ODEs and the proof follows by looking at the corresponding proof for ODEs in [SP02]. □

We now present the relation between solutions of the adjoint boundary value problem and the sensitivity problem.

### 4.2.1. Sensitivities in the Open Interval $(t_0, t_f)$

First, we show what happens if we wish to evaluate the sensitivities at values strictly inside $\mathbb{I}$, i. e., $\tau \in (t_0, t_f)$.

For this, we need the following assumption.

**Assumption 4.5.** *Let the dependence of the coefficient matrices in Equations (4.1) and (4.2) on $\theta$ be sufficiently smooth, such that for solutions of Equation (4.1) the forward system (4.3) also has a solution.*

*Further, assume that (4.1) is uniquely solvable in the sense of Theorem 3.15 pointwise in $\theta$.* ▷

We present results for the sensitivities $y_\theta(\tau)$ and *integrated sensitivities* of the form

$$\int_{t_0}^{\tau} y_\theta \mathrm{d}t = \int_{t_0}^{\tau} C x_\theta + \tilde{G} \mathrm{d}t.$$

**Theorem 4.6.** *Consider the original system (4.1) and (4.2) and the corresponding forward system (4.3). Let Assumptions 4.2 and 4.5 hold and define $\check{\Gamma}_0, \check{\Gamma}_f$ as in (4.6). Let a fixed $\tau \in (t_0, t_f)$ be given. Then, we have the following.*

1. *There exists a unique solution $M \in \mathcal{C}^1_{pw}(\mathbb{I}, \mathbb{K}^{n\times r}, \{\tau\})$ of*

$$\check{E}\dot{M} = \check{A}M + \mathbb{1}_{[t_0,\tau)} C^{\mathsf{H}}, \tag{4.12a}$$
$$0 = \check{\Gamma}_0(\check{E}M)(t_0) + \check{\Gamma}_f(\check{E}M)(t_f), \tag{4.12b}$$
$$0 = \check{E}(\tau)(M(\tau+) - M(\tau-)), \tag{4.12c}$$

   *where $(\check{E}, \check{A})$ is defined as in (4.4).*

2. *Set*

$$\Xi := \left(\Gamma_0\Gamma_0^{\mathsf{H}} + \Gamma_f\Gamma_f^{\mathsf{H}}\right)^+ \left(\Gamma_0(\check{E}M)(t_0) - \Gamma_f(\check{E}M)(t_f)\right) \tag{4.13}$$

   *and let $M$ be the unique solution of (4.12). Then, the integrated sensitivities are given by*

$$\int_{t_0}^{\tau} y_\theta \mathrm{d}t = -\Xi^{\mathsf{H}}\tilde{\Gamma} + \int_{t_0}^{t_f} \mathbb{1}_{[t_0,\tau)} \tilde{G} + M^{\mathsf{H}}\tilde{F} \mathrm{d}t. \tag{4.14}$$

▷

*Proof.* By Lemma 4.1 and Remark 4.1, Item 5 we conclude that the assumptions of Lemma 4.3 are fulfilled, which in turn implies that the assumptions of Lemma 3.17 are fulfilled. This shows Statement 1.

From the boundary condition (4.12b) we obtain that

$$\begin{bmatrix} \check{\Gamma}_{0,11} & \check{\Gamma}_{\mathrm{f},11} \end{bmatrix} \begin{bmatrix} (E_{11}^{\mathsf{H}} Z_1^{\mathsf{H}} M)(t_0) \\ (E_{11}^{\mathsf{H}} Z_1^{\mathsf{H}} M)(t_\mathrm{f}) \end{bmatrix} = Z_\Gamma^{\mathsf{H}} \begin{bmatrix} \check{\Gamma}_0 & \check{\Gamma}_\mathrm{f} \end{bmatrix} \begin{bmatrix} (E^{\mathsf{H}} M)(t_0) \\ (E^{\mathsf{H}} M)(t_\mathrm{f}) \end{bmatrix} = 0.$$

Hence, from the definition of $\check{\Gamma}_0$ and $\check{\Gamma}_\mathrm{f}$ as in (4.6) we obtain existence of $\Xi_{11} \in \mathbb{K}^{n_d \times n_d}$ such that

$$\begin{bmatrix} -T_2^{\mathsf{H}}(t_0)\Gamma_0^{\mathsf{H}} \\ T_2^{\mathsf{H}}(t_\mathrm{f})\Gamma_\mathrm{f}^{\mathsf{H}} \end{bmatrix} \underbrace{Z_\Gamma \Xi_{11}}_{=:\Xi} = \begin{bmatrix} -\check{\Gamma}_{0,11}^{\mathsf{H}} \\ \check{\Gamma}_{0,11}^{\mathsf{H}} \end{bmatrix} \Xi_{11} = \begin{bmatrix} (E_{11}^{\mathsf{H}} Z_1^{\mathsf{H}} M)(t_0) \\ (E_{11}^{\mathsf{H}} Z_1^{\mathsf{H}} M)(t_\mathrm{f}) \end{bmatrix}.$$

After premultiplication with $\mathrm{diag}(T_2(t_0), T_2(t_\mathrm{f})$ and using Assumption 4.2 this leads to

$$\begin{bmatrix} -\Gamma_0^{\mathsf{H}} \\ \Gamma_\mathrm{f}^{\mathsf{H}} \end{bmatrix} \Xi = \begin{bmatrix} (E^{\mathsf{H}} M)(t_0) \\ (E^{\mathsf{H}} M)(t_\mathrm{f}) \end{bmatrix}. \tag{4.15}$$

Indeed, $\Xi$ given by formula (4.13) is a possible solution.

Let $x_\theta \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^n)$ be the solution of (4.3). It follows, that

$$\begin{aligned} & M^{\mathsf{H}} E x_\theta|_{t_0}^{\tau-} + M^{\mathsf{H}} E x_\theta|_{\tau+}^{t_\mathrm{f}} \\ &= \left(E^{\mathsf{H}}(\tau)(M(\tau-) - M(\tau+))\right)^{\mathsf{H}} x_\theta(\tau) - (M^{\mathsf{H}} E x_\theta)(t_0) + (M^{\mathsf{H}} E x_\theta)(t_\mathrm{f}) \\ &= \Xi^{\mathsf{H}}(\Gamma_\mathrm{f} x_\theta(t_\mathrm{f}) + \Gamma_0 x_\theta(t_0)) = \Xi^{\mathsf{H}} \tilde{\Gamma}. \end{aligned} \tag{4.16}$$

Thus, one obtains for any fixed $\tau \in (t_0, t_\mathrm{f})$ that

$$\begin{aligned} 0 &= \int_{t_0}^{\tau} M^{\mathsf{H}}(E\dot{x}_\theta - A x_\theta - \tilde{F})\mathrm{d}t + \int_{\tau}^{t_\mathrm{f}} M^{\mathsf{H}}(E\dot{x}_\theta - A x_\theta - \tilde{F})\mathrm{d}t \\ &= M^{\mathsf{H}} E x_\theta|_{t_0}^{\tau-} + M^{\mathsf{H}} E x_\theta|_{\tau+}^{t_\mathrm{f}} - \int_{t_0}^{\tau} (\dot{M}^{\mathsf{H}} E + M^{\mathsf{H}} \dot{E} + M^{\mathsf{H}} A) x_\theta \mathrm{d}t \\ &\qquad - \int_{\tau}^{t_\mathrm{f}} (\dot{M}^{\mathsf{H}} E + M^{\mathsf{H}} \dot{E} + M^{\mathsf{H}} A) x_\theta \mathrm{d}t - \int_{t_0}^{t_\mathrm{f}} M^{\mathsf{H}} \tilde{F} \mathrm{d}t \\ &= \Xi^{\mathsf{H}} \tilde{\Gamma} - \int_{t_0}^{t_\mathrm{f}} -\mathbb{1}_{[t_0,\tau)} C x_\theta + M^{\mathsf{H}} \tilde{F} \mathrm{d}t \end{aligned} \tag{4.17}$$

which leads to

$$\int_{t_0}^{\tau} y_\theta \mathrm{d}t = \int_{t_0}^{\tau} Cx_\theta + \tilde{G}\,\mathrm{d}t = -\Xi^{\mathsf{H}}\tilde{\Gamma} + \int_{t_0}^{t_\mathrm{f}} \mathbb{1}_{[t_0,\tau)}\tilde{G} + M^{\mathsf{H}}\tilde{F}\mathrm{d}t. \quad \square$$

If we are interested in computing the derivatives $y_\theta(\tau)$ at single timepoints $\tau$ we have to differentiate the quantities in Theorem 4.6 and solve a slightly different system, now including the delta distribution as introduced in Subsection 3.4.3 and the projection $\mathcal{P}_z^\sharp$ introduced in (3.23).

**Theorem 4.7.** *Consider the original system (4.1) and (4.2) and the corresponding forward system (4.3). Let Assumptions 4.2 and 4.5 hold and define $\check{\Gamma}_0, \check{\Gamma}_\mathrm{f}$ as in (4.6). Let a fixed $\tau \in (t_0, t_\mathrm{f})$ be given. Then, we have the following.*

1. *There exists a unique solution $N \in \mathcal{C}^1_{pw}(\mathbb{I}, \mathbb{K}^{n\times r}, \{\tau\})$ of*

$$\check{E}\dot{N} = \check{A}N + \delta_\tau (C\mathcal{P}_z^\sharp)^{\mathsf{H}}, \tag{4.18a}$$
$$0 = \check{\Gamma}_0(\check{E}N)(t_0) + \check{\Gamma}_\mathrm{f}(\check{E}N)(t_\mathrm{f}), \tag{4.18b}$$
$$(C\mathcal{P}_z^\sharp)^{\mathsf{H}}(\tau) = \check{E}(N(\tau+) - N(\tau-)), \tag{4.18c}$$

   *where $(\check{E}, \check{A})$ is defined as in (4.4).*

2. *Set*

$$\Xi_\tau := (\Gamma_0\Gamma_0^{\mathsf{H}} + \Gamma_\mathrm{f}\Gamma_\mathrm{f}^{\mathsf{H}})^+(-\Gamma_0 E^{\mathsf{H}}N(t_0) + \Gamma_\mathrm{f}E^{\mathsf{H}}N(t_\mathrm{f})) \tag{4.19}$$

   *and let $\tilde{F}_a$ denote the projected quantity of $\tilde{F}$ as defined in (3.23). Then*

$$y_\theta(\tau) = (\tilde{G} - C\tilde{F}_a)(\tau) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} + \int_{t_0}^{t_\mathrm{f}} N^{\mathsf{H}}\tilde{F}\mathrm{d}t. \tag{4.20}$$

$\triangleright$

*Proof.* The first part of the proof is analogous to the first part of the proof of Theorem 4.6, where we note that at time $t = \tau$ the coefficient $(C\mathcal{P}_z^\sharp)^{\mathsf{H}}(\tau)$ of the delta distribution lies in $\operatorname{im}\check{E}(\tau)$. Then, similar to (4.16) we have that

$$N^{\mathsf{H}}Ex_\theta|_{t_0}^{t_\mathrm{f}} = \Xi_\tau^{\mathsf{H}}\tilde{\Gamma}. \tag{4.21}$$

Then, similar to (4.17), by using (4.18a), we obtain that

$$\begin{aligned}0 &= \int_{t_0}^{t_\mathrm{f}} N^\mathsf{H}(E\dot{x}_\theta - Ax_\theta - \tilde{F})\mathrm{d}t \\ &= N^\mathsf{H}Ex_\theta|_{t_0}^{t_\mathrm{f}} - \int_{t_0}^{t_\mathrm{f}} (\dot{N}^\mathsf{H}E + N^\mathsf{H}\dot{E} + N^\mathsf{H}A)x_\theta \mathrm{d}t - \int_{t_0}^{t_\mathrm{f}} N^\mathsf{H}\tilde{F}\mathrm{d}t \\ &= \Xi_\tau^\mathsf{H}\tilde{\Gamma} + (C\mathcal{P}_z x_\theta)(\tau) - \int_{t_0}^{t_\mathrm{f}} N^\mathsf{H}\tilde{F}\mathrm{d}t.\end{aligned}$$

Note, that we can rewrite $Cx_\theta$ as

$$Cx_\theta = C\mathcal{P}_z x_\theta - C\tilde{F}_\mathrm{a}.$$

Thus,

$$\begin{aligned}y_\theta(\tau) &= \tilde{G}(\tau) + (Cx_\theta)(\tau) = \tilde{G}(\tau) + (C\mathcal{P}_z x_\theta)(\tau) - (C\tilde{F}_\mathrm{a})(\tau) \\ &= (\tilde{G} - C\tilde{F}_\mathrm{a})(\tau) - \Xi_\tau^\mathsf{H}\tilde{\Gamma} + \int_{t_0}^{t_\mathrm{f}} N^\mathsf{H}\tilde{F}\mathrm{d}t.\end{aligned}$$

□

*Remark* 4.2. There is another interpretation of Equation (4.18). The main idea is to differentiate the respective formulas in Theorem 4.6 with respect to $\tau$, in particular Equations (4.12a) to (4.12c) and (4.13). Note, that $M$ depends on the chosen timepoint $\tau$ and is related to $N$ via $N = \frac{\mathrm{d}}{\mathrm{d}\tau}M$. This holds in a distributional sense in the space of piecewise smooth functions. Thus, differentiating the boundary condition (4.12c) with respect to $\tau$ leads us to

$$\begin{aligned}0 &= \dot{E}^\mathsf{H}(\tau)(M(\tau+) - M(\tau-)) + E^\mathsf{H}(\tau)(\dot{M}(\tau+) - \dot{M}(\tau-) + N(\tau+) - N(\tau-)) \\ &= -A^\mathsf{H}(M(\tau+) - M(\tau-)) + C^\mathsf{H}(\tau) + E^\mathsf{H}(\tau)(N(\tau+) - N(\tau-)).\end{aligned}$$

Since $M$ fulfills Equation (4.12c) we obtain from $\check{D}_\mathrm{a}, \check{f}_\mathrm{a}$ defined for (4.12a) as in (3.25) that for the projected quantities $M_\mathrm{a}$, $M_\mathrm{d}$ we have

$$\begin{aligned}M_\mathrm{a}(\tau+) - M_\mathrm{a}(\tau-) &= -\left(\check{D}_\mathrm{a}(E^+)^\mathsf{H}E^\mathsf{H}\right)(\tau)(M_\mathrm{d}(\tau+) - M_\mathrm{d}(\tau-)) - \check{f}_\mathrm{a}(\tau+) + \check{f}_\mathrm{a}(\tau-) \\ &= 0 + \left((P_z^\perp A^\mathsf{H}\check{P}_z^\perp)^+ C^\mathsf{H}\right)(\tau)\end{aligned}$$

and thus

$$E^\mathsf{H}(\tau)(N(\tau+) - N(\tau-)) = \left(A^\mathsf{H}(P_z^\perp A^\mathsf{H}\check{P}_z^\perp)^+ C^\mathsf{H} - C^\mathsf{H}\right)(\tau) = ((D_\mathrm{a} - I)^\mathsf{H}C^\mathsf{H})(\tau),$$

or, equivalently, after premultiplication with $P_z$

$$E^{\mathsf{H}}(\tau)(N(\tau+)-N(\tau-)) = -(C\mathcal{P}_z)^{\mathsf{H}}(\tau). \qquad \triangleright$$

*Remark* 4.3. The quantity $C\tilde{F}_{\mathrm{a}}$ in (4.20) does not appear in (4.14). This term essentially recovers the sensitivities of the algebraic variables directly from the flow formulation via (3.22b). If we choose to compute only sensitivities of the differential variables, then we have $C^{\mathsf{H}}(\theta,t) \in \operatorname{im} P_z(\theta,t)$ for all $(\theta,t) \in \Theta \times \mathbb{I}$ and thus $C\mathcal{P}_z = C$ and $C\tilde{F}_{\mathrm{a}} = 0$. $\triangleright$

### 4.2.2. Sensitivities at the Boundary $t_0$ or $t_{\mathrm{f}}$

Let us now consider the special case of $\tau \in \{t_0, t_{\mathrm{f}}\}$ being at the boundary of $\mathbb{I}$. This can be seen as a limit process of the previous subsection letting $\tau$ go to $t_0$ or $t_{\mathrm{f}}$. Thus, one immediately obtains the following result for the integrated sensitivities.

**Corollary 4.8.** *Consider the original system (4.1) and (4.2) and the corresponding forward system (4.3). Let Assumptions 4.2 and 4.5 hold and define $\check{\Gamma}_0, \check{\Gamma}_{\mathrm{f}}$ as in (4.6). Let $\tau$ be fixed to $\tau = t_{\mathrm{f}}$. Then, we have the following.*

1. *There exists a unique solution $M \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times r})$ of*

$$\check{E}\dot{M} = \check{A}M + C^{\mathsf{H}}, \tag{4.22a}$$
$$0 = \check{\Gamma}_0(\check{E}M)(t_0) + \check{\Gamma}_{\mathrm{f}}(\check{E}M)(t_{\mathrm{f}}), \tag{4.22b}$$

   *where $(\check{E},\check{A})$ is defined as in (4.4).*

2. *Set*

$$\Xi := (\Gamma_0\Gamma_0^{\mathsf{H}} + \Gamma_{\mathrm{f}}\Gamma_{\mathrm{f}}^{\mathsf{H}})^{+}(\Gamma_0(\check{E}M)(t_0) - \Gamma_{\mathrm{f}}(\check{E}M)(t_{\mathrm{f}}))$$

   *and let $M$ be the unique solution of (4.22). Then, the integrated sensitivities are given by*

$$\int_{t_0}^{t_{\mathrm{f}}} y_\theta \mathrm{d}t = -\Xi^{\mathsf{H}}\tilde{\Gamma} + \int_{t_0}^{t_{\mathrm{f}}} \tilde{G} + M^{\mathsf{H}}\tilde{F}\mathrm{d}t. \tag{4.23}$$

$\triangleright$

*Proof.* The proof follows along the lines of the proof of Theorem 4.6 by setting $\tau- = \tau+ = t_{\mathrm{f}}$. $\square$

*Remark* 4.4. The case $\tau = t_0$ is not of much interest, since in that case

$$\int_{t_0}^{t_0} y_{\theta \mathrm{d}t} = 0.$$

Note, that also in this case, taking the limit in (4.12a) gives the correct result. Equation (4.12a) is turned into a homogeneous equation with only the trivial solution. ▷

As it has been remarked already in [SP02] for the ODE case, we cannot easily differentiate the integral output equation (4.23) with respect to either of the boundary points $t_0$ or $t_\mathrm{f}$, since this would involve differentiating $x$ with respect to either of these points as well.

Without loss of generality, let us choose $\tau = t_\mathrm{f}$. We can again let $\tau \to t_\mathrm{f}$ and take the limits of the Equations (4.18a) to (4.18b) and (4.19) in Theorem 4.7 for the open interval. This is feasible, since the adjoint solution $N$ is only needed in the integration of the inhomogeneity $\tilde{F}$, and we can insert the result of Equation (4.18c) into Equation (4.18b). After all, we obtain a solution, where $N(t_\mathrm{f})$ is replaced by $N(t_\mathrm{f}-)$.

**Corollary 4.9.** *Consider the original system (4.1) and (4.2) and the corresponding forward system (4.3). Let Assumptions 4.2 and 4.5 hold and define $\check{\Gamma}_0, \check{\Gamma}_f$ as in (4.6). Let $\tau$ be fixed to $\tau = t_\mathrm{f}$. Then, we have the following.*

1. *There exists a unique solution $N \in \mathcal{C}^1(\mathbb{I}, \mathbb{K}^{n\times r})$ of*

$$\check{E}\dot{N} = \check{A}N, \tag{4.24a}$$
$$-\check{\Gamma}_\mathrm{f}(C\mathcal{P}_z)^\mathsf{H}(t_\mathrm{f}) = \check{\Gamma}_0(\check{E}N)(t_0) + \check{\Gamma}_\mathrm{f}(\check{E}N)(t_\mathrm{f}), \tag{4.24b}$$

   *where $(\check{E}, \check{A})$ is defined as in (4.4).*

2. *Set*

$$\Xi_\tau := (\Gamma_0\Gamma_0^\mathsf{H} + \Gamma_\mathrm{f}\Gamma_\mathrm{f}^\mathsf{H})^+(\Gamma_0(\check{E}N)(t_0) + \Gamma_\mathrm{f}(-\check{E}N + (C\mathcal{P}_z)^\mathsf{H})(t_\mathrm{f})) \tag{4.25}$$

   *and let $N$ be the unique solution of (4.24). Then, the sensitivities are given by*

$$y_\theta(\tau) = (\tilde{G} - C\tilde{F}_a)(\tau) - \Xi_\tau^\mathsf{H}\tilde{\Gamma} + \int_{t_0}^{t_\mathrm{f}} N^\mathsf{H}\tilde{F}\mathrm{d}t. \tag{4.26}$$

▷

*Proof.* The proof is analogous to the proof of Theorem 4.7. First note, that the linear system (4.15) transforms to the linear system

$$\begin{bmatrix} -\Gamma_0^{\mathsf{H}} \\ \Gamma_{\mathrm{f}}^{\mathsf{H}} \end{bmatrix} \Xi_\tau = \begin{bmatrix} (E^{\mathsf{H}} N)(t_0) \\ (E^{\mathsf{H}} N - (C\mathcal{P}_z)^{\mathsf{H}})(t_{\mathrm{f}}) \end{bmatrix}$$

for the given quantities in Equations (4.24b) and (4.25). Hence, we obtain that (4.21) transforms to

$$\begin{aligned} & N^{\mathsf{H}} E x_\theta|_{t_0}^{t_{\mathrm{f}}} = (N^{\mathsf{H}} E x_\theta)(t_{\mathrm{f}}) - (N^{\mathsf{H}} E x_\theta)(t_0) \\ = {} & ((N^{\mathsf{H}} E - C\mathcal{P}_z) x_\theta)(t_{\mathrm{f}}) - (N^{\mathsf{H}} E x_\theta)(t_0) + (C\mathcal{P}_z x_\theta)(t_{\mathrm{f}}) \\ = {} & \Xi^{\mathsf{H}} \tilde{\Gamma} + (C\mathcal{P}_z x_\theta)(t_{\mathrm{f}}) \end{aligned}$$

by using (4.24b). Then, (4.26) follows along the lines of the proof of Theorem 4.7, by noting that all terms that included $\delta_\tau$ in the proof of Theorem 4.7 vanish in the context of the present theorem. □

*Remark* 4.5. The statement for $\tau = t_0$ is completely analogous. ▷

## 4.3. Weaker Assumptions on Time Differentiability

In the previous subsections we assumed that time derivatives of $x$ exist also in the kernel of $E(t)$. This is not feasible in all occasions. For instance, we have seen in Subsection 3.4.4 that for non-differentiable inputs and inhomogeneities the algebraic components of $x$ may not be differentiable for all possible values of $u$ and $f$.

Throughout this section we assume that the DAE (4.1a) is strangeness-free pointwise in $\theta$. If we want to apply the results from Sections 4.1 and 4.2 in this case, we have to rewrite the original DAE (4.1a) as

$$E \frac{\mathrm{d}}{\mathrm{d}t}(E^+ E x) = (A + E \frac{\mathrm{d}}{\mathrm{d}t}(E^+ E))x + f, \tag{4.27}$$

where now only $E^+ E x$ has to be differentiable with respect to $t$. Recall, that a solution $x(\cdot, \theta)$ of (4.27) lies in $\mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n)$.

### 4.3.1. Forward Sensitivities

For writing the sensitivity forward system, we need sufficient differentiability of $E, A, f, x$, and $E^+Ex$ with respect to $\theta$.

**Assumption 4.10.** *The matrix functions $E, A$ fulfill $E \in \mathcal{C}^2(\mathbb{I} \times \Theta, \mathbb{K}^{n\times n})$, $A \in \mathcal{C}^1(\mathbb{I} \times \Theta, \mathbb{K}^{n\times n})$. The inhomogeneity $f$ is continuous in the first argument, and continuously differentiable in the second argument. Moreover, the DAE (4.27) is strangeness-free and the characteristic quantities $n_d$ and $n_a$ of Hypothesis 3.3 are constant on $\mathbb{I} \times \Theta$.* ▷

Then we can formulate the following lemma.

**Lemma 4.11.** *Consider the parameter-dependent DAE (4.27) and let Assumption 4.10 hold. Let $(\hat{t}, \hat{\theta}) \in \mathbb{I} \times \Theta$ be given. Further, let pointwise isometric $U, V \in \mathcal{C}^1(\mathbb{A}_0, \mathbb{K}^{n\times n})$ on a sufficiently small subset $\mathbb{A}_0 \subseteq \mathbb{I} \times \Theta$ with $(\hat{t}, \hat{\theta}) \in \mathbb{A}_0$ be given and set $(\tilde{E}, \tilde{A}, \tilde{f}) = (U^H E V, U^H A V - U^H E \dot{V}, U^H f)$.*

*Then the solutions $x$ and $\tilde{x} = V^H x$ of (4.27) for the coefficients $(E, A, f)$ and $(\tilde{E}, \tilde{A}, \tilde{f})$ and their respective time derivatives are differentiable with respect to $\theta$. In particular, we have*

$$\frac{\partial^2}{\partial t \partial \theta}(E^+Ex) = \frac{\partial^2}{\partial \theta \partial t}(E^+Ex). \tag{4.28}$$

*Further, we have the relation*

$$\tilde{x}_\theta = V_\theta^H x + V^H x_\theta. \tag{4.29}$$

▷

*Proof.* Existence of the derivatives with respect to $\theta$ follow as in the proof of Lemma 4.1, by noting that the flow equations (3.22) are identical for systems in the form (4.1a) and (4.27), see Proposition 3.13.

Since, $x_d = E^+Ex$, we also obtain (4.28). Moreover, relation (4.29) follows by direct differentiation of $\tilde{x} = V^H x$. □

Lemma 4.11 shows, that we can safely differentiate (4.1b), (4.2) and (4.27) with respect to $\theta$ when Assumption 4.10 is fulfilled. In that case the forward sensitivity

system (4.3) is given by

$$E\frac{\partial^2}{\partial t\partial\theta}(E^+Ex) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x_\theta + \tilde{F}^{(2)}, \tag{4.30a}$$

$$y_\theta = Cx_\theta + \tilde{G}, \tag{4.30b}$$

$$\tilde{\Gamma} = \Gamma_0 x_\theta(t_0) + \Gamma_\mathrm{f} x_\theta(t_\mathrm{f}), \tag{4.30c}$$

where $\tilde{F}^{(2)} = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))_\theta x - E_\theta\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) + f_\theta$, $\tilde{G} \coloneqq g_\theta + C_\theta x$, and $\tilde{\Gamma} \coloneqq \gamma_\theta - (\Gamma_0)_\theta x(t_0) - (\Gamma_\mathrm{f})_\theta x(t_\mathrm{f})$. Note, that $\tilde{F}^{(2)}$, $\tilde{G}$, and $\tilde{\Gamma}$ depend on the solution $x(\theta)$ of the original system (4.27) for a given $\theta$, and except for $\tilde{F}^{(2)}$ they correspond to the respective quantities in (4.3).

The first equation (4.30a) is not a DAE anymore since the time derivative is not explicitly formed on $E^+Ex_\theta$ or $x_\theta$ and thus in this form not computable via existing numerical methods. The problem is, that the term in front of the time derivative in (4.30a)

$$(E^+Ex)_\theta = (E^+E)_\theta x + E^+Ex_\theta$$

cannot be separated into two terms dependent only on $x$ and $x_\theta$, respectively, which are separately differentiable with respect to time.

However, if we assume that

$$\ker(E^+E) \subseteq \ker(E^+E)_{\theta_i} \tag{4.31}$$

for $i = 1, \dots p$, we can reformulate the system. Condition (4.31) implies that

$$(E^+E)_\theta = (E^+E)_\theta E^+E.$$

Hence, it follows that (4.30a) is equivalent to

$$\begin{aligned} &E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex_\theta) \\ &= (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x_\theta + \tilde{F}^{(2)} - E\frac{\mathrm{d}}{\mathrm{d}t}((E^+E)_\theta E^+Ex) \\ &= (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x_\theta + \tilde{F}^{(2)} - E\frac{\mathrm{d}}{\mathrm{d}t}((E^+E)_\theta)E^+Ex - E(E^+E)_\theta\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex). \end{aligned} \tag{4.32}$$

Thus, we have proven the following lemma.

**Lemma 4.12.** *Let Assumption 4.10 and condition (4.31) hold. Then if for a fixed $\hat{\theta}$ there is a solution $x$ of (4.1b), (4.2) and (4.27), then $x_\theta$ solves (4.30), where (4.30a) can be equivalently replaced by (4.32).* ▷

*Proof.* The proof follows analogously to Lemma 4.1 by using Lemma 4.11 and the discussion preceding this lemma. □

If, however, condition (4.31) does not hold, then we can use the flow equations (3.22) directly. Note, that for any quantity in the flow representation (3.23), e. g., $x_\mathrm{d}$, we denote by $x_{\mathrm{d},\theta}$ the derivative of $x_\mathrm{d}$ with respect to $\theta$, whereas $x_{\theta,\mathrm{d}}$ denotes the projected quantity $P_z x_\theta$.

For the remainder of this section, we make the following assumption, which is analogous to Assumption 4.2.

**Assumption 4.13.** *For all $\theta \in \Theta$ it holds that*

1. *the DAE (4.27) is regular and strangeness-free;*
2. *the boundary coefficients (4.1b) satisfy*

$$\Gamma_0 E^+ E = \Gamma_0$$

*and*

$$\Gamma_\mathrm{f} E^+ E = \Gamma_\mathrm{f};$$

3. *there exists isometric $Z_\Gamma \in \mathbb{K}^{n \times n_d}$ such that*

$$\operatorname{rk} Z_\Gamma^\mathsf{H} \begin{bmatrix} \Gamma_0 & \Gamma_\mathrm{f} \end{bmatrix} = n_d = \operatorname{rk} \begin{bmatrix} \Gamma_0 & \Gamma_\mathrm{f} \end{bmatrix}$$

*and the boundary inhomogeneity $\gamma$ lies in $\operatorname{im} Z_\Gamma$;* ▷

**Lemma 4.14.** *Let Assumptions 4.10 and 4.13 hold, and consider the forward system (4.30) and the corresponding flow equations (3.22). Further, assume that the original system (4.27) is uniquely solvable in the sense of Theorem 3.15. Then, $x_\theta = x_{\mathrm{d},\theta} + x_{\mathrm{a},\theta}$ can be computed via the unique solution of*

$$\dot{x}_{\mathrm{d},\theta} = D_\mathrm{d} x_{\mathrm{d},\theta} + D_{\mathrm{d},\theta} x_\mathrm{d} + f_{\mathrm{d},\theta}, \tag{4.33a}$$

$$\tilde{\Gamma} + \Gamma_0\big((E^+E)_\theta x\big)(t_0) + \Gamma_\mathrm{f}\big((E^+E)_\theta x\big)(t_\mathrm{f}) = \Gamma_0 x_{\mathrm{d},\theta}(t_0) + \Gamma_\mathrm{f} x_{\mathrm{d},\theta}(t_\mathrm{f}), \tag{4.33b}$$

$$(P_z^\perp x_{\mathrm{d},\theta})(t_0) = (P_z^\perp (E^+E)_\theta x)(t_0), \tag{4.33c}$$

$$x_{\mathrm{a},\theta} = -D_\mathrm{a} x_{\mathrm{d},\theta} - D_{\mathrm{a},\theta} x_\mathrm{d} + f_{\mathrm{a},\theta}. \tag{4.33d}$$

▷

*Proof.* In the proof of Lemma 4.11 we already remarked, that the derivatives $x_{\mathrm{d},\theta}$ and $x_{\mathrm{a},\theta}$ exist and $x_\theta = x_{\mathrm{d},\theta} + x_{\mathrm{a},\theta}$ solves (4.30). By noting that

$$x_{\mathrm{d},\theta} = (E^+E)_\theta x + (E^+E)x_\theta = (E^+E)_\theta x + x_{\theta,\mathrm{d}}, \tag{4.34a}$$
$$x_{\mathrm{a},\theta} = -(E^+E)_\theta x + (I_n - E^+E)x_\theta = -(E^+E)_\theta x + x_{\theta,\mathrm{a}}, \tag{4.34b}$$

we conclude, that also (4.33b) and (4.33c) are fulfilled.

Now we show that (4.33) indeed is (uniquely) solvable by considering $x_{\mathrm{d},\theta}$ and $x_{\mathrm{a},\theta}$ as independent variables. It suffices to prove that Equations (4.33b) and (4.33c) uniquely fix the initial value $x_{\mathrm{d},\theta}(t_0)$.

By Assumption 4.13 and Theorem 3.15 we conclude that (4.33b) uniquely fixes $(T_2^{\mathsf{H}} x_{\mathrm{d},\theta})(t_0)$, thus also $(P_z x_{\mathrm{d},\theta})(t_0)$. Hence, together, with (4.33c) this proves unique solvability of (4.33a) and thus of (4.33d) and $x_\theta$. □

*Remark* 4.6. If condition (4.31) is fulfilled, then we can differentiate (4.34a) with respect to time and the system (4.33) can be reformulated in terms of the independent variables $x_{\theta,\mathrm{d}}$ and $x_{\theta,\mathrm{a}}$ which is the flow representation of the DAE (4.32). ▷

We present an example where condition (4.31) is fulfilled.

**Example 4.1.** Let a system $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ with

$$E(t,\theta) = \begin{bmatrix} E_{11}(t,\theta) & 0 \\ E_{21}(t,\theta) & E_{22}(t) \end{bmatrix}$$

with pointwise invertible $E_{11}(t,\theta) \in \mathbb{K}^{r\times r}$ be given. Then, the derivative $E_\theta$ is given by

$$E_\theta(t,\theta) = \begin{bmatrix} E_{11,\theta}(t,\theta) & 0 \\ E_{21,\theta}(t,\theta) & 0 \end{bmatrix}.$$

Let $x \in \ker(E^+E)(t,\theta) = \ker E(t,\theta)$ be given. Partitioning $x$ as $x = (x_1^T, x_2^T)^T$, we then obtain that $E_{11}(t,\theta)x_1 = 0$. Thus, omitting arguments,

$$(E^+E)_\theta x = E^+E_\theta x + E_\theta^+ E x = E^+ \begin{bmatrix} E_{11,\theta}E_{11}^{-1} & 0 \\ E_{21,\theta}E_{11}^- 1 & 0 \end{bmatrix} \begin{pmatrix} E_{11}x_1 \\ 0 \end{pmatrix} + 0 = 0.$$

Hence, condition (4.31) is fulfilled and the approach of Lemma 4.12 is applicable. ▷

An example, where condition (4.31) is not fulfilled is given as follows.

**Example 4.2.** Let the system $(E, A, B, f)$ be defined by the coefficients

$$E(t,\theta) = \begin{bmatrix} 1 & \theta \\ 0 & 0 \end{bmatrix}, \qquad A(t,\theta) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad f(t,\theta) = \begin{pmatrix} 0 \\ f_2(t) \end{pmatrix}. \tag{4.35}$$

We have

$$E^+(t,\theta) = \begin{bmatrix} 1 & 0 \\ \theta & 0 \end{bmatrix} \frac{1}{1+\theta^2}$$

and

$$\ker E = \operatorname{im} \begin{pmatrix} \theta \\ -1 \end{pmatrix}$$

Thus, in this case the kernel of

$$E_\theta(t,\theta) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

is not a superset of the kernel of $E(t,\theta)$ for any $(t,\theta) \in \mathbb{R} \times \mathbb{R}$. In particular, for $x \in \ker E = \ker E^+ E$ we have

$$(E^+E)_\theta x = E_\theta^+ E x + E^+ E_\theta x = \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix} \frac{1}{1+\theta^2} \alpha \begin{pmatrix} \theta \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ -\theta \end{pmatrix} \frac{\alpha}{1+\theta^2} \neq 0,\ \alpha \neq 0.$$

A parameterization of solutions of the system $(E, A, B, f)$ in the weak setting is given by

$$x(t,\theta) = \begin{pmatrix} x_1(t,\theta) \\ x_2(t,\theta) \end{pmatrix} = \begin{pmatrix} \theta f_2(t) + c(\theta) \\ -f_2(t) \end{pmatrix}.$$

Thus, $E^+Ex$ is differentiable with respect to time, even if $f_2$ is continuous only. However, $(E^+E)_\theta x$ is not differentiable with respect to time and neither of the approaches of Lemmas 4.1 and 4.12 are applicable.

On the other hand, assuming that $c(\theta)$ is continuously differentiable, the solution $x(t,\theta)$ *is* differentiable with respect to the parameters and the $x_\theta(t,\theta)$ is even continuous in time and parameters, if $f$ is continuous. ▷

### 4.3.2. Adjoint Sensitivities

The situation for the adjoint approach is not as bad as for the forward sensitivities. It is not necessary to make the assumption (4.31). This can be seen by inspecting the proof of Theorem 4.6 and in particular Equation (4.17), where we do a partial integration, and thus, the term $(E^+E)_\theta x$ does not need to be differentiable with respect to time.

We can state the following theorem, where we define $\mathcal{C}^1_{EE^+}(\mathbb{I},\mathbb{K}^n,\{\tau\})$ as the set of functions $\lambda$, for which $EE^+\lambda$ is a piecewise smooth function with $EE^+\lambda \in \mathcal{C}^1_{\mathrm{pw}}(\mathbb{I},\mathbb{K}^n,\{\tau\})$ and $\lambda \in \mathcal{C}^0_{\mathrm{pw}}(\mathbb{I},\mathbb{K}^n,\{\tau\})$.

**Theorem 4.15.** *Consider the original system (4.1b), (4.2) and (4.27) and the corresponding forward system (4.30). Let Assumptions 4.10 and 4.13 hold and define $\check{\Gamma}_0, \check{\Gamma}_\mathrm{f}$ as in (4.6). Let a fixed $\tau \in (t_0, t_\mathrm{f})$ be given. Then, we have the following.*

1. *There exists a unique solution $M \in \mathcal{C}^1_{E^+E}(\mathbb{I},\mathbb{K}^{n\times r},\{\tau\})$ of*

$$-E^\mathsf{H}\frac{\mathrm{d}}{\mathrm{d}t}(EE^+M) = (A + EE^+\dot{E})^\mathsf{H}M + \mathbb{1}_{[t_0,\tau)}C^\mathsf{H}, \tag{4.36a}$$

$$0 = \check{\Gamma}_0(E^\mathsf{H}M)(t_0) + \check{\Gamma}_\mathrm{f}(E^\mathsf{H}M)(t_\mathrm{f}), \tag{4.36b}$$

$$0 = E^\mathsf{H}(\tau)(M(\tau+) - M(\tau-)). \tag{4.36c}$$

2. *Set*

$$\Xi := \left(\Gamma_0\Gamma_0^\mathsf{H} + \Gamma_\mathrm{f}\Gamma_\mathrm{f}^\mathsf{H}\right)^+\left(-\Gamma_0(E^\mathsf{H}M)(t_0) + \Gamma_\mathrm{f}(E^\mathsf{H}M)(t_\mathrm{f})\right) \tag{4.37}$$

*and let $M$ be the unique solution of (4.36). Then, the integrated sensitivities are given by*

$$\int_{t_0}^{\tau} y_\theta \mathrm{d}t = -\Xi^\mathsf{H}\tilde{\Gamma} - M^\mathsf{H}E(E^+E)_\theta x|_{t_0}^{t_\mathrm{f}} + \int_{t_0}^{t_\mathrm{f}} \mathbb{1}_{[t_0,\tau)}(\tilde{G} - C(E^+E)_\theta x) + M^\mathsf{H}\tilde{F}^{(3)}\mathrm{d}t,$$

*where*

$$\tilde{F}^{(3)} := \tilde{F}^{(2)} - A(E^+E)_\theta x \tag{4.38}$$

*and $\tilde{F}^{(2)}, \tilde{\Gamma}$ are as in (4.30).* ▷

*Proof.* The existence and uniqueness of solutions follows analogously to the proof of Theorem 4.6 by noting that (4.36a) and (4.12a) share the same flow representation, see Proposition 3.13, because

$$EE^+\dot{E}E^+E + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E) = EE^+\dot{E}. \tag{4.39}$$

See also [KM11a] for the derivation of (4.39).

Thus, let us assume that $M$ is as solution of (4.36) and that $x$ solves (4.30). In that case we can get rid of time differentiation of the $(E^+Ex)_\theta$ tensor in (4.30a) and separate the $E^+Ex_\theta$ part by partial integration. Hence, we obtain

$$\begin{aligned}
0 &= \int_{t_0}^{\tau} M^{\mathsf{H}}\left(E\frac{\partial^2}{\partial t\partial\theta}(E^+Ex) - (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x_\theta - \tilde{F}^{(2)}\right)\mathrm{d}t \\
&= M^{\mathsf{H}}E(E^+Ex)_\theta|_{t_0}^{\tau} \\
&\quad - \int_{t_0}^{\tau}\left(\frac{\mathrm{d}}{\mathrm{d}t}(EE^+M)^{\mathsf{H}}E + M^{\mathsf{H}}EE^+\dot{E}E^+E + M^{\mathsf{H}}A + M^{\mathsf{H}}E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E)\right)x_\theta\mathrm{d}t \\
&\quad - \int_{t_0}^{\tau}\left(\frac{\mathrm{d}}{\mathrm{d}t}(EE^+M)^{\mathsf{H}}E + M^{\mathsf{H}}EE^+\dot{E}\right)(E^+E)_\theta x\mathrm{d}t \\
&\quad - \int_{t_0}^{\tau} M^{\mathsf{H}}\tilde{F}^{(2)}\mathrm{d}t.
\end{aligned} \tag{4.40}$$

Again, using (4.39) and that $M$ solves (4.36) we deduce that (4.40) is equivalent to

$$\begin{aligned}
0 &= M^{\mathsf{H}}E(E^+E)_\theta x|_{t_0}^{\tau} + M^{\mathsf{H}}Ex_\theta|_{t_0}^{\tau} \\
&\quad + \int_{t_0}^{\tau} Cx_\theta\mathrm{d}t \\
&\quad + \int_{t_0}^{\tau} M^{\mathsf{H}}\left(A + EE^+\dot{E} - EE^+\dot{E}\right)(E^+E)_\theta x + C(E^+E)_\theta x\mathrm{d}t \\
&\quad - \int_{t_0}^{\tau} M^{\mathsf{H}}\tilde{F}^{(2)}\mathrm{d}t.
\end{aligned}$$

Repeating these steps on the interval $(\tau, t_{\mathrm{f}}]$ and adding the results we obtain that

$$
\begin{aligned}
0 &= \int_{t_0}^{t_f} M^H \left( E \frac{\partial^2}{\partial t \partial \theta} (E^+ E x) - (A + E \frac{\mathrm{d}}{\mathrm{d}t}(E^+ E)) x_\theta - \tilde{F}^{(2)} \right) \mathrm{d}t \\
&= M^H E (E^+ E x)_\theta |_{t_0}^{t_f} + M^H E (E^+ E x)_\theta |_{\tau^-}^{\tau^+} \\
&\quad + \int_{t_0}^{\tau} C x_\theta \mathrm{d}t \\
&\quad - \int_{t_0}^{t_f} M^H \tilde{F}^{(2)} - M^H A (E^+ E)_\theta x - \mathbb{1}_{[t_0,\tau)} C (E^+ E)_\theta x \mathrm{d}t.
\end{aligned}
$$

Recall, that under Assumption 4.13 on the boundary coefficients we have existence of $\Xi$ such that

$$
\begin{bmatrix} -\Gamma_0^H \\ \Gamma_f^H \end{bmatrix} \Xi = \begin{bmatrix} (E^H M)(t_0) \\ (E^H M)(t_f) \end{bmatrix}
$$

and $\Xi$ given by formula (4.37) is a possible solution, see the proof of Theorem 4.6 for more details.

Then, for the boundary terms we compute

$$
M^H E (E^+ E x)_\theta |_{\tau^-}^{\tau^+} = \left( M^H(\tau+) - M^H(\tau-) \right) \left( E \left( E^+ E x \right)_\theta \right) (\tau) = 0
$$

by (4.36c) and

$$
M^H E (E^+ E x)_\theta |_{t_0}^{t_f} = M^H E x_\theta |_{t_0}^{t_f} + M^H E (E^+ E)_\theta x |_{t_0}^{t_f} = \Xi^H \tilde{\Gamma} + M^H E (E^+ E)_\theta x |_{t_0}^{t_f}
$$

analogously to the proof of Theorem 4.6.

Hence, in total we obtain that the sensitivities are given by

$$
\begin{aligned}
\int_{t_0}^{\tau} y_\theta \mathrm{d}t &= \int_{t_0}^{\tau} \tilde{G} \mathrm{d}t + \int_{t_0}^{\tau} C x_\theta \mathrm{d}t \qquad (4.41) \\
&= -\Xi^H \tilde{\Gamma} - M^H E (E^+ E)_\theta x |_{t_0}^{t_f} + \int_{t_0}^{t_f} \mathbb{1}_{[t_0,\tau)} (\tilde{G} - C (E^+ E)_\theta x) + M^H \tilde{F}^{(3)} \mathrm{d}t. \square
\end{aligned}
$$

**Theorem 4.16.** *Consider the original system (4.1b), (4.2) and (4.27) and the corresponding forward system (4.30). Let Assumptions 4.10 and 4.13 hold and define $\check{\Gamma}_0, \check{\Gamma}_f$ as in (4.6). Let a fixed $\tau \in (t_0, t_f)$ be given. Then, we have the following.*

1. *There exists a unique solution* $N \in \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^{n\times r}, \{\tau\})$ *of*

$$-E^{\mathsf{H}}\frac{\mathrm{d}}{\mathrm{d}t}(EE^+N) = (A + EE^+\dot{E})^{\mathsf{H}}N + \delta_\tau(C\mathcal{P}_z)^{\mathsf{H}} \tag{4.42a}$$

$$0 = \check{\Gamma}_0(E^{\mathsf{H}}N)(t_0) + \check{\Gamma}_{\mathrm{f}}(E^{\mathsf{H}}N)(t_{\mathrm{f}}), \tag{4.42b}$$

$$-(C\mathcal{P}_z)^{\mathsf{H}}(\tau) = E^{\mathsf{H}}(\tau)(N(\tau+) - N(\tau-)). \tag{4.42c}$$

2. *Set*

$$\Xi_\tau := \left(\Gamma_0\Gamma_0^{\mathsf{H}} + \Gamma_{\mathrm{f}}\Gamma_{\mathrm{f}}^{\mathsf{H}}\right)^+\left(-\Gamma_0(E^{\mathsf{H}}N)(t_0) + \Gamma_{\mathrm{f}}(E^{\mathsf{H}}N)(t_{\mathrm{f}})\right). \tag{4.43}$$

*Then the sensitivities* $y_\theta(\tau)$ *are given by*

$$y_\theta(\tau) = \left(\tilde{G} + C(\mathcal{P}_z)_\theta P_z x - Cf_{\mathrm{a},\theta}\right)(\tau) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} - N^{\mathsf{H}}E(E^+E)_\theta x|_{t_0}^{t_{\mathrm{f}}} + \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}}\tilde{F}^{(3)}\mathrm{d}t. \tag{4.44}$$

*with* $\tilde{F}^{(3)}$ *given as in* (4.38). ▷

*Proof.* For the proof of this theorem we combine ideas from the proofs of Theorems 4.7 and 4.15. The existence and uniqueness discussion of Statement 1 follows analogously to the proofs of Theorems 4.7 and 4.15.

For Statement 2 let $N$ be a solution of (4.42) and let $x$ solve (4.30). Then,

$$\begin{aligned}
0 &= \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}}\left(E\frac{\partial^2}{\partial t\partial\theta}(E^+Ex) - (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x_\theta - \tilde{F}^{(2)}\right)\mathrm{d}t \\
&= N^{\mathsf{H}}E(E^+Ex)_\theta|_{t_0}^{t_{\mathrm{f}}} \\
&\quad - \int_{t_0}^{t_{\mathrm{f}}}\left(\frac{\mathrm{d}}{\mathrm{d}t}(EE^+N)^{\mathsf{H}}E + N^{\mathsf{H}}EE^+\dot{E}E^+E + N^{\mathsf{H}}A + N^{\mathsf{H}}E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E)\right)x_\theta\mathrm{d}t \\
&\quad - \int_{t_0}^{t_{\mathrm{f}}}\left(\frac{\mathrm{d}}{\mathrm{d}t}(EE^+N)^{\mathsf{H}}E + N^{\mathsf{H}}EE^+\dot{E}\right)(E^+E)_\theta x\mathrm{d}t \\
&\quad - \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}}\tilde{F}^{(2)}\mathrm{d}t.
\end{aligned} \tag{4.45}$$

Again, using the fact (4.39) and by using that $N$ solves (4.42) we deduce that

relation (4.45) is equivalent to

$$\begin{aligned}
0 &= N^{\mathsf{H}} E (E^{+} E x)_{\theta} |_{t_0}^{t_{\mathrm{f}}} \\
&\quad + \int_{t_0}^{t_{\mathrm{f}}} C \mathcal{P}_z x_{\theta} \delta_{\tau} \mathrm{d}t \\
&\quad + \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}} \left( A + E E^{+} \dot{E} - E E^{+} \dot{E} \right) (E^{+} E)_{\theta} x + C \mathcal{P}_z (E^{+} E)_{\theta} x \delta_{\tau} \mathrm{d}t \\
&\quad - \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}} \tilde{F}^{(2)} \mathrm{d}t \\
&= N^{\mathsf{H}} E (E^{+} E x)_{\theta} |_{t_0}^{t_{\mathrm{f}}} \\
&\quad + (C \mathcal{P}_z x_{\theta})(\tau) \\
&\quad + \left( C \mathcal{P}_z (E^{+} E)_{\theta} x \right)(\tau) - \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}} \tilde{F}^{(2)} - N^{\mathsf{H}} A (E^{+} E)_{\theta} x \mathrm{d}t.
\end{aligned}$$

Recall, that under the Assumption 4.13 on the boundary coefficients we have existence of $\Xi_\tau$ such that

$$\begin{bmatrix} -\Gamma_0^{\mathsf{H}} \\ \Gamma_{\mathrm{f}}^{\mathsf{H}} \end{bmatrix} \Xi_\tau = \begin{bmatrix} (E^{\mathsf{H}} N)(t_0) \\ (E^{\mathsf{H}} N)(t_{\mathrm{f}}) \end{bmatrix}$$

and $\Xi_\tau$ given by formula (4.43) is a possible solution. See the proof of Theorem 4.6 for more details.

Then, for the boundary term we compute

$$N^{\mathsf{H}} E (E^{+} E x)_{\theta} |_{t_0}^{t_{\mathrm{f}}} = N^{\mathsf{H}} E x_{\theta} |_{t_0}^{t_{\mathrm{f}}} + N^{\mathsf{H}} E (E^{+} E)_{\theta} x |_{t_0}^{t_{\mathrm{f}}} = \Xi_\tau^{\mathsf{H}} \tilde{\Gamma} + N^{\mathsf{H}} E (E^{+} E)_{\theta} x |_{t_0}^{t_{\mathrm{f}}}.$$

Note, that using the notation from (4.34) we obtain

$$\begin{aligned}
x_\theta &= x_{\mathrm{d},\theta} + x_{\mathrm{a},\theta} = \left( \mathcal{P}_z x_{\mathrm{d}} - f_{\mathrm{a}} \right)_\theta = \mathcal{P}_z P_z x_\theta + (\mathcal{P}_z P_z)_\theta x - f_{\mathrm{a},\theta} \\
&= \mathcal{P}_z x_\theta + (\mathcal{P}_z)_\theta x - f_{\mathrm{a},\theta}.
\end{aligned}$$

Hence, in total we obtain that the sensitivities $y_\theta$ at time $\tau$ are given by

$$\begin{aligned}
y_\theta(\tau) &= (\tilde{G} + Cx_\theta)(\tau) \\
&= \left(\tilde{G} + C\mathcal{P}_z x_\theta + C(\mathcal{P}_z)_\theta x - Cf_{\mathrm{a},\theta}\right)(\tau) \\
&= \left(\tilde{G} + C(\mathcal{P}_z)_\theta x - Cf_{\mathrm{a},\theta}\right)(\tau) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} - N^{\mathsf{H}}E(E^+E)_\theta x|_{t_0}^{t_\mathrm{f}} \\
&\quad - \left(C\mathcal{P}_z(E^+E)_\theta x\right)(\tau) + \int_{t_0}^{t_\mathrm{f}} N^{\mathsf{H}}\tilde{F}^{(2)} - N^{\mathsf{H}}A(E^+E)_\theta x \mathrm{d}t \\
&= \left(\tilde{G} + C(\mathcal{P}_z)_\theta P_z x - Cf_{\mathrm{a},\theta}\right)(\tau) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} - N^{\mathsf{H}}E(E^+E)_\theta x|_{t_0}^{t_\mathrm{f}} + \int_{t_0}^{t_\mathrm{f}} N^{\mathsf{H}}\tilde{F}^{(3)}\mathrm{d}t. \square
\end{aligned}$$

*Remark* 4.7. Comparing the result of Theorem 4.16 with the result of Theorem 4.7 we note the following.

1. The approach in Theorem 4.16 does not require time differentiability of the algebraic variables for both, the original system (4.30) and the adjoint system (4.42).
2. In the formula for the sensitivities (4.41) we need the quantity $f_{\mathrm{a},\theta}$ instead of $\tilde{F}_\mathrm{a}$ in (4.20). This is because the equivalent of $\tilde{F}_\mathrm{a}$ is not available in the former case as we cannot use rewrite the forward system is in the form (4.30).
3. However, we could have used the quantity $f_{\mathrm{a},\theta}$ in formula (4.20) instead of $\tilde{F}_\mathrm{a}$.
4. In particular, the approach of Theorem 4.16 also works in the case where (4.31) is not fulfilled. ▷

**Example 4.3.** Let the system of Example 4.2 be given by the coefficients (4.35). We impose the initial condition

$$E(t_0,\theta)x(t_0) = \begin{pmatrix} x_0(\theta) \\ 0 \end{pmatrix}, \quad x_0 \in \mathcal{C}^1(\mathbb{R}^p, \mathbb{R})$$

which translates to the boundary coefficients

$$\Gamma_0(\theta) = E(t_0,\theta), \qquad \Gamma_\mathrm{f}(\theta) = 0, \qquad \gamma(\theta) = \begin{pmatrix} x_0(\theta) \\ 0 \end{pmatrix}.$$

Note, that

$$T_2(t,\theta) = \begin{bmatrix} 1 \\ \theta \end{bmatrix} \frac{1}{1+\theta^2}.$$

Thus, the corresponding adjoint boundary coefficients are given according to (4.6) by

$$\check{\Gamma}_0 = 0, \qquad \check{\Gamma}_\mathrm{f} = \begin{bmatrix} 1 & \theta \\ 0 & 0 \end{bmatrix} \frac{1}{1+\theta^2}.$$

Further we have, omitting arguments,

$$P_z = E^+E = \begin{bmatrix} 1 & \theta \\ \theta & \theta^2 \end{bmatrix} \frac{1}{1+\theta^2}, \qquad \check{P}_z = EE^+ = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$
$$P_z^\perp = I_2 - P_z \begin{bmatrix} \theta^2 & -\theta \\ -\theta & 1 \end{bmatrix} \frac{1}{1+\theta^2}, \qquad \check{P}_z^\perp = I_2 - \check{P}_z = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, for the quantities $D_\mathrm{a}$ and $\mathcal{P}_z$ in (3.26) we obtain that

$$D_\mathrm{a}(t,\theta) = \begin{bmatrix} 0 & -\theta \\ 0 & 1 \end{bmatrix}, \qquad \mathcal{P}_z = \begin{bmatrix} 1 & \theta \\ 0 & 0 \end{bmatrix} P_z = E.$$

Hence, for a fixed $\tau \in (t_0, t_\mathrm{f})$ and setting $C = I_2$, the adjoint boundary value problem (4.42) is given by

$$\begin{bmatrix} -1 & 0 \\ -\theta & 0 \end{bmatrix} \begin{bmatrix} \dot{N}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} + \delta_\tau \begin{bmatrix} 1 & 0 \\ \theta & 0 \end{bmatrix},$$
$$0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N_1(t_\mathrm{f}) \\ N_2(t_\mathrm{f}) \end{bmatrix},$$
$$\begin{bmatrix} -1 & 0 \\ -\theta & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \theta & 0 \end{bmatrix} \begin{bmatrix} N_1(\tau+) - N_1(\tau-) \\ N_2(\tau+) - N_2(\tau-) \end{bmatrix}.$$

It has the unique solution

$$N_1(t,\theta) = \begin{bmatrix} \mathbb{1}_{[t_0,\tau]}(t) & 0 \end{bmatrix}, \qquad N_2(t,\theta) = 0.$$

For the computation of the sensitivities (4.44) we note that, omitting arguments,

$$\Xi_\tau = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{G} = 0, \quad \tilde{\Gamma} = \begin{bmatrix} x_{0,\theta}(\theta) \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \theta f_2(t_0) + x_0(\theta) \\ -f_2(t_0) \end{pmatrix},$$
$$(P_z)_\theta = (E^+E)_\theta = \begin{bmatrix} -2\theta & 1-\theta^2 \\ 1-\theta^2 & 2\theta \end{bmatrix} \frac{1}{(1+\theta^2)^2}, \quad \tilde{F}^{(2)} = 0,$$
$$\tilde{F}^{(3)} = -\begin{bmatrix} 0 \\ (1-\theta^2)\theta f_2 + (1-\theta^2)x_0(\theta) - 2\theta f_2 \end{bmatrix} \frac{1}{(1+\theta^2)^2},$$
$$(\mathcal{P}_z)_\theta = E_\theta = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad f_{\mathrm{a},\theta} = \begin{bmatrix} -f_2 \\ 0 \end{bmatrix}.$$

Thus, we have

$$y_\theta(\tau) = 0 + \begin{bmatrix} \frac{\theta}{1+\theta^2} x_0(\theta) \\ 0 \end{bmatrix} + \begin{bmatrix} f_2(\tau) \\ 0 \end{bmatrix} + \begin{bmatrix} x_{0,\theta}(\theta) + f_2(t_0) \\ 0 \end{bmatrix}$$
$$+ \begin{bmatrix} -f_2(t_0) - \frac{\theta}{1+\theta^2} x_0(\theta) \\ 0 \end{bmatrix} + \int_{t_0}^{t_\mathrm{f}} 0 \mathrm{d}t = \begin{bmatrix} f_2(\tau) + x_{0,\theta}(\theta) \\ 0 \end{bmatrix} \tag{4.46}$$

as expected. Note, that we did not need any time derivatives of $f_2$. ▷

Examples with non-trivial boundary conditions are presented in Section 4.5. The analogous result of Corollary 4.9 for values $\tau$ on the boundary of $[t_0, t_\mathrm{f}]$ is given as follows.

**Corollary 4.17.** *Consider the original system (4.1b), (4.2) and (4.27) and the corresponding forward system (4.30). Let Assumptions 4.10 and 4.13 hold and define $\check{\Gamma}_0, \check{\Gamma}_\mathrm{f}$ as in (4.6). Let $\tau$ be fixed to $\tau = t_\mathrm{f}$. Then, we have the following.*

1. *There exists a unique solution $N \in \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^{n\times r})$ of*

$$-E^\mathsf{H} \frac{\mathrm{d}}{\mathrm{d}t}(EE^+N) = (A + EE^+\dot{E})^\mathsf{H} N \tag{4.47a}$$
$$\check{\Gamma}_\mathrm{f}(C\mathcal{P}_z)^\mathsf{H}(t_\mathrm{f}) = \check{\Gamma}_0(E^\mathsf{H}N)(t_0) + \check{\Gamma}_\mathrm{f}(E^\mathsf{H}N)(t_\mathrm{f}). \tag{4.47b}$$

2. *Set*

$$\Xi_\tau := \left(\Gamma_0\Gamma_0^{\mathsf{H}} + \Gamma_{\mathrm{f}}\Gamma_{\mathrm{f}}^{\mathsf{H}}\right)^{+}\left(\Gamma_0(\check{E}N)(t_0) + \Gamma_{\mathrm{f}}(-\check{E}N + (C\mathcal{P}_z)^{\mathsf{H}})(t_{\mathrm{f}})\right) \tag{4.48}$$

*and let $N$ be the unique solution of (4.47). Then the sensitivities $y_\theta(\tau)$ are given by*

$$y_\theta(t_{\mathrm{f}}) = \left(\tilde{G} + C(\mathcal{P}_z)_\theta x - Cf_{\mathrm{a},\theta}\right)(t_{\mathrm{f}}) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} - N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} + \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}}\tilde{F}^{(3)}\mathrm{d}t \tag{4.49}$$

*with $\tilde{F}^{(3)}$ given as in (4.38).* ▷

*Proof.* The proof is analogous to the proofs of Corollary 4.9 and Theorem 4.16. Note again, that the linear system (4.15) transforms to the linear system

$$\begin{bmatrix} -\Gamma_0^{\mathsf{H}} \\ \Gamma_{\mathrm{f}}^{\mathsf{H}} \end{bmatrix} \Xi_\tau = \begin{bmatrix} (E^{\mathsf{H}}N)(t_0) \\ (E^{\mathsf{H}}N - (C\mathcal{P}_z)^{\mathsf{H}})(t_{\mathrm{f}}) \end{bmatrix}$$

for the given quantities in Equations (4.47b) and (4.48). Hence, we obtain that (4.21) transforms to

$$\begin{aligned} & N^{\mathsf{H}}E(E^{+}Ex)_\theta|_{t_0}^{t_{\mathrm{f}}} \\ &= N^{\mathsf{H}}Ex_\theta|_{t_0}^{t_{\mathrm{f}}} + N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} \\ &= (N^{\mathsf{H}}Ex_\theta)(t_{\mathrm{f}}) - (N^{\mathsf{H}}Ex_\theta)(t_0) + N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} \\ &= ((N^{\mathsf{H}}E - C\mathcal{P}_z)x_\theta)(t_{\mathrm{f}}) - (N^{\mathsf{H}}Ex_\theta)(t_0) + (C\mathcal{P}_z x_\theta)(t_{\mathrm{f}}) + N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} \\ &= \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} + (C\mathcal{P}_z x_\theta)(t_{\mathrm{f}}) + N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} \end{aligned}$$

by using (4.47b). Then, by noting that all terms that included $\delta_\tau$ in the proof of Theorem 4.16 vanish in the context of this theorem, we obtain

$$\begin{aligned} y_\theta(t_{\mathrm{f}}) &= \left(\tilde{G} + C\mathcal{P}_z x_\theta + C(\mathcal{P}_z)_\theta x - Cf_{\mathrm{a},\theta}\right)(t_{\mathrm{f}}) \\ &= \left(\tilde{G} + C(\mathcal{P}_z)_\theta x - Cf_{\mathrm{a},\theta}\right)(t_{\mathrm{f}}) - \Xi_\tau^{\mathsf{H}}\tilde{\Gamma} - N^{\mathsf{H}}E(E^{+}E)_\theta x|_{t_0}^{t_{\mathrm{f}}} + \int_{t_0}^{t_{\mathrm{f}}} N^{\mathsf{H}}\tilde{F}^{(3)}\mathrm{d}t. \end{aligned}$$ □

*Remark* 4.8. Note, that the formulas (4.44) and (4.49) slightly differ in addition to the evaluation point, which is $\tau \in (t_0, t_\mathrm{f})$ and $\tau = t_\mathrm{f}$, respectively. Equation (4.44) contains the term $(C\mathcal{P}_z P_z x)(\tau)$ whereas the respective term in (4.49) reads $(C\mathcal{P}_z x)(t_\mathrm{f})$. Moreover, the result of Corollary 4.17 can be analogously stated at the initial time with $\tau = t_0$. ▷

**Example 4.4** (Example 4.3 revisited)**.** We revisit Example 4.3 and compute the sensitivities at $\tau = t_\mathrm{f}$.

Setting $C = I_2$, the adjoint boundary value problem (4.47) is given by

$$\begin{bmatrix} -1 & 0 \\ -\theta & 0 \end{bmatrix} \begin{bmatrix} \dot{N}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N_1(t_\mathrm{f}) \\ N_2(t_\mathrm{f}) \end{bmatrix}.$$

It has the unique solution

$$N_1(t,\theta) = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad N_2(t,\theta) = 0.$$

For the computation of the sensitivities (4.49) we note that (4.48) is given by

$$\Xi_\tau = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus, we have

$$\begin{aligned} y_\theta(t_\mathrm{f}) = 0 &+ \begin{bmatrix} -f_2(t_\mathrm{f}) \\ 0 \end{bmatrix} + \begin{bmatrix} f_2(t_\mathrm{f}) \\ 0 \end{bmatrix} + \begin{bmatrix} x_{0,\theta}(\theta) + f_2(t_0) \\ 0 \end{bmatrix} \\ &+ \begin{bmatrix} f_2(t_\mathrm{f}) + \frac{\theta}{1+\theta^2} x_0(\theta) \\ 0 \end{bmatrix} + \begin{bmatrix} -f_2(t_0) - \frac{\theta}{1+\theta^2} x_0(\theta) \\ 0 \end{bmatrix} + \int_{t_0}^{t_\mathrm{f}} 0 \mathrm{d}t \\ &= \begin{bmatrix} f_2(t_\mathrm{f}) + x_{0,\theta}(\theta) \\ 0 \end{bmatrix} \end{aligned} \tag{4.50}$$

as expected. Note, that the resulting value of $y_\theta(t_\mathrm{f})$ is the limit of $y_\theta(\tau)$ for $\tau \to t_\mathrm{f}$. The respective quantities in the computation of (4.46) and (4.50) differ though. ▷

## 4.4. Summary and Comparison of the Approaches

Let us briefly summarize the results of Sections 4.1 to 4.3. We have seen several possibilities for computing the sensitivities of the system (4.1) or its weaker version (4.1b), (4.2) and (4.27). Depending on which system we consider and whether we choose to use the forward system approach or the adjoint approach, different assumptions are made. Also, certain terms may change depending on the method we choose. Some of those differences are mentioned in Remarks 4.1, 4.3 and 4.6 to 4.8. Further, note the following.

1. The adjoint equations (4.12), (4.18), (4.24), (4.36), (4.42) and (4.47) do *not* depend on the number of parameters $p$, however, a new solution has to be computed for every timepoint $\tau$ of interest, as their respective boundary conditions change.

   In contrast, the number of variables in the forward systems (4.3) and (4.30) grow linearly with the number of parameters $p$, but solving the forward system, immediately gives access to the sensitivity values at every timepoint of the solution.

   Hence, regarding this point, one should favor the adjoint approach, when, roughly, the number of parameters is higher than the number of sensitivity evaluation points.

2. It is not always favorable to use the approaches for the weaker formulation of the forward system (4.30). If condition (4.31) does not hold, we can only use the forward system approach based on the flow formulation, see Lemma 4.14. However, if the system coefficients and the inhomogeneity in (4.30) are smooth enough, then we can reformulate the problem in the stronger form (4.3), for which the forward problem can be solved without condition (4.31).

3. All approaches require that the data is sufficiently smooth with respect to the parameters $\theta$. However, the different approaches partially require derivatives of different quantities. For example, the forward approach based on the flow formulation in Lemma 4.14 in particular requires derivatives of $D_\mathrm{d}$, $D_\mathrm{a}$ and thus also of $E^+$, while the forward approach (4.3) only needs derivatives of the original data $E, A, f$. Derivatives of $E^+, D_\mathrm{a}$ and

Table 4.1.: Comparison of selected assumptions of the different theorems and lemmas for the computation of the sensitivities. The first block describes hard assumptions, that are necessary for the theory to work. The second block are soft assumptions in terms of usefulness for a numerical method, where $n_\tau$ denotes the number of evaluation points of the sensitivity function $y(\tau)$ and $p$ is the number of parameters.

| Assumption | L. 4.1 | L. 4.12 | L. 4.14 | Th. 4.7 | Th. 4.15 |
|---|---|---|---|---|---|
| $\dot{x}$ exists | ✓ | ✘ | ✘ | ✓ | ✘ |
| condition (4.31) holds | ✘ | ✓ | ✘ | ✘ | ✘ |
| compute $(E^+)_\theta$ | ✘ | ✘ | ✓ | ✘ | ✘ |
| compute $D_{\mathrm{a},\theta}$ | ✘ | ✘ | ✓ | ✘ | ✘ |
| $n_\tau > p$ | ✓ | ✓ | ✓ | ✘ | ✘ |
| $n_\tau < p$ | ✘ | ✘ | ✘ | ✓ | ✓ |

derivatives of all further quantities are available in terms of the original data $E, A, f$ through Lemma 3.2. This is computationally hard though as those formulas have to be applied at every single timepoint of the integration.

4. For strangeness-free systems all statements hold globally on the whole set $\mathbb{I} \times \Theta$, if the respective assumptions 4.2, 4.5, 4.10 and 4.13 are fulfilled. In particular, the approaches do not rely on existence of the global version of the local smooth full rank decomposition of Lemma 3.26. An exception is Lemma 4.1 for higher index systems, where we explicitly require, that the reduced system (3.14) is sufficiently smooth globally on $\mathbb{I} \times \Theta$.

The assumptions are also summarized in Table 4.1.

## 4.5. Application to Optimal Control Problems

In this section, we apply the results of Section 4.3 to boundary values coming from a parameter-dependent optimal control problem.

We consider sensitivities of parameter-dependent optimal control problems, which pointwise in the parameter $\theta$ correspond to the optimal control problem (3.38). Hence, we look for the sensitivities of solutions of the optimization problem

$$\begin{cases} \min\limits_{u\in\mathcal{C}^0(\mathbb{I},\mathbb{K}^m)} & x(t_\mathrm{f},\theta)^\mathsf{H} K x(t_\mathrm{f},\theta) + \displaystyle\int_{t_0}^{t_\mathrm{f}} \begin{pmatrix} x(t,\theta) \\ u(t,\theta) \end{pmatrix}^\mathsf{H} \begin{bmatrix} Q(t,\theta) & S(t,\theta) \\ S^\mathsf{H}(t,\theta) & R(t,\theta) \end{bmatrix} \begin{pmatrix} x(t,\theta) \\ u(t,\theta) \end{pmatrix} \mathrm{d}t \\ \text{s.t.} & \big(E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex)\big)(t,\theta) = \big(Ax+Bu+f\big)(t,\theta), \\ & x(t_0,\theta) = x_0(\theta), \end{cases} \tag{4.51}$$

where we assume, that the following assumptions are satisfied.

**Assumption 4.18.**

1. *Assume that the system $(E(\cdot,\theta), A(\cdot,\theta), B(\cdot,\theta), f(\cdot,\theta)) \in \Sigma_{m,n}(\mathbb{K})$ is strangeness-free pointwise in $\theta$, and all the rank conditions of Hypothesis 3.3 are constant.*

2. *Assume, that also the system of necessary conditions (3.40) is strangeness-free pointwise in $\theta$, uniquely solvable, and all characteristic quantities stay constant for all $(t,\theta) \in \mathbb{I}\times\Theta$. The former is fulfilled, if and only if the assumptions of Theorem 3.20 are fulfilled.*

3. *Assume, that the final weight matrix $K$ fulfills $KE^+E = K$ and thus due to its self-adjointness also $E^+EK = K$.* ▷

The necessary conditions of the optimal control problem are given pointwise for every parameter $\theta \in \Theta$ by Equation (3.40), i. e., we have

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))x + Bu + f, \tag{4.52a}$$

$$-E^\mathsf{H}\frac{\mathrm{d}}{\mathrm{d}t}(EE^+\lambda) = -Qx - Su + (A + EE^+\dot{E})^\mathsf{H}\lambda, \tag{4.52b}$$

$$0 = -S^\mathsf{H}x - Ru + B^\mathsf{H}\lambda, \tag{4.52c}$$

where by setting $z = (\lambda^T, x^T, u^T)^T$ the boundary condition is given by

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & (E^+E)(t_0) & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{=:\Gamma_0} z(t_0) + \underbrace{\begin{bmatrix} (EE^+)(t_f) & (E^+)^H(t_f)K & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{=:\Gamma_f} z(t_f) = \underbrace{\begin{pmatrix} 0 \\ x_0 \\ 0 \end{pmatrix}}_{=:\gamma}. \quad (4.53)$$

Using relation (4.39), the system (4.52) can be written as

$$\mathcal{E}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^+\mathcal{E}z) = \left(\mathcal{A} + \mathcal{E}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^+\mathcal{E})\right)z + \begin{pmatrix} f \\ 0 \\ 0 \end{pmatrix}$$

with

$$\mathcal{E} = \begin{bmatrix} 0 & E & 0 \\ -E^H & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{A} = \begin{bmatrix} 0 & A & B \\ (A+\dot{E})^H & -Q & -S \\ B^H & -S^H & -R \end{bmatrix}.$$

Noting that $\mathcal{E}^H = -\mathcal{E}$ and $(\mathcal{A} + \dot{\mathcal{E}})^H = \mathcal{A}$, we conclude by Remark 3.6 that the system (4.52) is self-adjoint according to Definition 3.8 pointwise in $\theta$.

In the following, we derive the boundary conditions of the adjoint problem as in Lemma 4.3 for the boundary coefficients (4.53). A possible choice for the selector matrices of Hypothesis 3.3 is given by

$$\mathcal{Z}_1 = \begin{bmatrix} Z_1 & 0 \\ 0 & -T_2 \\ 0 & 0 \end{bmatrix}, \qquad \mathcal{T}_2 = \begin{bmatrix} Z_1 & 0 \\ 0 & T_2 \\ 0 & 0 \end{bmatrix},$$

where $Z_1$ and $T_2$ are the respective quantities of Hypothesis 3.3 for the system (4.52a).

Then, by setting

$$Z_\Gamma = \begin{bmatrix} Z_1(t_f) & 0 \\ 0 & T_2(t_0) \\ 0 & 0 \end{bmatrix},$$

and recalling Equations (4.5) and (4.6) we obtain that

$$\operatorname{im}\begin{bmatrix}\check{\Gamma}_{0,11}^{\mathsf{H}}\\ \check{\Gamma}_{\mathrm{f},11}^{\mathsf{H}}\end{bmatrix} = \ker Z_\Gamma^{\mathsf{H}}\begin{bmatrix}-\Gamma_0\mathcal{T}_2(t_0) & \Gamma_\mathrm{f}\mathcal{T}_2(t_\mathrm{f})\end{bmatrix} \tag{4.54a}$$

$$= \ker\begin{bmatrix}0 & 0 & I_{n_d} & E_{11}^{-\mathsf{H}}(t_\mathrm{f})K_{11}\\ 0 & -I_{n_d} & 0 & 0\end{bmatrix} \tag{4.54b}$$

$$= \operatorname{im}\begin{bmatrix}0 & -E_{11}^{-\mathsf{H}}(t_\mathrm{f})\\ 0 & 0\\ -(E_{11}^{-\mathsf{H}}K_{11}E_{11}^{-1})(t_\mathrm{f}) & 0\\ E_{11}^{-1}(t_\mathrm{f}) & 0\end{bmatrix}. \tag{4.54c}$$

There is certain freedom in the actual representation of the spaces in (4.54) as the image representation in (4.54c) of (4.54b) is invariant under transformations from the right. We chose this particular representation for simplifying the upcoming result.

We then deduce that

$$\check{\Gamma}_0 = Z_\Gamma\check{\Gamma}_{0,11}\mathcal{T}_2^{\mathsf{H}}(t_0) = \begin{bmatrix}0 & 0 & 0\\ -E^+(t_0) & 0 & 0\\ 0 & 0 & 0\end{bmatrix},$$

$$\check{\Gamma}_\mathrm{f} = Z_\Gamma\check{\Gamma}_{\mathrm{f},11}\mathcal{T}_2^{\mathsf{H}}(t_\mathrm{f}) = \begin{bmatrix}-((E^+)^{\mathsf{H}}KE^+)(t_\mathrm{f}) & (E^+)^{\mathsf{H}}(t_\mathrm{f}) & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix},$$

and consequently

$$\begin{aligned}\check{\Gamma}_0\mathcal{E}^{\mathsf{H}}(t_0) &= \begin{bmatrix}0 & 0 & 0\\ 0 & (E^+E)(t_0) & 0\\ 0 & 0 & 0\end{bmatrix} = \Gamma_0,\\ \check{\Gamma}_\mathrm{f}\mathcal{E}^{\mathsf{H}}(t_\mathrm{f}) &= \begin{bmatrix}(EE^+)(t_\mathrm{f}) & E^{+\mathsf{H}}(t_\mathrm{f})K & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} = \Gamma_\mathrm{f}.\end{aligned} \tag{4.55}$$

Hence, also the boundary conditions can be considered self-adjoint, in the sense that the boundary conditions of the adjoint system (4.7a) correspond to the original boundary conditions (4.55). This relation was also mentioned in [BL05] for a different index notion.

Thus, the adjoint system of, e. g., Theorem 4.16 is given by

$$\mathcal{E}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^{+}\mathcal{E}N) = \left(\mathcal{A} + \mathcal{E}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^{+}\mathcal{E})\right)N + (C\mathcal{P}_z)^{\mathsf{H}}\delta_\tau \tag{4.56a}$$

$$\mathcal{E}(\tau)(N(\tau+) - N(\tau-)) = (C\mathcal{P}_z)^{\mathsf{H}}(\tau) \tag{4.56b}$$

$$0 = \Gamma_0 N(t_0) + \Gamma_{\mathrm{f}} N(t_{\mathrm{f}}) \tag{4.56c}$$

for a fixed $\tau \in (t_0, t_{\mathrm{f}})$. Partitioning $N$ and $C$ according to the structure of the original variable $z$ in the form

$$N =: \begin{bmatrix} N^\lambda \\ N^x \\ N^u \end{bmatrix}, \qquad C\mathcal{P}_z =: \begin{bmatrix} C^\lambda \\ C^x \\ 0 \end{bmatrix}^{\mathsf{H}} \tag{4.57}$$

and after removing redundant equations and the algebraic variables, we obtain the boundary condition

$$(E^{+}EN^x)(t_0) = 0, \qquad (EE^{+}N^\lambda)(t_{\mathrm{f}}) = -E^{+\mathsf{H}}(t_{\mathrm{f}})KN^x(t_{\mathrm{f}}) \tag{4.58}$$

and the jump condition

$$E(\tau)N^x(\tau+) = E(\tau)N^x(\tau-) + C^\lambda(\tau), \qquad E^{\mathsf{H}}(\tau)N^\lambda(\tau+) = E^{\mathsf{H}}(\tau)N^\lambda(\tau-) - C^x(\tau). \tag{4.59}$$

For completeness, the quantities $\tilde{\Gamma}$ and $\Xi_\tau$ in Theorem 4.16 for the computation of the sensitivities are given by

$$\tilde{\Gamma} = \begin{bmatrix} -((EE^{+})_\theta x)(t_{\mathrm{f}}) - ((E^{+\mathsf{H}}K)_\theta x)(t_{\mathrm{f}}) \\ x_{0,\theta} - (E^{+}E)_\theta(t_0)x(t_0) \\ 0 \end{bmatrix},$$

$$\Xi_\tau = \begin{bmatrix} \left((EE^{+} + E^{+\mathsf{H}}K^2E^{+})^{+}(E^{+\mathsf{H}}KE^{\mathsf{H}}N^\lambda - EN^x)\right)(t_{\mathrm{f}}) \\ -(E^{\mathsf{H}}N^\lambda)(t_0) \\ 0 \end{bmatrix}.$$

**Example 4.5.** Let the data in the optimal control problem (4.51) defined on $\mathbb{I} \times \Theta = \mathbb{I} \times (1, \infty)$ be given by

$$E(t,\theta) = \begin{bmatrix} 1 \end{bmatrix}, \quad A(t,\theta) = \begin{bmatrix} -1 \end{bmatrix}, \quad B(t,\theta) = \begin{bmatrix} 1 \end{bmatrix}, \quad f(t,\theta) = 0,$$

$$Q(t,\theta) = \begin{bmatrix} \theta^2 - 1 \end{bmatrix}, \quad S(t,\theta) = \begin{bmatrix} 0 \end{bmatrix}, \quad R(t,\theta) = \begin{bmatrix} 1 \end{bmatrix}, \quad K = \begin{bmatrix} 0 \end{bmatrix},$$

and $x_0(\theta) \in \mathcal{C}^1(\Theta, \mathbb{K}^n)$. Then the necessary conditions (4.52) are given by

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{\lambda} \\ \dot{x} \\ 0 \end{pmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ -1 & \theta^2 - 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda \\ x \\ u \end{pmatrix}, \tag{4.60}$$
$$x(t_0) = x_0, \qquad \lambda(t_\mathrm{f}) = 0.$$

System (4.60) has the analytic solution

$$\begin{pmatrix} \lambda \\ x \\ u \end{pmatrix}(t,\theta) = \frac{\begin{pmatrix} -e^{-(t+t_0)\theta}\left(e^{2t\theta} - e^{2t_\mathrm{f}\theta}\right) x_0 \left(\theta^2 - 1\right) \\ e^{-(t+t_0)\theta} x_0 \left(e^{2t\theta}(\theta - 1) + e^{2t_\mathrm{f}\theta}(\theta + 1)\right) \\ e^{-(t+t_0)\theta}\left(e^{2t\theta} - e^{2t_\mathrm{f}\theta}\right) x_0 \left(\theta^2 - 1\right) \end{pmatrix}}{\theta + e^{2(t_\mathrm{f} - t_0)\theta}(\theta + 1) - 1}. \tag{4.61}$$

Suppose now, we would like to compute the sensitivities of the solution (4.61) inside the interval $\mathbb{I} = [t_0, t_\mathrm{f}]$. Thus, let a fixed $\tau \in (t_0, t_\mathrm{f})$ and $\theta \in \Theta$ be given and assume, that we want sensitivity information of the whole state $z$, i. e., we set $C = I_3$. Note, that the projection $\mathcal{P}_z$ as in (3.23) is pointwise given by

$$\mathcal{P}_z(t,\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

Thus, the partitioning for $C\mathcal{P}_z$ in (4.57) is given by

$$C^\lambda = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, \qquad C^x = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}.$$

The adjoint boundary value problem (4.42) in Theorem 4.16 then has the form

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{N}^\lambda \\ \dot{N}^x \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ -1 & \theta^2 - 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} N^\lambda \\ N^x \\ N^u \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \delta_\tau,$$
$$N^x(\tau+) = N^x(\tau-) + \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, \qquad N^\lambda(\tau+) = N^\lambda(\tau-) - \begin{bmatrix} 0 & 1 & 0 \end{bmatrix},$$
$$N^x(t_0) = 0, \qquad N^\lambda(t_\mathrm{f}) = 0. \qquad \triangleright$$

## 4.6. Higher Index Cases

For the forward system approach in Section 4.1 we already noted in Remark 4.1 that it is also feasible for an arbitrary strangeness index. In Sections 4.2 to 4.5 we assumed that the system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{f})$ or the system of necessary conditions (4.52) is strangeness-free pointwise in $\theta$.

If we drop that assumption, we can proceed with a formal adjoint sensitivity approach similar to (3.50). In the general boundary value problem case of Theorem 4.7 we need the adjoint boundary coefficients $\check{\Gamma}_0, \check{\Gamma}_\mathrm{f}$ as in (4.6) which are defined based on the strangeness-free coefficients. In the case of the necessary conditions (4.52), we can use the fact that the boundary conditions of the true adjoint system (4.56) correspond to the boundary conditions of the necessary conditions (4.53).

We focus on the optimal control case for a fixed $\tau \in (t_0, t_\mathrm{f})$. Let some $C^x \in \mathcal{C}^1(\mathbb{I} \times \Theta, \mathbb{K}^{n\times r})$ and $C^\lambda \in \mathcal{C}^1(\mathbb{I} \times \Theta, \mathbb{K}^{n\times r})$ be given such that $C^x(\tau) \in \operatorname{im} E^\mathsf{H}(\tau)$ and $C^\lambda(\tau) \in \operatorname{im} E(\tau)$. Then, define the formal adjoint sensitivity system by

$$E\dot{N}^x = (A + \dot{E})N^x + BN^u + C^\lambda \delta_\tau, \tag{4.62a}$$
$$-E^\mathsf{H}\dot{N}^\lambda = -QN^x - SN^u + (A + \dot{E})^\mathsf{H} N^\lambda + C^x \delta_\tau, \tag{4.62b}$$
$$0 = -S^\mathsf{H} N^x - RN^u + B^\mathsf{H} N^\lambda, \tag{4.62c}$$

with corresponding boundary and jump conditions

$$N^x(t_0) = 0, \qquad (E^\mathsf{H} N^\lambda)(t_\mathrm{f}) = -KN^x(t_\mathrm{f}), \tag{4.62d}$$
$$E(\tau)N^x(\tau+) = E(\tau)N^x(\tau-) + C^\lambda(\tau), \qquad E^\mathsf{H}(\tau)N^\lambda(\tau+) = E^\mathsf{H}(\tau)N^\lambda(\tau-) - C^x(\tau).$$

In the strangeness-free case, assuming that all coefficients are sufficiently smooth and $C^x$ and $C^\lambda$ are defined by (4.57), the formal sensitivity adjoint (4.62) corresponds to the true sensitivity adjoint (4.56).

We thus have the following possibilities for the treatment of higher index cases, where we need to distinguish between the strangeness index of the original system (4.52a) and the strangeness index of the necessary conditions (4.52), see the discussion in Subsection 3.4.6 and also compare with Figure I.2.

1. Let us assume that the necessary conditions (4.52) are given in closed form, are not necessarily strangeness-free, and all data is sufficiently smooth

with respect to parameters and time, such that the quantities of Hypothesis 3.3 are available and sufficiently smooth with respect to parameters and time globally on $\mathbb{I} \times \Theta$. Also assume, that the original system (4.52a) is strangeness-free as a controlled system.

If we want to avoid performing the index reduction with smooth versions of $T_2$ and $Z_1$ we can proceed as in Subsection 3.4.6 for solving the adjoint equations (4.18) and (4.42). The actual sensitivities given by formulas (4.20) and (4.44), however, contain terms like $C(\mathcal{P}_z)_\theta$, $C\tilde{F}_\mathrm{a}$, and $Cf_{\mathrm{a},\theta}$ which are flow quantities as in (3.23) that are only defined for strangeness-free systems. Thus, we need to make sure, that these quantities are either available, or to choose $C$ in such a way, that they vanish.

2. If the original system (4.52a) is not strangeness-free as a controlled system, then the discussion of Subsection 3.4.6 applies, where in addition we require, that all obtained data is sufficiently smooth, in particular with respect to the parameters.

3. If we opt for reducing the formal necessary conditions, we must ensure, that all boundary conditions in the parameter-dependent version of Equation (3.50) are fulfilled such that we have a chance to apply Theorem 3.24. Obtaining those needs knowledge of $T_2(t_0)$, $Z_1(t_\mathrm{f})$.

4. On the other hand, we can directly use the parameter-dependent version of the formal necessary conditions (3.50) to obtain the corresponding formal adjoint sensitivity system (4.62). To fix the correct jump conditions (4.59) we need the quantity $C\mathcal{P}_z$, in particular we need global knowledge of $T_2$.

   The formal sensitivity adjoint approach is problematic due to the following. The parameter-dependent formal necessary conditions are formally self-adjoint in the sense of Section 4.5 and thus Theorem 3.24 applies accordingly to the adjoint equation (4.56). This implies, that using the partitioning (4.57), $N$ may solve the formal adjoint sensitivity system (4.62), but the quantity $N^\lambda$ does not necessarily solve the true adjoint system (4.42).

   The quantity $N^\lambda$, though, is necessary in the computation of the sensitivities (4.20) or (4.44). However, if in (4.52) only the weight functions $Q, S, R$ depend on the parameters, then all multiplicative terms with $N^\lambda$ and also $\lambda$ in $\tilde{F}$ or $\tilde{F}^{(3)}$ vanish and the approach may be feasible.

**Example 4.6.** Let us consider the following parameter-dependent optimal control problem with dynamical system given by

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{4.63}$$

and corresponding objective function

$$\int_0^2 \theta x_1^2(t,\theta) + u^2(t,\theta) \mathrm{d}t,$$

where the parameter space $\Theta$ is given by $\Theta = [0,\infty)$. This is a parameter-dependent version of [KM11b, Example 1.11]. We assume, that $f_1$ and $f_2$ are sufficiently smooth, such that we can pointwise in $\theta$ use the strong setting (3.8). The system (4.63) is independent of the parameter $\theta$ and not strangeness-free, even as a controlled system. Further, the reduced system (3.14) is given by

$$0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{pmatrix} f_1 + \dot{f}_2 \\ f_2 \end{pmatrix},$$

which constitutes a purely algebraic equation. The corresponding adjoint equation (4.52b) is given by

$$0 = -\begin{bmatrix} \theta & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \tag{4.64}$$

and the optimality condition (4.52c) by

$$0 = -u + \lambda_1.$$

Note that $\hat{R}$ in (3.45) is given by

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

is invertible, and equals $\mathcal{A}$ in (3.43). We thus obtain the unique solution

$$x_1 = -\frac{1}{1+\theta}(f_1 + \dot{f}_2), \quad u = \lambda_1 = -\frac{\theta}{1+\theta}(f_1 + \dot{f}_2), \quad x_2 = -f_2, \quad \lambda_2 = 0.$$

Note, that $P_z = 0$ and thus also $\mathcal{P}_z = 0$ as well as the projected quantity

$$\tilde{F}_{\mathrm{a}} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -\theta & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ -x_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{-1}{\theta+1} x_1 \\ 0 \\ \frac{1}{\theta+1} x_1 \\ 0 \\ \frac{-1}{\theta+1} x_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{(\theta+1)^2}(f_1 + \dot{f}_2) \\ 0 \\ \frac{-1}{(\theta+1)^2}(f_1 + \dot{f}_2) \\ 0 \\ \frac{1}{(\theta+1)^2}(f_1 + \dot{f}_2) \end{bmatrix}$$

which is defined as in (4.20). Then, the adjoint sensitivity system (4.56) is given by

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{N}_1^{\lambda} \\ \dot{N}_2^{\lambda} \\ \dot{N}_1^{x} \\ \dot{N}_2^{x} \\ \dot{N}^{u} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -\theta & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} N_1^{\lambda} \\ N_2^{\lambda} \\ N_1^{x} \\ N_2^{x} \\ N^{u} \end{bmatrix}, \tag{4.65}$$

which is purely algebraic without any prescribed boundary conditions or jump conditions. It has 0 as unique solution. Thus, the sensitivities $y_\theta(\tau)$ with $C = I_5$ can be computed via (4.20) to obtain

$$y_\theta(\tau) = -C\tilde{F}_{\mathrm{a}} = \begin{bmatrix} \frac{-1}{(\theta+1)^2}(f_1 + \dot{f}_2) \\ 0 \\ \frac{1}{(\theta+1)^2}(f_1 + \dot{f}_2) \\ 0 \\ \frac{-1}{(\theta+1)^2}(f_1 + \dot{f}_2) \end{bmatrix} \tag{4.66}$$

as expected.

On the other hand, considering the formal adjoint approach we can replace the adjoint equation (4.64) by its formal adjoint equation (3.50b) given by

$$-\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \end{pmatrix} = -\begin{bmatrix} \theta & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

and corresponding boundary condition (3.50d) which reads $\lambda_1(2) = 0$. The formal system of necessary conditions has the unique solution

$$x_1 = -\frac{1}{1+\theta}(f_1 + \dot{f}_2), \quad u = \lambda_1 = -\frac{\theta}{1+\theta}(f_1 + \dot{f}_2), \quad x_2 = -f_2, \quad \lambda_2 = \frac{\theta}{1+\theta}(\dot{f}_1 + \ddot{f}_2),$$

assuming that the boundary condition on $\lambda_1$ is fulfilled, which, depending on the data, may or may not be the case. However, in comparison with (4.64) it is clear that this condition should not be present. Note, that the solutions of the true and the necessary conditions only differ in $\lambda_2$ and that we need more smoothness of the inhomogeneity. Also, compare with Theorem 3.24.

Moreover, we can state the formal adjoint system (4.62) for the computation of the sensitivities. Choosing some $C^\lambda, C^x \in \mathcal{C}^1(\mathbb{I} \times \Theta, \mathbb{K}^{2\times 1})$ it is given by

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{N}_1^\lambda \\ \dot{N}_2^\lambda \\ \dot{N}_1^x \\ \dot{N}_2^x \\ \dot{N}^u \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -\theta & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} N_1^\lambda \\ N_2^\lambda \\ N_1^x \\ N_2^x \\ N^u \end{bmatrix} + \begin{bmatrix} C^\lambda \\ 0 \\ 0 \\ C^x \\ 0 \end{bmatrix} \delta_\tau, \tag{4.67}$$

$$N_2^x(0) = 0, \qquad N_1^\lambda(2) = 0,$$
$$N_2^x(\tau+) = N_2^x(\tau-) + C^\lambda(\tau), \qquad N_1^\lambda(\tau+) = N_1^\lambda(\tau-) - C^x(\tau).$$

This system is only solvable, if $C^x(\tau) = C^\lambda(\tau) = 0$. In that case it also has 0 as unique solution and corresponds to the solution of the true adjoint sensitivity system (4.65).

Note, that the derivatives with respect to $\theta$ of the respective matrix coefficients in (4.65) and (4.67) coincide and in particular the first two columns and rows vanish such that $\tilde{F}$ defined as in (4.3) does not depend on $\lambda$ and $N^\lambda$. Thus, provided that we set $C = I_5$ and have knowledge of $\tilde{F}_a$, which is only defined for strangeness-free systems, the formal approach gives the correct solution (4.66) for the sensitivities. Nevertheless, choosing $C^x = C^\lambda = 0$ corresponds to choosing $C = 0$ in (4.57) in the strangeness-free case and no sensitivity information is obtained. ▷

In summary, we have stated that using a given original higher index system directly is beneficial in the sense that we do not have to blow up the system dimension. Several difficulties arise and we have seen several possibilities of overcoming individuals of these issues. However, the concrete procedure has to be decided case by case. We conclude that, whenever possible, one should avoid higher index models or provide full information about the derivative array including smooth selector matrices $Z_1$ and $T_2$ and possible derivatives with

respect to the parameters $\theta$. Also, if the strangeness index is higher than 1, one should consider a different modeling approach, e. g., using port-Hamiltonian modeling, for which it has been proved in the linear case that the strangeness index is always at most 1, see [MMW18].

## 4.7. Numerical Treatment

We present different possibilities for computing numerical solutions of the sensitivity approaches presented in Sections 4.1 to 4.3 and 4.5. In particular, we only consider the strangeness-free case, i. e., all DAEs in this section are assumed to be strangeness-free.

Note, that for a given fixed value $\hat{\theta}$ all of these systems require a solution of the original boundary value problem (4.1) or of its weaker version, where (4.1a) is replaced by (4.27). Thus, we first need to compute such a solution with any feasible method. An overview of possible methods is given in [CK13; KM06].

The forward system approaches of Sections 4.1 and 4.3 do not require special treatment as the homogeneous parts of the original DAE (4.1a) or (4.27) coincides with the respective homogeneous part of (4.3a) and (4.32). We just need to make sure, that $\tilde{F}$ is available at all integration points of the forward sensitivity system or use a simple interpolation scheme. See, e. g., [Kie99; LP00; PLCS06] for implementational considerations in the case of DAE initial value problems. In the case of optimal control problems, there is more structure that can be exploited, and we are able to use a differential Riccati approach.

This section is structured as follows. In Subsection 4.7.1 we introduce a multiple shooting approach for solving the adjoint sensitivity equations in Sections 4.2 and 4.3. In Subsection 4.7.2 we analyze how Riccati equations can be used to solve the adjoint equations, when the original boundary value problem is given by the parameter-dependent necessary conditions (4.52).

### 4.7.1. Multiple Shooting Approach

The adjoint equations (4.18) and (4.42) both include a jump condition at $t = \tau$. Lemma 3.17 provides a possible way of rewriting that problem in a suitable form.

Thus, the system (4.42) can be rewritten as

$$\begin{bmatrix} -E^{\mathsf{H}} & 0 \\ 0 & -E^{\mathsf{H}} \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} EE^{+}N_1 \\ EE^{+}N_2 \end{bmatrix} = \begin{bmatrix} (A+EE^{+}\dot{E})^{\mathsf{H}} & 0 \\ 0 & (A+EE^{+}\dot{E})^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}, \tag{4.68a}$$

with boundary condition

$$\begin{aligned} \begin{bmatrix} 0 \\ -(C\mathcal{P}_z)^{\mathsf{H}}(\tau) \end{bmatrix} = \begin{bmatrix} \check{\Gamma} E^{\mathsf{H}}(t_0) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N_1(t_0) \\ N_2(t_0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ E^{\mathsf{H}}(\tau) & -E^{\mathsf{H}}(\tau) \end{bmatrix} \begin{bmatrix} N_1(\tau) \\ N_2(\tau) \end{bmatrix} \\ + \begin{bmatrix} 0 & \check{\Gamma}_{\mathrm{f}} E^{\mathsf{H}}(t_{\mathrm{f}}) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N_1(t_{\mathrm{f}}) \\ N_2(t_{\mathrm{f}}) \end{bmatrix}, \end{aligned} \tag{4.68b}$$

Note, that we are only interested in $N_1$ on $[t_0,\tau)$ and $N_2$ on $[\tau,t_{\mathrm{f}}]$ and that $N = \mathbb{1}_{[t_0,\tau)}N_1 + \mathbb{1}_{[\tau,t_{\mathrm{f}}]}N_2$ solves (4.42). The augmented system for (4.18) is developed analogously.

We can directly interpret Equation (4.68) as a multiple shooting approach with one intermediate node at $t=\tau$, where at this node we impose a jump condition. For stability reasons, see, e. g., [KM06; KMS05], we should also include intermediate nodes in the intervals $(t_0,\tau)$ and $(\tau,t_{\mathrm{f}})$. If we want to add additional nodes at timepoints $t_0 < t_i < \tau =: t_{i_\tau} < t_j < t_{\mathrm{f}}$, $i = 1,\dots,i_\tau - 1, j = i_\tau + 1,\dots,i_2$ with corresponding solutions $N_i$ on $[t_i,t_{i+1}]$ , we have to impose the additional conditions

$$E^{\mathsf{H}}(t_i)N_i(t_i) = E^{\mathsf{H}}(t_i)N_{i+1}(t_i), i = 0,\dots,i_2-1, i \neq i_\tau.$$

Using Lemmas 3.14 and 3.17 iteratively and setting $(\check{E},\check{A}) = (-E^{\mathsf{H}},(A+\dot{E})^{\mathsf{H}})$, $\Phi_i := \Phi^{t_i}_{(\check{E},\check{A})}$, $i = 0,\dots,i_2$, we infer that solvability of (4.68) implies unique solvability of the *shooting system*

$$\begin{aligned} &\begin{bmatrix} \Phi_0(t_1) & -\check{P}_z(t_1) & & & \\ & \Phi_1(t_2) & \ddots & & \\ & & \ddots & \ddots & \\ & & & \Phi_{t_{i_2-1}}(t_{\mathrm{f}}) & -\check{P}_z(t_{i_2-1}) \\ \check{\Gamma}_0 & 0 & \dots & 0 & \check{\Gamma}_{\mathrm{f}} \end{bmatrix} \begin{bmatrix} N_0(t_0) \\ N_1(t_1) \\ \vdots \\ N_{i_2-1}(t_{i_2-1}) \\ N_{i_2}(t_{\mathrm{f}}) \end{bmatrix} \\ &\qquad = e_{i_\tau} \otimes (C\mathcal{P}_z)^{\mathsf{H}}(\tau). \end{aligned} \tag{4.69}$$

Once, correct node values $N_i(t_i)$ are fixed, we can obtain solution values at intermediate points by either forward or backward integration from the closest node point using the corresponding node value.

For the computation of the sensitivity term (4.44) we need a solution $x$ of the original boundary value problem (4.27) and a solution $N$ of the adjoint boundary value problem (4.42). Except for the integral term

$$\int_{t_0}^{t_f} N^H \tilde{F}^{(3)} \mathrm{d}t \tag{4.70}$$

with $\tilde{F}^{(3)}$ given as in (4.38), the terms in (4.44) depend on data and evaluations of $N$ and $x$ at $t_0$, $t_f$, and $\tau$ only.

In the following we discuss, whether and how we can control the absolute and relative error of the computed sensitivity $y_\theta(\tau)$ based on formula (4.44). Let $\nu_a$ and $\nu_r$ denote the absolute and relative tolerance, respectively. Then, it is well-known that the errors of the evaluation of $N(t)$ and $x(t)$ at any timepoint $t$ – not necessarily a shooting node – can be controlled to obey $\nu_a$ and $\nu_r$. The same holds for the evaluation of any integral of the form

$$\int_{t_0}^{t_f} \tilde{g} \mathrm{d}t$$

for any continuous $g \in \mathcal{C}^0(\mathbb{I}, \mathbb{K}^{r\times p})$. The difficulty in the latter case is that in the case of (4.70) the timepoints at which $g$ is evaluated are determined dynamically and not known beforehand. Thus, for the accurate integration of (4.70) we cannot start by computing $N$ and $x$ on a prescribed time grid. Instead we can do the following.

1. Compute the solutions of $N$ and $x$ with the multiple shooting approach at prescribed shooting nodes including $t_0$, $t_f$, and $\tau$.

2. Compute the integral

$$\int_{t_0}^{t_f} N^H \tilde{F}^{(3)} \mathrm{d}t = \int_{t_0}^{\tau} N^H \tilde{F}^{(3)} \mathrm{d}t + \int_{\tau}^{t_f} N^H \tilde{F}^{(3)} \mathrm{d}t,$$

   where each summand is an integral over a continuous function. At each intermediate timepoint $t$, where $x$ or $N$ is unknown yet, compute $N(t)$ and $x(t)$ via forward or backward integration from the nearest node value.

Set $N_{\mathfrak{E}} \coloneqq N - \tilde{N}$ and $x_{\mathfrak{E}} \coloneqq x - \tilde{x}$. Then, for the absolute error of $N^{\mathsf{H}} E(E^+E)_\theta x|_{t_0}^{t_\mathrm{f}}$ we can use Lemma 3.4 to obtain

$$\begin{aligned}\|N^{\mathsf{H}} E(E^+E)_\theta x|_{t_0}^{t_\mathrm{f}} - \tilde{N}^{\mathsf{H}} E(E^+E)_\theta \tilde{x}|_{t_0}^{t_\mathrm{f}}\|_2 \dot{\leqq}& \\ \left(\|\tilde{N}(t_0)\|_2 \|N_{\mathfrak{E}}\|_2 + \|\tilde{x}(t_0)\|_2 \|x_{\mathfrak{E}}\|_2\right) \|(E(E^+E)_\theta)(t_0)\|_2 +& \\ \left(\|\tilde{N}(t_\mathrm{f})\|_2 \|N_{\mathfrak{E}}\|_2 + \|\tilde{x}(t_\mathrm{f})\|_2 \|x_{\mathfrak{E}}\|_2\right) \|(E(E^+E)_\theta)(t_\mathrm{f})\|_2 \quad (4.71)\end{aligned}$$

after neglecting higher order terms.

Note, that the absolute error $N_{\mathfrak{E}}$ or $x_{\mathfrak{E}}$ is multiplied by norms of the solution $\|\tilde{N}(t)\|_2$ or $\|\tilde{x}(t)\|_2$, $t \in \{t_0, t_\mathrm{f}\}$, respectively. These are both unknown terms before the computation, and thus can only be used as estimates for another computation close to the current one.

In summary, all summands in formula (4.44) have a bounded absolute error for prescribed tolerances $\nu_\mathrm{a}$ and $\nu_\mathrm{r}$ and thus also $y_\theta(\tau)$ has a bounded absolute error.

On the other hand, this does not necessarily mean, that the 2-norm of the error $y_{\theta,\mathfrak{E}}(\tau) \coloneqq y_\theta(\tau) - \tilde{y}_\theta(\tau)$ between the correct sensitivities $y_\theta(\tau)$ and the approximation $\tilde{y}_\theta(\tau)$ is also bounded by $\nu_\mathrm{a}$. We need, that in addition the norms of all involved coefficients like $E(E^+E)_\theta$ in (4.71) are sufficiently small and to comprise for the fact, that in (4.44) we sum over 6 terms, e. g., by scaling $\nu_\mathrm{a}$ by $\frac{1}{6}$ for the computation of $\tilde{x}$ and $\tilde{N}$. Then, if the norms of the errors of the solutions $x$ and $N$ are rather small, then $\|y_{\theta,\mathfrak{E}}\|_2$ does approximately obey the absolute tolerance $\nu_\mathrm{a}$.

The relative error of $y_\theta(\tau)$, however, cannot be controlled in general as cancellation effects may occur in formula (4.44).

In the case of optimal control problems and the necessary conditions (4.52) we already noted in Section 4.5 that the homogeneous parts of (4.52) and of the respective adjoint equation (4.56) coincide. Hence, their respective flows and shooting matrices (4.69) coincide and thus the computation of solutions can be simplified.

**Example 4.7** (Example 4.5 revisited)**.** The sensitivities of the solution (4.61) can be computed analytically, e. g., by direct differentiation with respect to $\theta$. Note, that the differentiation of the terms in (4.61) needs applications of the quotient rule and multiple applications of the product rule. The resulting term is very lengthy and can be easily computed with symbolic mathematics packages like

Table 4.2.: Displays absolute and relative error of the computed sensitivity and additional evaluation points of $N$ and $x$ inside the integral term (4.70) at $\tau = 1$ for equal prescribed tolerances $\nu_a = \nu_r$ in Example 4.7.

| tolerance | relative error | absolute error | additional integral points |
|---|---|---|---|
| $10^{0}$ | $5.533 \cdot 10^{-1}$ | $1.520 \cdot 10^{-1}$ | 35 |
| $10^{-1}$ | $6.346 \cdot 10^{-1}$ | $1.744 \cdot 10^{-1}$ | 35 |
| $10^{-2}$ | $2.289 \cdot 10^{-2}$ | $6.290 \cdot 10^{-3}$ | 54 |
| $10^{-3}$ | $3.877 \cdot 10^{-3}$ | $1.065 \cdot 10^{-3}$ | 54 |
| $10^{-4}$ | $9.364 \cdot 10^{-4}$ | $2.573 \cdot 10^{-4}$ | 119 |
| $10^{-5}$ | $1.509 \cdot 10^{-4}$ | $4.146 \cdot 10^{-5}$ | 170 |
| $10^{-6}$ | $2.317 \cdot 10^{-5}$ | $6.367 \cdot 10^{-6}$ | 158 |
| $10^{-7}$ | $4.491 \cdot 10^{-6}$ | $1.234 \cdot 10^{-6}$ | 228 |

`Mathematica`. The exact term is thus omitted here. Let us fix $t_0 = 0$, $t_f = 2$, $\theta = 2$, $x_0(t,\theta) = 2$, $\tau = 1$.

Solving the adjoint system (4.42) at timepoint $t = 1$ for different tolerances $\nu_a = \nu_r$, we obtain the errors listed in Table 4.2. To accommodate for the fact, that formula (4.44) comprises 6 individual summands, we multiply the allowed tolerance for the computation of $N$ and $x$ by $\frac{1}{6}$.

Table 4.2 shows that in the case of this example the absolute and relative error roughly fulfill the required tolerances. Note, that the number of additional integral evaluations does not increase with the same rate as the tolerance decreases. ▷

We consider another example based on Example 4.5, where condition (4.31) is not fulfilled.

**Example 4.8** (Example 4.5 revisited)**.** Let the data in the optimal control prob-

lem (4.51) defined on $\mathbb{I}\times\Theta=\mathbb{I}\times(1,\infty)\times\mathbb{R}$ be given by

$$E(t,\theta,\omega)=\begin{bmatrix}1&0\\ \omega&0\end{bmatrix},\quad A(t,\theta,\omega)=\begin{bmatrix}-1&0\\ -\omega&1\end{bmatrix},\quad B(t,\theta,\omega)=\begin{bmatrix}1\\ \omega\end{bmatrix},\quad f(t,\theta,\omega)=0,$$
$$Q(t,\theta,\omega)=\begin{bmatrix}\theta^2-1&0\\ 0&0\end{bmatrix},\quad S(t,\theta,\omega)=\begin{bmatrix}0\\ 0\end{bmatrix},\quad R(t,\theta,\omega)=\begin{bmatrix}1\end{bmatrix}\quad K=0_{2\times 2},$$

with $x_0(\theta,\omega)\in\mathcal{C}^1(\Theta,\mathbb{K}^2)$. Note, that

$$(E^+E)(t,\theta,\omega)=\begin{bmatrix}1&0\\ 0&0\end{bmatrix},\qquad (EE^+)(t,\theta,\omega)=\begin{bmatrix}1&\omega\\ \omega&\omega^2\end{bmatrix}\frac{1}{1+\omega^2},$$

and thus

$$\ker(EE^+)(t,\theta,\omega)=\operatorname{im}\begin{bmatrix}\omega\\ -1\end{bmatrix},\qquad \ker(EE^+)_\omega(t,\theta,\omega)\begin{bmatrix}\omega\\ -1\end{bmatrix}=\begin{bmatrix}-1\\ -\omega\end{bmatrix}\neq 0,$$

with

$$\ker(EE^+)\cap\ker(EE^+)_\omega=\{0\}.$$

In particular, condition (4.31) does not hold.

Then, by setting

$$x=\begin{pmatrix}x_1\\ x_2\end{pmatrix},\qquad \begin{pmatrix}\tilde\lambda_1\\ \tilde\lambda_2\end{pmatrix}=EE^+\lambda=\begin{bmatrix}1&\omega\\ \omega&\omega^2\end{bmatrix}\frac{1}{1+\omega^2}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix},$$

the necessary conditions (4.52) are given by

$$\begin{bmatrix}0&0&1&0&0\\ 0&0&\omega&0&0\\ -1&-\omega&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\end{bmatrix}\begin{pmatrix}\dot{\tilde\lambda}_1\\ \dot{\tilde\lambda}_2\\ \dot x_1\\ 0\\ 0\end{pmatrix}=\begin{bmatrix}0&0&-1&0&1\\ 0&0&-\omega&1&\omega\\ -1&-\omega&\theta^2-1&0&0\\ 0&1&0&0&0\\ 1&\omega&0&0&1\end{bmatrix}\begin{pmatrix}\lambda_1\\ \lambda_2\\ x_1\\ x_2\\ u\end{pmatrix},$$
$$x(t_0)=x_{0,1},\qquad \lambda(t_\mathrm{f})=0.$$

From the 4th equation, we deduce that $\lambda_2=0$ and hence

$$-\dot{\tilde\lambda}_1-\omega\dot{\tilde\lambda}_2=-\dot\lambda_1.$$

Inserting the first equation into the second, we conclude that $x_2 = 0$. The DAE for the remaining equations exactly reads like (4.60) and thus the solution $\lambda_1, x_1, u$ is given by (4.61) and independent of $\omega$. In summary, we have the solution

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ x_1 \\ x_2 \\ u \end{pmatrix}(t,\theta,\omega) = \frac{1}{\theta + e^{2(t_\mathrm{f}-t_0)\theta}(\theta+1) - 1} \begin{pmatrix} -e^{-(t+t_0)\theta}\left(e^{2t\theta} - e^{2t_\mathrm{f}\theta}\right) x_0 \left(\theta^2 - 1\right) \\ 0 \\ e^{-(t+t_0)\theta} x_0 \left(e^{2t\theta}(\theta-1) + e^{2t_\mathrm{f}\theta}(\theta+1)\right) \\ 0 \\ e^{-(t+t_0)\theta}\left(e^{2t\theta} - e^{2t_\mathrm{f}\theta}\right) x_0 \left(\theta^2 - 1\right) \end{pmatrix}.$$

Thus, we conclude that the sensitivities $\lambda_\omega, x_\omega, u_\omega$ all equal zero. Using the adjoint approach of Theorem 4.16 we obtain the following results for the single terms in (4.44) using a tolerance of $10^{-6}$ on $\mathbb{I} = [0,2]$ and for $C = I_5$:

$$(C_\omega x)(1.) = 0., \quad (C(\mathcal{P}_z)_\omega P_z x)(1.) = \begin{bmatrix} 1.3284 \cdot 10^{-1} \\ 7.3742 \cdot 10^{-17} \\ 3.0231 \cdot 10^{-17} \\ 6.2662 \cdot 10^{-17} \\ -1.3284 \cdot 10^{-1} \end{bmatrix}, \quad Cf_{\mathrm{a},\omega}(\tau) = 0.$$

$$\Xi_\tau^\mathsf{H} \tilde{\Gamma} = 0., \quad N^\mathsf{H} E(E^+E)_\omega x|_{t_0}^{t_\mathrm{f}} = \begin{bmatrix} -2.9497 \cdot 10^{-17} \\ 0. \\ -3.0231 \cdot 10^{-17} \\ 0. \\ 2.9497 \cdot 10^{-17} \end{bmatrix},$$

$$\int_{t_0}^{t_\mathrm{f}} N^\mathsf{H} \tilde{F}^{(3)} \mathrm{d}t = \begin{bmatrix} -1.3284 \cdot 10^{-1} \\ 0. \\ 2.7790 \cdot 10^{-7} \\ 0. \\ 1.3284 \cdot 10^{-1} \end{bmatrix}.$$

Thus, the computed sensitivity

$$y_\omega(1.) = \begin{bmatrix} -1.9406 \cdot 10^{-6} \\ 7.3742 \cdot 10^{-17} \\ 2.7790 \cdot 10^{-7} \\ 6.2662 \cdot 10^{-17} \\ 1.9406 \cdot 10^{-6} \end{bmatrix}$$

is indeed close to 0. Note, that cancellation occurs in the first and last component when adding the second and last term. Further, the relative error is $\infty$ as the exact solution is 0. ▷

### 4.7.2. Differential Riccati Equations

In the optimal control setting of Section 4.5 we have more structure which can be exploited. In particular, the system (4.52) is self-adjoint and the corresponding sensitivity adjoint system is given by system (4.56).

Differential Riccati equations are a well-known tool for solving the finite-time optimal control problems under consideration, see [KM11b]. The goal of this subsection is to extend these Riccati equations to the adjoint system (4.56). Let us first recall the result for differential Riccati equations for the necessary conditions (4.52).

**Theorem 4.19** ([KM11b])**.** *Consider the necessary conditions* (4.52) *for a fixed parameter* $\theta \in \Theta$ *and let Assumption 4.18 hold. In addition, assume that* $R$ *is positive definite. Set*

$$\begin{aligned} F &:= A - BR^{-1}S^{\mathsf{H}}, \\ G &:= BR^{-1}B^{\mathsf{H}}, \\ H &:= Q - SR^{-1}S^{\mathsf{H}}, \end{aligned}$$

*and assume that*

$$HE^{+}E = H. \tag{4.72}$$

*If* $X \in \mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^{n\times n})$, $v \in \mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^n)$ *solves*

$$\frac{\mathrm{d}}{\mathrm{d}t}(E^{\mathsf{H}}XE) + E^{\mathsf{H}}XF + F^{\mathsf{H}}XE + E^{\mathsf{H}}XGXE - H = 0, \tag{4.73a}$$

$$(E^{\mathsf{H}}XE)(t_{\mathrm{f}}) = -K \tag{4.73b}$$

*and*

$$\frac{\mathrm{d}}{\mathrm{d}t}(E^{\mathsf{H}}v) + E^{\mathsf{H}}XGv + F^{\mathsf{H}}v + E^{\mathsf{H}}Xf = 0, \tag{4.74a}$$

$$(E^{\mathsf{H}}v)(t_{\mathrm{f}}) = 0, \tag{4.74b}$$

*then the solution* $z = (\lambda^T, x^T, u^T)^T$ *of the necessary condition* (4.52) *fulfills*

$$\lambda = XEx + v. \qquad \triangleright$$

Note, that already in the ODE case (4.73a) may not have a solution on the whole interval $\mathbb{I}$, even if the necessary conditions (4.52) are solvable for a fixed parameter $\theta$ [KM11b]. See, e. g., [AFIJ03; Lib12] for a discussion under which conditions both approaches admit global solutions.

In view of the jump condition (4.59), we must allow for non-zero final condition on $\nu$. However, inspecting the steps in [KM11a] for the derivation of the Riccati approach, the final condition on $\nu$ is only fixed to zero to fulfill the final condition $\lambda(t_\mathrm{f}) = -E^{+\mathsf{H}}(t_\mathrm{f})Kx(t_\mathrm{f}) + 0$. Let us assume, that we have a solution $N$ of (4.56) which is partitioned according to (4.57). Thus, $N$ fulfills the boundary conditions (4.58). We now apply the ansatz

$$N^\lambda = XEN^x + \nu, \tag{4.75}$$

on the two subintervals $[\tau, t_\mathrm{f}]$ and $[t_0, \tau]$ separately. On the subinterval $[\tau, t_\mathrm{f}]$ we are exactly in the situation of Theorem 4.19 and thus for $(E^\mathsf{H}\nu)(t_\mathrm{f}) = 0$ and a solution $X_2$ of the Riccati equation (4.73a) we obtain that (4.75) holds. Note, that $X_2$ corresponds to the solution of the Riccati equation for the original optimal control problem (4.52) as the adjoint system (4.56) and the necessary conditions (4.52) share the same homogeneous dynamics.

We thus infer that

$$N^\lambda(t) = (X_2E)(t)N^x(t)$$

for all $t \in (\tau, t_\mathrm{f}]$. Hence, using the jump conditions (4.59) we also have

$$\begin{aligned}(E^\mathsf{H}N^\lambda)(\tau-) &= (E^\mathsf{H}N^\lambda)(\tau+) + C^x(\tau) = (E^\mathsf{H}X_2E)(\tau)N^x(\tau+) + C^x(\tau)\\ &= (E^\mathsf{H}X_2E)(\tau)N^x(\tau-) + (E^\mathsf{H}X_2C^\lambda)(\tau) + C^x(\tau).\end{aligned} \tag{4.76}$$

Note, that also on $[t_0, \tau]$ the adjoint system (4.56) and the necessary conditions (4.52) share the same homogeneous dynamics. Thus, we can view the condition (4.76) as a final condition compliant with the ansatz function (4.75) by using the Riccati solution $X_1$ from the optimal control problem (4.52) and setting the final conditions

$$(E^\mathsf{H}X_1E)(\tau) = (E^\mathsf{H}X_2E)(\tau), \qquad (E^\mathsf{H}\nu)(\tau) = (E^\mathsf{H}X_2C^\lambda)(\tau) + C^x(\tau).$$

Hence, we have shown the following theorem.

**Theorem 4.20.** *Consider the necessary conditions* (4.52) *and the corresponding adjoint equation* (4.56) *for a fixed parameter* $\theta \in \Theta$ *and let Assumption 4.18 hold. In addition, assume that* $R$ *is positive definite and assume that the consistency condition* (4.72) *is fulfilled. Let* $X \in \mathcal{C}^1_{EE^+}(\mathbb{I}, \mathbb{K}^{n\times n})$ *solve the Riccati equation* (4.73). *Further, for a fixed* $\tau \in (t_0, t_f)$ *let* $v_1$ *be a solution of*

$$\frac{\mathrm{d}}{\mathrm{d}t}(E^{\mathsf{H}}v_1) + E^{\mathsf{H}}XGv_1 + F^{\mathsf{H}}v_1 = 0, \tag{4.77a}$$

$$(E^{\mathsf{H}}v_1)(\tau) = (E^{\mathsf{H}}XC^{\lambda})(\tau) + C^{x}(\tau). \tag{4.77b}$$

*Then, setting*

$$v(t) = \begin{cases} v_1(t), & t \in [t_0, \tau) \\ 0, & t \in [\tau, t_f], \end{cases}$$

*we obtain that*

$$N^{\lambda} = XEN^{x} + v. \tag{4.78}$$

▷

*Proof.* See the discussion preceding the theorem. □

Once the relation (4.78) is established, we can compute $N^x$ by inserting (4.78) and

$$N^u = R^{-1}(B^{\mathsf{H}}N^{\lambda} - S^{\mathsf{H}}N^x) = R^{-1}(B^{\mathsf{H}}XE - S^{\mathsf{H}})N^x + R^{-1}B^{\mathsf{H}}v$$

into the first block equation in (4.56a), i. e., into

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+EN^x) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E))N^x + BN^u + \delta_\tau C^{\lambda},$$

which leads to

$$E\frac{\mathrm{d}}{\mathrm{d}t}(E^+EN^x) = (A + E\frac{\mathrm{d}}{\mathrm{d}t}(E^+E) + BR^{-1}(B^{\mathsf{H}}XE - S^{\mathsf{H}}))N^x + BR^{-1}B^{\mathsf{H}}v + \delta_\tau C^{\lambda}. \tag{4.79}$$

Together with the initial condition (4.58) equation (4.79) constitutes an initial value problem on the interval $[t_0, \tau)$. When $N^x(\tau-)$ is fixed, we can proceed by inserting the solution into the jump condition (4.59), yielding an initial condition for $N^x$ and system (4.79) on $(\tau, t_f]$.

*Remark* 4.9. Assuming sufficient memory storage we can proceed as follows.

1. Compute a solution of the original necessary conditions (4.52) for a fixed parameter with the help of differential-algebraic Riccati equation (4.73) with corresponding solution $X(t)$.

2. For every $\tau \in (t_0, t_f)$ of interest compute a solution of the Riccati inhomogeneity $\nu_1$ as in (4.77) on the interval $[t_0, \tau]$.

3. As the final condition (4.77b) consists of multiple columns, depends on the varying final point $\tau$, and the dynamics (4.77a) are homogeneous, it may be beneficial to compute and withhold the corresponding flow representation (3.23).

4. Using the solution $N^\lambda$ propagate forward $N^x$ from $t_0$ to $t_f$.

5. For values $\tau$ close to $t_0$ this approach may save considerable amount of computation as the computation of $\nu_1$ on $[t_0, \tau]$ is relatively cheap. ▷

# 5. A Bilevel Problem with Nonlinear Least Squares Upper Level

Here we consider a bilevel optimal control problem according to Definition 2.2, where the dynamics in the lower level optimal control problem is given by a differential-algebraic equation (DAE) as introduced in Subsection 3.4.4 and the upper level is a nonlinear least squares problem as introduced in Section 3.3. In this case, comparing with the notation in Definition 2.2, $\Theta \coloneqq \mathbb{Z}_{\mathrm{u}} = \mathbb{X}_{\mathrm{u}} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$, $\mathbb{Z}_{\mathrm{l}} = \mathbb{X}_{\mathrm{l}} \times \mathbb{U}_{\mathrm{l}} = \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n) \times \mathcal{C}^0(\mathbb{I}, \mathbb{K}^m)$, and

$$\mathbb{X}'_{\mathrm{l}} = \left\{ \frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex) \mid x \in \mathcal{C}^1_{E^+E}(\mathbb{I}, \mathbb{K}^n) \right\}.$$

Overall, the bilevel problem can be formulated in a more readable way as follows. Given data points $(t_i, \xi_i) \in \mathbb{I} \times \mathbb{R}^n$, $i = 1, \ldots, q$, solve

$$\left\{ \begin{array}{ll} \min\limits_{\theta \in \Theta} & \sum\limits_{k=1}^{q} \left\| \begin{pmatrix} x^*(t_k, \theta) \\ u^*(t_k, \theta) \end{pmatrix} - \xi_k \right\|_2^2 \\ \text{s.t.} & (x^*(\theta), u^*(\theta)) \text{ uniquely solves} \\ & \left\{ \begin{array}{ll} \min\limits_{u \in \mathbb{U}_{\mathrm{l}}} & (x^{\mathsf{H}} K x)(t_{\mathrm{f}}, \theta) + \int_{t_0}^{t_{\mathrm{f}}} \begin{pmatrix} x(t,\theta) \\ u(t,\theta) \end{pmatrix}^{\mathsf{H}} \begin{bmatrix} Q(t,\theta) & S(t,\theta) \\ S(t,\theta)^{\mathsf{H}} & R(t,\theta) \end{bmatrix} \begin{pmatrix} x(t,\theta) \\ u(t,\theta) \end{pmatrix} \mathrm{d}t \\ \text{s.t.} & \left(E \frac{\mathrm{d}}{\mathrm{d}t}(E^+Ex)\right)(t,\theta) = \big(Ax + Bu + f\big)(t,\theta), \\ & x(t_0, \theta) = x_0(\theta). \end{array} \right. \end{array} \right. \tag{5.1}$$

For the inner optimal control problem, we assume that Assumption 4.18 is fulfilled. The function $(x^*(\theta), u^*(\theta)) = \mathcal{L}_{\mathrm{l}}(\theta)$ is a time dependent function that is evaluated at different time points $t_i$ for the objective function of the nonlinear least squares problem. The residuals $r_k(\theta)$ in the nonlinear least squares

approach are given by

$$r_k(\theta) = \begin{pmatrix} x^*(t_k,\theta) \\ u^*(t_k,\theta) \end{pmatrix} - \xi_k$$

and we set $r(\theta)$ according to (3.3).

We now analyze the numerical solution of the Optimization Problem 5.1 when using a Gauss-Newton approach of the form (3.4) for this nonlinear least squares problem. As a first step we replace the inner optimal control problem by its necessary conditions (4.52).

Further, we denote by $J(\theta)$ the Jacobian of $r(\theta)$. For the computation of $J(\theta)$ we essentially need sensitivity information of the solution $(x^*(t_i,\theta), u^*(t_i,\theta))$, $i = 1,\dots,q$ of the necessary conditions (4.52). Computation of the sensitivities of the solution $(x^*(\theta), u^*(\theta))$ is discussed in Chapter 4.

Thus, the Gauss-Newton method for a given starting value $\theta_0$ reads as

$$J(\theta_k)^{\mathsf{H}} J(\theta_k)\Delta_k = -J^{\mathsf{H}}(\theta_k) r(\theta_k), \tag{5.2a}$$

$$\theta_{k+1} = \theta_k + \Delta_k, \tag{5.2b}$$

where in every new iteration a new solution $(x^*(\theta_k), u^*(\theta_k))$ and sensitivities $J(\theta_k)$ have to be computed.

In the sequel we want to analyze the convergence behavior of the Gauss-Newton method, when we only have approximations $\tilde{r}(\theta)$ and $\tilde{J}(\theta)$ of $r(\theta)$ and $J(\theta)$, respectively. Since the approximation errors $r_{\mathfrak{E}}(\theta)$ and $J_{\mathfrak{E}}(\theta)$ can in some way be controlled by the approximation methods, we can hope for estimates that give convergent solutions of the bilevel optimal control method with in some sense minimal computational effort.

We already noted in Section 4.7, that we can compute and control the approximation errors $r_{\mathfrak{E}}(\theta) = r(\theta) - \tilde{r}(\theta)$ and $J_{\mathfrak{E}}(\theta) = J(\theta) - \tilde{J}(\theta)$, where $\tilde{r}$ and $\tilde{J}$ are approximated values, e. g., by using the methods discussed in Section 4.7.

The next result shows that in this case the maximal allowed errors such that the Gauss-Newton method converges are essentially depending on the squared condition number of $\tilde{J}(\theta)$ and the residual $\tilde{J}(\theta)^{\mathsf{H}}\tilde{r}(\theta)$.

**Theorem 5.1.** *Let $\theta^* \in \Theta$ be a minimizer of Optimization Problem 5.1 and let the assumptions of Theorem 3.6 hold. Further, let $c \in \mathbb{R}$ be given such that $1 < c < \lambda/\sigma$. Choose $\eta_k$ and a fixed $\hat{\eta}$ such that*

$$0 \le \eta_k \le \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}.$$

*Then the bilevel nonlinear least squares method (5.2) converges to a (local) solution of the exact problem (5.1) if in iteration $k$ the errors $\|r_{\mathfrak{E}_k}\|_2$ and $\|J_{\mathfrak{E}_k}\|_2$ fulfill the inequality*

$$\begin{aligned}\left(\operatorname{cond}(\tilde{J}(\theta_k))^2 + 2\eta_k\right)\|\tilde{J}(\theta_k)\|_2\|r_{\mathfrak{E}_k}\|_2 + \\ \left(\operatorname{cond}(\tilde{J}(\theta_k))^2 + \eta_k \operatorname{cond}(\tilde{J}(\theta_k)) + 1\right)\|\tilde{r}(\theta_k)\|_2\|J_{\mathfrak{E}_k}\|_2 \\ \leq \eta_k\|\tilde{J}(\theta_k)^{\mathsf{H}}\tilde{r}(\theta_k)\|_2 + \mathcal{O}\Big(\left(\|r_{\mathfrak{E}_k}\|_2 + \|J_{\mathfrak{E}_k}\|_2\right)^2\Big). \quad \triangleright\end{aligned}$$

*Proof.* For brevity, we omit all dependencies on $\theta$ and the current iteration $k$, whenever this is clear from context. First note that with $e_k := -\tilde{J}(\theta_k)^{\mathsf{H}} r_{\mathfrak{E}}(\theta_k)$ we are exactly in the situation of Theorem 3.9. Now let us find bounds for the $\beta_k$ of Theorem 3.9. Note, that $\|\cdot\|_2 = \sigma_{\max}(\cdot)$. Then, for

$$\left\|\tilde{J}^{\mathsf{H}} r_{\mathfrak{E}}\right\|_2 \leq \beta_k \left\|\tilde{J}^{\mathsf{H}} r\right\|_2$$

to hold, it is sufficient that

$$\beta_k \geq \frac{\sigma_{\max}(\tilde{J})\|r_{\mathfrak{E}}\|_2}{\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J})\|r_{\mathfrak{E}}\|_2}.$$

Here we assume that $\|r_{\mathfrak{E}}\|_2$ is sufficiently small.

Now, we derive sufficient conditions for the inequality

$$0 \leq \beta_k \leq \frac{\eta_k \left\|J^{\mathsf{H}} r\right\|_2 - \left\|J^{\mathsf{H}} J (J^+ - \tilde{J}^+) r\right\|_2}{\left\|\tilde{J}^{\mathsf{H}} r\right\|_2 \left\|J^{\mathsf{H}} J (\tilde{J}^{\mathsf{H}}\tilde{J})^{-1}\right\|_2}.$$

Using the results of Lemmas 3.3 to 3.5 we have that

$$\begin{aligned}\left\|J^{\mathsf{H}} r\right\|_2 &\geq \|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J})\left\|r_{\mathfrak{E}}\right\|_2 - \|\tilde{r}\|_2\|J_{\mathfrak{E}}\|_2 - \|J_{\mathfrak{E}}\|_2\|r_{\mathfrak{E}}\|_2, \\ \left\|J^{\mathsf{H}} J\right\|_2 &\leq \sigma_{\max}(\tilde{J})^2 + 2\sigma_{\max}(\tilde{J})\|J_{\mathfrak{E}}\|_2 + \|J_{\mathfrak{E}}\|_2^2, \\ \left\|\tilde{J}^{\mathsf{H}} r\right\|_2 &\leq \|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 + \sigma_{\max}(\tilde{J})\|r_{\mathfrak{E}}\|_2, \\ \|r\|_2 &\leq \|\tilde{r}\|_2 + \|r_{\mathfrak{E}}\|_2, \\ \|J^+ - \tilde{J}^+\|_2 &\leq \frac{\sigma_{\min}(\tilde{J})^{-2}\|J_{\mathfrak{E}}\|_2}{1 - \sigma_{\min}(\tilde{J})^{-1}\|J_{\mathfrak{E}}\|_2}.\end{aligned}$$

Thus it is sufficient that $\beta_k$ fulfills

$$\begin{aligned}
&\frac{\sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2}{\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2} \\
&\leq \beta_k \\
&\leq \Bigg(\eta_k \left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J}) \left\|r_{\mathfrak{E}}\right\|_2 - \|\tilde{r}\|_2\|J_{\mathfrak{E}}\|_2 - \|J_{\mathfrak{E}}\|_2\|r_{\mathfrak{E}}\|_2\right) \\
&\qquad - \left(\sigma_{\max}(\tilde{J})^2 + 2\sigma_{\max}(\tilde{J})\,\|J_{\mathfrak{E}}\|_2 + \|J_{\mathfrak{E}}\|_2^2\right)\left(\|\tilde{r}\|_2 + \|r_{\mathfrak{E}}\|_2\right)\frac{\sigma_{\min}(\tilde{J})^{-2}\|J_{\mathfrak{E}}\|_2}{1-\sigma_{\min}(\tilde{J})^{-1}\|J_{\mathfrak{E}}\|_2}\Bigg)\cdot \\
&\qquad \left(\left(\sigma_{\max}(\tilde{J})^2 + 2\sigma_{\max}(\tilde{J})\,\|J_{\mathfrak{E}}\|_2 + \|J_{\mathfrak{E}}\|_2^2\right)\left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 + \sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2\right)\sigma_{\min}(\tilde{J})^{-2}\right)^{-1}.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
&\left(\sigma_{\max}(\tilde{J})^2 + 2\sigma_{\max}(\tilde{J})\,\|J_{\mathfrak{E}}\|_2 + \|J_{\mathfrak{E}}\|_2^2\right)\left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 + \sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2\right)\sigma_{\min}(\tilde{J})^{-2}\cdot \\
&\qquad \sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2\left(1-\sigma_{\min}(\tilde{J})^{-1}\|J_{\mathfrak{E}}\|_2\right) \\
&\qquad \leq \left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J})\,\|r_{\mathfrak{E}}\|_2\right)\cdot \\
&\qquad \left(\left(1-\sigma_{\min}(\tilde{J})^{-1}\|J_{\mathfrak{E}}\|_2\right)\cdot \right. \\
&\qquad \eta_k\left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2 - \sigma_{\max}(\tilde{J})\left\|r_{\mathfrak{E}}\right\|_2 - \|\tilde{r}\|_2\|J_{\mathfrak{E}}\|_2 - \|J_{\mathfrak{E}}\|_2\|r_{\mathfrak{E}}\|_2\right) \\
&\qquad \left. - \left(\sigma_{\max}(\tilde{J})^2 + 2\sigma_{\max}(\tilde{J})\,\|J_{\mathfrak{E}}\|_2 + \|J_{\mathfrak{E}}\|_2^2\right)\left(\|\tilde{r}\|_2 + \|r_{\mathfrak{E}}\|_2\right)\sigma_{\min}(\tilde{J})^{-2}\|J_{\mathfrak{E}}\|_2\right).
\end{aligned}$$

And, after neglecting all higher order terms, we deduce that

$$\begin{aligned}
&\left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2\sigma_{\min}(\tilde{J})^{-2}\sigma_{\max}(\tilde{J})^3 + 2\eta_k\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2\sigma_{\max}(\tilde{J})\right)\|r_{\mathfrak{E}}\|_2 + \\
&\quad \left(\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2\sigma_{\min}(\tilde{J})^{-2}\sigma_{\max}(\tilde{J})^2\|\tilde{r}\|_2 + \eta_k\sigma_{\min}(\tilde{J})^{-1}\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2^2 + \|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2\|\tilde{r}\|_2\right)\|J_{\mathfrak{E}}\|_2 \\
&\qquad\qquad \dot{\leq} \eta_k\|\tilde{J}^{\mathsf{H}}\tilde{r}\|_2^2.
\end{aligned}$$

Thus, since $\mathrm{cond}(\cdot) = \frac{\sigma_{\max}(\cdot)}{\sigma_{\min}(\cdot)}$ the assertion follows. □

*Remark* 5.1. We discuss conditions, under which the assumptions of Theorem 5.1 on $J(\theta)$ and $r(\theta)$ are fulfilled.

Since $J(\theta)$ essentially needs evaluations of the sensitivities $x_\theta$, it is sufficient to look at the corresponding statements for $x_\theta$. Following the discussion of Lemmas 4.1 and 4.11 and Theorem 3.27, we conclude that if the coefficients $E, A, f$ are sufficiently smooth, then $J(\theta)$ is differentiable with respect to $\theta$ and thus, also (locally) Lipschitz continuous.

Condition 3 of Theorem 3.6 requires that in a neighborhood of $\theta^*$ we have

$$\|(J(\theta^*) - J(\theta))r(\theta^*)\|_2 \leq \sigma\|\theta^* - \theta\|_2 \tag{5.3}$$

for sufficiently small $\sigma$. By the submultiplicativity of the 2-norm this condition is fulfilled if

$$\|J(\theta^*) - J(\theta)\|_2\|r(\theta^*)\|_2 \leq \sigma\|\theta^* - \theta\|_2.$$

On the other hand, since $J(\theta)$ is locally Lipschitz continuous with constant $\gamma_J$, we know that

$$\|J(\theta^*) - J(\theta)\|_2\|r(\theta^*)\|_2 \leq \gamma_J\|r(\theta^*)\|_2\|\theta^* - \theta\|_2.$$

Thus, (5.3) is fulfilled, if $\|r(\theta^*)\|_2$ is sufficiently small. If all $\xi_i$ correspond to evaluations of a solution $x^*(\theta^*)$, then $r(\theta^*) = 0$. Since all quantities are continuous, condition (5.3) is still fulfilled, if the $\xi_i$ are sufficiently close to a solution $x^*(\theta^*)$ at the corresponding time points $t_i$.

We have seen in Example 4.7, that the sensitivities may be 0. In such a case, the condition of $J(\theta)$ having full column rank is violated and Theorem 5.1 cannot be applied.

Following the discussion in [GLN07], we note that Theorem 5.1 requires, that also the perturbed quantity $\tilde{J}(\theta^*)r(\theta^*)$ has to fulfill $\tilde{J}(\theta^*)r(\theta^*) = 0$ at the optimal solution $\theta^*$ of the *exact* problem (5.1). This is always guaranteed, if we let $J_{\mathfrak{E}_k} \to 0$ and thus also $r_{\mathfrak{E}_k} \to 0$ as $k \to \infty$. ▷

We can use Theorem 5.1 to control the maximum of the errors $\|r_{\mathfrak{E}_k}\|_2$ and $\|J_{\mathfrak{E}_k}\|_2$.

**Corollary 5.2.** *Let $\theta^* \in \Theta$ be a minimizer of Optimization Problem 5.1 and let the assumptions of Theorem 3.6 hold. Further, let $c \in \mathbb{R}$ be given such that $1 < c < \lambda/\sigma$. Choose $\eta_k$ and a fixed $\hat{\eta}$ such that*

$$0 \leq \eta_k \leq \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}.$$

*Then the bilevel nonlinear least squares method* (5.2) *converges to a (local) solution of the exact problem* (5.1) *if in iteration k the error* $\max(\|r_{\mathfrak{E}_k}\|_2, \|J_{\mathfrak{E}_k}\|_2)$ *fulfills the inequality*

$$\max(\|r_{\mathfrak{E}_k}\|_2, \|J_{\mathfrak{E}_k}\|_2) \leq \frac{\eta_k \|\tilde{J}(\theta_k)^{\mathsf{H}} \tilde{r}(\theta_k)\|_2}{\left(\operatorname{cond}(\tilde{J}(\theta_k))^2 + 2\eta_k\right)\|\tilde{J}(\theta_k)\|_2 + \left(\operatorname{cond}(\tilde{J}(\theta_k))^2 + \eta_k \operatorname{cond}(\tilde{J}(\theta_k)) + 1\right)\|\tilde{r}(\theta_k)\|_2} + \mathcal{O}\left(\left(\|r_{\mathfrak{E}_k}\|_2 + \|J_{\mathfrak{E}_k}\|_2\right)^2\right). \quad (5.4)$$

▷

*Proof.* The assertion is an immediate consequence of Theorem 5.1. □

*Remark* 5.2. Inequality (5.4) shows, that the maximal allowed error directly depends on the condition number $\operatorname{cond}(\tilde{J}(\theta_k))^2$ and the norm of the residual $\|\tilde{J}(\theta_k)^{\mathsf{H}} \tilde{r}(\theta_k)\|_2$. In particular, we can choose less accurate solutions if we are far away from the optimal point, or if we are interested in less accurate solutions. This also justifies the use of the linear differential-algebraic setting, which often arises from linearization process around a solution of a nonlinear differential-algebraic and which is prone to an approximation error, see e. g., [KM06] for more details. ▷

**Example 5.1** (Example 4.5 continued)**.** Let us consider the optimal control problem of Example 4.5 with corresponding sensitivity equations and solution (4.61). For the given time points $t_0 = 0, t_1 = 1, t_2 = 1.3, t_f = t_3 = 2$ we set the corresponding data points $\xi_i$ to

$$\xi_i = \begin{pmatrix} x \\ u \end{pmatrix}(t_i, 2), \quad i = 0, \ldots, 3.$$

We want to compute a (local) solution of the bilevel optimization problem (5.1), which is given by $\theta^* = 2$. Since we know the correct solution, we can compute $\lambda_{\min}\left(J(\theta^*)^{\mathsf{H}} J(\theta^*)\right) = 0.37$. To accommodate for the uncertainty in the remaining unknown quantities in the definition of $\eta_k$ and $\hat{\eta}$ we choose $\hat{\eta} = 0.01$ and $\eta_k = 0.9\hat{\eta}$.

As starting value we choose $\theta_0 = 1$. If we start by prescribing a rather large absolute tolerance of 0.1, then the evolution of solution $\theta_k$ and corresponding

Table 5.1.: Evolution of the solution $\theta_k$, corresponding residual norm $\|r_k\|_2$, Jacobian norm $\|J_k\|_2$, norm of the Gauss-Newton residual $\|J_k^T r_k\|_2$, and new prescribed absolute tolerance $\nu_{a,k}$ of Example 5.1 according to (5.4) for an initial absolute tolerance $\nu_{a,0} = 10^{-1}$.

| $k$ | $\theta_k$ | $\|r_k\|_2$ | $\|J_k\|_2$ | $\|J_k^T r_k\|_2$ | $\nu_{a,k}$ |
|---|---|---|---|---|---|
| 0 | 1.00 | $6.47 \cdot 10^{-1}$ | $1.03 \cdot 10^{0}$ | $6.66 \cdot 10^{-1}$ | $1.00 \cdot 10^{-1}$ |
| 1 | 1.63 | $1.71 \cdot 10^{-1}$ | $5.59 \cdot 10^{-1}$ | $9.55 \cdot 10^{-2}$ | $2.33 \cdot 10^{-2}$ |
| 2 | 1.93 | | | | $8.45 \cdot 10^{-3}$ |

residual terms $\|r_k\|_2$, $\|J_k\|_2$, and $\|J_k^T r_k\|_2$ with new prescribed absolute tolerance according to (5.4) are described in Table 5.1. For a smaller prescribed tolerance of $10^{-6}$, the evolution of solution $\theta_k$ and corresponding residual terms $\|r_k\|_2$, $\|J_k\|_2$, and $\|J_k^T r_k\|_2$ with new prescribed absolute tolerance according to (5.4) are described in Table 5.2.

Note, that in both cases the chosen absolute tolerances at each iteration according to (5.4) are decreased after the first step. Also, the further away from the solution $\theta^* = 2$ we are, the larger the absolute tolerance is allowed to be. Vice versa, when the iterates $\theta_k$ get close to the optimal solution $\theta^*$ we need much smaller absolute errors in the computation of $r(\theta)$ and $J(\theta)$. Also, the prescribed tolerance gets small, because eventually $\|r_k\|_2$ gets small, and not $\|J_k\|_2$. Further note, that for both runs the first 3 iterations lead to similar solutions $\theta_k$ and residual norms $\|r_k\|_2$. The first run stops after 2 iterations, as the Gauss-Newton residual $\|J_k^T r_k\|_2$ is below the prescribed absolute tolerance. The second run needs 5 more iterations for convergence. ▷

Table 5.2.: Evolution of the solution $\theta_k$, corresponding residual norm $\|r_k\|_2$, Jacobian norm $\|J_k\|_2$, norm of the Gauss-Newton residual $\|J_k^T r_k\|_2$, and new prescribed absolute tolerance $\nu_{\mathrm{a},k}$ of Example 5.1 according to (5.4) for an initial absolute tolerance $\nu_{\mathrm{a},0} = 10^{-6}$.

| $k$ | $\theta_k$ | $\|r_k\|_2$ | $\|J_k\|_2$ | $\|J_k^T r_k\|_2$ | $\nu_{\mathrm{a},k}$ |
|---|---|---|---|---|---|
| 0 | 1.0000000 | $6.472 \cdot 10^{-1}$ | $1.032 \cdot 10^{0}$ | $6.660 \cdot 10^{-1}$ | $1.000 \cdot 10^{-6}$ |
| 1 | 1.6256578 | $1.654 \cdot 10^{-1}$ | $5.456 \cdot 10^{-1}$ | $9.022 \cdot 10^{-2}$ | $2.550 \cdot 10^{-3}$ |
| 2 | 1.9286950 | $2.623 \cdot 10^{-2}$ | $3.821 \cdot 10^{-1}$ | $1.002 \cdot 10^{-2}$ | $9.146 \cdot 10^{-4}$ |
| 3 | 1.9973584 | $9.540 \cdot 10^{-4}$ | $3.523 \cdot 10^{-1}$ | $3.360 \cdot 10^{-4}$ | $2.043 \cdot 10^{-4}$ |
| 4 | 2.0000661 | $2.157 \cdot 10^{-5}$ | $3.507 \cdot 10^{-1}$ | $7.554 \cdot 10^{-6}$ | $8.388 \cdot 10^{-6}$ |
| 5 | 2.0000046 | $1.644 \cdot 10^{-6}$ | $3.507 \cdot 10^{-1}$ | $5.705 \cdot 10^{-7}$ | $1.904 \cdot 10^{-7}$ |
| 6 | 2.0000000 | | | | $1.438 \cdot 10^{-8}$ |

# Part II.

# Other Parameter Optimizations

In the second part we investigate problems, that do not fit into the multilevel optimal control framework of Definition 2.2. Nevertheless, they have in common, that they contain at least two levels of optimization in the context of optimal control problems.

In contrast to Part I of this thesis we restrict ourselves to linear time-invariant control systems with dynamics described by ordinary differential equations (ODEs) and their discrete-time counterparts. The system coefficients given by $\mathcal{M} = \{A, B, C, D\}$ or $\mathcal{M} = \{A, B, Q, S, R\}$ represent the continuous-time systems

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx + Du, \end{aligned}$$

with outputs and

$$\dot{x} = Ax + Bu,$$

respectively, where in the latter case an optimal control problem with infinite time-horizon and objective function of the form

$$\mathcal{J}(x,u) \coloneqq \int_{t_0}^{\infty} \begin{pmatrix} x \\ u \end{pmatrix}^{\mathsf{H}} \begin{bmatrix} Q & S \\ S^{\mathsf{H}} & R \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \mathrm{d}t \tag{II.1}$$

similar to (I.2) with time and parameter independent matrices $Q, S$, and $R$ is associated.

In Chapter 6 we derive formulas for the analytic center of the solution set of linear matrix inequalities (LMIs) defining passive transfer functions. In the context of this chapter, the matrix function

$$W(X,\mathcal{M}) = \begin{bmatrix} -A^{\mathsf{H}}X - XA & C^{\mathsf{H}} - XB \\ C - B^{\mathsf{H}}X & D + D^{\mathsf{H}} \end{bmatrix}, \qquad X = X^{\mathsf{H}},$$

plays an important role. The analytic center is defined as the minimizer of the minimization problem

$$\min_X \quad -\log\det W(X,\mathcal{M}) \tag{II.2}$$

and proves to be a good choice regarding computation of the passivity radius, in particular see [BMV19], which can be defined as the solution to the bilevel optimization problem

$$\rho = \sup_X \inf_{\Delta_{\mathcal{M}}} \{\|\Delta_{\mathcal{M}}\| \mid \det W(X,\mathcal{M}+\Delta_{\mathcal{M}}) = 0\}. \tag{II.3}$$

Assuming that the supremum and the infimum are actually attained, the definition of $\rho$ indeed fits Definition 2.1. This can be seen by setting $\mathbb{Z}_1 = \{X \in \mathbb{K}^{n\times n} \mid X = X^{\mathsf{H}}\}$, $\mathbb{Z}_2 = \mathbb{K}^{(n+m)\times(n+m)}$, $\mathcal{G}_2(X,\Delta_{\mathcal{M}}) = \det(W(X,\mathcal{M}+\Delta_{\mathcal{M}}))$, and the objective functions $\mathcal{K}_1(X,\Delta^*_{\mathcal{M}}(X)) = -\|\Delta^*_{\mathcal{M}}(X)\|$, $\mathcal{K}_2(X,\Delta_{\mathcal{M}}) = \|\Delta_M\|$. Instead of solving the bilevel optimization (II.3) we can replace the upper level optimization by Optimization Problem II.2, which is independent of a concrete choice of $\Delta_{\mathcal{M}}$. We show why this is reasonable in Section 6.4.

In Chapter 7 we return to the computation of algebraic Riccati equations in the more general context of optimal control with system coefficients $\mathcal{M} = \{A,B,Q,S,R\}$. Solutions of the algebraic Riccati equations can be used to compute a stabilizing feedback $F$ such that the resulting closed-loop matrix $A-BF$ has all eigenvalues in the left complex half plane. We consider robust stabilization problems, where we additionally require that the absolute values of the real parts of the eigenvalues of $A-BF$ are less than some $\varrho > 0$.

In the context of those robust stabilization problems certain robustness criteria were introduced in [MX00] subject to optimization with respect to the weight matrices $Q,S$, and $R$. We show how a certain condition number appearing in the computation of a solution of the algebraic Riccati equation can be parametrized in terms of the weights $Q,S$, and $R$ and how this may improve the overall performance of the optimization of the robustness criteria introduced in [MX00].

# 6. The Analytic Center of the Passivity Linear Matrix Inequality

This chapter including Appendices A and B is essentially a copy of the article [BMNV20b]. We consider realizations of linear dynamical systems that are denoted as *positive real or passive* and their associated transfer functions. In particular, we study positive transfer functions which play a fundamental role in systems and control theory: they represent e. g., spectral density functions of stochastic processes, show up in spectral factorizations, are the Hermitian part of a positive real transfer function, characterize port-Hamiltonian systems, and are also related to algebraic Riccati equations.

Positive transfer functions form a convex set, and this property has led to the extensive use of convex optimization techniques in this area (especially for so-called linear matrix inequalities (LMIs) [BEFB94]). In order to optimize a certain scalar function $f(X)$ over a convex set, one often defines a barrier function $b(X)$ that becomes infinite near the boundary of the set, and then finds the minimum of $c \cdot f(X) + b(X)$, $c \geq 0$, as $c \to +\infty$. These minima (which are functions of the parameter $c$) are called the points of the *central path*. The starting point of this path ($c = 0$) is called the *analytic center* of the set. Notice that the analytic center depends as well on the barrier function as on the corresponding convex set.

In this work we present an explicit set of equations that define the analytic center of the solution set of the LMI defining a passive transfer function. We also show how these equations relate to the algebraic Riccati equations that typically arise in the spectral factorization of transfer functions. We discuss transfer functions both on the imaginary axis (i. e., the continuous-time case), as well as on the unit circle (i. e., the discrete-time case). In the continuous-time setting the transfer function arises from the *Laplace transform* of the system

$$\begin{aligned} \dot{x} &= Ax + Bu,\ x(0) = 0, \\ y &= Cx + Du, \end{aligned} \tag{6.1}$$

where $u : \mathbb{R} \to \mathbb{C}^m$, $x : \mathbb{R} \to \mathbb{C}^n$, and $y : \mathbb{R} \to \mathbb{C}^m$ are vector-valued functions denoting, respectively, the *input*, *state*, and *output* of the system. Denoting real and complex $n$-vectors ($n \times m$ matrices) by $\mathbb{R}^n$, $\mathbb{C}^n$ ($\mathbb{R}^{n\times m}$, $\mathbb{C}^{n\times m}$), respectively, the coefficient matrices satisfy $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{m\times n}$, and $D \in \mathbb{C}^{m\times m}$.

In the discrete-time setting the transfer function arises from the *z-transform* applied to the system

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k,\ x_0 = 0, \\ y_k &= Cx_k + Du_k, \end{aligned}$$

with state, input, and output sequences $\{x_k\}$, $\{u_k\}$, $\{y_k\}$. In both cases, we usually denote these systems by four-tuples of matrices $\mathcal{M} := \{A, B, C, D\}$ and the associated transfer functions by

$$\mathcal{T}_c(s) := D + C(sI_n - A)^{-1}B, \qquad \mathcal{T}_d(z) := D + C(zI_n - A)^{-1}B, \tag{6.2}$$

respectively.

We restrict ourselves to systems which are *minimal*, i. e., the pair $(A, B)$ is *controllable* (for all $\lambda \in \mathbb{C}$, $\operatorname{rk}[\lambda I - A\ B] = n$), and the pair $(A, C)$ is *reconstructable* (i. e., $(A^{\mathsf{H}}, C^{\mathsf{H}})$ is controllable). Here, the conjugate transpose (transpose) of a vector or matrix $V$ is denoted by $V^{\mathsf{H}}$ ($V^{\mathsf{T}}$) and the identity matrix is denoted by $I_n$ or $I$ if the dimension is clear. We furthermore require that input and output port dimensions are equal to $m$ and assume that $\operatorname{rk} B = \operatorname{rk} C = m$.

*Passive* systems and their relationships with *positive-real transfer functions* are well studied, starting with the works [Kal63; Pop73; Wil71; Wil72a; Wil72b; Yak62] and the topic has recently received a revival in the work on *port-Hamiltonian (pH) systems*, [MV20; Sch04; SJ14]. For a summary of the relationships see [BMV19; Wil71], where also the characterization of passivity via the solution set of an associated LMI is highlighted.

The chapter is organized as follows. After some preliminaries in Section 6.1, in Section 6.2 we study the analytic centers of the solution sets of LMIs associated with the continuous- and discrete-time case. In Section 6.3 we discuss numerical methods to compute the analytic centers using steepest descent as well as Newton methods and show that the analytic centers can be computed efficiently. In Section 6.4, lower bounds for the distance to non-passivity (the passivity radius) are derived using smallest eigenvalues of the Hermitian matrices associated with the LMIs evaluated at the analytic center. The results are illustrated with some

simple examples where the analytic center can be calculated analytically. In Appendix A we derive formulas for the computation of the gradients and the Hessian of the functions that we optimize and in Appendix B we clarify some of the differences that arise between the continuous- and the discrete-time case.

## 6.1. Preliminaries

Throughout this chapter we use the following notation. We denote the set of Hermitian matrices in $\mathbb{C}^{n\times n}$ by $\mathbb{H}_n$. Positive definiteness (semidefiniteness) of $A \in \mathbb{H}_n$ is denoted by $A \succ 0$ ($A \succeq 0$). For a positive semi-definite matrix $A$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote, respectively, the smallest and largest eigenvalues of $A$. The real and imaginary parts of a complex matrix $Z$ are written as $\Re(Z)$ and $\Im(Z)$, respectively, and $\imath$ is the imaginary unit. We consider functions over $\mathbb{H}_n$, which is a vector space if considered as a *real* subspace of $\mathbb{R}^{n\times n} + \imath\mathbb{R}^{n\times n}$. We identify $\mathbb{C}^{m\times n}$ with $\mathbb{R}^{m\times n} + \imath\mathbb{R}^{m\times n}$, but we note that this has implications when one is carrying out differentiations, see Appendix A. The *Frobenius scalar product* for matrices $X, Y \in \mathbb{R}^{n\times n} + \imath\mathbb{R}^{n\times n}$ is given by

$$\langle X, Y\rangle_{\mathbb{R}} := \Re(\operatorname{tr}(Y^{\mathsf{H}}X)) = \operatorname{tr}(Y_r^T X_r + Y_i^T X_i),$$

where we have partitioned $X, Y$ as $X = X_r + \imath X_i$, $Y = Y_r + \imath Y_i$ with real and imaginary parts in $\mathbb{R}^{n\times n}$. As we are mainly concerned with this scalar product, we drop the subscript $\mathbb{R}$. We make frequent use of the following properties of this inner product given by

$$\langle X, Y\rangle = \langle Y, X\rangle, \ \|X\|_F = \langle X, X\rangle^{\frac{1}{2}}, \ \langle X, YZ\rangle = \left\langle Y^{\mathsf{H}}X, Z\right\rangle = \left\langle XZ^{\mathsf{H}}, Y\right\rangle.$$

The concepts of *positive-realness* and *passivity* are well studied. In the following subsections we briefly recall some important properties following [GNV99; Wil71], where we repeat a few observations from [BMV19]. See also [Wil71] for a more detailed survey.

### 6.1.1. Positive-realness and Passivity, Continuous-Time

Consider a continuous-time system $\mathcal{M}$ as in (6.1) and the transfer function $\mathcal{T}_c$ as in (6.2). The transfer function $\mathcal{T}_c(s)$ is called *positive real* if the matrix-valued

rational function

$$\Phi_c(s) \coloneqq \mathcal{T}_c^{\mathsf{H}}(-s) + \mathcal{T}_c(s)$$

is positive semidefinite for $s$ on the imaginary axis, i. e., $\Phi_c(\imath\omega) \succeq 0$ for all $\omega \in \mathbb{R}$ and it is called *strictly positive real* if $\Phi_c(\imath\omega) \succ 0$ for all $\omega \in \mathbb{R}$.

We associate with $\Phi_c$ a system pencil

$$S_c(s) \coloneqq \left[\begin{array}{cc|c} 0 & A - sI_n & B \\ A^{\mathsf{H}} + sI_n & 0 & C^{\mathsf{H}} \\ \hline B^{\mathsf{H}} & C & R \end{array}\right], \tag{6.3}$$

where here and in the following $R$ is an abbreviation for $R \coloneqq D + D^{\mathsf{H}}$. Also, equation (6.3) has a Schur complement which is the transfer function $\Phi_c(s)$ and, under the condition of minimality, the finite generalized eigenvalues of $S_c(s)$ are the finite zeros of $\Phi_c(s)$.

For $X \in \mathbb{H}_n$ we introduce the matrix function

$$W_c(X) \coloneqq \left[\begin{array}{cc} -XA - A^{\mathsf{H}}X & C^{\mathsf{H}} - XB \\ C - B^{\mathsf{H}}X & R \end{array}\right]. \tag{6.4}$$

If $\mathcal{T}_c(s)$ is positive real, then the LMI

$$W_c(X) \succeq 0 \tag{6.5}$$

has a solution $X \in \mathbb{H}_n$ and we have the sets

$$\mathbb{X}_c^{\succ} \coloneqq \left\{ X \in \mathbb{H}_n \,\middle|\, W_c(X) \succeq 0,\ X \succ 0 \right\},$$
$$\mathbb{X}_c^{\succ\succ} \coloneqq \left\{ X \in \mathbb{H}_n \,\middle|\, W_c(X) \succ 0,\ X \succ 0 \right\}.$$

An important subset of $\mathbb{X}_c^{\succ}$ are those solutions to (6.5) for which the rank $r$ of $W_c(X)$ is minimal (i. e., for which $r = \operatorname{rk}\Phi_c(s)$). If $R$ is invertible, then the minimum rank solutions in $\mathbb{X}_c^{\succ}$ are those for which $\operatorname{rk} W_c(X) = \operatorname{rk}(R) = m$, which in turn is the case if and only if the Schur complement of $R$ in $W_c(X)$ is zero. This Schur complement is associated with the continuous-time *algebraic Riccati equation (ARE)*

$$\mathsf{Ricc}_c(X) \coloneqq -XA - A^{\mathsf{H}}X - (C^{\mathsf{H}} - XB)R^{-1}(C - B^{\mathsf{H}}X) = 0. \tag{6.7}$$

Solutions $X$ to (6.7) produce a spectral factorization of $\Phi_c(s)$, and each solution corresponds to a *Lagrangian invariant subspace* spanned by the columns of $U_c := \begin{bmatrix} I_n & -X^\mathsf{T} \end{bmatrix}^\mathsf{T}$ that remains invariant under the action of the *Hamiltonian matrix*

$$\mathcal{H}_c := \begin{bmatrix} A - BR^{-1}C & -BR^{-1}B^\mathsf{H} \\ C^\mathsf{H}R^{-1}C & -(A - BR^{-1}C)^\mathsf{H} \end{bmatrix}, \tag{6.8}$$

i. e., $U_c$ satisfies $\mathcal{H}_c U_c = U_c A_{F_c}$ for a *closedhyphenloop matrix* $A_{F_c} = A - BF_c$ with $F_c := R^{-1}(C - B^\mathsf{H}X)$ (see e.g., [FMX02]). Each solution $X$ of (6.7) can also be associated with an *extended Lagrangian invariant subspace* for the pencil $S_c(s)$ (see [BLMV15]), spanned by the columns of $\widehat{U}_c := \begin{bmatrix} -X^\mathsf{T} & I_n & -F_c^\mathsf{T} \end{bmatrix}^\mathsf{T}$. In particular, $\widehat{U}_c$ satisfies

$$\begin{bmatrix} 0 & A & B \\ A^\mathsf{H} & 0 & C^\mathsf{H} \\ B^\mathsf{H} & C & R \end{bmatrix} \widehat{U}_c = \begin{bmatrix} 0 & I_n & 0 \\ -I_n & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \widehat{U}_c A_{F_c}.$$

The sets $\mathbb{X}_c^{\succ}, \mathbb{X}_c^{\succcurlyeq}$ are related to the concepts of *passivity and strict passivity* see [Wil71]. If for the system $\mathcal{M} := \{A, B, C, D\}$ of (6.3) the LMI (6.5) has a solution $X \in \mathbb{X}_c^{\succ}$ then $\mathcal{M}$ is *(Lyapunov) stable* (i. e., all eigenvalues are in the closed left half plane with any eigenvalues occurring on the imaginary axis being semisimple), and *passive*, and if there exists a solution $X \in \mathbb{X}_c^{\succcurlyeq}$ then $\mathcal{M}$ is *asymptotically stable* (i. e., all eigenvalues are in the open left half plane) and *strictly passive*. Furthermore, if $\mathcal{M}$ is passive, then there exist maximal and minimal solutions $X_- \preccurlyeq X_+$ of (6.5) in $\mathbb{X}_c^{\succ}$ such that all solutions $X$ of $W_c(X) \succcurlyeq 0$ satisfy

$$0 \prec X_- \preccurlyeq X \preccurlyeq X_+,$$

which implies that $\mathbb{X}_c^{\succ}$ is bounded. For more details on the different concepts discussed in this section, see the extended preprint version of [BMV19].

### 6.1.2. Positive-Realness and Passivity, Discrete-Time

For each of the results of the previous subsection there are discrete-time versions which we briefly recall in this section, see [IOW99; Pop73]. Note, that these results can be obtained by applying a bilinear transform (see Appendix B) to the continuous-time counterparts.

The transfer function $\mathcal{T}_d(z)$ in (6.2) is called *positive real* if the matrix-valued rational function

$$\Phi_d(z) \coloneqq \mathcal{T}_d^{\mathsf{H}}(z^{-1}) + \mathcal{T}_d(z)$$

satisfies $\Phi_d(e^{\imath\omega}) = \Phi_d^{\mathsf{H}}(e^{\imath\omega}) \succeq 0$ for $0 \le \omega \le 2\pi$, and it is called *strictly positive real* if $\Phi_d(e^{\imath\omega}) \succ 0$ for $0 \le \omega \le 2\pi$.

We consider an associated the matrix function

$$W_d(X) = \begin{bmatrix} X - A^{\mathsf{H}}XA & C^{\mathsf{H}} - A^{\mathsf{H}}XB \\ C - B^{\mathsf{H}}XA & R - B^{\mathsf{H}}XB \end{bmatrix},$$

where again $R = D + D^{\mathsf{H}}$, the sets

$$\mathbb{X}_d^{\succeq} \coloneqq \left\{ X \in \mathbb{H}_n \,\middle|\, W_d(X) \succeq 0,\ X \succ 0 \right\},$$
$$\mathbb{X}_d^{\succ} \coloneqq \left\{ X \in \mathbb{H}_n \,\middle|\, W_d(X) \succ 0,\ X \succ 0 \right\},$$

and the system pencil

$$S_d(z) = \left[\begin{array}{cc|c} 0 & A - zI_n & B \\ zA^{\mathsf{H}} - I_n & 0 & C^{\mathsf{H}} \\ \hline zB^{\mathsf{H}} & C & R \end{array}\right]$$

whose Schur complement is $\Phi_d(z)$.

If the system is positive real then, see [Wil71], there exists $X \in \mathbb{H}_n$ such that $W_d(X) \succeq 0$. If $W_d(X) \succeq 0$, a transfer function $\mathcal{T}_d(z) \coloneqq C(zI_n - A)^{-1}B + D$ is called *passive* and if $W_d(X) \succ 0$ it is said to be *strictly passive*. We again have an associated discrete-time Riccati equation defined as

$$\mathsf{Ricc}_d(X) \coloneqq -A^{\mathsf{H}}XA + X - (C^{\mathsf{H}} - A^{\mathsf{H}}XB)(R - B^{\mathsf{H}}XB)^{-1}(C - B^{\mathsf{H}}XA) = 0 \quad (6.10)$$

from which one directly obtains a spectral factorization of $\Phi_d(z)$. The solutions of the discrete-time Riccati equation can be obtained by computing a Lagrangian invariant subspace spanned by the columns of $U_d \coloneqq \begin{bmatrix} I_n & -X^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ of the *symplectic matrix*

$$\mathcal{S}_d \coloneqq \begin{bmatrix} I & BR^{-1}B^{\mathsf{H}} \\ 0 & A^{\mathsf{H}} - C^{\mathsf{H}}R^{-1}B^{\mathsf{H}} \end{bmatrix}^{-1} \begin{bmatrix} A - BR^{-1}C & 0 \\ C^{\mathsf{H}}R^{-1}C & I \end{bmatrix},$$

satisfying $\mathcal{S}_d U_d = U_d A_{F_d}$, where $A_{F_d} \coloneqq A - BF_d$ with $F_d \coloneqq (R - B^{\mathsf{H}} X B)^{-1}(C - B^{\mathsf{H}} X A)$.

Each solution $X$ of (6.10) can also be associated with an *extended Lagrangian invariant subspace* for the pencil $S_d(z)$ (see [BLMV15]), spanned by the columns of $\widehat{U}_d \coloneqq \begin{bmatrix} -X^{\mathsf{T}} & I_n & -F_d^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$. In particular, $\widehat{U}_d$ satisfies

$$\begin{bmatrix} 0 & A & B \\ I_n & 0 & C^{\mathsf{H}} \\ 0 & C & R \end{bmatrix} \widehat{U}_d = \begin{bmatrix} 0 & I_n & 0 \\ A^{\mathsf{H}} & 0 & 0 \\ B^{\mathsf{H}} & 0 & 0 \end{bmatrix} \widehat{U}_d A_{F_d}.$$

Again, if the system is passive, then there exist maximal and minimal solutions $X_- \preceq X_+$ in $\mathbb{X}_d^{\succ}$, such that all solutions $X$ of $W_d(X) \succeq 0$ satisfy

$$0 \prec X_- \preceq X \preceq X_+,$$

which implies that $\mathbb{X}_d^{\succ}$ is bounded.

## 6.2. The Analytic Center

If the sets $\mathbb{X}_c^{\succ}$, $\mathbb{X}_d^{\succ}$ in (6.6), respectively (6.9), are non-empty, then we can define their respective analytic center. Following the discussion in [GNV99], we first consider the continuous-time case, the discrete-time case is derived in an analogous way. We choose a scalar barrier function

$$b(X) \coloneqq -\ln \det W_c(X), \quad X \in \mathbb{H}_n,$$

which is bounded from below but becomes infinitely large when $W_c(X)$ becomes singular. We define the analytic center of the domain $\mathbb{X}_c^{\succ}$ as the minimizer of this barrier function.

### 6.2.1. The Continuous-Time Case

Since $\mathbb{X}_c^{\succ}$ is non-empty, $R$ is invertible and the Riccati equation $\mathrm{Ricc}_c(X) = 0$ in (6.7) is well defined. Their solutions $X_+$ and $X_-$ are both on the boundary of $\mathbb{X}_c^{\succ}$, and hence are not in $\mathbb{X}_c^{\succ}$. Since we assume that $\mathbb{X}_c^{\succ}$ is non-empty, the analytic center is well defined, see, e. g., Section 4.2 in [Nes04].

To characterize the analytic center, we need to consider the variation of the *gradient* $b_X$ of the barrier function $b$ at point $X$ along a direction $\Delta_X \in \mathbb{H}_n$. As explained in Appendix A, this is equal to

$$-\left\langle W_c(X)^{-1}, \Delta W_c(X)[\Delta_X] \right\rangle, \tag{6.11}$$

where $b_X = -W_c(X)^{-1}$ and $\Delta W_c(X)[\Delta_X]$ is the incremental step in the direction $\Delta_X$. It appears that $X$ is an extremal point of the barrier function if and only if

$$-\left\langle W_c(X)^{-1}, \Delta W_c(X)[\Delta_X] \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n.$$

The increment of $W_c(X)$ corresponding to an incremental direction $\Delta_X \in \mathbb{H}_n$ of $X$ is given by

$$\Delta W_c(X)[\Delta_X] = -\begin{bmatrix} A^{\mathsf{H}}\Delta_X + \Delta_X A & \Delta_X B \\ B^{\mathsf{H}}\Delta_X & 0 \end{bmatrix}.$$

The equation for the extremal point then becomes

$$\left\langle W_c(X)^{-1}, \begin{bmatrix} A^{\mathsf{H}}\Delta_X + \Delta_X A & \Delta_X B \\ B^{\mathsf{H}}\Delta_X & 0 \end{bmatrix} \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n. \tag{6.12}$$

Defining

$$F_c := R^{-1}(C - B^{\mathsf{H}}X), \quad P_c := -A^{\mathsf{H}}X - XA - F_c^{\mathsf{H}}RF_c,$$

then

$$W_c(X) = \begin{bmatrix} I & F_c^{\mathsf{H}} \\ 0 & I \end{bmatrix} \begin{bmatrix} P_c & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} I & 0 \\ F_c & I \end{bmatrix}.$$

For a point $X \in \mathbb{X}_c^{\succ}$ it is obvious that we also have $P_c = \mathsf{Ricc}_c(X) \succ 0$, and hence (6.12) is equivalent to

$$\left\langle \begin{bmatrix} P_c^{-1} & 0 \\ 0 & R^{-1} \end{bmatrix}, \begin{bmatrix} I & -F_c^{\mathsf{H}} \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{\mathsf{H}}\Delta_X + \Delta_X A & \Delta_X B \\ B^{\mathsf{H}}\Delta_X & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ -F_c & I \end{bmatrix} \right\rangle = 0,$$

or

$$\left\langle P_c^{-1}, A^{\mathsf{H}}\Delta_X + \Delta_X A - F_c^{\mathsf{H}}B^{\mathsf{H}}\Delta_X - \Delta_X B F_c \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n,$$

which is equivalent to

$$\left\langle P_c^{-1}A_{F_c}^{\mathsf{H}} + A_{F_c}P_c^{-1}, \Delta_X \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n,$$

where we have set $A_{F_c} = A - BF_c$. This implies

$$P_c^{-1} A_{F_c}^{\mathsf{H}} + A_{F_c} P_c^{-1} = 0. \tag{6.13}$$

We emphasize that $P_c$ is nothing but the Riccati operator $\mathsf{Ricc}_c(X)$ defined in (6.7), and that $A_{F_c}$ is the corresponding closedhyphenloop matrix. For the classical Riccati solutions we have $P_c = \mathsf{Ricc}_c(X) = 0$ and the corresponding closedhyphenloop matrix is well-known to have its eigenvalues equal to a subset of the eigenvalues of the corresponding Hamiltonian matrix (6.8).

Since $P_c = \mathsf{Ricc}_c(X) \succ 0$, it follows that $P_c$ has a Hermitian square root $T_c$ satisfying $P_c = T_c^2$. Transforming (6.13) with the invertible matrix $T_c$, we obtain

$$T_c^{-1} A_{F_c}^{\mathsf{H}} T_c + T_c A_{F_c} T_c^{-1} = 0.$$

Hence $\hat{A}_{F_c} \coloneqq T_c A_{F_c} T_c^{-1}$ is skew-Hermitian and has all its eigenvalues on the imaginary axis, and so does $A_{F_c}$. Therefore, the closedhyphenloop matrix $A_{F_c}$ of the analytic center has a spectrum that is also central.

It is important to also note that

$$\det W_c(X) = \det \mathsf{Ricc}_c(X) \det R,$$

which implies that we are also finding a stationary point of $\det \mathsf{Ricc}_c(X)$, since $\det R$ is constant and non-zero.

Since the matrix $P_c$ is positive definite and invertible, we can rewrite the equations defining the analytic center as

$$\begin{aligned} RF_c &= C - B^{\mathsf{H}} X, \\ P_c &= -A^{\mathsf{H}} X - XA - F_c^{\mathsf{H}} R F_c, \\ 0 &= P_c(A - BF_c) + (A^{\mathsf{H}} - F_c^{\mathsf{H}} B^{\mathsf{H}}) P_c, \end{aligned}$$

where $X = X^{\mathsf{H}}$ and $P_c = P_c^{\mathsf{H}} \succ 0$. We can compute the analytic center by solving these three equations which actually form a cubic equation in $X$. Note that due to the convexity of the problem, the analytic center is the only solution of these equations where the conditions $X \in \mathbb{X}_c^{\succ}$ and $P_c \succ 0$ are both met.

Note that even though the eigenvalues of the closedhyphenloop matrix $A_{F_c}$ associated with the analytic center are all purely imaginary, the eigenvalues of the original system and the poles of the transfer function stay invariant under the state space transformation $T_c$.

*Remark* 6.1 (Interpretation of the analytic center). For strictly positive real systems, the set of strictly positive LMI solutions $\mathbb{X}_c^{\succ}$ contains infinitely many elements. Every solution $X \in \mathbb{X}_c^{\succ}$ defines a port-Hamiltonian realization of which the analytic center is most robust in terms of conditioning, see [BMV19] for more details. ▷

### 6.2.2. The Discrete-Time Case

For discrete-time systems, the increment of $W_d(X)$ equals

$$\Delta W_d(X)[\Delta_X] = -\begin{bmatrix} A^{\mathsf{H}}\Delta_X A - \Delta_X & A^{\mathsf{H}}\Delta_X B \\ B^{\mathsf{H}}\Delta_X A & B^{\mathsf{H}}\Delta_X B \end{bmatrix},$$

for all $\Delta_X \in \mathbb{H}_n$. Defining $F_d := (R - B^{\mathsf{H}}XB)^{-1}(C - B^{\mathsf{H}}XA)$, $A_{F_d} := A - BF_d$, and $P_d := X - A^{\mathsf{H}}XA - F_d^{\mathsf{H}}(R - B^{\mathsf{H}}XB)F_d$, then $W_d(X)$ factorizes as

$$W_d(X) = \begin{bmatrix} I & F_d^{\mathsf{H}} \\ 0 & I \end{bmatrix}\begin{bmatrix} P_d & 0 \\ 0 & R - B^{\mathsf{H}}XB \end{bmatrix}\begin{bmatrix} I & 0 \\ F_d & I \end{bmatrix},$$

and the equation for the extremal point becomes

$$\left\langle \begin{bmatrix} P_d^{-1} & 0 \\ 0 & (R - B^{\mathsf{H}}XB)^{-1} \end{bmatrix}, \begin{bmatrix} I & -F_d^{\mathsf{H}} \\ 0 & I \end{bmatrix}\begin{bmatrix} A^{\mathsf{H}}\Delta_X A - \Delta_X & A^{\mathsf{H}}\Delta_X B \\ B^{\mathsf{H}}\Delta_X A & B^{\mathsf{H}}\Delta_X B \end{bmatrix}\begin{bmatrix} I & 0 \\ -F_d & I \end{bmatrix} \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n,$$

or

$$\left\langle P_d^{-1}, A_{F_d}^{\mathsf{H}}\Delta_X A_{F_d} - \Delta_X \right\rangle + \left\langle (R - B^{\mathsf{H}}XB)^{-1}, B^{\mathsf{H}}\Delta_X B \right\rangle = 0 \quad \text{for all } \Delta_X \in \mathbb{H}_n.$$

This is equivalent to

$$A_{F_d}P_d^{-1}A_{F_d}^{\mathsf{H}} - P_d^{-1} + B(R - B^{\mathsf{H}}XB)^{-1}B^{\mathsf{H}} = 0, \tag{6.14}$$

which can be seen as a discrete-time Lyapunov equation if $X$ was fixed and independent of $P_d$. Since $(A, B)$ is controllable (by assumption), so is $(A_{F_d}, B)$ and it follows then from (6.14) that the eigenvalues of $A_{F_d}$ are now strictly inside the unit circle. This is *clearly different from the continuous-time case*, where

the spectrum of $A_{F_c}$ was on the boundary of the stability region. The equations defining the discrete-time analytic center then become

$$\begin{aligned}(R - B^{\mathsf{H}} X B) F_d &= C - B^{\mathsf{H}} X A,\\ P_d &= X - A^{\mathsf{H}} X A - F_d^{\mathsf{H}} (R - B^{\mathsf{H}} X B) F_d,\\ 0 &= (A - B F_d) P_d^{-1} (A^{\mathsf{H}} - F_d^{\mathsf{H}} B^{\mathsf{H}})\\ &\quad - P_d^{-1} + B (R - B^{\mathsf{H}} X B)^{-1} B^{\mathsf{H}}.\end{aligned}$$

*Remark* 6.2. Note that the solution of the discrete-time problem does not coincide with the one obtained via a bilinear transformation of the continuous-time problem, since this would yield a feedback $F_d$ that puts all eigenvalues on the unit circle. The bilinear transformation does not preserve determinants, and therefore the solution of the minimization problem can be expected to be different (see also Appendix B). ▷

## 6.3. Numerical Computation of the Analytic Center

In this section we present methods for the numerical computation of the analytic center.

Suppose that we are at a point $X_0 \in \mathbb{X}_c^{\succ}$ ($\mathbb{X}_d^{\succ}$) and want to perform the next step using an increment $\Delta_X$. We discuss a steepest descent and a Newton method to obtain that increment.

### 6.3.1. A Steepest Descent Method

In order to formulate an optimization scheme to compute the analytic center, we can use the gradient of the barrier function $b(X)$ with respect to $X$ in a point $X_0$ to obtain a steepest descent method.

In the continuous-time case, we then need to take a step $\Delta_X$ for which

$$\langle b(X_0), \Delta W_c(X_0)[\Delta_X] \rangle$$

is minimized, which is equivalent to

$$\Delta_X \coloneqq \operatorname*{arg\,min}_{\langle \Delta_X, \Delta_X \rangle = 1} \left\langle \Delta_X, P_c^{-1}(X_0) A_{F_c}(X_0)^{\mathsf{H}} + A_{F_c}(X_0) P_c^{-1}(X_0) \right\rangle.$$

The minimum is obtained by choosing $\Delta_X$ proportional to the gradient

$$P_c^{-1}(X_0)A_{F_c}(X_0)^{\mathsf{H}} + A_{F_c}(X_0)P_c^{-1}(X_0).$$

The corresponding optimal stepsize $\alpha$ for the increment $\Delta_X$ can be obtained from the determinant of the incremented LMI $W_c(X_0 + \alpha\Delta_X) \succ 0$.

In the discrete-time case, we obtain the increment from

$$\Delta_X := \underset{\langle \Delta_X, \Delta_X \rangle = 1}{\arg\min} \left\langle \Delta_X, A_{F_d}(X_0)P_d^{-1}(X_0)A_{F_d}^{\mathsf{H}}(X_0) - P_d^{-1}(X_0) + B(R - B^{\mathsf{H}}X_0B)^{-1}B^{\mathsf{H}} \right\rangle.$$

The minimum is obtained by choosing $\Delta_X$ proportional to the gradient

$$A_{F_d}(X_0)P_d^{-1}(X_0)A_{F_d}^{\mathsf{H}}(X_0) - P_d^{-1}(X_0) + B(R - B^{\mathsf{H}}X_0B)^{-1}B^{\mathsf{H}},$$

and the stepsize $\alpha$ for the increment $\Delta_X$ can again be obtained from the determinant of the incremented LMI $W_d(X_0 + \alpha\Delta_X) \succ 0$.

*Remark* 6.3. The detailed explanation how to compute the stepsize $\alpha$ is done later as a special case of the derivation of the Newton step, see Subsection 6.3.2. The idea is to find the second order Taylor expansion of the function $b(X_0 + \alpha\Delta_X) = -\ln\det W(X_0 + \alpha\Delta_X)$ and then to minimize this quadratic function in the scalar $\alpha$. This is a one dimensional Newton step and only yields an inexact line-search. ▷

## 6.3.2. Newton Method

For the computation of a Newton step $\Delta_X$ we also need the Hessian of the barrier function $b$. In order to simplify the derivation of the Hessian, we first reformulate the determinant of $W(X_0 + \Delta_X)$ into a more suitable form. We also point out that minimizing $-\ln\det W(X)$ is equivalent to maximizing $\det W(X)$.

### 6.3.2.1. The Continuous-Time Case

In the continuous-time case, we have that

$$W_c(X_0 + \Delta_X) = \begin{bmatrix} Q_0 & C_0^{\mathsf{H}} \\ C_0 & R_0 \end{bmatrix} - \begin{bmatrix} \Delta_X \\ 0 \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix} - \begin{bmatrix} A^{\mathsf{H}} \\ B^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \Delta_X & 0 \end{bmatrix},$$

where

$$\begin{bmatrix} Q_0 & C_0^{\mathsf{H}} \\ C_0 & R_0 \end{bmatrix} \coloneqq W_c(X_0).$$

By taking Schur complements and applying congruence transformations, it follows that the product $\det W_c(X_0+\Delta_X)(-1)^n$ is equal to

$$\det \left[\begin{array}{cc|cc} 0 & I_n & \Delta_X & 0 \\ I_n & 0 & A & B \\ \hline \Delta_X & A^{\mathsf{H}} & Q_0 & C_0^{\mathsf{H}} \\ 0 & B^{\mathsf{H}} & C_0 & R_0 \end{array}\right] = \det \left[\begin{array}{cc|cc} 0 & I_n & \Delta_X & 0 \\ I_n & 0 & A_{F_c} & B \\ \hline \Delta_X & A_{F_c}^{\mathsf{H}} & P_0 & 0 \\ 0 & B^{\mathsf{H}} & 0 & R_0 \end{array}\right],$$

where $A_{F_c} \coloneqq A - BR_0^{-1}C_0$ and $P_0 \coloneqq Q_0 - C_0^{\mathsf{H}}R_0^{-1}C_0$ are associated with the current point $X_0$. Carrying out an additional congruence transformation with

$$Z_c \coloneqq \begin{bmatrix} P_0^{-\frac{1}{2}} & 0 & 0 & 0 \\ 0 & P_0^{\frac{1}{2}} & 0 & -\hat{B}R_0^{-1} \\ 0 & 0 & P_0^{-\frac{1}{2}} & 0 \\ 0 & 0 & 0 & R_0^{-\frac{1}{2}} \end{bmatrix},$$

we obtain

$$\begin{bmatrix} 0 & I_n & \hat{\Delta}_X & 0 \\ I_n & -\hat{B}R_0^{-1}\hat{B}^{\mathsf{H}} & \hat{A}_{F_c} & 0 \\ \hat{\Delta}_X & \hat{A}_{F_c}^{\mathsf{H}} & I_n & 0 \\ 0 & 0 & 0 & I_m \end{bmatrix} \coloneqq Z_c \begin{bmatrix} 0 & I_n & \Delta_X & 0 \\ I_n & 0 & A_{F_c} & B \\ \Delta_X & A_{F_c}^{\mathsf{H}} & P_0 & 0 \\ 0 & B^{\mathsf{H}} & 0 & R_0 \end{bmatrix} Z_c^{\mathsf{H}}, \tag{6.16}$$

where $\hat{B} = P_0^{\frac{1}{2}}B$, $\hat{A}_{F_c} \coloneqq P_0^{\frac{1}{2}}A_{F_c}P_0^{-\frac{1}{2}}$, and $\hat{\Delta}_X = P_0^{-\frac{1}{2}}\Delta_X P_0^{-\frac{1}{2}}$. It is clear that the determinant of the congruence transformation $Z_c$ is given by

$$(\det Z_c)^2 = 1/(\det P_0 \cdot \det R_0) = 1/\det W_c(X_0). \tag{6.17}$$

The above transformations finally lead to the following lemma.

**Lemma 6.1.** *The change of variables*

$$\hat{B} = P_0^{\frac{1}{2}}B, \quad \hat{A}_{F_c} \coloneqq P_0^{\frac{1}{2}}A_{F_c}P_0^{-\frac{1}{2}}, \quad \hat{\Delta}_X = P_0^{-\frac{1}{2}}\Delta_X P_0^{-\frac{1}{2}}, \quad \hat{X} = P_0^{-\frac{1}{2}}XP_0^{-\frac{1}{2}},$$

*yields the following determinant identity*

$$\det W_c(X_0+\Delta_X) = \det \begin{bmatrix} 0 & I_n & \hat{\Delta}_X \\ I_n & -\hat{B}R_0^{-1}\hat{B}^{\mathsf{H}} & \hat{A}_{F_c} \\ \hat{\Delta}_X & \hat{A}_{F_c}^{\mathsf{H}} & I_n \end{bmatrix} \tag{6.18}$$

$$= \det\left[I_n - \hat{\Delta}_X\hat{A}_{F_c} - \hat{A}_{F_c}^{\mathsf{H}}\hat{\Delta}_X - \hat{\Delta}_X\hat{B}R_0^{-1}\hat{B}^{\mathsf{H}}\hat{\Delta}_X\right]. \tag{6.19}$$

*Proof.* The determinant of the right hand side of (6.16) equals $\det W_c(X_0+\Delta_X)$ because of (6.17). The equalities (6.18), (6.19) then easily follow. □

We thus have an equivalent minimization problem $\min_{X\in\mathbb{H}_n} f(X)$ in the new 'translated' variable $X = \hat{\Delta}_X$ corresponding to an initial point at the origin of the barrier function

$$\begin{aligned} f(X) &:= -\ln\det(G(X)), \\ Q_c &:= \hat{B}R_0^{-1}\hat{B}^{\mathsf{H}}, \\ G(X) &:= I_n - X\hat{A}_{F_c} - \hat{A}_{F_c}^{\mathsf{H}}X - XQ_cX. \end{aligned}$$

In the set of Hermitian matrices (over the reals), the gradient of $f(X)$ then is given by

$$f_X(X)[\Delta] = \langle -G(X)^{-1}, -(\Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta + \Delta Q_cX + XQ_c\Delta)\rangle$$

and the Hessian is given by

$$\begin{aligned} f_{XX}(X)[\Delta,\Delta] = \Big\langle & -G(X)^{-1}(\Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta + \Delta Q_cX + XQ_c\Delta)G(X)^{-1}, \\ & -(\Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta + \Delta Q_cX + XQ_c\Delta)\Big\rangle \\ & + \langle -G(X)^{-1}, -2\Delta Q_c\Delta\rangle. \end{aligned}$$

A second order approximation of $f$ (at $X=0$) is given by

$$\begin{aligned} f(\Delta) \approx T_f^{(2)}(\Delta) &= f(0) + f_X(0)[\Delta] + \frac{1}{2}f_{XX}(0)[\Delta,\Delta] \\ &= \langle I_n, \Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta\rangle + \frac{1}{2}\langle \Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta, \Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta\rangle \\ &\quad + \langle I_n, \Delta Q_c\Delta\rangle. \end{aligned}$$

Remember, that in order to minimize $f(X)$, we want the gradient of $f$ to be 0. Thus, for the Newton step we want to determine $\Delta = \Delta^{\mathsf{H}}$ such that $\frac{\partial T_f^{(2)}}{\partial \Delta}(\Delta)[Y] = 0$ for all $Y \in \mathbb{H}_n$, i. e., we require that

$$\langle I_n, Y\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}} Y\rangle + \langle \Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta, Y\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}} Y\rangle + 2\langle I_n, Y Q_c \Delta\rangle = 0$$

for all $Y \in \mathbb{H}_n$. Using the properties of the scalar product, we obtain that this is equivalent to

$$\langle Y, \hat{A}_{F_c}^{\mathsf{H}} + \hat{A}_{F_c} + \hat{A}_{F_c}\Delta\hat{A}_{F_c} + \hat{A}_{F_c}\hat{A}_{F_c}^{\mathsf{H}}\Delta + \hat{A}_{F_c}^{\mathsf{H}}\Delta\hat{A}_{F_c}^{\mathsf{H}} + \Delta\hat{A}_{F_c}\hat{A}_{F_c}^{\mathsf{H}} + Q_c\Delta + \Delta Q_c\rangle = 0$$

for all $Y \in \mathbb{H}_n$, or equivalently

$$\hat{A}_{F_c}\Delta\hat{A}_{F_c} + \hat{A}_{F_c}\hat{A}_{F_c}^{\mathsf{H}}\Delta + \hat{A}_{F_c}^{\mathsf{H}}\Delta\hat{A}_{F_c}^{\mathsf{H}} + \Delta\hat{A}_{F_c}\hat{A}_{F_c}^{\mathsf{H}} + Q_c\Delta + \Delta Q_c = -\hat{A}_{F_c}^{\mathsf{H}} - \hat{A}_{F_c}. \quad (6.20)$$

If we fix a direction $\Delta$ and look for $\alpha$ such that $f(\alpha\Delta)$ is minimal, then the one-dimensional Newton step corresponds to an inexact line search. It can be computed in an analogous way. With $g(\alpha) = f(\alpha\Delta)$, we then have

$$g(\alpha) \approx f(0) + \alpha f_X(0)[\Delta] + \frac{1}{2}\alpha^2 f_{XX}(0)[\Delta,\Delta]$$

and thus the one-dimensional Newton correction in $\alpha$ is given by

$$\delta_\alpha = -\frac{\langle I_n, \Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta\rangle}{\langle I_n, \Delta Q_c\Delta\rangle + \frac{1}{2}\|\Delta\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}\Delta\|_F^2}.$$

#### 6.3.2.2. The Discrete-Time Case

For the discrete-time case, we have that

$$W_d(X_0 + \Delta_X) = \begin{bmatrix} Q_0 & C_0^{\mathsf{H}} \\ C_0 & R_0 \end{bmatrix} - \begin{bmatrix} A^{\mathsf{H}} \\ B^{\mathsf{H}} \end{bmatrix} \Delta_X \begin{bmatrix} A & B \end{bmatrix} + \begin{bmatrix} I_n \\ 0 \end{bmatrix} \Delta_X \begin{bmatrix} I_n & 0 \end{bmatrix},$$

where

$$\begin{bmatrix} Q_0 & C_0^{\mathsf{H}} \\ C_0 & R_0 \end{bmatrix} := W_d(X_0).$$

By taking Schur complements and applying congruence transformations, it follows again that the product $\det W_d(X_0+\Delta_X)(-1)^n$ is equal to

$$\det\left[\begin{array}{cc|cc} -I_n & 0 & \Delta_X & 0 \\ 0 & I_n & A & B \\ \hline I_n & A^{\mathsf{H}}\Delta_X & Q_0 & C_0^{\mathsf{H}} \\ 0 & B^{\mathsf{H}}\Delta_X & C_0 & R_0 \end{array}\right] = \det\left[\begin{array}{cc|cc} -I_n & 0 & \Delta_X & 0 \\ 0 & I_n & A_{F_d} & B \\ \hline I_n & A_{F_d}^{\mathsf{H}}\Delta_X & P_0 & 0 \\ 0 & B^{\mathsf{H}}\Delta_X & 0 & R_0 \end{array}\right],$$

where $R_0 = R - B^{\mathsf{H}}X_0B$, $C_0 = C - B^{\mathsf{H}}X_0A$, $Q_0 = X_0 - A^{\mathsf{H}}X_0A$, $A_{F_d} \coloneqq A - BR_0^{-1}C_0^{\mathsf{H}}$ and $P_0 \coloneqq Q_0 - C_0R_0^{-1}C_0^{\mathsf{H}}$ are associated with the current point $X_0$. Setting

$$Z_\ell \coloneqq \begin{bmatrix} P_0^{-\frac{1}{2}} & 0 & 0 & 0 \\ 0 & P_0^{\frac{1}{2}} & 0 & -\hat{B}R_0^{-1} \\ 0 & 0 & P_0^{-\frac{1}{2}} & 0 \\ 0 & 0 & 0 & R_0^{-\frac{1}{2}} \end{bmatrix},\ Z_r \coloneqq \begin{bmatrix} P_0^{\frac{1}{2}} & 0 & 0 & 0 \\ 0 & P_0^{-\frac{1}{2}} & 0 & 0 \\ 0 & 0 & P_0^{-\frac{1}{2}} & 0 \\ 0 & -R_0^{-1}\hat{B}^{\mathsf{H}}\hat{\Delta}_X & 0 & R_0^{-\frac{1}{2}} \end{bmatrix},$$

transforming with $Z_\ell$ from the left and $Z_r$ from the right, and substituting $\hat{B} = P_0^{\frac{1}{2}}B$, $\hat{A}_{F_d} \coloneqq P_0^{\frac{1}{2}}A_{F_d}P_0^{-\frac{1}{2}}$, and $\hat{\Delta}_X = P_0^{-\frac{1}{2}}\Delta_X P_0^{-\frac{1}{2}}$, we obtain the matrix

$$\begin{bmatrix} -I_n & 0 & \hat{\Delta}_X & 0 \\ 0 & I_n - \hat{B}R_0^{-1}\hat{B}^{\mathsf{H}}\hat{\Delta}_X & \hat{A}_{F_d} & 0 \\ I_n & \hat{A}_{F_d}^{\mathsf{H}}\hat{\Delta}_X & I_n & 0 \\ 0 & 0 & 0 & I_m \end{bmatrix} \coloneqq Z_\ell \begin{bmatrix} -I_n & 0 & \Delta_X & 0 \\ 0 & I_n & A_{F_d} & B \\ I_n & A_{F_d}^{\mathsf{H}}\Delta_X & P_0 & 0 \\ 0 & B^{\mathsf{H}}\Delta_X & 0 & R_0 \end{bmatrix} Z_r.$$

Using $\det Z_\ell \cdot \det Z_r = 1/(\det P_0.\det R_0) = 1/\det W_d(X_0)$, we obtain a similar lemma to the continuous-time case.

**Lemma 6.2.** *The change of variables* $\hat{B} = P_0^{\frac{1}{2}}B$, $\hat{A}_{F_d} = P_0^{\frac{1}{2}}A_{F_d}P_0^{-\frac{1}{2}}$, *and* $\hat{\Delta}_X = P_0^{-\frac{1}{2}}\Delta_X P_0^{-\frac{1}{2}}$, *yields the following determinant identity*

$$\det W_d(X_0+\Delta_X) = \det\begin{bmatrix} I_n - \hat{B}R_0^{-1}\hat{B}^{\mathsf{H}}\hat{\Delta}_X & \hat{A}_{F_d} \\ \hat{A}_{F_d}^{\mathsf{H}}\hat{\Delta}_X & I_n + \hat{\Delta}_X \end{bmatrix}. \qquad \triangleright$$

*Proof.* The proof is analogous to the continuous-time case. □

We have again an equivalent minimization problem $\min_{X\in\mathbb{H}_n} f(X)$ in the 'translated' variable $X = \hat{\Delta}_X$ with the barrier function

$$\begin{aligned} f(X) &:= -\ln \Re \det(G(X)), \\ G(X) &:= \begin{bmatrix} I_n - Q_d X & \hat{A}_{F_d} \\ \hat{A}_{F_d}^{\mathsf{H}} X & I_n + X \end{bmatrix}, \\ Q_d &:= \hat{B} R_0^{-1} \hat{B}^{\mathsf{H}}, \end{aligned}$$

and compute the gradient and the Hessian of $f(X)$. The computation of the gradient is not as straight-forward as in the continuous-time case, since we consider non-Hermitian matrices. It is given by

$$f_X(X)[\Delta] = \left\langle -\frac{\overline{\det G(X)}}{\Re \det G(X)} G(X)^{-\mathsf{H}}, \begin{bmatrix} -Q_d\Delta & 0 \\ \hat{A}_{F_d}^{\mathsf{H}}\Delta & \Delta \end{bmatrix} \right\rangle,$$

see Appendix A for more details. It follows from the derivation of $G(X)$ in Lemma 6.2 that $\det(G(X))$ is positive and real and the solution of the minimization problem is still unique and Hermitian. Moreover, $\overline{\det G(X)} = \Re \det G(X)$ and the Hessian is then given by

$$f_{XX}(X)[\Delta,\Delta] = \left\langle G(X)^{-\mathsf{H}} \begin{bmatrix} -Q_d\Delta & 0 \\ \hat{A}_{F_d}^{\mathsf{H}}\Delta & \Delta \end{bmatrix}^{\mathsf{H}} G(X)^{-\mathsf{H}}, \begin{bmatrix} -Q_d\Delta & 0 \\ \hat{A}_{F_d}^{\mathsf{H}}\Delta & \Delta \end{bmatrix} \right\rangle,$$

and a second order approximation of $f$ (at $X = 0$) is given by

$$\begin{aligned} f(\Delta) &\approx T_f^{(2)}(\Delta) \\ &= f(0) + f_X(0)[\Delta] + \frac{1}{2} f_{XX}(0)[\Delta,\Delta] \\ &= -\left\langle \begin{bmatrix} I_n & 0 \\ -\hat{A}_{F_d}^{\mathsf{H}} & I_n \end{bmatrix}, \begin{bmatrix} -Q_d Y & 0 \\ \hat{A}_{F_d}^{\mathsf{H}} Y & Y \end{bmatrix} \right\rangle \\ &\quad + \frac{1}{2} \left\langle \begin{bmatrix} I_n & 0 \\ -\hat{A}_{F_d}^{\mathsf{H}} & I_n \end{bmatrix} \begin{bmatrix} -\Delta Q_d & \Delta\hat{A}_{F_d} \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} I_n & 0 \\ -\hat{A}_{F_d}^{\mathsf{H}} & I_n \end{bmatrix}, \begin{bmatrix} -Q_d\Delta & 0 \\ \hat{A}_{F_d}^{\mathsf{H}}\Delta & \Delta \end{bmatrix} \right\rangle \\ &= -\left\langle I_n - Q_d - \hat{A}_{F_d}\hat{A}_F^{\mathsf{H}}, \Delta \right\rangle \\ &\quad + \frac{1}{2} \left\langle Q_d\Delta Q_d\Delta - 2\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta Q_d\Delta, I_n \right\rangle \\ &\quad + \frac{1}{2} \left\langle \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta + 2\hat{A}_{F_d}\Delta\hat{A}_{F_d}^{\mathsf{H}}\Delta - \Delta^2, I_n \right\rangle. \end{aligned}$$

We want the gradient of $f$ to be 0, so for the Newton step we determine $\Delta = \Delta^{\mathsf{H}}$ such that $\frac{\partial T_f^{(2)}}{\partial \Delta}(\Delta)[Y] = 0$ for all $Y \in \mathbb{H}_n$, or equivalently

$$\begin{aligned} 0 = &\langle I_n - Q_d - \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}, Y\rangle + \langle Q_d \Delta Q_d + \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} \Delta Q_d, Y\rangle \\ &+ \langle Q_d \Delta \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} + \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} - \hat{A}_{F_d}\Delta\hat{A}_{F_d}^{\mathsf{H}} - \hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d} + \Delta, Y\rangle \end{aligned}$$

for all $Y \in \mathbb{H}_n$. Using the properties of the scalar product, we obtain that

$$\begin{aligned} I_n - Q_d - \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} &= Q_d\Delta Q_d + \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta Q_d + Q_d\Delta\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} + \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}} \\ &\quad - \hat{A}_{F_d}\Delta\hat{A}_{F_d}^{\mathsf{H}} - \hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d} + \Delta. \end{aligned} \tag{6.21}$$

If we fix a direction $\Delta$ and look for $\alpha$ such that $f(\alpha\Delta)$ is minimal, then the one-dimensional Newton step corresponds to an inexact line search. With $g(\alpha) = f(\alpha\Delta)$, we then have

$$\delta_\alpha = \frac{2\left\langle I_n - Q_d - \hat{A}_{F_d}\hat{A}_F^{\mathsf{H}}, \Delta\right\rangle}{\left\langle Q_d\Delta Q_d\Delta - 2\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta Q_d\Delta - \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta\hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}\Delta + 2\hat{A}_{F_d}\Delta\hat{A}_{F_d}^{\mathsf{H}}\Delta - \Delta^2, I_n\right\rangle}.$$

*Remark* 6.4. To carry out the Newton step, we have to solve equation (6.20) in the continuous-time case or (6.21) in the discrete-time case. This can be done via Kronecker products (for the cost of increasing the system dimension to $n^2$), i. e., via

$$\begin{aligned} &\left((I_n \otimes \hat{A}_{F_c} + \overline{\hat{A}_{F_c}} \otimes I_n)(\hat{A}_{F_c}^T \otimes I_n + I_n \otimes \hat{A}_{F_c}^{\mathsf{H}}) + I_n \otimes Q_c + \overline{Q}_c \otimes I_n\right)\operatorname{vec}\Delta \\ &\quad = \operatorname{vec}(\hat{A}_{F_c} + \hat{A}_{F_c}^{\mathsf{H}}) \end{aligned}$$

in the continuous-time case, or

$$\begin{aligned} &\left((\overline{\hat{A}_{F_d}} \otimes \hat{A}_{F_d} - I_n \otimes I_n)(\hat{A}_{F_d}^T \otimes \hat{A}_{F_d}^{\mathsf{H}} - I_n \otimes I_n) + \overline{Q}_d \otimes \hat{A}_{F_c}\hat{A}_{F_c}^{\mathsf{H}}\right. \\ &\qquad \left. + \overline{\hat{A}_{F_c}}\hat{A}_{F_c}^T \otimes Q_d + \overline{Q_d} \otimes Q_d\right)\operatorname{vec}\Delta = \operatorname{vec}(I_n - Q_d - \hat{A}_{F_d}\hat{A}_{F_d}^{\mathsf{H}}) \end{aligned}$$

in the discrete-time case. ▷

#### 6.3.2.3. Convergence

In this subsection, we show that the functions that we consider here actually have a globally converging Newton method. For this we have to analyze some more properties of our functions and refer to [BV04; Nes04] for more details. Recall that a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ is *self-concordant* if it is a closed and convex function with open domain and

$$|f^{(3)}(x)| \leq 2\left(f^{(2)}(x)\right)^{\frac{3}{2}}$$

in the case $n = 1$, and if $n > 1$, then $f$ is *self-concordant* if it is self-concordant along every direction in its domain. In particular, if $n = 1$ then $f(x) = -\ln(x)$ is self-concordant and in general, if $f$ is self-concordant and in addition $A \in \mathbb{C}^{n\times m}$, $b \in \mathbb{R}^n$, then $f(Ax + b)$ is also self-concordant. These results can be easily extended to the real space of complex matrices showing that the function $b(X) = -\ln\det(W(X))$ is self-concordant. Let $b_X$ and $b_{XX}$ denote the gradient and the Hessian of the barrier function $b(X)$, and let

$$\lambda(X) \coloneqq \left\langle \left(b_{XX}\right)^{-1} b_X,\, b_X \right\rangle,$$

where $\Delta_N \coloneqq \left(b_{XX}\right)^{-1} b_X$ is the Newton step, i. e.,

$$\lambda(X) = \left\langle \Delta_N,\, P_c^{-1} A_{F_c}^{\mathsf{H}} + A_{F_c} P_c^{-1} \right\rangle$$

in the continuous-time case, and

$$\lambda(X) = \left\langle \Delta_N,\, A_{F_d} P_d^{-1} A_{F_d}^{\mathsf{H}} - P_d^{-1} + B(R - B^{\mathsf{H}} X B)^{-1} B^{\mathsf{H}} \right\rangle$$

in the discrete-time case, respectively. In both cases $\lambda(X)$ can be easily computed during the Newton step and gives an estimate of the residual of the current approximation of the solution.

Furthermore, note, that for each $X \in \mathbb{X}_c^{\succ}(\mathbb{X}_d^{\succ})$ we have, that the incremental step $\Delta W(X)[\Delta_X]$ appearing in the directional derivative (6.11) is independent of $X$. Thus, the quadratic form of the Hessian can be expressed as

$$\begin{aligned}\langle b_{XX}\Delta_X, \Delta_X\rangle &= \langle W^{-1}\Delta W[\Delta_X]W^{-1}, \Delta W[\Delta_X]\rangle \\ &= \operatorname{tr}\left(W^{-\frac{1}{2}}\Delta W[\Delta_X]W^{-1}\Delta W[\Delta_X]W^{-\frac{1}{2}}\right).\end{aligned}$$

Using the Courant-Fischer theorem twice, see e. g. [Bel97], this implies that

$$\begin{aligned}\operatorname{tr}\left(W^{-\frac{1}{2}}\Delta W[\Delta_X]W^{-1}\Delta W[\Delta_X]W^{-\frac{1}{2}}\right) &\geq \frac{1}{\lambda_{\max}(W(X))}\operatorname{tr}\left(\Delta W[\Delta_X]W^{-1}\Delta W[\Delta_X]\right)\\ &\geq \frac{1}{\lambda^2_{\max}(W(X))}\operatorname{tr}\left(\Delta W[\Delta_X]\Delta W[\Delta_X]\right).\end{aligned}$$

Note that $\|\Delta W[\Delta_X]\|_F \neq 0$ for controllable $(A,B)$ and $\Delta_X \neq 0$. Minimizing the left-hand side over all $\Delta_X$ with $\|\Delta_X\|_F^2 = 1$ yields uniform positivity of the Hessian, since the spectrum of $W(X)$ is bounded.

Hence, it follows, see e. g., [Nes04, Theorem 4.1.14], that the Newton method is quadratically convergent, whenever $\lambda(X) < .25$ in some intermediate step. Once this level is reached, the methods stays in the quadratically convergent regime. If the condition does not hold, then one has to take a smaller stepsize $(1+\lambda(X))^{-1}\Delta_X$ in order to obtain convergence.

#### 6.3.2.4. Initialization

Note that for the reformulations of the Newton step we have to assume that the starting value $X_0$ is in the interior of the domain. In this section, we show how to compute an initial point $X_0 \in \mathbb{X}_c^{\succ}$ (or $X_0 \in \mathbb{X}_d^{\succ}$), which therefore satisfies the LMI $W_c(X_0,\mathcal{M}) \succ 0$ (or $W_d(X_0,\mathcal{M}) \succ 0$) for the model $\mathcal{M} = \{A,B,C,D\}$. Since the reasoning for both the continuous-time case and the discrete-time case are very similar, we first focus on the continuous-time case.

We start the optimization from a model $\mathcal{M}$ that is minimal and strictly passive. It then follows that the solution set of $W_c(X_0,\mathcal{M}) \succ 0$ has an interior point $X_0 \succ 0$ such that

$$W_c(X_0,\mathcal{M}) \succ 0, \quad 0 \prec X_- \preceq X_0 \preceq X_+,$$

where $X_-$ and $X_+$ are the Riccati solutions corresponding to this LMI. To construct such an $X_0$, let $\alpha := \lambda_{\min} W_c(X_0) > 0$ and $\beta := \max(\|X_0\|_2, 1) > 0$. Then, for $0 < 2\xi \leq \alpha/\beta$, we have the inequality

$$W_c(X_0,\mathcal{M}) - 2\xi \begin{bmatrix} X_0 & 0 \\ 0 & I_m \end{bmatrix} \succeq 0. \tag{6.22}$$

In order to compute a solution $X_0$ for this LMI, we rewrite it as

$$W_c(X_0,\mathcal{M}_\xi) := \begin{bmatrix} -(A+\xi I_n)^{\mathsf{H}}X_0 - X_0(A+\xi I_n) & C^{\mathsf{H}} - X_0 B \\ C - B^{\mathsf{H}}X_0 & R - 2\xi I_m \end{bmatrix} \succeq 0$$

for the modified model $\mathcal{M}_\xi := \{A + \xi I_n, B, C, D - \xi I_m\}$ and $R = D + D^{\mathsf{H}}$. It then follows from (6.22) that $\mathcal{M}_\xi$ is passive. Therefore we have the following lemma.

**Lemma 6.3.** *Let $\mathcal{M} := \{A, B, C, D\}$ be strictly passive. Then there exists a sufficiently small $\xi > 0$, such that the modified model $\mathcal{M}_\xi := \{A + \xi I_n, B, C, D - \xi I_m\}$ is passive. Then the extremal solutions $X_-(\xi)$ and $X_+(\xi)$ of the model $\mathcal{M}_\xi$ are interior points of $\mathbb{X}_c^{\succ}$.* ▷

*Proof.* Let $X_0$ be any point such that $W_c(X_0, \mathcal{M}_\xi) \succeq 0$ and $\xi > 0$. Then it follows from (6.22) that $W_c(X_0, \mathcal{M}) \succ 0$ and hence it is an interior point of $\mathbb{X}_c^{\succ}$. This also applies to the Riccati solutions $X_-(\xi)$ and $X_+(\xi)$. □

The reasoning for the discrete-time case is very similar. Starting from a strictly passive and minimal model $\mathcal{M}$, we have the inequality

$$W_d(\mathcal{M}) - 2\xi \begin{bmatrix} X_0 & 0 \\ 0 & I_m \end{bmatrix} \succeq 0, \quad \text{for} \quad 0 < 2\xi \le \alpha/\beta = \lambda_{\min} W_d(X_0) / \max(\|X_0\|_2, 1).$$

In order to compute a solution $X_0$ for this LMI, we rewrite it as the scaled LMI

$$W_d(X_0, \mathcal{M}_\xi) := (1 - 2\xi) \begin{bmatrix} X_0 - A_\xi^{\mathsf{H}} X_0 A_\xi & C_\xi^{\mathsf{H}} - A_\xi^{\mathsf{H}} X_0 B_\xi \\ C_\xi - B_\xi^{\mathsf{H}} X_0 A_\xi & R_\xi - B_\xi^{\mathsf{H}} X_0 B_\xi \end{bmatrix} \succeq 0$$

for the modified model $\mathcal{M}_\xi := \{A_\xi, B_\xi, C_\xi, D_\xi\} := \{A/\sqrt{1-2\xi}, B/\sqrt{1-2\xi}, C/(1-2\xi), (D - \xi I_m)/(1-2\xi)\}$ and $R_\xi = D_\xi + D_\xi^{\mathsf{H}}$. The solutions $X_-(\xi)$ and $X_+(\xi)$ of this scaled LMI are again strictly included in the original solution set.

The procedure to find an inner point is thus to choose one of the Riccati solutions $X_-(\xi)$ or $X_+(\xi)$ of shifted or scaled problems, respectively, or some kind of average of both, since they are then guaranteed to be an interior point of the original problem. An upper bound for $2\xi$ is $\lambda_{\min}(R)$. If the Riccati solutions of $\mathcal{M}_\xi$ indicate that the shifted model is not passive, $\xi$ is divided by 2.

Another possibility to compute an initial point is to take the *geometric mean* of the minimal and maximal solution of the Riccati equations (6.7), respectively (6.10), denoted by $X_-$ and $X_+$, which is defined by $X_0 = X_-(X_-^{-1}X_+)^{\frac{1}{2}}$, see [Moa05]. However, e. g., if $X_-$ and $X_+$ are multiples of the identity matrix, then the geometric mean is a convex combination of $X_-$ and $X_+$ and is not necessarily in the interior.

### 6.3.3. Numerical Results

We have implemented the steepest descent method of Subsection 6.3.1 and the Newton method introduced in Subsection 6.3.2. The software package is written in `python 3.6`. The code and all the examples can be downloaded under [BMNV20a].

We have performed several experiments to test convergence for the different methods developed in this work. All of them present qualitatively similar convergence behavior.

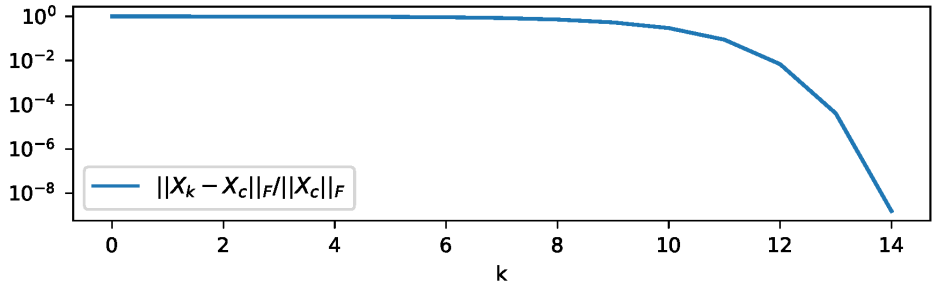
**Example 6.1.** As a prototypical example consider a randomly generated continuous-time example with real coefficients and $n = 30$ and $m = 10$, i. e., the overall dimension of the matrix $W_c(X)$ is $40 \times 40$ and we have a total of 465 unknowns. As one would expect, the steepest descent method shows linear convergence behavior, whereas the Newton method has quadratic convergence as soon as one is close enough to the analytic center.

Figure 6.1 shows the convergence behavior using the Newton method. Note, that the barrier function $\det(W(X))$ increases monotonously, whereas the distance of the argument $X$ to the analytic center $X_c$ slightly increases in the linearly convergent phase. The number of steps required in the steepest descent approach, however, is much higher than in the Newton approach.

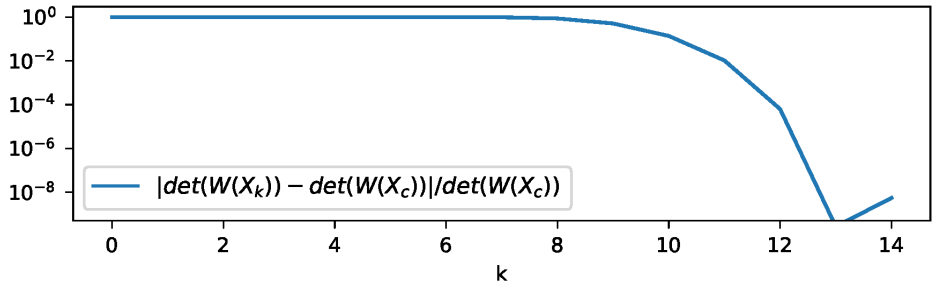
Table 6.1 shows the convergence behavior of the steepest descent method after starting the algorithm at a point well inside the feasible region, which has been obtained from a previous run with the Newton method. One can see, that even after 10,000 steps, there is no significant improvement for the residual in the determinant of $W(X)$. Though, one can at least confirm, that the values are monotonously decreasing as expected. ▷

Also, the initial point computed by the geometric mean approach turns out to be much better in all the practical examples, even though one cannot guarantee positivity in some extreme cases.

Note that one has to be extremely careful with the implementation of the algorithm. Without explicitly forcing the intermediate solutions $X_k$ to be Hermitian in finite precision arithmetic, the intermediate Riccati residuals $P_k$ may diverge from the Hermitian subspace.




(a) Convergence of the relative error between the current solution $X_k$ and the analytic center $X_c$




(b) Convergence of the relative error between the current value of the objective function $\det(W_c(X_k))$ and the value $\det(W_c(X_c))$ at the analytic center

Figure 6.1.: Convergence behavior for the Newton method applied to Example 6.1

Table 6.1.: Convergence of the relative error of the current value of the objective function $\det(W_c(X_k))$ and the intermediate solutions $X_k$ for the steepest descent method applied to Example 6.1, where $c_1 = 0.86808$, $c_2 = 0.72171$.

| $k$ | 1 | 10 | 100 | 1,000 | 10,000 |
|---|---|---|---|---|---|
| $\left(\frac{\lvert\det(W_c(X_k))-\det(W_c(X_c))\rvert}{\det(W_c(X_c))} + c_1\right)10^8$ | 524 | 522 | 512 | 512 | 428 |
| $\left(\frac{\lVert X_k - X_c\rVert}{\lVert X_c\rVert} + c_2\right)10^8$ | 198 | 198 | 197 | 194 | 160 |

## 6.4. Computation of Bounds for the Passivity Radius

Once we have found a solution $X \in \mathbb{X}_c^{\succ}$, respectively $X \in \mathbb{X}_d^{\succ}$, we can use this solution to find an estimate of the *passivity radius* of our system, i. e., the smallest perturbation $\Delta_{\mathcal{M}}$ to the system coefficients $\mathcal{M} = \{A, B, C, D\}$ that puts the system on the boundary of the set of passive systems, so that an arbitrary small further perturbation makes the system non-passive. In this section we derive *lower bounds* for the passivity radius in terms of the smallest eigenvalue of a scaled version of the matrices $W_c(X, \mathcal{M})$ or $W_d(X, \mathcal{M})$, respectively. Since the analytic center is central to the solution set of the LMI, we choose it for the realization of the transfer function, since then we expect to maximize a very good lower bound for the passivity radius.

### 6.4.1. The Continuous-time Case

As soon as we fix $X \in \mathbb{X}_c^{\succ}$, the matrix

$$W_c(X, \mathcal{M}) = \begin{bmatrix} -A^{\mathsf{H}}X - XA & C^{\mathsf{H}} - XB \\ C - B^{\mathsf{H}}X & D + D^{\mathsf{H}} \end{bmatrix}$$

is linear as a function of the coefficients $A, B, C, D$. When perturbing the coefficients, we thus preserve strict passivity, as long as

$$W_c(X, \mathcal{M} + \Delta_{\mathcal{M}}) \coloneqq \begin{bmatrix} -(A+\Delta_A)^{\mathsf{H}}X - X(A+\Delta_A) & (C+\Delta_C)^{\mathsf{H}} - X(B+\Delta_B)) \\ (C+\Delta_C) - (B+\Delta_B)^{\mathsf{H}}X & (D+\Delta_D) + (D+\Delta_D)^{\mathsf{H}} \end{bmatrix} \succ 0.$$

We thus suppose that $W_c(X, \mathcal{M}) \succ 0$ and look for the smallest perturbation $\Delta_{\mathcal{M}}$ to our model $\mathcal{M}$ that makes $\det W_c(X, \mathcal{M} + \Delta_{\mathcal{M}}) = 0$. To measure the model perturbation, we propose to use the norm of the perturbation of the system pencil

$$\|\Delta_{\mathcal{M}}\| \coloneqq \left\| \begin{bmatrix} 0 & \Delta_A & \Delta_B \\ \Delta_A^{\mathsf{H}} & 0 & \Delta_C^{\mathsf{H}} \\ \Delta_B^{\mathsf{H}} & \Delta_C & \Delta_D + \Delta_D^{\mathsf{H}} \end{bmatrix} \right\|_2 \approx \left\| \begin{bmatrix} \Delta_A & \Delta_B \\ \Delta_C & \Delta_D \end{bmatrix} \right\|_2,$$

which holds when $\Delta_D$ is Hermitian and where $\|\cdot\|_2$ denotes the matrix 2-norm. We have the following lower bound in terms of the smallest eigenvalue $\lambda_{\min}$ of a scaled version of $W_c(X, \mathcal{M})$.

**Lemma 6.4.** *The* $X$-passivity radius*, defined for a given* $X \in \mathbb{X}_c^{\succ}$ *as*

$$\rho_{\mathcal{M}}^c(X) \coloneqq \inf_{\Delta_{\mathcal{M}}} \{\|\Delta_{\mathcal{M}}\| \,|\, \det W_c(X, \mathcal{M} + \Delta_{\mathcal{M}}) = 0\},$$

*satisfies*

$$\lambda_{\min}(Y_c W_c(X, \mathcal{M}) Y_c) \leq \rho_{\mathcal{M}}^c(X),$$

*for*

$$Y_c \coloneqq \begin{bmatrix} I_n + X^2 & 0 \\ 0 & I_m \end{bmatrix}^{-\frac{1}{2}} \preceq I_{n+m}.$$

▷

*Proof.* See [BMNV20b, Lemma 4]. □

### 6.4.2. The Discrete-time Case

In the discrete-time case, for a fixed $X$ the LMI takes the form

$$W_d(X) = \begin{bmatrix} -A^{\mathsf{H}}XA + X & C^{\mathsf{H}} - A^{\mathsf{H}}XB \\ C - B^{\mathsf{H}}XA & D + D^{\mathsf{H}} - B^{\mathsf{H}}XB \end{bmatrix} \succeq 0,$$

and its perturbed version is

$$\begin{aligned} &W_d(X, \mathcal{M} + \Delta_{\mathcal{M}}) \\ &\quad := \begin{bmatrix} -(A+\Delta_A)^{\mathsf{H}} X (A+\Delta_A) + X & (C+\Delta_C)^{\mathsf{H}} - (A+\Delta_A)^{\mathsf{H}} X (B+\Delta_B) \\ C+\Delta_C - (B+\Delta_B)^{\mathsf{H}} X (A+\Delta_A) & R+\Delta_R - (B+\Delta_B)^{\mathsf{H}} X (B+\Delta_B) \end{bmatrix} \\ &\quad \succ 0, \end{aligned}$$

where again $R := D + D^{\mathsf{H}}$ and $\Delta_R := \Delta_D + \Delta_D^{\mathsf{H}}$.

Note that, in contrast to the continuous-time case, for given $X \in \mathbb{X}_d^{\succ}$ , the quantity $W_d(X, \mathcal{M} + \Delta_{\mathcal{M}})$ is not linear in the perturbations. Nevertheless, we have an analogous bound as in Lemma 6.4 also in the discrete-time case.

**Lemma 6.5.** *The* $X$-passivity radius, *defined for a given* $X \in \mathbb{X}_d^{\succ}$ *as*

$$\rho_{\mathcal{M}}^d(X) := \inf_{\Delta_{\mathcal{M}}} \{ \|\Delta_{\mathcal{M}}\| \,|\, \det W_d(X, \mathcal{M} + \Delta_{\mathcal{M}}) = 0 \},$$

*satisfies*

$$\begin{aligned} &\lambda_{\min}\Big(Y_d \Big( W_d(X,\mathcal{M}) - \begin{bmatrix} A^{\mathsf{H}} + I_n \\ B^{\mathsf{H}} \end{bmatrix} \frac{X}{2} \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} - \begin{bmatrix} \Delta_A^{\mathsf{H}} \\ \Delta_B^{\mathsf{H}} \end{bmatrix} \frac{X}{2} \begin{bmatrix} A + I_n & B \end{bmatrix} \Big) Y_d \Big) \\ &\quad \leq \rho_{\mathcal{M}}^d(X), \end{aligned}$$

*where*

$$Y_d := \begin{bmatrix} I_{n+m} + Z_d^{\mathsf{H}} Z_d \end{bmatrix}^{-\frac{1}{2}} \preceq I_{n+m},\ Z_d := -\begin{bmatrix} \frac{X}{2}(A + \Delta_A - I_n) & \frac{X}{2}(B + \Delta_B) \end{bmatrix}.$$

▷

*Proof.* See [BMNV20b, Lemma 5]. □

### 6.4.3. Examples with Analytic Solution

In this subsection, to illustrate the results, we present simple examples of scalar transfer functions ($m = 1$) of first degree ($n = 1$).

Consider first an asymptotically stable continuous-time system and transfer function $T(s) = d + \frac{cb}{s-a}$ i. e., with $a < 0$. Then

$$W_c(x) = \begin{bmatrix} -2ax & c - bx \\ c - bx & 2d \end{bmatrix}$$

and its determinant is $\det(W_c(x)) = -4adx - (c - bx)^2$, which is maximal at the central point $x_a = \frac{c}{b} - \frac{2ad}{b^2}$. We then get

$$W_c(x_a) = \begin{bmatrix} 4d\frac{a^2}{b^2} - 2c\frac{a}{b} & 2d\frac{a}{b} \\ 2d\frac{a}{b} & 2d \end{bmatrix} = \begin{bmatrix} 1 & \frac{a}{b} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p & 0 \\ 0 & 2d \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{a}{b} & 1 \end{bmatrix},$$

with $p = 2d\frac{a^2}{b^2} - 2c\frac{a}{b}$, which implies that $\det(W_c(x_a)) = 2d \cdot p$. For the transfer function to be strictly passive, it must be asymptotically stable and positive on the imaginary axis and hence also at 0 and $\infty$. Thus, we have the conditions

$$a < 0,\ d > 0,\ \frac{da - cb}{a} > 0. \tag{6.23}$$

The function $\Phi_c(\imath\omega) = 2d - \frac{2acb}{a^2+\omega^2}$ is a unimodal function, which reaches its minimum either at 0 (namely $\Phi_c(0) = p\frac{b^2}{a^2}$) or at $\infty$ (namely $\Phi_c(\infty) = 2d$) and hence the conditions in (6.23) are sufficient to check passivity. Thus, for the model $\mathcal{M}$, strict passivity gets lost when either one of the following happens

$$d + \delta_d = 0, \quad a + \delta_a = 0, \quad \begin{bmatrix} c + \delta_c & d + \delta_d \end{bmatrix} \begin{bmatrix} -b - \delta_b \\ a + \delta_a \end{bmatrix} = 0.$$

Therefore, it follows that

$$\rho = \min\left(d, a, \sigma_2 \begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \sigma_2 \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

At the analytic center $x_a$ we have

$$\det W_c(x_a) = 2dp = 4\frac{ad}{b^2}(ad - bc)$$

and the smallest perturbation of the parameters that makes this determinant go to 0, yields exactly the same conditions as (6.23). This illustrates that the $X$-passivity radius at the analytic center yields a very good condition for strict passivity of the model.

In the discrete-time case the transfer function is $T(z) = d + \frac{cb}{z-a}$ and for it to be asymptotically stable we need $a^2 < 1$, when we assume the coefficients to be real. Then

$$W_d(x) = \begin{bmatrix} x - a^2 x & c - abx \\ c - abx & 2d - b^2 x \end{bmatrix}$$

and the analytic center, where $\det W_d(x) = (1-a^2)x(2d - b^2 x) - (c - abx)^2$ is maximal, is given by $x_a = \frac{d - a^2 d + abc}{b^2}$ with

$$\det W_d(x_a) = \frac{\left(a^2 - 1\right)(bc - (a-1)d)(bc - (a+1)d)}{b^2}.$$

The function $\Phi_d(z) = \frac{bc}{\frac{1}{z} - a} + \frac{bc}{z-a} + 2d$ is minimal on the unit circle at $z = 1$ or $z = -1$. Thus positivity gets lost, when either $a$ reaches 1 or $-1$, or $bc - (a-1)d = 0$ or $bc - (a+1)d = 0$. This is exactly the condition also reflected in the determinant of $W(x_c)$ at the analytic center $x_a$. This again illustrates that the $X$-passivity radius at the analytic center gives a good bound for the passivity radius of the system.

# 7. Condition Number Optimization in Algebraic Riccati Equations

In this chapter we analyze how the infinite time-horizon optimal control problem given by the objective function (II.1) can be used for optimization of certain stability quantities of linear time-invariant systems with ordinary differential equations (ODEs). Thus, we consider systems $(E, A, B, f) \in \Sigma_{m,n}(\mathbb{K})$ with $E = I_n$ and $f = 0$, and $A \in \mathbb{K}^{n\times n}$, $B \in \mathbb{K}^{n\times m}$ are time-invariant constant matrices.

Stabilization in the context of these homogeneous systems $(I_n, A)$ can be viewed as computing a feedback matrix $F \in \mathbb{K}^{n\times m}$ such that the resulting *closed-loop matrix* $A_F = A - BF$ is stable, i. e., the spectrum of $A_F$ is contained in the closed complex left-half plane. However, with eigenvalues of $A_F$ close or on the imaginary axis, stability may be very sensitive to perturbations of the data.

Here, we thus consider the problem of *robust stabilization* which aims at moving all eigenvalues of $A_F$ to the region

$$\mathcal{R}_\varrho = \{z \in \mathbb{C} \mid \Re z \le \varrho\}, \tag{7.1}$$

for a fixed $\varrho < 0$, while minimizing a certain objective function. Possible choices for the objective function are the robustness criteria

1. $\|F\|_F$,
2. $\kappa_F := \operatorname{cond}(A - BF)$, or
3. $\kappa_F\sqrt{1 + \|F\|_F^2}$.

We have a bilevel optimization problem according to Definition 2.1. The lower level is given by the linear-quadratic optimal control problem, where instead of the optimal solution $x^*$ we want the optimal feedback $F$. The upper level optimizes one of the Robustness Criteria 1 to 3 over the weight coefficients $Q, S$, and $R$ subject to the inequality constraint (7.1).

In [MX00] it is analyzed, how different methods can be used to minimize Robustness Criteria 1 to 3. Here, we consider the approach of computing solutions of the algebraic Riccati equation via the deflating subspaces approach, see e. g., [Meh91]. This approach is to be favored over the subspace computation via Hamiltonian matrix pencils, as the feedback $F$ is computed directly without forming a solution $X$ explicitly [MX00].

## 7.1. Subspace Computations for Algebraic Riccati equations

The algebraic Riccati equation is given by

$$Q + A^{\mathsf{H}}X + XA - (S + XB)R^{-1}(S^{\mathsf{H}} + B^{\mathsf{H}}X) = 0, \quad X^{\mathsf{H}} = X, \tag{7.2}$$

where $Q \in \mathbb{K}^{n\times n}$, $Q \succeq 0$, $S \in \mathbb{K}^{n\times m}$, $R \in \mathbb{K}^{m\times m}$, $R \succ 0$, and

$$\begin{bmatrix} Q & S \\ S^{\mathsf{H}} & R \end{bmatrix} \succeq 0.$$

We call $X$ a *robustly stabilizing solution* of (7.2), if for $F = R^{-1}(B^{\mathsf{H}}X + S^{\mathsf{H}})$, we have that the eigenvalues of the closedhyphenloop matrix $A_F = A - BF$ lie in $\mathcal{R}_\varrho$ as defined in (7.1) for some $\varrho < 0$. Whether or not the eigenvalues of $A_F$ fulfill that condition depends on the particular choices of $Q, S$, and $R$.

A solution of (7.2) can be computed via a stable deflating subspace of the pencil

$$s\mathcal{E} + \mathcal{A} = \begin{bmatrix} 0 & -sI_n + A & B \\ sI_n + A^{\mathsf{H}} & Q & S \\ B^{\mathsf{H}} & S^{\mathsf{H}} & R \end{bmatrix}, \tag{7.3}$$

see [Meh91]. In that case, $X$ and the corresponding feedback $F$ fulfill

$$(s\mathcal{E} + \mathcal{A}) \begin{bmatrix} X \\ I_n \\ -F \end{bmatrix} = \begin{bmatrix} I_n \\ -X \\ 0 \end{bmatrix} \left(-sI_n + A - BF\right). \tag{7.4}$$

The matrix pencil (7.3) is called an *even* pencil as it fulfills the defining properties $\mathcal{E}^{\mathsf{H}} = -\mathcal{E}$ and $\mathcal{A}^{\mathsf{H}} = \mathcal{A}$. This structural property is preserved under unitary

congruence transformations and can be exploited in numerical methods, see e. g., [KSW09].

The deflating subspace

$$\begin{bmatrix} X \\ I_n \\ -F \end{bmatrix} \tag{7.5}$$

is then computed in the following way. First, unitary congruence transformations on the pencil $(\mathcal{E},\mathcal{A})$ are applied such that the deflating subspace is given by the isometric matrix

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix}. \tag{7.6}$$

In the final step, one recovers $X$ and $F$ in (7.4) by multiplying $U_1$ and $U_3$ by the inverse of $U_2$, i. e., $X = U_1(U_2)^{-1}$ and $F = -U_3(U_2)^{-1}$.

Note, that the isometric representation (7.6) is unique up to unitary transformations from the right.

Since both subspaces (7.5) and (7.6) span the same stabilizing subspace and $U$ is isometric it follows that

$$U_2^{\mathsf{H}}(I_n + X^2 + F^{\mathsf{H}}F)U_2 = U_2^{\mathsf{H}}\begin{bmatrix} X & I_n & -F^{\mathsf{H}} \end{bmatrix}\begin{bmatrix} X \\ I_n \\ -F \end{bmatrix} U_2 = I_n.$$

Thus, up to a unitary scaling $V$ the matrix $U_2$ is given by

$$U_2 = \left(I_n + X^2 + F^{\mathsf{H}}F\right)^{-\frac{1}{2}} V.$$

## 7.2. Optimization of the Condition Number of $U_2$

The condition number of $U_2$ is optimal, if, up to unitary scaling, $U_2$ is a multiple of the identity matrix. Thus, in order to optimize the condition number, we want to find $Q, R, S$ such that

$$I_n + X^2 + F^{\mathsf{H}}F = \delta I_n \tag{7.7}$$

for some $\delta \geq 1$.

First, we consider the case $S = 0$. Since $F = R^{-1}B^{\mathsf{H}}X$ we need to find $X$ such that

$$X(I_n + BR^{-2}B^{\mathsf{H}})X = \varepsilon I_n \tag{7.8}$$

with some $\varepsilon = \delta - 1 \geq 0$. Hence,

$$\hat{X} = \pm\sqrt{\varepsilon}(I_n + BR^{-2}B^{\mathsf{H}})^{-\frac{1}{2}} \tag{7.9}$$

are the two solutions of (7.8). Note, that $\hat{X}$ is independent of a concrete choice of $Q$. The aim now is to modify $Q$ and $R$ such that this $\hat{X}$ becomes the solution of the Riccati equation (7.2).

Inserting $\hat{X}$ from (7.9) into the Riccati equation (7.2) we obtain

$$Q + A^{\mathsf{H}}\hat{X} + \hat{X}A - \hat{X}BR^{-1}B^{\mathsf{H}}\hat{X} = 0 \tag{7.10}$$

and we can choose $Q$ depending on $R$. We still need to make sure, that the resulting $Q$ is positive semi-definite.

**Lemma 7.1.** *Assume, that $B \in \mathbb{K}^{n\times m}$ has full column rank. Consider Equation (7.10) with corresponding solution $\hat{X}$ given by (7.9). Then, there exists $\varepsilon > 0$ such that $Q$ is positive semi-definite and fulfills (7.10).* ▷

*Proof.* First note, that $Q$ is 0 for $\varepsilon = 0$. For positive $\varepsilon$ we can assume that $B$ has full column rank. Thus, there exists a unitary transformation $V \in \mathbb{K}^{n\times n}$ such that

$$V^{\mathsf{H}}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}.$$

Hence, $V^{\mathsf{H}}\hat{X}V$ is of the form

$$V^{\mathsf{H}}\hat{X}V = \pm\sqrt{\varepsilon}\begin{bmatrix} X_{11} & 0 \\ 0 & I_m \end{bmatrix}. \tag{7.11}$$

Setting

$$V^{\mathsf{H}}AV = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad V^{\mathsf{H}}RV = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

we note, that the matrix

$$\begin{bmatrix} X_{11}B_1R_{11}^{-1}B_1^{\mathsf{H}}X_{11} & 0 \\ 0 & I_m \end{bmatrix}$$

is positive definite, since $\hat{X}$ in (7.9) is positive definite and thus by construction also $X_{11}$. Consequently, we obtain that

$$\sqrt{\varepsilon}\begin{bmatrix} A_{11}^{\mathsf{H}}X_{11} & 0 \\ A_{12}^{\mathsf{H}}X_{11} & A_{22}^{\mathsf{H}} \end{bmatrix} + \sqrt{\varepsilon}\begin{bmatrix} X_{11}A_{11} & X_{11}A_{12} \\ 0 & A_{22} \end{bmatrix} - \varepsilon \begin{bmatrix} X_{11}B_1R_{11}^{-1}B_1^{\mathsf{H}}X_{11} & 0 \\ 0 & I_m \end{bmatrix} \prec 0$$

for sufficiently large $\varepsilon$. Hence, $Q$ can be chosen positive definite for sufficiently large $\varepsilon$. □

*Remark* 7.1. We can slightly weaken condition (7.7) such that we prescribe arbitrary eigenvalues instead of all being identical.

This can be achieved by requiring that instead of (7.7) the solutions $X$ and $F$ fulfill

$$X^2 + F^{\mathsf{H}}F = Y \tag{7.12}$$

for some $Y \succeq I_n$ that commutes with $BR^{-2}B^{\mathsf{H}}$. As the coordinates of $Y$ are fixed by $BR^{-2}B^{\mathsf{H}}$, we are left to choose the eigenvalues of $Y$ and possibly their order.

A solution of (7.12) then is given by

$$\hat{X} = Y^{\frac{1}{2}}(I_n + BR^{-2}B^{\mathsf{H}})^{-\frac{1}{2}}$$

and the statement of Lemma 7.1 holds analogously. ▷

In [MX00] it has been shown, that parametrizations with $S = 0$ are not always sufficient to optimize the Robustness Criteria 1 to 3. If $S \neq 0$, however, the situation is much more complicated. The feedback $F$ is given by $F = R^{-1}(B^{\mathsf{H}}X + S^{\mathsf{H}})$. Then, in order to fulfill (7.7), we have to find $\hat{X}$ such that

$$SR^{-2}S^{\mathsf{H}} + SR^{-2}B^{\mathsf{H}}\hat{X} + \hat{X}BR^{-2}S^{\mathsf{H}} + \hat{X}(I_n + BR^{-2}B^{\mathsf{H}})\hat{X} = \varepsilon I_n. \tag{7.13}$$

Compared to (7.8) we are not able to state the solution explicitly anymore. Equation (7.13) rather corresponds to finding a solution of the algebraic Riccati equation (7.2) corresponding to the new system

$$\tilde{A} = BR^{-2}S^{\mathsf{H}}, \quad \tilde{B} = I_n, \quad \tilde{Q} = -SR^{-2}S^{\mathsf{H}} + \varepsilon I_n, \quad \tilde{S} = 0, \quad \tilde{R} = (I_n + BR^{-2}B^{\mathsf{H}})^{-1}.$$

After changing the signs by setting $\tilde{X} = -\hat{X}$ we get

$$\tilde{Q} + \tilde{A}^{\mathsf{H}}\tilde{X} + \tilde{X}\tilde{A} - \tilde{X}\tilde{R}^{-1}\tilde{X} = 0. \tag{7.14}$$

For the computation of $\tilde{X}$ in (7.14) we need to solve another algebraic Riccati equation and thus possibly face the same conditioning problem as for the original Riccati equation. One idea is to revert back to the case $S = 0$ to obtain the solutions

$$\tilde{X} = \pm\sqrt{\varepsilon}\left(I_n + (I_n + BR^{-2}B^{\mathsf{H}})^2\right)^{-\frac{1}{2}}. \tag{7.15}$$

However, with this approach we also need to modify $\tilde{Q}$ and in general it is not guaranteed, that there exist $S$ and $\varepsilon$ such that $\tilde{Q} = -SR^{-2}S^{\mathsf{H}} + \varepsilon I_n$ can be fulfilled.

If, on the other hand, such $R$, $S$, and $\varepsilon$ are given with $\tilde{X}$ as in (7.15) and $B$ with full column rank, then we can compute the corresponding $Q$ as in Lemma 7.1.

## 7.3. Improving the Robustness Criteria

The Robustness Criteria 1 to 3 do not constrain the condition number of $U_2$. Thus, one could add the condition $\operatorname{cond} U_2 = 1$ to the Robustness Criteria 1 to 3. However, since adding constraints makes the feasible set smaller, analytically, one can only expect worse or equal results.

On the other hand, if we have a bad initial guess for the parameters the error introduced by a bad conditioning of $U_2$ may be large enough, such that the constrained solution may behave better than the unconstrained approach of [MX00]. Or, the constrained solution may at least serve as a better initial guess. Let us consider the following example.

**Example 7.1.** Let $U \in \mathbb{R}^{2\times 2}$ be an arbitrary orthogonal matrix, $\gamma > 0$, $\delta \geq 0$ two parameters, and set

$$A = U\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}U^T, \quad B = U, \quad Q = U\begin{bmatrix} 6 & 0 \\ 0 & \delta \end{bmatrix}U^T, \quad R = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \gamma \end{bmatrix}, \quad S = 0. \tag{7.16}$$

Then, as solution of (7.2), we obtain

$$X = U\begin{bmatrix} 3 & 0 \\ 0 & \sqrt{\delta\gamma} \end{bmatrix}U^T, \quad F = R^{-1}B^TX = \begin{bmatrix} 6 & 0 \\ 0 & \sqrt{\delta\gamma^{-1}} \end{bmatrix}U^T$$

and

$$A - BF = U\begin{bmatrix} -4 & 0 \\ 0 & -\sqrt{\delta\gamma^{-1}} \end{bmatrix}U^T.$$

The eigenvalues of $A - BF$ are given by $\lambda_1 = -4$ and $\lambda_2 = -\sqrt{\delta\gamma^{-1}}$. Let $0 < \varrho \leq 4$ be given. Then $\lambda_1, \lambda_2 \in \mathcal{R}_\varrho$ if and only if

$$\delta \geq \varrho^2\gamma. \tag{7.17}$$

Robustness Criterion 1 is minimizing the Frobenius norm of the feedback matrix $F$. The Frobenius norm is given by

$$\|F\|_F^2 = 36 + \delta\gamma^{-1} \tag{7.18}$$

and, using (7.17), is bounded below by $36 + \varrho^2$. Thus, the Frobenius norm (7.18) is minimized by choosing $\delta$ and $\gamma$ such that equality holds in (7.17).

Robustness Criterion 2 is minimizing the condition number of $A - BF$. For $\delta\gamma^{-1} \leq 16$ it is given by

$$\operatorname{cond}(A + BF) = \frac{4}{\delta\gamma^{-1}}. \tag{7.19}$$

If $\delta = \varrho^2\gamma$, then (7.19) is independent of $\delta$ and $\gamma$.

According to (7.11) the condition number of $U_2$ is given by the condition number of

$$\left(I_2 + \begin{bmatrix} 9 & 0 \\ 0 & \delta\gamma \end{bmatrix} + \begin{bmatrix} 36 & 0 \\ 0 & \delta\gamma^{-1} \end{bmatrix}\right)^{\frac{1}{2}}.$$

Note, that $\gamma + \gamma^{-1}$ is bounded below by 2. Thus, for $\delta \geq 22.5$ the condition number of $U_2$ equals

$$\operatorname{cond} U_2 = \sqrt{\frac{1 + \delta\gamma + \delta\gamma^{-1}}{46}}.$$

Note, that for $\delta = \varrho^2\gamma$ all robustness criteria are independent of the value of $\delta$ and $\gamma$, where at the same time, $\operatorname{cond} U_2$ is unbounded for $\gamma \to \infty$. This suggests, that the numerical computation of the optimal robustness criteria is depending on the values of $\gamma$ and whether the parametrization results in a reasonably sized condition number of $U_2$. Although, for the exact solution the value of $\gamma$ is irrelevant as long as $\delta = \varrho^2\gamma$.

Instead of leaving $\delta$ as a free parameter, we can use Lemma 7.1 to compute $Q$ differently and enforce $\operatorname{cond} U_2 = 1$. Choose

$$\hat{X} = \sqrt{\varepsilon}\begin{bmatrix} 5 & 0 \\ 0 & 1 + \gamma^{-2} \end{bmatrix}^{-\frac{1}{2}}$$

and therefore set

$$Q = -\sqrt{\varepsilon}\begin{bmatrix} \frac{4}{\sqrt{5}} & 0 \\ 0 & 0 \end{bmatrix} + \varepsilon \begin{bmatrix} \frac{2}{5} & 0 \\ 0 & \frac{1}{\gamma+\gamma^{-1}} \end{bmatrix},$$

which is positive definite for $\varepsilon > 20$.

The corresponding feedback matrix $\hat{F}$ is given by

$$\hat{F} = R^{-1}B^T\hat{X} = \sqrt{\varepsilon}\begin{bmatrix} \frac{2}{\sqrt{5}} & 0 \\ 0 & \frac{\gamma^{-1}}{\sqrt{1+\gamma^{-2}}} \end{bmatrix} U^T \tag{7.20}$$

and

$$A - B\hat{F} = U\begin{bmatrix} 2 - 2\sqrt{\frac{\varepsilon}{5}} & 0 \\ 0 & -\sqrt{\varepsilon}\frac{\gamma^{-1}}{\sqrt{1+\gamma^{-2}}} \end{bmatrix} U^T.$$

For $\varepsilon = 45$, this new $Q$ corresponds to setting $\delta = 45(\gamma + \gamma^{-1})^{-1}$ in (7.16). The eigenvalues of $A - B\hat{F}$ are given by $\hat{\lambda}_1 = -4 = \lambda_1$ and $\hat{\lambda}_2 = -45^{\frac{1}{2}}(\gamma^2 + \gamma^{-2})^{-\frac{1}{2}}$.

From (7.20) it can be seen that the Frobenius norm $\|\hat{F}\|_F$ is minimized, if either $\gamma$ is maximized or minimized such that $\hat{\lambda}_2$ hits the $-\varrho$ bound.

In summary, we have seen that both approaches, keeping $\delta$ as a free parameter or setting $\delta$ such that $\operatorname{cond} U_2 = 1$, lead to optimal solutions for appropriate values of $\gamma$. However, in the first case, we may face the problem that a bad choice of $\gamma$ leads to a bad conditioning of the matrix $U_2$. Thus, in a numerical optimization process for Robustness Criteria 1 to 3 we may obtain more robust results by constraining $\operatorname{cond} U_2$. ▷

# 8. Conclusion & Outlook

We have seen several examples of multilevel optimization and multilevel optimal control problems.

## Conclusion

In Part I we have analyzed the computation of sensitivities for boundary value problems in the context of differential-algebraic control systems and an application to multilevel optimal control problem with nonlinear least squares upper level. We applied the adjoint approach in the weak setting to the necessary conditions of the optimal control problem and we discussed possibilities for the computation of the sensitivities, a multiple shooting approach and a Riccati like approach.

We concluded Part I by analyzing a bilevel optimal control problem with nonlinear least squares upper level, also known as parameter estimation. We used perturbation results for the Gauss-Newton method to develop an estimate for the maximal allowed error in the computation of solutions of the optimal control problem and the corresponding sensitivity computation. A numerical example confirmed convergence using the error estimator with prescribed tolerances. Also recall Figure I.2 for a graphical summary of the procedure.

In Part II we analyzed two more examples of multilevel optimization problems. In Chapter 6 we derived conditions for the analytic center of the linear matrix inequalities (LMIs) associated with the passivity of linear continuous-time or discrete-time systems. We presented numerical methods to compute these analytic centers with steepest descent and Newton methods and we presented lower bounds for the passivity radii associated with the LMIs evaluated at the respective analytic center.

In Chapter 7 we have seen possible ways of constraining the weights $Q, S$, and $R$ depending on the system coefficients $A$ and $B$ such that in the computation of

the respective Riccati equation (7.2) the condition number of $U_2$ is 1 or any other desired value larger than 1. We have seen an example, where constraining the weights does not lead to suboptimal solutions while enforcing that cond $U_2 = 1$.

However, practical usefulness is limited so far for computing solutions with cond $U_2 = 1$. Instead of computing $X$ and $F$ with the Riccati equation (7.2), one would directly use the parametrization of $X$ given by (7.9) or (7.15).

## Synopsis

In compact form, the following statements were established.

1. The domains of linear parameter-dependent differential-algebraic equations (DAEs) can be decomposed into pairwise disjoint domains, where the characteristic quantities of the DAE (3.52) are constant, see Theorem 3.25.

2. On each of those domains, there exist local smooth full rank decompositions, see Lemma 3.26.

3. The sensitivities of a parameter-dependent DAE (4.1) can be uniquely determined via a forward system approach, see Lemma 4.1.

4. The sensitivities of (4.1) can also be computed via an adjoint approach.

   a) We developed a general adjoint equation for boundary value problems of strangeness-free DAEs and analyzed solvability and uniqueness properties, see Lemma 4.3.

   b) We recovered a version of the Lagrange identity in Lemma 4.4 in the strangeness-free case.

   c) We developed formulas for the computation of the integrated sensitivities based on a solution of a particular adjoint equation, see Theorem 4.6 and Corollary 4.8.

5. The sensitivities of a parameter-dependent DAE (4.27) in the weaker setting can be uniquely determined via a forward system approach, see Lemma 4.1, provided that condition (4.31) holds.

6. We developed a formula for the computation of the sensitivities of (4.27) in the weaker setting based on a solution of a particular adjoint equation, see Corollaries 4.8 and 4.9.

7. Based on the results we developed an adjoint equation for the necessary conditions in optimal control problems with strangeness-free DAEs.

8. We discussed possibilities for dealing with higher index DAEs in Section 4.6. We have seen an example with a formal sensitivity adjoint equation based on the formal necessary conditions, where we observed additional regularity requirements on the inhomogeneity and the boundary and jump conditions. We concluded that, whenever possible, higher index models should be avoided.

9. We derived a multiple shooting approach and analyzed the convergence behavior, see Subsection 4.7.1

10. We derived a differential Riccati type approach for the computation of the sensitivities, see Subsection 4.7.2.

11. We developed an error estimator for a bilevel parameter estimation problem (5.1) and illustrated the results with a numerical example, see Chapter 5.

12. We derived formulas for the analytic center of the passivity LMI for continuous-time and discrete-time linear time invariant systems, see Section 6.2.

13. We developed a steepest descent approach, see Subsection 6.3.1, and a Newton method for the computation of the analytic center, see Subsection 6.2.2.

14. We proved quadratic convergence of the Newton method, when the iterates are close enough to the analytic center. We still have linear convergence, otherwise, see Subsubsection 6.3.2.3.

15. The Newton method needs the solution of a linear system which we solve by a vectorization procedure, see Remark 6.4.

16. Numerical experiments confirmed quadratic convergence. The steepest descent method produces very small improvements and thus is not feasible in general, see Subsection 6.3.3.

17. We developed a parametrization of the condition number of the matrix $U_2$ involved in the computation of a solution to the algebraic Riccati equation in terms of a solution $X$ and the corresponding feedback $F$.

18. We showed how that condition number can be optimized by choosing according weight matrices and thus fixing a corresponding solution of the Riccati equation, see Lemma 7.1.

19. We provided an example which confirms that enforcing the condition number of $U_2$ to 1 still may yield optimal results for the upper level optimization of certain robustness criteria in the context of robust stabilization, see Example 7.1.

## Outlook

Possible future work includes the following.

1. In Subsection 3.4.7 we have seen that the characteristic quantities of the DAE (3.52) may change for different values of the parameters. In particular, Example 3.9 showed that a deeper analysis is necessary. The regularization techniques presented in [KM18] could be helpful for finding solutions in such cases.

2. So far, we restricted ourselves to parameters $\theta \in \mathbb{R}^p$. An extension to time dependent parameter functions $\theta \in \mathcal{C}^0(\mathbb{I}, \mathbb{R}^p)$ is desirable.

3. We presented methods for the numerical computation in Section 4.7. It would be beneficial to extend the analysis to collocation methods and a comparison of all approaches. In particular in comparison with direct approaches not using the necessary conditions. We suspect, that convergence issues, similar to what is described in [CK13], are noticeable, especially for time and possibly parameter-dependent kernels of the leading matrix $E$.

4. Throughout Part I we assumed that the lower level optimal control problem is uniquely solvable for every fixed parameter. Dropping that assumption requires techniques presented in [Meh16] and a derivation of subdifferentials for DAEs similar to [SB16; SB17].

5. In Chapter 6 the analysis and computation of the analytic center is restricted to dynamics based on ordinary differential equations (ODEs). In the DAE case we have to adapt certain quantities as the matrix function $W(X)$, to accommodate for the algebraic constraints. This can be achieved, e. g., by using projected LMIs as developed in [RRV15] for the continuous-time case or [BV18] in the discrete-time case.

6. In the computation of the analytic center via the steepest descent method we noted, that the convergence speed is very slow. A good preconditioner matrix may improve the situation.

7. The Newton method relies on the solution of a linear system, which by now is solved by a vectorization process, see Remark 6.4. An iterative approach, without blowing up the dimension due to vectorization, is desirable.

8. Also, in Chapter 7 many problems remain open and are subject to further research, for example:
   a) What is the distance between the optimal value of the optimization of the Robustness Criteria 1 to 3 with the unconstrained optimization to the constrained case with small cond $U_2$?
   b) Is the distance always small or even 0 as in Example 7.1?
   c) How are those distances affected by allowing $S \neq 0$?
   d) Can we improve the overall performance of the optimization of the robustness criteria by constraining the condition number of $U_2$?
   e) An extension to DAEs with analysis of limiting situations.

# A. Derivatives of Functions of Complex Matrices

In this appendix we present a precise derivation of the formulas for the differentiation of a matrix function with respect to a complex matrix. Here we distinguish between complex vector spaces $\mathbb{C}^n$ and the corresponding real vector space $\mathbb{R}^n + \mathrm{i}\mathbb{R}^n$. Both spaces can be identified by $\mathbf{c} : \mathbb{C}^n \to \mathbb{R}^n + \mathrm{i}\mathbb{R}^n$, $\mathbf{c}(v) = \Re(v) + \mathrm{i}\Im(v)$. For matrix spaces of dimension $m \times n$ we use the usual identification with the vector spaces $\mathbb{C}^n$ and $\mathbb{R}^n + \mathrm{i}\mathbb{R}^n$. The space $\mathbb{C}^n$ is equipped with the standard scalar product $\langle x, y\rangle_{\mathbb{C}} := x^{\mathsf{H}} y$. By $\frac{\partial}{\partial X}$ we denote the differentiation in a real vector space, whereas the differentiation of a holomorphic function $g$ is denoted by $g'$. Note that if we write $\mathbf{c} \circ g(x) = u(x_r + \mathrm{i}x_i) + \mathrm{i}v(x_r + \mathrm{i}x_i)$, then by the Cauchy-Riemann equations, see e. g. [FB05], we have $\mathbf{c} \circ g'(x) = \frac{\partial}{\partial x_r} u(x_r + \mathrm{i}x_i) - \mathrm{i}\frac{\partial}{\partial x_i} u(x_r + \mathrm{i}x_i)$. Then we have the following result:

**Lemma A.1.** *Assume that $g : \mathbb{C}^{n\times n} \to \mathbb{C}$ is holomorphic. Then $f : \mathbb{R}^{n\times n} + \mathrm{i}\mathbb{R}^{n\times n} \to \mathbb{R}$ defined by*

$$f(X_r + \mathrm{i}X_i) := \Re g(X)$$

*is differentiable over $\mathbb{R}$ with*

$$\frac{\partial}{\partial X} f(X_r + \mathrm{i}X_i) = \Re(\overline{g'(X)} \circ \mathbf{c}^{-1})$$

*and*

$$\left\langle \frac{\partial}{\partial X} f(X_r + \mathrm{i}X_i), \Delta \right\rangle_{\mathbb{R}} = \Re\langle \overline{g'(X)}, \mathbf{c}^{-1}(\Delta)\rangle_{\mathbb{C}}, \quad \Delta = \Delta_r + \mathrm{i}\Delta_i. \qquad \triangleright$$

For the holomorphic function $g(X) = \det(X)$ the following fact is well-known, see e. g. [MN99] for a proof in the real case, that easily extends to the complex case.

**Lemma A.2** (Jacobi's formula)**.** *Let* $g(X) = \det(X)$ *and* $X \in \mathbb{C}^{n\times n}$*. Then* $g'(X) = \operatorname{adj}(X^T)$ *and the directional derivative of* $g$ *in the direction* $\Delta \in \mathbb{C}^{n\times n}$ *equals*

$$g'(X) \circ \Delta = \operatorname{tr}(\operatorname{adj}(X)\Delta) = \langle \operatorname{adj}(X)^{\mathsf{H}}, \Delta \rangle_{\mathbb{C}}. \qquad \triangleright$$

Applying the chain-rule we finally obtain the differentiation formula, which is used throughout Chapter 6.

**Corollary A.3.** *Let* $f : \mathbb{R}^{n\times n} + \mathrm{i}\mathbb{R}^{n\times n} \to \mathbb{R}$ *with* $f(X_r + \mathrm{i}X_i) = \ln \Re \det(X)$ *and* $X \in \mathbb{C}^{n\times n}$ *with* $\Re \det(X) > 0$*. Then*

$$\frac{\partial}{\partial X} f(X_r + \mathrm{i}X_i) = \mathbf{c} \circ \left( \frac{\overline{\det(X)}}{\Re \det(X)} X^{-\mathsf{H}} \right).$$

*Moreover, if* $X \in \mathbb{H}_n$ *then*

$$\frac{\partial}{\partial X} f(X_r + \mathrm{i}X_i) = \mathbf{c} \circ \left( X^{-\mathsf{H}} \right). \qquad \triangleright$$

# B. Differences Between Continuous-Time and Discrete-Time Systems

Usually, statements for a continuous linear time-invariant system can be transformed back and forth to discrete-time systems using some bilinear transform. However, the equations determining the analytic center in both cases are cubic in $X$, which suggests that there might not be a one-to-one correspondence. We have shown that the eigenvalues of the feedback system matrix $A_{F_c}$ at the analytic center lie on the imaginary axis in the continuous-time case, whereas they lie inside the unit disk in the discrete-time setting. In this appendix we show that it is indeed necessary to consider the continuous-time and discrete-time case separately by showing that the three equations determining the analytic center are not preserved under the usual bilinear transformations. In the first subsection we show that the domains of the LMIs are the same for the continuous-time and discrete-time cases, but in the second subsection we show that the feedbacks associated with the analytic center are not related via a bilinear transformation.

## B.1. Bilinear Transformations

The bilinear transformation $s = (z-1)/(z+1)$ maps every asymptotically stable continuous-time system $\{A_c, B_c, C_c, D_c\}$ to a corresponding asymptotically stable discrete-time system $\{A_d, B_d, C_d, D_d\}$.

Let us start with an asymptotically stable continuous-time system given by $\{A_c, B_c, C_c, D_c\}$. For some $Q_c \in \mathbb{C}^{n\times n}$ and $R_c = D_c + D_c^{\mathsf{H}}$ set

$$Z_c := \begin{bmatrix} \sqrt{2}(I - A_c)^{-1} & (I - A_c)^{-1}B_c \\ 0 & I \end{bmatrix}, \qquad \tilde{W}_c := \begin{bmatrix} Q_c & C_c^{\mathsf{H}} \\ C_c & R_c \end{bmatrix}, \tag{B.1}$$

and

$$\tilde{W}_c(X_c) \coloneqq \begin{bmatrix} Q_c & C_c^{\mathsf{H}} \\ C_c & R_c \end{bmatrix} - \begin{bmatrix} A_c^{\mathsf{H}} & I \\ B_c^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} 0 & X_c \\ X_c & 0 \end{bmatrix} \begin{bmatrix} A_c & B_c \\ I & 0 \end{bmatrix}.$$

Note, that $\tilde{W}_c(X_c)$ differs from $W_c(X_c)$ as defined in (6.4) by a constant summand, i. e.,

$$\tilde{W}_c(X_c) = W_c(X_c) + \begin{bmatrix} Q_c & 0 \\ 0 & 0 \end{bmatrix}.$$

Then we obtain a transformed discrete-time system by setting

$$\begin{aligned} A_d &\coloneqq (-I + A_c)^{-1}(I + A_c) \\ B_d &\coloneqq \sqrt{2}(-I + A_c)^{-1} B_c \\ \tilde{W}_d &\coloneqq \begin{bmatrix} Q_d & C_d^{\mathsf{H}} \\ C_d & R_d \end{bmatrix} \coloneqq Z_c^{\mathsf{H}} \tilde{W}_c Z_c \\ \tilde{W}_d(X_d) &\coloneqq Z_c^{\mathsf{H}} \tilde{W}_c(X_d) Z_c, \end{aligned} \tag{B.2}$$

where $Q_d, C_d$ and $R_d$ are obtained from $\tilde{W}_d$ and we choose some $D_d$ such that $R_d = D_d + D_d^{\mathsf{H}}$. This defines a mapping $\mathcal{C}$ with $\mathcal{C}(A_c, B_c, C_c, R_c) = (A_d, B_d, C_d, R_d)$. Note, that the transformation $\mathcal{C}$ can also be reversed.

Bilinear transformations preserve asymptotic stability, and they also relate the domains of the continuous-time and discrete-time linear matrix inequalities. To see this, we express the two LMIs as

$$\begin{aligned} \tilde{W}_c(X_c) &= \begin{bmatrix} Q_c & C_c^{\mathsf{H}} \\ C_c & R_c \end{bmatrix} - \begin{bmatrix} A_c^{\mathsf{H}} & I \\ B_c^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} 0 & X_c \\ X_c & 0 \end{bmatrix} \begin{bmatrix} A_c & B_c \\ I & 0 \end{bmatrix} \succeq 0, \\ \tilde{W}_d(X_d) &= \begin{bmatrix} Q_d & C_d^{\mathsf{H}} \\ C_d & R_d \end{bmatrix} - \begin{bmatrix} A_d^{\mathsf{H}} & I \\ B_d^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} X_d & 0 \\ 0 & -X_d \end{bmatrix} \begin{bmatrix} A_d & B_d \\ I & 0 \end{bmatrix} \succeq 0, \end{aligned}$$

respectively. Since

$$\begin{bmatrix} 0 & X_c \\ X_c & 0 \end{bmatrix} = \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \begin{bmatrix} \frac{X_c}{2} & 0 \\ 0 & -\frac{X_c}{2} \end{bmatrix} \begin{bmatrix} I & I \\ I & -I \end{bmatrix},$$

we can also express $\tilde{W}_c(X_c)$ as

$$\tilde{W}_c(X_c) = \begin{bmatrix} Q_c & C_c^{\mathsf{H}} \\ C_c & R_c \end{bmatrix} - \begin{bmatrix} A_c + I & B_c \\ A_c - I & B_c \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \frac{X_c}{2} & 0 \\ 0 & -\frac{X_c}{2} \end{bmatrix} \begin{bmatrix} A_c + I & B_c \\ A_c - I & B_c \end{bmatrix}.$$

Applying the congruence transformation $Z_c$ defined in (B.2), then

$$Z_c^{\mathsf{H}} \tilde{W}_c(X_c) Z_c = \begin{bmatrix} Q_d & C_d^{\mathsf{H}} \\ C_d & R_d \end{bmatrix} - \begin{bmatrix} A_d^{\mathsf{H}} & I \\ B_d^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} X_d & 0 \\ 0 & -X_d \end{bmatrix} \begin{bmatrix} A_d & B_d \\ I & 0 \end{bmatrix},$$

with $A_d, B_d, C_d, R_d$ and $Q_d$ defined as in (B.2). For the transformation of the matrices $C_c$, $R_c$, and $Q_c$ we obtain

$$\begin{bmatrix} Q_d & C_d^{\mathsf{H}} \\ C_d & R_d \end{bmatrix} = Z_c^{\mathsf{H}} \begin{bmatrix} Q_c & C_c^{\mathsf{H}} \\ C_c & R_c \end{bmatrix} Z_c = Z_c^{\mathsf{H}} \begin{bmatrix} \sqrt{2} Q_c (I - A_c)^{-1} & Q_c (I - A_c)^{-1} B_c + C_c^{\mathsf{H}} \\ \sqrt{2} C_c (I - A_c)^{-1} & C_c (I - A_c)^{-1} B_c + R_c \end{bmatrix}, \tag{B.3}$$

where the respective quantities are given by

$$\begin{aligned} Q_d &= 2(I - A_c)^{-\mathsf{H}} Q_c (I - A_c)^{-1}, \\ C_d^{\mathsf{H}} &= \sqrt{2}(I - A_c)^{-\mathsf{H}} Q_c (I - A_c)^{-1} B_c + \sqrt{2}(I - A_c)^{-\mathsf{H}} C_c^{\mathsf{H}}, \\ R_d &= (I - A_c)^{-1} B_c + B_c^{\mathsf{H}} (I - A_c)^{-\mathsf{H}} Q_c (I - A_c)^{-1} B_c + B_c^{\mathsf{H}} (I - A_c)^{-\mathsf{H}} C_c^{\mathsf{H}} + R_c. \end{aligned}$$

This shows, that maximizing $\det W_d(X_d)$ over $X_d$ and maximizing $\det W_c(X_c)$ over $X_c$ is equivalent. In particular, this holds when $Q_c = 0$, which is equivalent to $Q_d = 0$. Thus, the respective continuous-time analytic center $X_{a,c}$ and the discrete-time analytic center $X_{a,d}$ coincide, i. e., $X_{a,c} = X_{a,d}$.

It is well-known, that the bilinear transformation also preserves the solution of the algebraic Riccati equation as well as the domain of the LMI $\tilde{W}_c(X_c) \succeq 0$.

## B.2. Transformation of the Deflating Subspaces

Following [BMMX09] we consider the pencils

$$s\mathcal{E}_c - \mathcal{A}_c := \begin{bmatrix} 0 & -sI + A_c & B_c \\ sI + A_c^{\mathsf{H}} & Q_c & C_c^{\mathsf{H}} \\ B_c^{\mathsf{H}} & C_c & R_c \end{bmatrix}$$

corresponding to the continuous-time case and

$$z\mathcal{A}_d^{\mathsf{H}} - \mathcal{A}_d := \begin{bmatrix} 0 & zI - A_d & -B_d \\ zA_d^{\mathsf{H}} - I & (z-1)Q_d & (z-1)C_d^{\mathsf{H}} \\ zB_d^{\mathsf{H}} & (z-1)C_d & (z-1)R_d \end{bmatrix}$$

corresponding to the discrete-time case, where

$$(A_d, B_d, C_d, R_d) = \mathcal{C}(A_c, B_c, C_c, R_c),$$

see (B.2). If $X_{r,d}$ is a solution of $\mathrm{Ricc}_d(X_{r,d}) = -Q_d$, then there is a deflating subspace of the form

$$\begin{bmatrix} 0 & A_d - zI & B_d \\ I - zA_d^{\mathsf{H}} & (z-1)Q_d & (z-1)C_d^{\mathsf{H}} \\ -zB_d^{\mathsf{H}} & (z-1)C_d & (z-1)R_d \end{bmatrix} \begin{bmatrix} -X_{r,d}\left(I - A_d + B_d F_{r,d}\right) \\ I \\ -F_{r,d} \end{bmatrix}$$
$$= \begin{bmatrix} I \\ (I - A_d^{\mathsf{H}})X_{r,d} \\ -B_d^{\mathsf{H}} X_{r,d} \end{bmatrix} \left(A_d - B_d F_{r,d} - zI\right).$$

Applying a generalized bilinear transformation to the pencil $s\mathcal{E}_c - \mathcal{A}_c$ gives

$$z\hat{\mathcal{A}}_d - \hat{\mathcal{A}}_d^{\mathsf{H}} := z(\mathcal{E}_c - \mathcal{A}_c) - (-\mathcal{E}_c - \mathcal{A}_c)$$
$$= \begin{bmatrix} 0 & z(A_c - I) - (I + A_c) & zB_c - B_c \\ -z(I + A_c^{\mathsf{H}}) - (A_c^{\mathsf{H}} - I) & (z-1)Q_c & (z-1)C_c^{\mathsf{H}} \\ -zB_c^{\mathsf{H}} - B_c^{\mathsf{H}} & (z-1)C_c & (z-1)R_c \end{bmatrix},$$

and then, performing a congruence transformation using $Z_c$ from equation (B.1), we obtain the new pencil

$$z\check{\mathcal{A}}_d - \check{\mathcal{A}}_d{}^{\mathsf{H}} := \begin{bmatrix} \frac{1}{\sqrt{2}}I & 0 \\ 0 & Z_c^{\mathsf{H}} \end{bmatrix} \left(z\hat{\mathcal{A}}_d - \hat{\mathcal{A}}_d{}^{\mathsf{H}}\right) \begin{bmatrix} \frac{1}{\sqrt{2}}I & 0 \\ 0 & Z_c \end{bmatrix}$$
$$= \begin{bmatrix} 0 & A_d - zI & B_d \\ I - zA_d^{\mathsf{H}} & (z-1)Q_d & (z-1)C_d^{\mathsf{H}} \\ -zB_d^{\mathsf{H}} & (z-1)C_d & (z-1)R_d \end{bmatrix}.$$

If, conversely, there is a continuous-time solution $X_{r,c}$ of $\mathrm{Ricc}_c(X_{r,c}) = -Q_c$, we have the deflating subspace

$$\begin{bmatrix} 0 & -sI + A_c & B_c \\ sI + A_c^{\mathsf{H}} & Q_c & C_c^{\mathsf{H}} \\ B_c^{\mathsf{H}} & C_c & R_c \end{bmatrix} \begin{bmatrix} -X_{r,c} \\ I \\ -F_{r,c} \end{bmatrix} = \begin{bmatrix} I \\ X_{r,c} \\ 0 \end{bmatrix} \left(A_c - B_c F_{r,c} - sI\right).$$

Then, using the same transformation we obtain

$$\left(z\breve{\mathcal{A}}_d - \breve{\mathcal{A}}_d^{\mathsf{H}}\right)\begin{bmatrix}\sqrt{2}I & 0\\ 0 & Z_c^{-1}\end{bmatrix}\begin{bmatrix}-X_{r,c}\\ I\\ -F_{r,c}\end{bmatrix}(I - A_c + B_cF_{r,c})^{-1}$$

$$= \begin{bmatrix}\frac{1}{\sqrt{2}}I & 0\\ 0 & Z_c^{\mathsf{H}}\end{bmatrix}\begin{bmatrix}I\\ X_{r,c}\\ 0\end{bmatrix}\left(z(-I + A_c - B_cF_{r,c}) - (I + A_c - B_cF_{r,c})\right)(I - A_c + B_cF_{r,c})^{-1},$$

which is equivalent to

$$\left(z\breve{\mathcal{A}}_d - \breve{\mathcal{A}}_d^{\mathsf{H}}\right)\begin{bmatrix}-X_{r,c}(I - A_{F_{r,d}})\\ I\\ -\sqrt{2}F_{r,c}(I - A_c + B_cF_{r,c})^{-1}\end{bmatrix} = \begin{bmatrix}I\\ (I - A_d)^{\mathsf{H}}X_{r,c}\\ -B_d^{\mathsf{H}}X_{r,c}\end{bmatrix}\left(A_{F_{r,d}} - zI\right),$$

where $A_{F_{r,d}}$ denotes the bilinear transform of the matrix $A_{F_{r,c}} = A_c - B_cF_{r,c}$. One can check that $A_{F_{r,d}}$ fulfills $A_{F_{r,d}} = A_d - B_dF_{r,d}$ with

$$F_{r,d} = \sqrt{2}F_{r,c}(I - A_c + B_cF_{r,c})^{-1}. \tag{B.4}$$

In summary, we have shown the following.

**Theorem B.1.** *Let $X_{r,c}$ solve the algebraic Riccati equation $\mathsf{Ricc}_c(X_{r,c}) = -Q_c$. Then, $X_{r,c}$ also solves the discrete-time Riccati equation $\mathsf{Ricc}_d(X_{r,c}) = -Q_d$, where the corresponding Riccati feedback $F_{r,d}$ is given by* (B.4).

*Also, the corresponding closed-loop matrices $A_{F_{r,c}}$ and $A_{F_{r,d}}$ are related via the bilinear transformation.* ▷

For the analytic center though, we have seen in Section 6.2, that the corresponding closed-loop matrices $A_{F_c}$ and $A_{F_d}$ are not related via a bilinear transform, as all eigenvalues of $A_{F_c}$ lie on the imaginary axis, while all eigenvalues of $A_{F_d}$ lie *inside* the unit circle.

In the remaining part of this section, we give some more explanation for that difference. In particular, we rewrite the continuous-time solution of the analytic center into an appropriately formulated Riccati equation approach, and, after applying the bilinear transformation and enforcing the discrete-time feedback $F_{r,d}$, compare this with its discrete counterpart.

Let us again denote the continuous-time analytic center by $X_{a,c}$. Thus, it fulfills the relation $P_c(X_{a,c}) = \mathrm{Ricc}_c(X_{a,c})$ and hence solves $\mathrm{Ricc}_c(X_{a,c}) = -Q_c$, where $Q_c := -P_c(X_{a,c})$. Consequently, using the bilinear transformation and since $X_{a,c} = X_{a,d}$, also $\mathrm{Ricc}_d(X_{a,d}) = -Q_d$ holds. Note, though, that the coefficients $C_d, R_d$ contained in $\mathrm{Ricc}_d(X)$ and $Q_d$, explicitly depend on $Q_c$. Thus, they do not represent the discrete-time system with $Q_c = Q_d = 0$ that is used for the computation of the analytic center.

Consequently, even though one could expect that also $P_d(X_{a,d}) = -Q_d$, this turns out not to be true. Furthermore, let us compute the quantity $\tilde{P}_d$, by enforcing the feedback $F_{r,d}$ given by (B.4) via the following relation (leaving out the explicit dependence on $X_{a,d}$)

$$\begin{bmatrix} \tilde{P}_d & 0 \\ 0 & R_d - B_d^{\mathsf{H}} X_{a,d} B_d \end{bmatrix} = (Z_P)^{\mathsf{H}} \begin{bmatrix} P_c & 0 \\ 0 & R_c \end{bmatrix} \underbrace{\begin{bmatrix} I & 0 \\ F_{r,c} & I \end{bmatrix} \begin{bmatrix} \sqrt{2}(I - A_c)^{-1} & (I - A_c)^{-1} B_c \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -F_{r,d} & I \end{bmatrix}}_{Z_P :=},$$

where we form

$$Z_P = \begin{bmatrix} \sqrt{2}(I - A_c + B_c F_{r,c})^{-1} & (I - A_c)^{-1} B_c \\ 0 & I + F_{r,c}(I - A_c)^{-1} B_c \end{bmatrix},$$

and have used (B.4) and that $\sqrt{2} F_{r,c}(I - A_c)^{-1} - F_{r,d} - F_{r,c}(I - A_c)^{-1} B_c F_{r,d} = 0$. We thus obtain that

$$\tilde{P}_d = 2(I - A_c + B_c F_{r,c})^{-\mathsf{H}} P_c (I - A_c + B_c F_{r,c})^{-1},$$

which, by considering that $P_c \succ 0$ and equation (B.3), only coincides with $-Q_d$ if $F_{r,c} = 0$.

In particular, we have shown, that if we enforce a discrete feedback $F_{r,d}$ as in (B.4), that keeps the eigenvalues of the closed-loop matrix $A_{F_d}$ at the analytic center on the unit circle, then $\tilde{P}_d \neq -Q_d$ and thus does not equal the discrete-time residual $P_d(X_{a,d})$. Indeed, as mentioned before, the eigenvalues of $A_{F_d}$ lie strictly inside the unit circle.

# C. Notes on Software

For performing the numerical experiments in this thesis, software packages and smaller scripts have been developed. In this appendix we introduce and briefly describe these software packages.

## C.1. Multilevel Optimizations and Optimal Control Problems

For the numerical experiments of Chapters 2, 4 and 5 we developed the `python` package `pymloc`. It is available at [Ban20] and the source code is published under `BSD-3` license. We use the `SciPy` stack [VGOH+20], the differential-algebraic integrator `GELDA` [KMRW95; KMRW97] and for the computation of parameter and time derivatives we use the automatic differentiation tool `jax` [BFHJ+18]. `Jupyter` notebooks [KRPG+16] are provided for easy verification of the numerical examples.

The purpose of this software package is to provide an interface for general multilevel optimization and optimal control problems as introduced in Chapter 2. One main goal of this package was to maintain a certain level of abstraction. Implementation of additional features as new optimization problems or different system classes as, e. g., parameter-dependent partial differential equations, parameter-dependent partial differential-algebraic equations (DAEs), or parameter-dependent port-Hamiltonian systems should be easily doable without touching the general structure of the code too much. This, however, should also increase maintainability of the code.

The main building blocks of a multilevel optimization object are variable containers and optimizations. Optimizations depend on lower level variables, higher level variables, and variables of the current level. Optimization problems can automatically be turned into *local optimizations* by fixing higher and lower level variables to their current values. Additionally, optimization objects may

possess a sensitivity method, which allows to compute sensitivities with respect to higher level variables.

At the time of handing in this thesis, there are two concrete implementations of optimization problems. First, parameter-dependent optimal control problems for strangeness-free differential-algebraic control systems, and, second, nonlinear least squares problems.

The general solution procedure of a multilevel optimal control problem is as follows. After initialization of all optimization problems and mapping variable containers appropriately, the lowermost optimization is turned into a local optimization by fixing higher level variables. A solution is then iteratively passed to the next optimizations, which are localized and solved again, until the uppermost optimization is reached. The localized optimizations are *solvable* objects, which are linked to available solvers through a generic interface.

We present exemplary code for the creation of the computation of Example 5.1 while leaving out parts of the code.

**Importing packages**

[1]:
```
import pymloc
import numpy as np
import jax.numpy as jnp
```

**Creating the variables object** The variables for the different levels are defined as follows.

[2]:
```
from pymloc.model.variables import InputStateVariables
from pymloc.model.variables import NullVariables
from pymloc.model.variables import ParameterContainer
from pymloc.model.variables.time_function import Time
from pymloc.model.domains import RNDomain


loc_vars = InputStateVariables(1, 1, time=Time(0., 2.))
hl_vars = ParameterContainer(1, domain=RNDomain(1))
variables2 = (hl_vars, loc_vars)
```

```
ll_vars = NullVariables()
```

**Creating the parameter optimal control object** First, we need to create a parameter-dependent optimal control problem. We need `objective` and `constraint`.

**Creating the control system** The parameter-dependent control system is defined by

[3]:
```
from pymloc.model.control_system.parameter_dae import␣
↪LinearParameterControlSystem

def e(p, t):
    return jnp.array([[1.]])


def a(p, t):
    return jnp.array([[-1.]])


def b(p, t):
    return jnp.array([[1.]])


def c(p, t):
    return jnp.array([[1.]])


def d(p, t):
    return jnp.array([[0.]])


def f(p, t):
    return jnp.array([0.])
```

```
param_control = LinearParameterControlSystem(ll_vars,␣
 ↪*variables2, e, a, b, c, d, f)
```

**Creating the constraint object** The constraint object is defined by

[4]:
```
from pymloc.model.optimization.parameter_optimal_control␣
 ↪import ParameterLQRConstraint


def initial_value(p):
    return jnp.array([2.])



time = Time(0., 2.)

pdoc_constraint = ParameterLQRConstraint(*variables2,␣
 ↪param_control, initial_value)
```

**Creating the objective function** The objective function is defined by

[5]:
```
from pymloc.model.optimization.parameter_optimal_control␣
 ↪import ParameterLQRObjective

def q(p, t):
    return jnp.array([[p**2. - 1.]])


def s(p, t):
    return jnp.zeros((1, 1))


def r(p, t):
    return jnp.array([[1.]])
```

```
def m(p):
    return jnp.array([[0.]])


time = Time(0., 2.)
pdoc_objective = ParameterLQRObjective(*variables2, time, q,␣
 ↪s, r, m)
```

**Create the parameter-dependent optimal control object**

[6]:
```
from pymloc.model.optimization.parameter_optimal_control␣
 ↪import ParameterDependentOptimalControl

pdoc_object = ParameterDependentOptimalControl(*variables2,␣
 ↪pdoc_objective, pdoc_constraint)
```

The necessary conditions can be obtained as follows

[7]:
```
neccessary_conditions = pdoc_object.get_bvp()
e = neccessary_conditions.dynamical_system.e(2., 3.)
a = neccessary_conditions.dynamical_system.a(2., 3.)
print("E =\n {},\nA =\n {}".format(e, a))
```

```
E =
 [[ 0.  1.  0.]
 [-1.  0.  0.]
 [-0.  0.  0.]],
A =
 [[ 0. -1.  1.]
 [-1.  3.  0.]
 [ 1.  0.  1.]]
```

### Setting up the nonlinear least squares problem

**Define the residual function** We define the residual function by using the true solution for the parameter value at 2 and comparing it with the current solution for the current parameter value.

[8]:
```
def compute_ref_sol(theta, time, x0, t):
    tf = time.t_f
    t0 = time.t_0
    exp0 = np.exp(2 * theta * (tf - t0))
    exp1 = np.exp(-(t + t0) * theta)
    exp2 = np.exp(2 * t * theta)
    exp3 = np.exp(2 * tf * theta)
    tmp1 = theta + exp0 * (theta + 1) - 1
    tmp2 = np.array([
        -(exp2 - exp3) * (theta**2 - 1),
        (exp2 * (theta - 1) + exp3 * (theta + 1)),
        (exp2 - exp3) * (theta**2 - 1)
    ])

    refsol = tmp1**-1 * tmp2 * exp1 * x0
    return refsol
```

[9]:
```
theta = 2.
t0 = 0.
tf = 2.
x01 = 2.
time = Time(t0, tf)
t2 = 1.3
t1 = 1.


def f_nlsq(ll_vars, hl_vars, loc_vars):
    sol1 = compute_ref_sol(theta, time, x01, t1)
    sol2 = compute_ref_sol(theta, time, x01, t2)
```

```
    sol0 = compute_ref_sol(theta, time, x01, t0)
    solf = compute_ref_sol(theta, time, x01, tf)
    f1 = ll_vars(t1)[1] - sol1[1]
    f2 = ll_vars(t2)[1] - sol2[1]
    f0 = ll_vars(t0)[1] - sol0[1]
    ff = ll_vars(tf)[1] - solf[1]
    return np.hstack((f0, f1, f2, ff))
```

**Define the NonlinearLeastSquares optimization object**

[10]:
```
import pymloc.model.optimization as optimization
variables = (variables2[1], NullVariables(), variables2[0])
nlsq_obj =  optimization.objectives.
 ↪NonLinearLeastSquares(*variables, f_nlsq)

nlsq =  optimization.NonLinearLeastSquares(nlsq_obj,␣
 ↪*variables)
```

### Setting up the bilevel optimal control problem

**Compute solution** We are now able to set up the multilevel optimal control problem by setting the optimizations and variables in the corresponding order.

[11]:
```
import logging
logger = logging.getLogger(__name__)
logging.getLogger().setLevel(logging.INFO)

optimizations = [nlsq, pdoc_object]
variables = (variables2[0], variables2[1])
variables[0].current_values = np.array([1.])
variables[1].current_values = np.array([])
variables[1].time.grid = np.array([1., 1.3])
```

We also only want sensitivities of the x component and thus, set the selector accordingly.

[12]:
```
pdoc_object.ll_sens_selector_shape = (1, 3)
pdoc_object.ll_sens_selector = lambda p: np.array([[0., 1.,␣
 ↪0.]])
```

Then, wen can initialize and run the multilevel optimal control problem

[13]:
```
from pymloc import MultiLevelOptimalControl

logger = logging.getLogger("pymloc.solvers.nonlinear.
 ↪gauss_newton")
logger.setLevel(logging.DEBUG)
logging.getLogger().handlers[0].filters[0].__class__.
 ↪max_level = 3
mloc = MultiLevelOptimalControl(optimizations, variables)

np.set_printoptions(precision=8)
mloc.init_solver(abs_tol=1e-6, rel_tol=1e-6)
gauss_newton = mloc.highest_opt.local_level_variables.
 ↪associated_problem.solver_instance
gauss_newton.upper_eta = 0.01
gauss_newton.save_intermediate = True
solution = mloc.solve()
logger.info("Solution: {}".format(solution.solution))
```

```
Starting solver MultiLevelIterativeSolver
Current option values:
     abs_tol: 1e-06
     rel_tol: 1e-06
     max_iter: 10
        Starting solver GaussNewton
        Current option values:
             abs_tol: 1e-06
             rel_tol: 1e-06
             max_iter: 20
        Starting iteration: 0
        Updating lower level variables…
```

```
          Current residual ||J^T r||_2: 0.6660321871331705
          Current ||r||_2: 0.6472009202931233
          Current ||J||_2: 1.0317611699564475
          Current cond(J): 1.0
          Current allowed lower level tolerance: 0.
↪0025501543923860965
          New x value:
               [1.62565779]
               New jac value:
               [[ 0.        ]
                [-0.69271902]
                [-0.64637977]
                [-0.40849072]]
          Starting iteration: 1
          Updating lower level variables…
          Current residual ||J^T r||_2: 0.09021829248969856
          Current ||r||_2: 0.16544194785738267
          Current ||J||_2: 0.5456313369531584
          Current cond(J): 1.0
          Current allowed lower level tolerance: 0.
↪0009145542277046758
          New x value:
               [1.92869503]
               New jac value:
               [[ 0.        ]
                [-0.40312896]
                [-0.32529563]
                [-0.17141571]]
          Starting iteration: 2
          Updating lower level variables…
          Current residual ||J^T r||_2: 0.010023113613176841
          Current ||r||_2: 0.026234802667399967
          Current ||J||_2: 0.38206638166805335
          Current cond(J): 1.0
```

```
          Current allowed lower level tolerance:
0.00020425261292175866
          New x value:
               [1.99735838]
               New jac value:
               [[ 0.        ]
                [-0.29575558]
                [-0.21925259]
                [-0.10213549]]
          Starting iteration: 3
          Updating lower level variables…
          Current residual ||J^T r||_2: 0.00033601046465835875
          Current ||r||_2: 0.000953972969659078
          Current ||J||_2: 0.35227148193310337
          Current cond(J): 1.0
          Current allowed lower level tolerance: 8.
↪387938329980655e-06
          New x value:
               [2.00006606]
               New jac value:
               [[ 0.        ]
                [-0.27553411]
                [-0.20008763]
                [-0.090228  ]]
          Starting iteration: 4
          Updating lower level variables…
          Current residual ||J^T r||_2: 7.553927904696164e-06
          Current ||r||_2: 2.15665120392612e-05
          Current ||J||_2: 0.35068348973442637
          Current cond(J): 1.0
          Current allowed lower level tolerance:
1.9041429186991604e-07
```

```
            New x value:
                 [2.00000464]
                 New jac value:
                 [[ 0.        ]
                  [-0.27440401]
                  [-0.19915694]
                  [-0.0895425 ]]
            Starting iteration: 5
            Updating lower level variables…
            Current residual ||J^T r||_2: 5.705364273345304e-07
            Current ||r||_2: 1.6437829701113807e-06
            Current ||J||_2: 0.3506677917787622
            Current cond(J): 1.0
            Current allowed lower level tolerance:
1.4383951618479082e-08
Solution: [2.]
```

**Perform the same computation for lower tolerances.**

[14]:
```
variables[0].current_values = np.array([1.])
variables =(variables[0], InputStateVariables(1, 1,␣
 ↪time=Time(0., 2.)))
variables[1].current_values = np.array([])
variables[1].time.grid = np.array([1., 1.3])

mloc = MultiLevelOptimalControl(optimizations, variables)
mloc.init_solver(abs_tol=1e-1, rel_tol=1e-1)

gauss_newton = mloc.highest_opt.local_level_variables.
 ↪associated_problem.solver_instance
gauss_newton.upper_eta = 0.1
gauss_newton.save_intermediate = True
solution_low = mloc.solve()
logger.info("Solution: {}".format(solution_low.solution))
```

```
Starting solver MultiLevelIterativeSolver
```

```
   Current option values:
        abs_tol: 0.1
        rel_tol: 0.1
        max_iter: 10
           Starting solver GaussNewton
           Current option values:
                abs_tol: 0.1
                rel_tol: 0.1
                max_iter: 20
           Starting iteration: 0
           Updating lower level variables…
           Current residual ||J^T r||_2: 0.6658138241615206
           Current ||r||_2: 0.6470057103540946
           Current ||J||_2: 1.0317489703858296
           Current cond(J): 1.0
           Current allowed lower level tolerance: 0.
 ↪02331910759868696
           New x value:
                [1.62546746]
                New jac value:
                [[ 0.        ]
                 [-0.69271174]
                 [-0.64637097]
                 [-0.40848617]]
           Starting iteration: 1
           Updating lower level variables…
           Current residual ||J^T r||_2: 0.09548903338269828
           Current ||r||_2: 0.17109116894885706
           Current ||J||_2: 0.5586616718068592
           Current cond(J): 1.0
           Current allowed lower level tolerance: 0.
 ↪008452008146137857
Solution: [1.93142118]
```

More examples and a documentation of the code is available at [Ban20].

## C.2. Computation of the Analytic Center

For the numerical experiments of Chapter 6 we developed the `python` package `analyticcenter`. It is available at [BMNV20a] and published under `BSD-3` license. We use the `SciPy` stack [VGOH+20] and `slycot`, a `python` wrapper for selected `SLICOT` [BMSH+99] routines. `Jupyter` notebooks [KRPG+16] are provided for easy verification of the numerical examples.

This python package is intended to be used for the computation of the analytic center of a strictly passive and stable discrete-time or continuous-time linear time-invariant system.

One can choose between several solvers, i. e., the Newton approach of Subsection 6.3.2 or the steepest descent approach of Subsection 6.3.1.

The following simple example can be found in `examples/example3.py`. Run the example with `python3 -O examples/example3.py` or with `python3 -O -m analyticcenter.examples.example3`. Ommiting the `-O` switch turns on some debugging information.

First one has to create a system object `sys`:

```
import numpy as np

from analyticcenter import WeightedSystem

A = np.matrix([[1, -1], [1, 1]])
B = np.matrix([[1], [1]])
C = B.T
D = C @ B
Q = np.zeros((2,2))
sys = WeightedSystem(A, B, C, D, Q, B, D + D.T)
```

Then one has to create an algorithm object, where one has to define the type of the system (discrete, continuous) and one can optionally provide some tolerances.

```
from analyticcenter import get_algorithm_object


alg = get_algorithm_object(sys, 'newton',
```

```
                    discrete_time=False, save_intermediate=True)
ac = alg()
```

The resulting *analytic center* object `ac` contains data and methods for analyzing the system at the analytic center. More details are available in the documentation of the code at [BMNV20a].

# Bibliography


[AFIJ03] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank. *Matrix Riccati Equations in Control and Systems Theory*. Systems & Control: Foundations & Applications. Birkhäuser Basel, 2003. ISBN: 978-3-7643-0085-2. DOI: 10.1007/978-3-0348-8081-7 (cit. on p. 100).

[Bac06] A. Backes. *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen*. Dissertation. Logos: Berlin, 2006. ISBN: 978-3-8325-1268-2 (cit. on p. 44).

[BM02] K. Balla and R. März. "A Unified Approach to Linear Differential Algebraic Equations and their Adjoints". *Z. Für Anal. Ihre Anwendungen* 21 (3): 783–802 (2002). DOI: 10.4171/ZAA/1108 (cit. on p. 55).

[BL05] K. Balla and V. H. Linh. "Adjoint pairs of differential-algebraic equations and Hamiltonian systems". *Appl. Numer. Math.* Tenth Seminar on Numerical Solution of Differential and Differential-Algebraic Equations (NUMDIFF-10) 53 (2): 131–148 (2005). DOI: 10.1016/j.apnum.2004.08.015 (cit. on pp. 2, 51, 84).

[BM00] K. Balla and R. März. "Linear differential algebraic equations of index 1 and their adjoint equations". *Results Math.* 37 (1-2): 13–35 (2000). DOI: 10.1007/BF03322509 (cit. on pp. 2, 51).

[Ban20] D. Bankmann. *Code and examples for the computation of solutions to multilevel optimal control problems with differential algebraic equations*. Zenodo, 2020. DOI: 10.5281/zenodo.3971868 (cit. on pp. 163, 174).

[BMNV20a] D. Bankmann, V. Mehrmann, Y. Nesterov, and P. Van Dooren. *Code and examples for the paper 'Computation of the analytic center of the solution set of the linear matrix inequality arising in con-*

*tinuous- and discrete-time passivity analysis'*. Zenodo, 2020. DOI: 10.5281/zenodo.3997097 (cit. on pp. 134, 175, 176).

[BMNV20b] D. Bankmann, V. Mehrmann, Y. Nesterov, and P. Van Dooren. "Computation of the Analytic Center of the Solution Set of the Linear Matrix Inequality Arising in Continuous- and Discrete-Time Passivity Analysis". *Vietnam J. Math.*: (2020). DOI: 10.1007/s10013-020-00427-x. arXiv: 1904.08202 (cit. on pp. 113, 137, 138).

[BV18] D. Bankmann and M. Voigt. "On linear-quadratic optimal control of implicit difference equations". *IMA J Math Control Info*: (2018). DOI: 10.1093/imamci/dny007 (cit. on p. 153).

[BL02] P. I. Barton and C. K. Lee. "Modeling, simulation, sensitivity analysis, and optimization of hybrid systems". *ACM Trans. Model. Comput. Simul.* 12 (4): 256–289 (2002). DOI: 10.1145/643120.643122 (cit. on p. 49).

[Bau14] A.-K. Baum. "A flow-on-manifold formulation of differential-algebraic equations. Application to positive systems." Dissertation. Technische Universität Berlin, 2014 (cit. on pp. 2, 27, 30).

[Bau17] A.-K. Baum. "A Flow-on-Manifold Formulation of Differential-Algebraic Equations". *J. Dyn. Differ. Equ.* 29 (4): 1259–1281 (2017). DOI: 10.1007/s10884-015-9511-5 (cit. on pp. 2, 27, 28).

[BMV19] C. Beattie, V. Mehrmann, and P. Van Dooren. "Robust port-Hamiltonian representations of passive systems". *Automatica* 100: 182–186 (2019). DOI: 10.1016/j.automatica.2018.11.013 (cit. on pp. 3, 112, 114, 115, 117, 122).

[Bel97] R. Bellman. *Introduction to matrix analysis, second edition*. Classics in applied mathematics. Society for Industrial and Applied Mathematics: Philadelphia, 1997. ISBN: 978-0-89871-399-2. DOI: 10.1137/1.9781611971170 (cit. on p. 132).

[BLMV15] P. Benner, P. Losse, V. Mehrmann, and M. Voigt. "Numerical linear algebra methods for linear differential-algebraic equations". In: *Surveys in differential-algebraic equations III*. Ed. by A. Ilchmann and T. Reis. Differential-algebaric equations forum. Springer-Verlag: Cham, Switzerland, 2015, pp. 117–175 (cit. on pp. 117, 119).

[BMSH+99] P. Benner, V. Mehrmann, V. Sima, S. V. Huffel, et al. “SLICOT - a subroutine library in systems and control theory”. *Appl. Comput. Control Signals Circuits* 1: 505–546 (1999) (cit. on p. 175).

[Boc87] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*. Bonner mathematische Schriften Nr. 183. Math. Inst. d. Univ., Bibliothek: Bonn, 1987 (cit. on p. 2).

[BP84] H. G. Bock and K.-J. Plitt. “A multiple shooting algorithm for direct solution of optimal control problems”. *IFAC Proc. Vol.* 17 (2): 1603–1608 (1984) (cit. on p. 2).

[Boy01] W. E. Boyce. *Elementary differential equations and boundary value problems*. Wiley: New York, 2001 (cit. on p. 56).

[BEFB94] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM: Philadelphia, PA, 1994 (cit. on p. 113).

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press: New York, NY, USA, 2004. ISBN: 978-0-521-83378-3 (cit. on p. 131).

[BFHJ+18] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.1.64. 2018 (cit. on p. 163).

[BMMX09] R. Byers, D. S. Mackey, V. Mehrmann, and X. Xu. “Symplectic, BVD, and palindromic eigenvalue problems and their relation to discrete-time control problems”. In: *Collection of papers dedicated to the 60-th anniversary of mihail konstantinov*. Publ. House RODINA, Sofia, 2009, pp. 81–102 (cit. on p. 159).

[Cam87] S. L. Campbell. “A general form for solvable linear time varying singular systems of differential equations”. *SIAM J. Math. Anal.* 18: 1101–1115 (1987) (cit. on p. 22).

[CM79] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Pitman: San Francisco, CA, 1979 (cit. on pp. 15, 45).

[CK13] S. L. Campbell and P. Kunkel. “On the numerical treatment of linear–quadratic optimal control problems for general linear time-varying differential-algebraic equations”. *J. Comput. Appl. Math.* 242: 213–231 (2013). DOI: 10.1016/j.cam.2012.10.011 (cit. on pp. 2, 92, 152).

[CKM12] S. L. Campbell, P. Kunkel, and V. Mehrmann. “Regularization of Linear and Nonlinear Descriptor Systems”. In: *Control and Optimization with Differential-Algebraic Constraints*. 0 vols. Advances in Design and Control. Society for Industrial and Applied Mathematics, 2012, pp. 17–36. ISBN: 978-1-61197-224-5. DOI: 10.1137/9781611972252.ch2 (cit. on p. 23).

[CLPS03] Y. Cao, S. Li, L. Petzold, and R. Serban. “Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution”. *SIAM J. Sci. Comput.* 24 (3): 1076–1089 (2003) (cit. on p. 51).

[CMS07] B. Colson, P. Marcotte, and G. Savard. “An overview of bilevel optimization”. *Ann. Oper. Res.* 153 (1): 235–256 (2007). DOI: 10.1007/s10479-007-0176-2 (cit. on p. 1).

[DG19] A. De Marchi and M. Gerdts. “Free finite horizon LQR: A bilevel perspective and its application to model predictive control”. *Automatica* 100: 299–311 (2019). DOI: 10.1016/j.automatica.2018.11.032 (cit. on p. 1).

[DF19] S. Dempe and S. Franke. “Solution of bilevel optimization problems using the KKT approach”. *Optimization* 68 (8): 1471–1489 (2019). DOI: 10.1080/02331934.2019.1581192 (cit. on p. 1).

[Dem02] S. Dempe. *Foundations of bilevel programming*. Nonconvex optimization and its applications. Kluwer: Dordrecht u.a., 2002. ISBN: 978-1-4020-0631-9 (cit. on pp. 1, 6).

[DKPK15] S. Dempe, V. Kalashnikov, G. A. Pérez-Valdés, and N. Kalashnykova. *Bilevel Programming Problems: Theory, Algorithms and Applications to Energy Networks*. Energy Systems. Springer-Verlag: Berlin Heidelberg, 2015. ISBN: 978-3-662-45826-6. DOI: 10.1007/978-3-662-45827-3 (cit. on p. 1).

[DS96] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1996. ISBN: 978-0-89871-364-0. DOI: 10.1137/1.9781611971200 (cit. on pp. 3, 12, 19).

[FMX02] G. Freiling, V. Mehrmann, and H. Xu. "Existence, uniqueness and parametrization of Lagrangian invariant subspaces". *SIAM J. Matrix Anal. Appl.* 23: 1045–1069 (2002) (cit. on p. 117).

[FB05] E. Freitag and R. Busam. *Complex analysis*. 1st ed. Springer: Berlin; New York, 2005. ISBN: 978-3-540-25724-0 (cit. on p. 155).

[GNV99] Y. Genin, Y. Nesterov, and P. Van Dooren. "The analytic center of LMI's and Riccati equations". In: *Control Conference (ECC), 1999 European*. IEEE, 1999, pp. 3483–3487 (cit. on pp. 115, 119).

[Ger12] M. Gerdts. *Optimal Control of ODEs and DAEs*. De Gruyter, 2012. ISBN: 978-3-11-024999-6 (cit. on p. 36).

[GP73] G. H. Golub and V. Pereyra. "The Differentiation of Pseudo-Inverses and Nonlinear Least Squares Problems Whose Variables Separate". *SIAM J. Numer. Anal.* 10 (2): 413–432 (1973). JSTOR: 2156365 (cit. on p. 16).

[GV96] G. H. Golub and C. F. Van Loan. *Matrix computations*. 3rd. Johns Hopkins Univ. Press: Baltimore, 1996 (cit. on p. 16).

[GLN07] S. Gratton, A. S. Lawless, and N. K. Nichols. "Approximate Gauss–Newton methods for nonlinear least squares problems". *SIAM J. Optim.* 18 (1): 106–132 (2007) (cit. on pp. 12, 19–21, 107).

[GJ16] L. Grüne and O. Junge. *Gewöhnliche Differentialgleichungen: Eine Einführung aus der Perspektive der dynamischen Systeme*. In collab. with M. Aigner, H. Faßbender, B. Gentz, D. Grieser, et al. 2., aktualisierte Aufl. 2016. Springer Studium Mathematik - Bachelor. Springer Fachmedien Wiesbaden: Wiesbaden, 2016. ISBN: 978-3-658-10240-1 (cit. on p. 47).

[Hag00] W. W. Hager. "Runge-Kutta methods in optimal control and the transformed adjoint system". *Numer Math (Heidelb)* 87 (2): 247–282 (2000) (cit. on p. 2).

[HW19] F. Harder and G. Wachsmuth. "Optimality conditions for a class of inverse optimal control problems with partial differential equations". *Optimization* 68 (2-3): 615–643 (2019). DOI: 10.1080/02331934.2018.1495205 (cit. on p. 1).

[HSB12] K. Hatz, J. P. Schlöder, and H. G. Bock. "Estimating Parameters in Optimal Control Problems". *SIAM J. Sci. Comput.* 34 (3): A1707–A1728 (2012). DOI: 10.1137/110823390 (cit. on p. 2).

[IM05] A. Ilchmann and V. Mehrmann. "A behavioral approach to time-varying linear systems. Part 1: General theory". *SIAM J. Control Optim.* 44 (5): 1725–1747 (2005) (cit. on p. 22).

[IOW99] V. Ionescu, C. Oara, and M. Weiss. *Generalized Riccati Theory and Robust Control: A Popov Function Approach*. John Wiley & Sons Ltd.: Chichester, 1999. ISBN: 978-0-471-97147-4 (cit. on p. 117).

[Jan71] L. Jantscher. *Distributionen*. Reprint 2013. De Gruyter: Berlin, 1971. ISBN: 978-3-11-001972-8 (cit. on p. 34).

[JWBJ15] Y. Jiang, Y. Wang, S. Bortoff, and Z.-P. Jiang. "Nonlinear optimal co-design based on a modified policy iteration method". *IEEE Trans. Neural Netw. Learn. Syst.* 26 (2): 409–414 (2015). DOI: 10.1109/TNNLS.2014.2382338 (cit. on p. 2).

[Kal63] R. Kalman. "Lyapunov functions for the problem of Lur'e in automatic control". *Proc Nat Acad Sci.* 49: 201–205 (1963) (cit. on p. 114).

[KB06] S. Kameswaran and L. T. Biegler. "Simultaneous dynamic optimization strategies: Recent advances and challenges". *Comput Chem Eng.* Papers form Chemical Process Control VII 30 (10): 1560–1575 (2006). DOI: 10.1016/j.compchemeng.2006.05.034 (cit. on p. 2).

[Kie99] M. Kiehl. "Sensitivity analysis of ODEs and DAEs — theory and implementation guide". *Optim. Methods Softw.* 10 (6): 803–821 (1999). DOI: 10.1080/10556789908805742 (cit. on p. 92).

[KRPG+16] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, et al. “Jupyter Notebooks – a publishing format for reproducible computational workflows”. In: *Positioning and power in academic publishing: Players, agents and agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90 (cit. on pp. 163, 175).

[KSW09] D. Kressner, C. Schröder, and D. S. Watkins. “Implicit QR algorithms for palindromic and even eigenvalue problems”. *Numer Algor* 51 (2): 209–238 (2009). DOI: 10.1007/s11075-008-9226-3 (cit. on p. 143).

[KMS14] P. Kunkel, V. Mehrmann, and L. Scholz. “Self-adjoint differential-algebraic equations”. *Math Control Signals Syst* 26: 47–76 (2014). DOI: 10.1007/s00498-013-0109-3 (cit. on pp. 2, 43).

[KMS02] P. Kunkel, V. Mehrmann, and I. Seufer. *GENDA: A software package for the numerical solution of General Nonlinear Differential-Algebraic equations*. Preprint. Str. des 17. Juni 136, D-10623 Berlin, FRG: Institut für Mathematik, TU Berlin, 2002 (cit. on p. 42).

[KM06] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solution*. European Mathematical Society Publishing House: Zürich, 2006. ISBN: 978-3-03719-017-3 (cit. on pp. 2, 17, 22–25, 30, 34, 42, 44, 46, 92, 93, 108).

[KM08] P. Kunkel and V. Mehrmann. “Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index”. *Math. Control Signals Syst.* 20 (3): 227–269 (2008). DOI: 10.1007/s00498-008-0032-1 (cit. on pp. 2, 25, 36–39, 42, 43).

[KM11a] P. Kunkel and V. Mehrmann. “Formal adjoints of linear DAE operators and their role in optimal control.” *ELA Electron. J. Linear Algebra Electron. Only* 22: 672–693 (2011) (cit. on pp. 2, 40–42, 44, 71, 100).

[KM11b] P. Kunkel and V. Mehrmann. “Optimal Control for Linear Descriptor Systems with Variable Coefficients”. In: *Numerical Linear Algebra in Signals, Systems and Control*. Ed. by P. Van Dooren, S. P. Bhattacharyya, R. H. Chan, V. Olshevsky, et al. Lecture Notes in Electrical Engineering. Springer Netherlands: Dordrecht, 2011,

pp. 313–339. ISBN: 978-94-007-0602-6. DOI: 10.1007/978-94-007-0602-6_15 (cit. on pp. 36, 89, 99, 100).

[KM18] P. Kunkel and V. Mehrmann. "Regular solutions of DAE hybrid systems and regularization techniques". *BIT Numer. Math.* 58 (4): 1049–1077 (2018). DOI: 10.1007/s10543-018-0712-2 (cit. on pp. 49, 152).

[KMRW95] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert. *GELDA: A software package for the solution of General Linear Differential Algebraic equations*. Zenodo, 1995. DOI: 10.5281/zenodo.3972144 (cit. on p. 163).

[KMRW97] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert. "A New Software Package for Linear Differential-Algebraic Equations". *SIAM J. Sci. Comput.* 18 (1): 115–138 (1997). DOI: 10.1137/S1064827595286347 (cit. on pp. 42, 163).

[KMS05] P. Kunkel, V. Mehrmann, and R. Stöver. "Multiple Shooting for Unstructured Nonlinear Differential-Algebraic Equations of Arbitrary Index". *SIAM J. Numer. Anal.* 42 (6): 2277–2297 (2005). DOI: 10.1137/S0036142902418904 (cit. on pp. 31, 93).

[KM04] G. A. Kurina and R. März. "On linear-quadratic optimal control problems for time-varying descriptor systems". *SIAM J. Control Optim.* 42 (6): 2062–2077 (2004) (cit. on p. 2).

[LBBS03] D. B. Leineweber, I. Bauer, H. G. Bock, and J. P. Schlöder. "An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part 1: theoretical aspects". *Comput. Chem. Eng.* 27 (2): 157–166 (2003) (cit. on p. 2).

[LSBS03] D. B. Leineweber, A. Schäfer, H. G. Bock, and J. P. Schlöder. "An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization: Part II: software aspects and applications". *Comput. Chem. Eng.* 27 (2): 167–174 (2003) (cit. on p. 2).

[LP00] S. Li and L. Petzold. “Software and algorithms for sensitivity analysis of large-scale differential algebraic systems”. *J. Comput. Appl. Math.* Numerical Analysis 2000. Vol. VI: Ordinary Differential Equations and Integral Equations 125 (1): 131–145 (2000). DOI: 10.1016/S0377-0427(00)00464-7 (cit. on p. 92).

[Lib12] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012. ISBN: 978-0-691-15187-8. DOI: 10.2307/j.ctvcm4g0s (cit. on pp. 36, 100).

[LT12] D. Liberzon and S. Trenn. “Switched nonlinear differential algebraic equations: Solution theory, Lyapunov functions, and stability”. *Automatica* 48 (5): 954–963 (2012). DOI: 10.1016/j.automatica.2012.02.041 (cit. on p. 49).

[LPS99] J. Lygeros, G. Pappas, and S. Sastry. “An introduction to hybrid system modeling, analysis, and control”. *Prepr. First Nonlinear Control Netw. Pedagog. Sch.*: 307–329 (1999) (cit. on p. 49).

[MN99] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. 2nd ed. John Wiley, 1999. ISBN: 0-471-98632-1 978-0-471-98632-4 0-471-98633-X 978-0-471-98633-1 (cit. on p. 155).

[MMW18] C. Mehl, V. Mehrmann, and M. Wojtylak. “Linear Algebra Properties of Dissipative Hamiltonian Descriptor Systems”. *SIAM J. Matrix Anal. Appl.* 39 (3): 1489–1519 (2018). DOI: 10.1137/18M1164275 (cit. on p. 92).

[Meh16] P. Mehlitz. “Bilevel programming problems with simple convex lower level”. *Optimization* 65 (6): 1203–1227 (2016). DOI: 10.1080/02331934.2015.1122006 (cit. on pp. 1, 153).

[Meh17a] P. Mehlitz. “Contributions to complementarity and bilevel programming in Banach spaces.” Dissertation. Technische Universität Freiberg, 2017 (cit. on pp. 1, 6).

[Meh17b] P. Mehlitz. “Necessary optimality conditions for a special class of bilevel programming problems with unique lower level solution”. *Optimization* 66 (10): 1533–1562 (2017). DOI: 10.1080/02331934.2017.1349123 (cit. on p. 1).

[MW16] P. Mehlitz and G. Wachsmuth. “Weak and strong stationarity in generalized bilevel programming and bilevel optimal control”. *Optimization* 65 (5): 907–935 (2016). DOI: 10.1080/02331934.2015.1122007 (cit. on p. 1).

[Meh91] V. Mehrmann. *The Autonomous Linear Quadratic Control Problem*. Ed. by M. Thoma and A. Wyner. Vol. 163. Lecture Notes in Control and Information Sciences. Springer: Heidelberg, 1991. ISBN: 978-3-540-54170-7 (cit. on p. 142).

[Meh15] V. Mehrmann. “Index Concepts for Differential-Algebraic Equations”. In: *Encyclopedia of Applied and Computational Mathematics*. Ed. by B. Engquist. Springer Berlin Heidelberg: Berlin, Heidelberg, 2015, pp. 676–681. ISBN: 978-3-540-70529-1. DOI: 10.1007/978-3-540-70529-1_120 (cit. on pp. 2, 51).

[MV20] V. Mehrmann and P. Van Dooren. “Optimal Robustness of Port-Hamiltonian Systems”. *SIAM J. Matrix Anal. Appl.* 41 (1): 134–151 (2020). DOI: 10.1137/19M1259092 (cit. on p. 114).

[MW09] V. Mehrmann and L. Wunderlich. “Hybrid systems of differential-algebraic equations – Analysis and numerical solution”. *J. Process Control*. Special Section on Hybrid Systems: Modeling, Simulation and Optimization 19 (8): 1218–1228 (2009). DOI: 10.1016/j.jprocont.2009.05.002 (cit. on p. 49).

[MX00] V. Mehrmann and H. Xu. “Numerical methods in control”. *J. Comput. Appl. Math.* 123 (1): 371–394 (2000). DOI: 10.1016/S0377-0427(00)00392-7 (cit. on pp. 3, 112, 142, 145, 146).

[MZ10] L. Meng and B. Zheng. “The optimal perturbation bounds of the Moore–Penrose inverse under the Frobenius norm”. *Linear Algebra Its Appl.* 432 (4): 956–963 (2010). DOI: 10.1016/j.laa.2009.10.009 (cit. on p. 17).

[Moa05] M. Moakher. “A Differential Geometric Approach to the Geometric Mean of Symmetric Positive-Definite Matrices”. *SIAM J. Matrix Anal. Appl.* 26: 735–747 (2005) (cit. on p. 133).

[MTL09] K. Mombaur, A. Truong, and J.-P. Laumond. “From human to humanoid locomotion—an inverse optimal control approach”. *Auton. Robots* 28 (3): 369–383 (2009). DOI: 10.1007/s10514-009-9170-7 (cit. on p. 2).

[Nes04] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Applied optimization. Kluwer, 2004. ISBN: 978-1-4419-8853-9 (cit. on pp. 119, 131, 132).

[PG16] K. Palagachev and M. Gerdts. “Exploitation of the Value Function in a Bilevel Optimal Control Problem”. In: *System Modeling and Optimization*. Ed. by L. Bociu, J.-A. Désidéri, and A. Habbal. IFIP Advances in Information and Communication Technology. Springer International Publishing: Cham, 2016, pp. 410–419. ISBN: 978-3-319-55795-3. DOI: 10.1007/978-3-319-55795-3_39 (cit. on p. 1).

[PG17] K. D. Palagachev and M. Gerdts. “Numerical Approaches Towards Bilevel Optimal Control Problems with Scheduling Tasks”. *Math Digit. Fact.*: 205–228 (2017). DOI: 10.1007/978-3-319-63957-4_10 (cit. on p. 1).

[PLCS06] L. Petzold, S. Li, Y. Cao, and R. Serban. “Sensitivity analysis of differential-algebraic equations and partial differential equations”. *Comput. Chem. Eng.* Papers form Chemical Process Control VII 30 (10): 1553–1559 (2006). DOI: 10.1016/j.compchemeng.2006.05.015 (cit. on p. 92).

[PW98] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory*. Red. by J. E. Marsden, L. Sirovich, M. Golubitsky, W. Jäger, et al. Vol. 26. Texts in Applied Mathematics. Springer: New York, 1998. ISBN: 978-1-4757-2953-5 (cit. on p. 22).

[Pop73] V. M. Popov. *Hyperstability of Control Systems*. Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 1973. ISBN: 978-0-387-06373-7 (cit. on pp. 114, 117).

[RR88] A. Ran and L. Rodman. “Stability of invariant Lagrangian subspaces I”. In: *Operator theory: Advances and applications*. Ed. by I. Gohberg. Vol. 32. Birkhäuser-Verlag: Basel, Switzerland, 1988, pp. 181–218 (cit. on p. 34).

[RR89] A. Ran and L. Rodman. “Stability of invariant Lagrangian subspaces II”. In: *Operator theory: Advances and applications*. Ed. by H.Dym, S. Goldberg, M. Kaashoek, and P. Lancaster. Vol. 40. Birkhäuser-Verlag: Basel, Switzerland, 1989, pp. 391–425 (cit. on p. 34).

[RRV15] T. Reis, O. Rendel, and M. Voigt. “The Kalman–Yakubovich–Popov inequality for differential-algebraic systems”. *Linear Algebra Appl* 485: 153–193 (2015). DOI: 10.1016/j.laa.2015.06.021 (cit. on p. 153).

[Sch04] A. J. van der Schaft. “Port-Hamiltonian systems: network modeling and control of nonlinear physical systems”. In: *Advanced dynamics and control of structures and machines*. CISM courses and lectures, vol. 444. Springer Verlag: New York, N.Y., 2004 (cit. on p. 114).

[SJ14] A. J. van der Schaft and D. Jeltsema. “Port-Hamiltonian systems theory: An introductory overview”. *Found. Trends Syst. Control* 1 (2-3): 173–378 (2014) (cit. on p. 114).

[SP02] R. Serban and L. Petzold. “Efficient Computation of Sensitivities for Ordinary Differential Equation Boundary Value Problems”. *SIAM J. Numer. Anal.* 40 (1): 220–232 (2002). DOI: 10.1137/S0036142900376870 (cit. on pp. 2, 51, 56, 57, 63).

[SB16] P. G. Stechlinski and P. I. Barton. “Generalized Derivatives of Differential–Algebraic Equations”. *J Optim Theory Appl* 171 (1): 1–26 (2016). DOI: 10.1007/s10957-016-0988-9 (cit. on p. 153).

[SB17] P. G. Stechlinski and P. I. Barton. “Dependence of solutions of nonsmooth differential-algebraic equations on parameters”. *J Differ Equ* 262 (3): 2254–2285 (2017). DOI: 10.1016/j.jde.2016.10.041 (cit. on p. 153).

[Tre09] S. Trenn. “Distributional differential algebraic equations”. Dissertation. Technische Universität Ilmenau, 2009 (cit. on p. 34).

[VGOH+20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. *Nat. Methods* 17 (3): 261–272 (3 2020). DOI: 10.1038/s41592-019-0686-2 (cit. on pp. 163, 175).

[Wal98] W. Walter. *Ordinary Differential Equations*. Vol. 182. Graduate Texts in Mathematics, Readings in Mathematics, Springer New York: New York, NY, 1998. ISBN: 978-1-4612-0601-9. DOI: 10.1007/978-1-4612-0601-9 (cit. on pp. 47, 48).

[WWB16] Y. Wang, Y.-S. Wang, and S. A. Bortoff. “On extension of a gradient-based co-design algorithm to linear descriptor systems”. In: *12th world congress on intelligent control and automation*. 2016, pp. 2388–2393. DOI: 10.1109/WCICA.2016.7578834 (cit. on p. 2).

[Wey12] H. Weyl. “Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)”. *Math. Ann.* 71 (4): 441–479 (1912). DOI: 10.1007/BF01456804 (cit. on p. 17).

[Wil71] J. C. Willems. “Least squares stationary optimal control and the algebraic Riccati equation”. *IEEE Trans. Automat. Control* 16 (6): 621–634 (1971) (cit. on pp. 114, 115, 117, 118).

[Wil72a] J. C. Willems. “Dissipative dynamical systems – Part I: General theory”. *Arch. Ration. Mech. Anal.* 45: 321–351 (1972) (cit. on p. 114).

[Wil72b] J. C. Willems. “Dissipative dynamical systems – Part II: Linear systems with quadratic supply rates”. *Arch. Ration. Mech. Anal.* 45: 352–393 (1972) (cit. on p. 114).

[Wun08] L. Wunderlich. “Analysis and Numerical Solution of Structured and Switched Differential-Algebraic Systems”. Dissertation. TU Berlin, 2008 (cit. on p. 49).

[Yak62] V. A. Yakubovich. “Solution of certain matrix inequalities in the stability theory of nonlinear control systems”. *Dokl. Akad. Nauk. SSSR* 143: 1304–1307 (1962) (cit. on p. 114).

[Ye97] J. Ye. “Optimal Strategies For Bilevel Dynamic Problems”. *SIAM J Control Optim* 35 (2): 512–531 (1997). DOI: 10.1137/S0363012993256150 (cit. on p. 1).

[Zei85] E. Zeidler. *Nonlinear Functional Analysis and its Applications: III: Variational Methods and Optimization*. Springer-Verlag: New York, 1985. ISBN: 978-0-387-90915-8. DOI: 10.1007/978-1-4612-5020-3 (cit. on p. 37).