Generalized Inverses of Differential-Algebraic Operators and their Discretization

vorgelegt von

Dipl.-Math. Ingo Seufer

von 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. Michael Scheutzow
Berichter:	Prof. Dr. Volker Mehrmann
Berichter:	Prof. Dr. Peter Kunkel
Berichter:	Prof. Dr. Caren Tischendorf

Tag der wissenschaftlichen Aussprache: 20.12.2005

Berlin 2006
D 83

Contents

Preface **v**

1 Preliminaries **1**
1.1 The strangeness index 1
1.2 Numerical computation of the strangeness index 10

2 Generalized Inverses of DAOs **17**
2.1 The orthogonal standard form 17
2.2 The Moore-Penrose pseudoinverse 20
2.2.1 Matrix functions 21
2.2.2 Differential-algebraic operators 24
2.3 (1,2,3)-inverses 27

3 Numerical Determination of Generalized Solutions **31**
3.1 BDF-methods and discretization operators 31
3.2 Local Minimization 33
3.3 Global Minimization 48
3.3.1 Systems in orthogonal standard form 49
3.3.2 General strangeness free systems 56
3.3.3 The main theorem 84

4 Numerical Computations **87**
4.1 Algorithms 87
4.1.1 Local minimization 87
4.1.2 Global minimization 88
4.2 Numerical experiments 90
4.2.1 Higher order BDF-methods for local minimization 92
4.2.2 Convergence of global minimization 92
4.2.3 Comparison with GELDA 94
4.2.4 Performance 95
4.2.5 A purely algebraic example 96

5 Conclusions and Outlook **97**

A Proof of Lemma 41 **99**

B Software **111**

Preface

Differential equations are omnipresent in the modeling of physical and chemical processes. Often, theses processes are subject to additional algebraic constraints. Examples for such constraints are Kirchhoff's laws in electrical circuits and position constraints in mechanical systems, see, e.g., [10, 18, 40, 41]. Traditionally, such constraints are resolved by variable substitutions, reducing the model to a system of ordinary differential equations (ODEs). However, such substitutions are not always possible and may be difficult to realize numerically. Furthermore, the increasing size of the models makes such an approach a tedious task.

An elegant alternative to substitutions is to consider the differential equations along with the algebraic constraints in their original form in one single system, which leads to the notion of so called *differential-algebraic equations* (DAEs). In the scope of this thesis, we are concerned with linear DAEs of the form

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad t \in [t_0, T],$$

where E and A are matrix functions. This definition includes linear ODEs (for square and point-wise nonsingular E) as well as purely algebraic systems (for $E \equiv 0$).

Several frameworks for the theoretical and numerical treatment of DAEs have been developed, see, e.g. [6, 16, 19, 31]. Here, we focus on the strangeness index concept developed by Kunkel and Mehrmann [31], as this framework naturally includes under- and overdetermined systems. With the advance of automatic modeling packages, such as ANSYS, COSMOS/M, Modelica and Simulink, redundant constraints and variables are likely to be contained in the model. Such redundancies lead to under- and overdetermined DAEs, which are the topic of this thesis. Representing another application, the design and analysis of linear control systems can be embedded in the framework of underdetermined DAEs using a behavioural approach [26, 32, 38].

This thesis is based on work by Kunkel and Mehrmann [28], which provides a theory that extends the concepts of generalized inverse and least squares solution of linear algebraic equations to linear DAEs. The numerical aspects of this extension, however, were not covered in [28] and are the main purpose of this thesis. We treat the major aspects of the numerical computation of generalized solutions of linear DAEs: discretization, convergence, efficient algorithms and software. Various numerical examples illustrate the obtained theoretical and algorithmic results.

Outline of the thesis

In Chapter 1, we first provide two examples demonstrating that the direct numerical treatment of DAEs may lead to wrong or misleading results. Further, the strangeness index concept is introduced, based on normal forms of linear time-variant DAEs. These normal forms also yield simple conditions for the existence and uniqueness of solutions. Reducing the strangeness index to zero is an important preprocessing step of our numerical methods and can be done numerically based on so called derivative arrays.

Chapter 2 summarizes the theoretical results in [28], which form the basis of our work. First, an

orthogonal standard form of strangeness free linear DAEs is introduced; it allows to distinguish between the differential, algebraic and undetermined parts in the solution. In the spirit of the concept of the Moore-Penrose pseudoinverse for linear algebraic systems, we introduce the notion of differential-algebraic operators and the corresponding generalized inverses, including the notion of a Moore-Penrose pseudoinverse and the least squares solution for differential algebraic equations. It turns out that the latter can be seen as the solution of a linear quadratic optimal control problem. Based on the orthogonal standard form, this connection is used to turn determining the least squares solution of a DAE into solving an equivalent boundary value problem. Another approach to obtain a generalized solution is to force the undetermined part of the solution to be zero. It is shown that these solutions induce a so called (1,2,3)-generalized inverse of the differential-algebraic operator. Let us emphasize, however, that the orthogonal standard form is difficult to realize numerically and consequently we have to consider other means for computing generalized solutions.

Being the main chapter of this thesis, Chapter 3 is concerned with the numerical computation of generalized solutions of linear DAEs. First, we briefly survey BDF-methods and their use for computing unique solutions of strangeness free DAEs. Our goal is to apply BDF-methods for computing generalized solutions of over- and underdetermined DAEs. For this purpose, we introduce a compact notation for the linear system arising from the BDF discretization, using certain restriction operators. Two major approaches are presented for solving this linear system: local and global minimization.

The basic idea of local minimization is to solve the linear systems arising in each step of the BDF-methods *independently* in a least squares sense. It turns out that this leads to an $O(h)$ approximation of a (1,2,3)-generalized solution of the DAE. To prove this result, we compare the solution computed by this approach with the solution of the discretization of a certain, uniquely solvable DAE.

Global minimization consists of solving the *full* linear system obtained from an implicit Euler discretization, again, in a least squares sense. This method leads to an $O(h)$ approximation of the least squares solution of the DAE. The rather technical proof of this result is done in two steps. First, we show this assertion for systems given in orthogonal standard form by exploiting the connection to a discrete linear quadratic optimal control problem that represents a convergent discretization of the underlying boundary value problem. In the second step, we extend this result to the case of general strangeness free DAEs. An important ingredient of the proof is to show that a certain part of the Moore-Penrose inverse of the discretization is uniformly bounded; this is done in Appendix A.

In Chapter 4, two algorithms realizing the derived numerical methods are presented. The computational cost of both algorithms scales linearly with the number of time steps. This desirable property can be directly achieved for local minimization. In the case of global minimization, we present a special-purpose algorithm that takes the particular structure of the discretization matrix into account to achieve the same goal.

The numerical behaviour of these algorithms is tested in several numerical experiments. First, the theoretical result that using higher order BDF-methods does not lead to higher order of convergence in our setting is confirmed. Using an implicit Euler discretization, the actual convergence rate of both, local and global minimization, is verified. It is demonstrated that the software package GELDA [33], which uses a similar local minimization technique for underdetermined DAEs, does not produce satisfying approximations to the (1,2,3)-solution, nor to the least squares solution. It is shown that the computational time needed by our methods scales linearly with the number of time steps, with GELDA being faster than local minimization and

local minimization being faster than global minimization. Using a purely algebraic example, it is confirmed that both solutions produced by local and global minimization coincide in this case. Finally, a real-world application is considered.

Appendix B contains a brief description of the software developed as part of the work on this thesis.

Acknowledgments

First and foremost, I thank my advisors Peter Kunkel and Volker Mehrmann. Introducing me to the world of DAEs, Peter Kunkel has initiated this thesis. I am grateful to him for providing me with scientific, financial as well as personal support throughout my work in Oldenburg and for continuing this support during my Berlin years. Volker Mehrmann has not only generously complemented this support, but he has also introduced me to his vivid working group in Berlin and widened my understanding of DAEs and numerical analysis. I am grateful to both advisors for their patience and trust; several bottles of champagne are soon to be personally delivered.

I would like to thank all members of the Berlin group for providing such a gentle surrounding, in particular my office and smoke mate Michael Karow as well as Christian Mehl, Christian Meyer, Michael Schmidt, Andreas Steinbrecher, Caren Tischendorf and Fredi Tröltzsch for helpful scientific discussions. I appreciate the assistance of Daniel Kressner, University of Zagreb, in deriving Algorithm 3 and the proof of Lemma 41.

I would like to express my deep gratitude to my parents for their encouragement and support over the years.

Last but not least, I thank the Deutsche Forschungsgemeinschaft for the financial support of this work.

Chapter 1

Preliminaries

In this chapter we introduce preliminary definitions and notions used throughout the rest of this thesis. Section 1.1 is concerned with the strangeness index for linear *time-variant* differential-algebraic equations (DAEs). Numerical computations usually require the preliminary reduction of this index to zero. One way to achieve this is to embed the DAE into a larger one, as described in Section 1.2. This also admits the numerical determination of the strangeness index. A much more important consequence is that from the large DAE one can extract a DAE of the size of the original DAE, such that both smaller DAEs have the same solution sets but the extracted DAE has strangeness index zero.

1.1 The strangeness index

Throughout this thesis we consider linear DAEs of the form

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad t \in [t_0, T], \tag{1.1}$$

where $E, A \in C([t_0, T], \mathbb{R}^{m,n})$ and $f \in C([t_0, T], \mathbb{R}^m)$, with an initial condition

$$x(t_0) = x_0. \tag{1.2}$$

If E is a square, point-wise nonsingular matrix function, then the system (1.1) can be transformed into an ordinary differential equation by multiplying both sides with $E(t)^{-1}$ from the left. In this case, it is well known that there exists a unique solution for every given initial value $x_0 \in \mathbb{R}^n$. Furthermore, the system (1.1) can be discretized directly, e.g., by BDF-methods [20].

If the matrix function E is singular or nonsquare then the situation is far more complicated. For example in the special case $E = 0$, the DAE (1.1) becomes a purely algebraic system, which may have several solutions or no solution at all. In this case, an initial value has to satisfy the condition

$$0 = A(t_0)x_0 + f(t_0)$$

in order for (1.1) with (1.2) to be solvable. In the general case, further problems can occur, because often the algebraic conditions of the DAE are not given explicitly but can be hidden in the system (1.1). This frequently causes problems if the DAE is discretized directly as it is usually done with ODEs. The following two examples demonstrate some of these difficulties.

Example 1 ([13, 24]) The DAE

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix} \dot{x}(t) = \begin{bmatrix} -1 & -\eta t \\ 0 & -(\eta + 1) \end{bmatrix} x(t) + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \tag{1.3}$$

with $t \in [t_0, T]$, $f_1, f_2 \in C^2([t_0, T], \mathbb{R})$, is uniquely solvable for any value $\eta \in \mathbb{R}$. Setting $x = [x_1, x_2]^T$ we can rewrite the DAE as

$$0 = -x_1(t) - \eta t x_2(t) + f_1(t), \quad \dot{x}_1(t) + \eta t \dot{x}_2(t) = -(\eta + 1)x_2(t) + f_2(t).$$

Differentiating the first equation gives $\dot{x}_1(t) + \eta t \dot{x}_2(t) = -\eta x_2 + \dot{f}_1(t)$, and together with the second equation we get the solution

$$x_2(t) = f_2(t) - \dot{f}_1(t), \quad x_1(t) = -\eta t x_2(t) + f_1(t) = -\eta t f_2(t) + \eta t \dot{f}_1(t) + f_1(t).$$

Note that it is not necessary to solve any differential equation to compute this solution and that an initial value is consistent if and only if it satisfies these equations for $t = t_0$.

Now we discretize (1.3) with the implicit Euler method with a fixed step size $h = \frac{T - t_0}{N}$. At each time step $t_i = t_0 + ih$, $i = 1, \dots N$ we have to solve the system

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t_i \end{bmatrix} \frac{x_i - x_{i-1}}{h} = \begin{bmatrix} -1 & -\eta t_i \\ 0 & -(\eta + 1) \end{bmatrix} x_i + \begin{bmatrix} f_1(t_i) \\ f_2(t_i) \end{bmatrix}$$

with respect to $x_i = [x_{1,i}, x_{2,i}]^T$. This system can be rewritten as

$$\begin{aligned} x_{1,i} + \eta t_i x_{2,i} &= f_1(t_i), \\ x_{1,i} + (\eta t_i + h(\eta + 1))x_{2,i} &= x_{1,i-1} + \eta t_i x_{2,i-1} + h f_2(t_i). \end{aligned}$$

For $\eta = -1$ this system is singular. For all other values of η we get

$$h(\eta + 1)x_{2,i} = x_{1,i-1} + \eta t_i x_{2,i-1} + h f_2(t_i) - f_1(t_i)$$

and after inserting $x_{1,i-1} = -\eta t_{i-1} x_{2,i-1} + f_1(t_{i-1})$

$$x_{2,i} = \frac{\eta}{\eta + 1} x_{2,i-1} + \frac{f_2(t_i)}{\eta + 1} - \frac{f_1(t_i) - f_1(t_{i-1})}{h(\eta + 1)}.$$

It is obvious that this discretization is not stable if $|\frac{\eta}{\eta+1}| > 1$, i.e., $\eta < -\frac{1}{2}$. In all other cases the solutions of this discretization converge to the correct result. ◊

Example 2 ([16, 30]) The DAE

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{x}(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \tag{1.4}$$

with $t \in [0, T]$, $f_1, f_2 \in C^1([0, T], \mathbb{R})$, can be rewritten as

$$-t\dot{x}_1(t) + t^2 \dot{x}_2(t) = -x_1(t) + f_1(t), \tag{1.5}$$

$$-\dot{x}_1(t) + t\dot{x}_2(t) = -x_2(t) + f_2(t), \tag{1.6}$$

again setting $x = [x_1, x_2]^T$. By multiplying (1.6) with t and subtracting it from (1.5) we get the relation

$$x_1(t) = t x_2(t) + f_1(t) - t f_2(t), \tag{1.7}$$

which gives, after differentiation,

$$\dot{x}_1(t) - t\dot{x}_2(t) = x_2(t) + \dot{f}_1(t) - t\dot{f}_2(t) - f_2(t) \tag{1.8}$$

and thus, by adding (1.6) and (1.8),

$$0 = \dot{f}_1(t) - t\dot{f}_2(t). \tag{1.9}$$

We can see that the DAE (1.4) is solvable if and only if the condition (1.9) is satisfied. If this is the case, then the initial condition has to satisfy (1.7) for $t = 0$ and every function $x = [x_1, x_2]^T \in C^1([0,T], \mathbb{R}^2)$ that satisfies (1.7) solves (1.4). Again we discretize (1.4) as in the previous example using the implicit Euler method, which leads to linear systems of the form

$$\begin{bmatrix} -t_i & t_i^2 \\ -1 & t_i \end{bmatrix} \frac{x_i - x_{i-1}}{h} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x_i + \begin{bmatrix} f_1(t_i) \\ f_2(t_i) \end{bmatrix}$$

for $i = 1, \ldots, N$ with respect to $x_i = [x_{1,i}, x_{2,i}]^T$. These systems possess the unique solution

$$\begin{aligned} x_{2,i} &= \frac{1}{h}(-x_{1,i-1} + t_i x_{2,i-1} + f_1(t_i) + (h - t_i) f_2(t_i)), \\ x_{1,i} &= (h + t_i) x_{2,i} + x_{1,i-1} - t_i x_{2,i-1} - h f_2(t_i). \end{aligned}$$

Thus the implicit Euler method always leads to a unique solution although the system may have several solutions or may not be solvable at all. ◊

Both examples show that a direct discretization of a DAE may lead to wrong or misleading results. They also demonstrate that certain algebraic constraints may be hidden in the DAE and that it is necessary to form derivatives of some parts of the system to detect these constraints. This is a distinctive feature of DAEs.

There are several theories for the analytical and numerical treatment of DAEs. Most of these concepts define an *index* of a DAE, such as the differentiation [6], perturbation [19], strangeness [27, 30] or tractability [16, 34] index. Such indices provide measures of the order of derivatives that have to be computed to extract a system which admits the explicit detection of the algebraic constraints as well as the differential equations of the given DAE. In general, however, these theories require that the system is *regular* in the sense that the DAE possesses a unique solution. Thus they cannot be applied to define and compute generalized solutions of a DAE. A theory that allows for an analytic treatment of nonsquare over- and underdetermined systems is the theory of the *strangeness index* [32], which will therefore suit our purposes.

In the following, we give a short summary of the strangeness index theory for linear DAE systems with variable coefficients as it was introduced in [30] and [29].

In a first step we have to transform the DAE system to a normal form which helps us to examine the behaviour of certain parts of the system. For this purpose we employ the following transformations. Given a point-wise nonsingular matrix function $P \in C([t_0, T], \mathbb{R}^{m,m})$ we can scale the system (1.1) by multiplying with P from the left. The solution space can be transformed by setting $x = Q\tilde{x}$ with $Q \in C^1([t_0, T], \mathbb{R}^{n,n})$ point-wise nonsingular. Because of $\dot{x} = Q\dot{\tilde{x}} + \dot{Q}\tilde{x}$ we obtain that (1.1) is equivalent to the system

$$PEQ\dot{\tilde{x}} = (PAQ - PE\dot{Q})\tilde{x} + Pf.$$

This leads to the following definition of an equivalence relation for pairs of matrix functions.

Definition 1 (global equivalence) *Two pairs (E, A) and $(\tilde{E}, \tilde{A})$ of matrix functions, with $E, A, \tilde{E}, \tilde{A} \in C([t_0, T], \mathbb{R}^{m,n})$, are called* (globally) equivalent *if there are point-wise nonsingular matrix functions $P \in C([t_0, T], \mathbb{R}^{m,m})$ and $Q \in C^1([t_0, T], \mathbb{R}^{n,n})$ such that*

$$\tilde{E} = PEQ, \quad \tilde{A} = PAQ - PE\dot{Q}$$

as equality of functions. We then write $(E, A) \sim (\tilde{E}, \tilde{A})$.

It is easy to verify that this relation is in fact an equivalence relation. Under certain constant rank assumptions a pair (E, A) of matrix functions corresponding to a DAE (1.1) can then be transformed to a normal form that allows for an analytical treatment of the corresponding differential-algebraic system. First we have to define the following quantities that can be computed at any time point $t \in [t_0, T]$.

Definition 2 *Let $(E, A) \in C([t_0, T], \mathbb{R}^{m,n})$ and $t \in [t_0, T]$. Then the quantities*

$$\begin{array}{ll} \text{(a)} & r(t) = \operatorname{rank} E(t), \\ \text{(b)} & a(t) = \operatorname{rank}(Z(t)^T A(t) T(t)), \\ \text{(c)} & s(t) = \operatorname{rank}(V(t)^T Z(t)^T A(t) T'(t))), \\ \text{(d)} & d(t) = r(t) - s(t), \\ \text{(e)} & u(t) = n - r(t) - a(t), \\ \text{(f)} & v(t) = m - r(t) - a(t) - s(t), \end{array}$$

where

$$\begin{array}{ll} \text{(a)} & T(t) \text{ is a basis of } \operatorname{kernel} E(t), \\ \text{(b)} & Z(t) \text{ is a basis of } \operatorname{corange} E(t) = \operatorname{kernel} E(t)^T, \\ \text{(c)} & T'(t) \text{ is a basis of } \operatorname{cokernel} E(t) = \operatorname{range} E(t)^T, \\ \text{(d)} & V(t) \text{ is a basis of } \operatorname{corange}(Z(t)^T A(t) T(t)), \end{array} \tag{1.10}$$

are called local characteristic values *of the pair (E, A) at the point t.*

Note that these local characteristic values can be computed numerically at any given $t \in [t_0, T]$ by three singular value decompositions [15]. To compute a pair $(\tilde{E}, \tilde{A})$ of matrix functions in the normal form that is globally equivalent to the given pair (E, A) we have to make sure that there exist smooth matrix functions T, T', Z and V on $[t_0, T]$ that satisfy (1.10) point-wise. The following theorems show that this is the case whenever E and A are sufficiently smooth and the local characteristic values are constant on $[t_0, T]$.

Theorem 3 *Let $E \in C^l([t_0, T], \mathbb{R}^{m,n})$, $l \in \mathbb{N}_0 \cup \{\infty\}$, with $\operatorname{rank} E(t) = r$ for all $t \in [t_0, T]$. Then there are point-wise orthogonal (and therefore nonsingular) functions $U \in C^\ell([t_0, T], \mathbb{R}^{m,m})$ and $V \in C^\ell([t_0, T], \mathbb{R}^{n,n})$, such that*

$$U^T E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tag{1.11}$$

with point-wise nonsingular $\Sigma \in C^\ell(\mathbb{I}, \mathbb{C}^{r,r})$.

Proof. A detailed proof of this result can be found in [36, 39]. □

Theorem 4 *Let $E, A \in C([t_0, T], \mathbb{R}^{m,n})$ be sufficiently smooth and suppose that*

$$r(t) \equiv r, \quad a(t) \equiv a, \quad s(t) \equiv s \tag{1.12}$$

for the local characteristic values of (E, A). Then (E, A) is globally equivalent to the normal form

$$(\tilde{E}, \tilde{A}) = \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s \\ d \\ a \\ s \\ v \end{matrix}. \tag{1.13}$$

Here, the block entries A_{12}, A_{14}, A_{24} are matrix functions on $[t_0, T]$ and the last block column consists of $u = n - s - d - a$ columns.

Proof. A constructive proof of the normal form (1.13) is given in [27]. □

For the examples we considered in the beginning of this section, we obtain the following normal forms.

Example 3 For Example 1 we get

$$\begin{aligned}
(E,A) &= \left(\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix}, \begin{bmatrix} -1 & -\eta t \\ 0 & -(\eta+1) \end{bmatrix}\right) \\
&\sim \left(\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} -1 & -\eta t \\ 0 & -(\eta+1) \end{bmatrix}\right) \\
&= \left(\begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -(\eta+1) \\ 1 & \eta t \end{bmatrix}\right) \\
&\sim \left(\begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & -\eta t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -(\eta+1) \\ 1 & \eta t \end{bmatrix}\begin{bmatrix} 1 & -\eta t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & -\eta \\ 0 & 0 \end{bmatrix}\right) \\
&= \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\right) \\
&=: (\tilde{E}, \tilde{A}),
\end{aligned}$$

and thus, $s = u = 1$ and $A_{14} = [-1]$. The pair $(\tilde{E}, \tilde{A})$ in normal form corresponds to a DAE

$$\begin{aligned}
\dot{\tilde{x}}_1 &= -\tilde{x}_2 + \tilde{f}_1 \\
0 &= \tilde{x}_1 + \tilde{f}_2,
\end{aligned} \tag{1.14}$$

where

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & -\eta t \\ 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & \eta t \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$\begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

An obvious way to solve the system (1.14) is to differentiate the second equation, subtract it from the first equation and achieve the system

$$\begin{aligned}
0 &= -\tilde{x}_2 + \tilde{f}_1 - \dot{\tilde{f}}_2, \\
0 &= \tilde{x}_1 + \tilde{f}_2,
\end{aligned}$$

which corresponds to the pair of constant matrix functions

$$\left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\right).$$

◊

Example 4 For Example 2 we get

$$\begin{aligned}
(E, A) &= \left(\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 0 & -1 \\ -1 & t \end{bmatrix} \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & t \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \\
&= \left(\begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right) \\
&= \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right) \\
&=: (\tilde{E}, \tilde{A})
\end{aligned}$$

and thus, $s = u = 1$ and $A_{14} = [0]$. The pair $(\tilde{E}, \tilde{A})$ is in normal form and corresponds to the DAE

$$\begin{aligned}
\dot{\tilde{x}}_1 &= \tilde{f}_1, \\
0 &= \tilde{x}_1 + \tilde{f}_2,
\end{aligned}$$

which can be solved again by inserting the derivative of the second equation into the first equation. This leads to the system

$$\begin{aligned}
0 &= \tilde{f}_1 - \dot{\tilde{f}}_2, \\
0 &= \tilde{x}_1 + \tilde{f}_2,
\end{aligned}$$

which can be represented by the pair

$$\left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right).$$

◊

In both examples, the simplification of the DAE relied on differentiating an equation that corresponds to the *strangeness block* I_s of the matrix function $\tilde{A}$ and thus removing the strangeness block in $\tilde{E}$. As will be shown in the following, we can proceed in a similar way with general DAE systems that satisfy the assumptions of Theorem 4. The pair $(\tilde{E}, \tilde{A})$ in (1.13) is associated with a DAE system

$$\begin{aligned}
&\text{(a)} && \dot{\tilde{x}}_1 = A_{12}(t)\tilde{x}_2 + A_{14}(t)\tilde{x}_4 + f_1(t), \\
&\text{(b)} && \dot{\tilde{x}}_2 = A_{24}(t)\tilde{x}_4 + f_2(t), \\
&\text{(c)} && 0 = \tilde{x}_3 + f_3(t), \\
&\text{(d)} && 0 = \tilde{x}_1 + f_4(t), \\
&\text{(e)} && 0 = f_5(t).
\end{aligned} \tag{1.15}$$

This system consists of the algebraic equation (1.15c) of size a for $\tilde{x}_3$, the consistency condition (1.15e) for the inhomogeneity of size v and the differential condition (1.15b) for $\tilde{x}_2$ of size d. Looking at the so-called strangeness equations (1.15a) and (1.15d) of size s we can recognize a coupling for $\tilde{x}_1$. As in the examples we can differentiate (1.15d) and insert it into (1.15a) to obtain the modified equation

$$\text{(a')} \qquad 0 = A_{12}(t)\tilde{x}_2 + A_{14}(t)\tilde{x}_4 + f_1(t) + \dot{f}_4(t). \tag{1.16}$$

If we now replace (1.15a) by (1.16a'), the modified differential algebraic system can be represented by the pair

$$(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}}) = \left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{1.17}$$

This elimination procedure is reversible because the modified system still contains the algebraic equation (1.15d). One can show that the structure of the pair $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ is invariant under global equivalence, i.e., if two pairs $(\tilde{E}^{(1)}, \tilde{A}^{(1)})$ and $(\tilde{E}^{(2)}, \tilde{A}^{(2)})$ are globally equivalent and in global normal form (1.13), then the corresponding modified pairs $(\tilde{E}^{(1)}_{\mathrm{mod}}, \tilde{A}^{(1)}_{\mathrm{mod}})$ and $(\tilde{E}^{(2)}_{\mathrm{mod}}, \tilde{A}^{(2)}_{\mathrm{mod}})$ are also globally equivalent, see [27].

This fact allows for the following inductive procedure. We start with the pair $(E_0, A_0) = (E, A)$ and define a sequence (E_i, A_i), $i \in \mathbb{N}_0$, by transforming (E_i, A_i) to the pair $(\tilde{E}_i, \tilde{A}_i)$ in global canonical form (1.13). Here we have to assume in each step of this procedure that the assumptions of Theorem 4 are satisfied by the pair (E_i, A_i). We then define $(E_{i+1}, A_{i+1}) = (\tilde{E}_{i\mathrm{mod}}, \tilde{A}_{i\mathrm{mod}})$ where $(\tilde{E}_{i\mathrm{mod}}, \tilde{A}_{i\mathrm{mod}})$ is computed from $(\tilde{E}_i, \tilde{A}_i)$ by passing from (1.13) to (1.17).

For every pair (E_i, A_i) we can compute the corresponding characteristic values (r_i, a_i, s_i) as defined in Definition 2. By comparing (1.13) and (1.17) one can see that $r_{i+1} = r_i - s_i$ for $i \in \mathbb{N}_0$ and because r_i cannot be negative, the strangeness must vanish after a finite number of steps, i.e., $s_\mu = 0$ for some $\mu \in \mathbb{N}_0$ and thus the sequence (r_i, a_i, s_i) must become stationary for $i \geq \mu$. The index μ is a characteristic value of the pair (E, A).

Definition 5 (strangeness index) *Let (E, A) be a pair of sufficiently smooth matrix functions. Let the sequence (r_i, a_i, s_i), $i \in \mathbb{N}_0$, be well-defined (in particular, let (1.12) hold for each entry (E_i, A_i) of the above sequence). Then, we call*

$$\mu = \min\{i \in \mathbb{N}_0 \mid s_i = 0\}$$

the strangeness index *of (E, A) and of (1.1). In the case of $\mu = 0$, we call (E, A) and the corresponding DAE (1.1)* strangeness free.

From the discussion above it follows that if the strangeness index is well-defined for a DAE of the form (1.1) then the inductive procedure leads to a strangeness free DAE system, where the associated pair (E_μ, A_μ) transformed to the normal form can be written as

$$(\tilde{E}_\mu, \tilde{A}_\mu) = \left(\begin{bmatrix} I_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & A_{13} \\ 0 & I_a & 0 \\ 0 & 0 & 0 \end{bmatrix} \right). \tag{1.18}$$

To compute the strangeness free DAE it is necessary that at least certain parts of the inhomogeneity are μ times differentiable.

Theorem 6 *Let the strangeness index μ of (E, A) be well-defined (i.e., let the assumptions of Definition 5 hold) and let $f \in C^\mu([t_0, T], \mathbb{C}^m)$. Then the differential-algebraic equation (1.1) is equivalent (in the sense that there is a one-to-one correspondence between the solution spaces via a point-wise nonsingular matrix function) to a differential-algebraic equation of the form*

$$\begin{aligned} &\text{(a)} && \dot{x}_1 = A_{13}(t)x_3 + f_1(t), && d_\mu \\ &\text{(b)} && 0 = x_2 + f_2(t), && a_\mu \\ &\text{(c)} && 0 = f_3(t), && v_\mu \end{aligned} \tag{1.19}$$

where $A_{13} \in C([t_0, T], \mathbb{C}^{d_\mu, u_\mu})$ and the inhomogeneity is determined from $f^{(0)}, \ldots, f^{(\mu)}$.

The existence and uniqueness of solutions of a linear DAE (1.1) can be easily examined after the strangeness free normal form (1.19) has been computed.

Corollary 7 *Let the strangeness index μ of (E, A) be well-defined and let $f \in C^{\mu+1}(\mathbb{I}, \mathbb{C}^m)$. Then we have:*

1. *The problem (1.1) is solvable if and only if the v_μ functional consistency conditions*

$$f_3 = 0$$

   are fulfilled.

2. *An initial condition (1.2) is consistent if and only if in addition the a_μ conditions*

$$x_2(t_0) = -f_2(t_0)$$

   are implied by (1.2).

3. *The corresponding initial value problem is uniquely solvable if and only if in addition*

$$u_\mu = 0$$

   holds.

Let us illustrate these results by applying them to our running examples.

Example 5 For Example 1 we have already computed in Example 3 the normal form

$$(\tilde{E}_0, \tilde{A}_0) = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right)$$

and the corresponding characteristic values

$$r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 1, \quad v_0 = 0.$$

For the modified pair we get

$$\begin{aligned} (E_1, A_1) &= (\tilde{E}_{0\text{mod}}, \tilde{A}_{0\text{mod}}) \\ &= \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right) \\ &\sim \left(\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right) \\ &= \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ &=: (\tilde{E}_1, \tilde{A}_1), \end{aligned}$$

and thus,

$$r_1 = 0, \quad a_1 = 2, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 0, \quad v_1 = 0,$$

and finally the strangeness index $\mu = 1$. The system consists of two algebraic equations and possesses a unique solution, provided that the initial values are consistent. $\Diamond$

Example 6 For Example 2 we have seen that

$$(\tilde{E}_0, \tilde{A}_0) = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right),$$

see Example 4, and thus,

$$r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 1, \quad v_0 = 0.$$

For the modified pair we get

$$\begin{aligned} (E_1, A_1) &= (\tilde{E}_{0\text{mod}}, \tilde{A}_{0\text{mod}}) \\ &= \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right) \\ &\sim \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right) \\ &= \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right) \\ &=: (\tilde{E}_1, \tilde{A}_1). \end{aligned}$$

The characteristic values for the pair $(\tilde{E}_1, \tilde{A}_1)$ are

$$r_1 = 0, \quad a_1 = 1, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 1, \quad v_1 = 1.$$

It follows that this system has strangeness index $\mu = 1$. It consists of an algebraic equation and a consistency condition for the inhomogeneity and it has one undetermined solution component. In particular, the homogeneous initial value problem (for which the consistency condition is satisfied) does not possess a unique solution. ◊

Both results agree with the results that we have already computed in Example 1 and Example 2.

As we have seen, the strangeness index μ is well-defined for a differential-algebraic system (1.1) on an interval $\mathbb{I}$, whenever the constant-rank assumptions of Theorem 4 are satisfied by all pairs (E_i, A_i), $i = 0, \ldots, \mu$, provided that they are sufficiently smooth on $\mathbb{I}$. Due to the fact that the rank of any continuous matrix function can at most change outside of a dense subset of open intervals in a given closed interval $\mathbb{I}$ (see, e. g., [7]), we get the following result.

Corollary 8 *Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $E, A \in C(\mathbb{I}, \mathbb{C}^{m,n})$ be sufficiently smooth. Then there exist open intervals $\mathbb{I}_j$, $j \in \mathbb{N}$, with*

$$\overline{\bigcup_{j \in \mathbb{N}} \mathbb{I}_j} = \mathbb{I}, \quad \mathbb{I}_i \cap \mathbb{I}_j = \emptyset \quad \textit{for} \quad i \neq j,$$

such that the strangeness index of (E, A) restricted to $\mathbb{I}_j$ is well-defined for every $j \in \mathbb{N}$.

We have seen so far that the strangeness index can be defined for a large class of linear differential-algebraic equations. The fact that the aforementioned theory can also be applied to underdetermined and even unsolvable systems leads to intuitive solvability and uniqueness results for differential-algebraic systems. In addition, this theory can also be applied to control problems, where the input variables can just be treated as undetermined solution components [32, 25, 26].

However, it is in general not clear how to turn the above procedure into a reliable numerical method for computing the strangeness index. On the one hand it is difficult to realize the smooth transformations to compute the normal form (1.13), on the other hand for systems with a higher strangeness index the derivatives needed here cannot be computed accurately. In the next section, a method will be presented that avoids these difficulties.

1.2 Numerical computation of the strangeness index

For the computation of the strangeness index of a linear DAE (1.1), one has to compute derivatives of certain parts of the pair (E, A) of matrix functions. In addition, if we want to solve a given differential algebraic equation by discretization, we have already seen that this might lead to wrong or misleading results if the discretization is applied directly to a higher-index system. Therefore, one wants to compute a system that has the same solution set as the original system but which is strangeness free. This, however, makes it necessary to differentiate certain parts of the inhomogeneity but it is well-known that the numerical computation of higher derivatives is not necessarily stable [17]. It is also difficult or even infeasible to employ automatic differentiation techniques [17]; as one has to compute derivatives of *transformed* data, this would require to differentiate the original data *and* the transformations.

Therefore, it makes sense to assume that the necessary derivatives are given in advance and defined solely in terms of the original data. A system that contains all the required derivatives of the pair (E, A) and of the inhomogeneity should also contain all the information that is needed to compute the strangeness index. We will see that such a system also allows the computation of a strangeness free differential-algebraic equation which has the same solution set as the given DAE.

The following idea was first introduced by Campbell [6]. If we differentiate the DAE (1.1) once with respect to t then we obtain

$$\dot{E}(t)\dot{x}(t) + E(t)\ddot{x}(t) = \dot{A}(t)x(t) + A(t)\dot{x}(t) + \dot{f}(t).$$

After sorting all the derivatives of x to the left-hand side of the equation, we can merge this system with the original DAE to obtain the differential-algebraic equation

$$\begin{bmatrix} E(t) & 0 \\ \dot{E}(t) - A(t) & E(t) \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \ddot{x}(t) \end{bmatrix} = \begin{bmatrix} A(t) & 0 \\ \dot{A}(t) & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} + \begin{bmatrix} f(t) \\ \dot{f}(t) \end{bmatrix}. \tag{1.20}$$

If all coefficients and f are sufficiently smooth then the solution of the DAE (1.1) coincides with the solution of the inflated system (1.20). The same method can also be realized for higher derivatives. If we build l derivatives of (1.1), then we can combine all these derivatives to the so-called *inflated differential-algebraic equation* or *derivative array*

$$M_l(t)\dot{z}(t) = N_l(t)z(t) + g_l(t), \quad t \in [t_0, T],$$

where the coefficients are given by

$$\begin{aligned} (M_l)_{i,j} &= \binom{i}{j} E^{(i-j)} - \binom{i}{j+1} A^{(i-j-1)}, \quad i, j = 0, \ldots, l, \\ (N_l)_{i,j} &= \begin{cases} A^{(i)} & \text{for } i = 0, \ldots, l, \quad j = 0, \\ 0 & \text{otherwise,} \end{cases} \\ (z_l)_j &= x^{(j)}, \quad j = 0, \ldots, l, \\ (g_l)_i &= f^{(i)}, \quad i = 0, \ldots, l. \end{aligned}$$

Here we use the convention

$$\binom{i}{j} = 0 \quad \text{for } i < 0, \quad j < 0 \quad \text{or } j > i.$$

As for the pair (E, A) we can compute the local characteristic values of the inflated pairs (M_l, N_l) according to Definition 2. If the strangeness index μ of (E, A) is well-defined, then the local characteristic values of the pairs (M_l, N_l), $l = 0, \ldots, \mu$, can be used to compute the sequence (r_i, a_i, s_i) of global characteristic values of the pairs (E_i, A_i) that have been defined in the previous section.

Theorem 9 ([29, 30]) *Let the strangeness index μ of (E, A) be well-defined and let $(\tilde{r}_l, \tilde{a}_l, \tilde{s}_l)$, $l = 0, \ldots, \mu$, be the sequence of the local characteristic values of $(M_l(t), N_l(t))$ for some $t \in \mathbb{I}$. Then the sequence (r_i, a_i, s_i) of global characteristic values of (E, A) can be obtained by*

$$
\begin{array}{ll}
\text{(a)} & c_0 = \tilde{a}_0 + \tilde{s}_0, \quad c_{i+1} = (\tilde{a}_{i+1} - \tilde{a}_i) + (\tilde{s}_{i+1} - \tilde{s}_i), \\
\text{(b)} & v_0 = m - c_0 - \tilde{r}_0, \quad v_{i+1} = m - c_{i+1} - (\tilde{r}_{i+1} - \tilde{r}_i), \\
\text{(c)} & s_i = c_i - \tilde{a}_i, \\
\text{(d)} & a_i = c_0 + \cdots + c_i - s_i, \\
\text{(e)} & r_i = m - a_i - s_i - v_i.
\end{array} \tag{1.21}
$$

This means that if the strangeness index μ of a given DAE is well-defined, then we can get the information about the complete sequence of global characteristic values of the pairs (E_i, A_i) as they were constructed in the previous chapter from the inflated pair (M_μ, N_μ). In particular, they can be computed numerically. The following theorem can then be used to construct a strangeness free system which has the same solution set as the original DAE.

Theorem 10 *Let the strangeness index μ be well-defined for the pair (E, A) and let (r_i, a_i, s_i), $i = 0, \ldots, \mu$, be the sequence of global characteristic values of the pairs (E_i, A_i). Setting*

$$\hat{a} = a_\mu, \quad \hat{d} = d_\mu, \quad \hat{v} = v_0 + \cdots + v_\mu,$$

where $d_\mu = r_\mu$ and $v_i = m - r_i - a_i - s_i$ for $i = 1, \ldots, \mu$, then the inflated pair (M_μ, N_μ) has the following properties.

1. *For all $t \in [t_0, T]$ we have $\operatorname{rank} M_\mu(t) = (\mu + 1)m - \hat{a} - \hat{v}$. This implies the existence of a smooth matrix function Z of size $((\mu + 1)m, \hat{a} + \hat{v}))$ and point-wise orthonormal columns, satisfying $Z^T M_\mu = 0$.*

2. *For all $t \in [t_0, T]$ we have $\operatorname{rank} Z(t)^T N_\mu(t)[I_n\, 0 \cdots 0]^T = \hat{a}$. This implies that, without loss of generality, Z can be partitioned as $Z = [Z_2\, Z_3]$ with Z_2 of size $((\mu + 1)m, \hat{a})$ and Z_3 of size $((\mu + 1)m, \hat{v})$, such that $\hat{A}_2 = Z_2^T N_\mu [I_n\, 0 \cdots 0]^T$ has full row rank $\hat{a}$ and $Z_3^T N_\mu [I_n\, 0 \cdots 0]^T = 0$. Furthermore, there exists a smooth matrix function T_2 of size $(n, \hat{d})$, $\hat{d} = m - \hat{a} - \hat{v}$, and point-wise orthonormal columns, satisfying $\hat{A}_2 T_2 = 0$.*

3. *For all $t \in [t_0, T]$ we have $\operatorname{rank} E(t) T_2(t) = \hat{d}$. This implies the existence of a smooth matrix function Z_1 of size $(m, \hat{d})$ and point-wise orthonormal columns such that $\hat{E}_1 = Z_1^T E$ has constant rank $\hat{d}$.*

Furthermore, the system

$$\begin{bmatrix} \hat{E}_1(t) \\ 0 \\ 0 \end{bmatrix} \dot{x}(t) = \begin{bmatrix} \hat{A}_1(t) \\ \hat{A}_2(t) \\ 0 \end{bmatrix} x(t) + \begin{bmatrix} \hat{f}_1(t) \\ \hat{f}_2(t) \\ \hat{f}_3(t) \end{bmatrix}, \tag{1.22}$$

with $\hat{A}_1 = Z_1^T A$, $\hat{f}_1 = Z_1^T f$, $\hat{f}_2 = Z_2^T g_\mu$ and $\hat{f}_3 = Z_3^T g_\mu$, is strangeness free and has the same solution set as the given DAE.

Proof. See [32] and [30]. □

Note that in general $\hat{v} > m - \hat{d} - \hat{a}$, which implies that in these cases the system (1.22) consists of more equations than the original DAE (1.1). The function $\hat{f}_3$ contains the function f_3 as defined in (1.19) and derivatives of some parts of it. Hence if $\hat{f}_3(t) = 0$ for all $t \in [t_0, T]$ then we know that the original system is solvable.

For the numerical treatment of linear differential-algebraic equations another fact is important. By means of three singular value decompositions we can compute matrix functions $\tilde{Z}_1$, $\tilde{Z}_2$ and $\tilde{Z}_3$ that can replace the functions Z_1, Z_2 and Z_3 in the sense that they satisfy the conditions given in Theorem 10 but without being smooth. The numerical realization of smooth transformations is extremely expensive and will generally be impossible here, see [5] and [36].

Without loss of generality we may assume the functions Z_1, Z_2 and Z_3 to be orthogonal due to Theorem 3. What we can compute numerically are point-wise evaluations of $\tilde{Z}_1 = Z_1Q_1$, $\tilde{Z}_2 = Z_2Q_2$ and $\tilde{Z}_3 = Z_3Q_3$, where Q_1, Q_2 and Q_3 are orthogonal but not necessarily smooth. Thus if we compute the system (1.22) using these matrices we may get a system where the coefficient functions are only smooth (apart from roundoff errors) after a multiplication with $\operatorname{diag}(Q_1, Q_2, Q_3)$ from the left. This scaling, however, does not change the solution space of the system and thus we can neglect the possibly nonsmooth realization of the functions Z_1, Z_2 and Z_3.

Due to Theorem 9 and Theorem 10 we are now able to treat linear differential-algebraic equations numerically, whenever their strangeness index is well-defined and the inflated pair (M_μ, N_μ) is given. At every point $t \in [t_0, T]$ we can compute the characteristic values for the system using only local information at this point. We can extract a strangeness free DAE that has the same solution space as the original system. In Chapter 3, we will see that this system can be discretized, e.g. with BDF-methods, and that we get the same convergence results as for ordinary differential equations.

Example 7 In Example 5 we have already computed the global (and hence local) characteristic values of the pair

$$(M_0, N_0) = (E, A) = \left(\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix}, \begin{bmatrix} -1 & -\eta t \\ 0 & -(\eta+1) \end{bmatrix} \right),$$

according to Example 1, and thus we have

$$\tilde{r}_0 = r_0 = 1, \quad \tilde{a}_0 = a_0 = 0, \quad \tilde{s}_0 = s_0 = 1, \quad \tilde{v}_0 = v_0 = 0.$$

After differentiating the DAE we get the inflated pair

$$(M_1, N_1) = \left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & \eta t & 0 & 0 \\ 1 & \eta t & 0 & 0 \\ 0 & 2\eta+1 & 1 & \eta t \end{bmatrix}, \begin{bmatrix} -1 & -\eta t & 0 & 0 \\ 0 & -(\eta+1) & 0 & 0 \\ 0 & -\eta & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right).$$

By means of the transformations

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 & -\eta t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -2\eta - 1 & -\eta t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

we can transform the pair (M_1, N_1) to the normal form

$$\begin{aligned} (\tilde{M}_1, \tilde{N}_1) &= (PM_1Q, PN_1Q - PM_1\dot{Q}) \\ &= \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \eta \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \right) \end{aligned}$$

and read off the global characteristic values

$$\tilde{r}_1 = 2, \quad \tilde{a}_1 = 1, \quad \tilde{s}_1 = 1, \quad \tilde{v}_1 = 0.$$

Using the recursion (1.21) we get

$$\begin{aligned} c_0 &= \tilde{a}_0 + \tilde{s}_0 = 1, \quad c_1 = (\tilde{a}_1 - \tilde{a}_0) + (\tilde{s}_1 - \tilde{s}_0) = 1, \\ v_1 &= m - c_1 - (\tilde{r}_1 - \tilde{r}_0) = 0, \\ s_1 &= c_1 - \tilde{a}_1 = 0, \\ a_1 &= c_0 + c_1 - s_1 = 2, \\ r_1 &= m - a_1 - s_1 - v_1 = 0, \end{aligned}$$

and this agrees with the result that we have obtained in Example 5. To obtain the strangeness free DAE system as defined in Theorem 10 we can choose

$$Z = Z_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix}.$$

as the matrix containing a basis of the corange of M_1. Now for

$$\hat{A}_2 = Z_2^T N_1 \begin{bmatrix} I_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} -1 & -\eta t \\ 0 & -(\eta+1) \\ 0 & -\eta \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -\eta t \\ 0 & -1 \end{bmatrix}$$

we have $\operatorname{rank}(\hat{A}_2) = 2$ and we immediately get the strangeness free system

$$0 = \hat{A}_2(t)x(t) + \hat{f}_2(t) = \begin{bmatrix} -1 & -\eta t \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} f_1(t) \\ f_2(t) - \dot{f}_1(t) \end{bmatrix}. \tag{1.23}$$

The unique solution of (1.3) can now be computed without any discretization by solving the equivalent algebraic system (1.23). ◊

Example 8 For Example 2 we have already computed the global and local characteristic values (see Example 6) for

$$(M_0, N_0) = (E, A) = \left(\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right)$$

as

$$\tilde{r}_0 = r_0 = 1, \quad \tilde{a}_0 = a_0 = 0, \quad \tilde{s}_0 = s_0 = 1, \quad \tilde{v}_0 = v_0 = 0.$$

The inflated pair

$$(M_1, N_1) = \left(\begin{bmatrix} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ 0 & 2t & -t & t^2 \\ 0 & 2 & -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right)$$

can be transformed to the global normal form

$$\begin{aligned}(\tilde{M}_1, \tilde{N}_1) &= (PM_1Q, PN_1Q - PM_1\dot{Q}) \\ &= \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right)\end{aligned}$$

by means of the transformations

$$P = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & t & 0 & 0 \\ 0 & 0 & 1 & -t \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 & t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & t \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This yields the global and local characteristic values

$$\tilde{r}_1 = 2, \quad \tilde{a}_1 = 0, \quad \tilde{s}_1 = 1, \quad \tilde{v}_1 = 1.$$

The recursion (1.21) then gives

$$\begin{aligned} c_0 &= \tilde{a}_0 + \tilde{s}_0 = 1, \quad c_1 = (\tilde{a}_1 - \tilde{a}_0) + (\tilde{s}_1 - \tilde{s}_0) = 0, \\ v_1 &= m - c_1 - (\tilde{r}_1 - \tilde{r}_0) = 1, \\ s_1 &= c_1 - \tilde{a}_1 = 0, \\ a_1 &= c_0 + c_1 - s_1 = 1, \\ r_1 &= m - a_1 - s_1 - v_1 = 0, \end{aligned}$$

again in accordance with the computation in Example 6.

The strangeness free system can be obtained from (M_1, N_1) as follows. The matrix functions

$$Z_2 = \begin{bmatrix} 1 \\ -t \\ 0 \\ 0 \end{bmatrix}, \quad Z_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -t \end{bmatrix}$$

span the corange of M_1, $Z_3^T N_1 [I_2\, 0] = 0$, and

$$\hat{A}_2 = Z_2^T N_1 \begin{bmatrix} I_2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & t \end{bmatrix}$$

has full row rank. Thus, by Theorem 10 we get

$$0 = \begin{bmatrix} \hat{A}_2(t) \\ 0 \end{bmatrix} x(t) + \begin{bmatrix} \hat{f}_2(t) \\ \hat{f}_3(t) \end{bmatrix} = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} f_1(t) - tf_2(t) \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix}. \tag{1.24}$$

As already observed in Example 2, the system is solvable if the condition $\dot{f}_1(t) - t\dot{f}_2(t) = 0$ is satisfied for all $t \in [t_0, T]$ but the solution is not unique. ◊

In the next chapters we will only consider strangeness free linear differential-algebraic equations. This restriction is justified by the aforementioned results, which show that we can compute equivalent strangeness free systems numerically for every given system at any time point t, whenever the strangeness index is well-defined. As already mentioned, the functions Z_1, Z_2 and Z_3 defined in Theorem 10 can be chosen such that they have orthogonal columns. Combined with the fact that the Euclidean norm for vector spaces is invariant under orthogonal transformations, this property will admit the definition and computation of least-square solutions for differential-algebraic systems of higher index.

Chapter 2

Generalized Inverses of Differential-Algebraic Operators

Being closely related to finding least-squares solutions of over/underdetermined linear systems, the notion of generalized inverses of differential-algebraic operators is introduced in this chapter. For this purpose, an orthogonal standard form based on the strangeness free form (1.22) is defined. Furthermore, it is shown that computing these generalized inverses is equivalent to solving a linear-quadratic optimal control problem or a boundary value problem; a connection that will be used to justify the numerical method proposed in Section 3.3. The exposition in this chapter is along the lines of the work in [28].

2.1 The orthogonal standard form

The strangeness free differential-algebraic system (1.22) defined in Theorem 10 can be computed by using orthogonal transformations from the left only. The transformations do not change the solution set of the given differential-algebraic system. The system (1.22) allows to distinguish between a differential equation

$$\hat{E}_1(t)\dot{x}(t) = \hat{A}_1(t)x(t) + \hat{f}_1(t)$$

of dimension $\hat{d}$ and a purely algebraic equation

$$0 = \hat{A}_2(t)x(t) + \hat{f}_2(t)$$

of dimension $\hat{a}$, but it does not distinguish between the parts of the solution that belong to differential or algebraic parts of the system or those components of the solution that are undetermined. This could be achieved by transforming the system to the normal form (1.19), but such an approach requires a non-orthogonal transformation of the solution space. The *orthogonal standard form* proposed in this section is derived from (1.22) solely by means of *orthogonal* transformations. This form will allow us to define least-square solutions of differential-algebraic equations.

Theorem 11 *Let the DAE (1.1) be strangeness free and (E, A) sufficiently smooth. Then there exist matrix functions $P \in C([t_0, T], \mathbb{R}^{m,m})$ and $Q \in C^1([t_0, T], \mathbb{R}^{n,n})$, both point-wise orthogonal, such that we can transform (1.1) to the* orthogonal standard form

$$\tilde{E}(t)\dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{f}(t), \tag{2.1}$$

where

$$
\begin{aligned}
\tilde{E}(t) &= P(t)E(t)Q(t) = \begin{bmatrix} \Sigma_E(t) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
\tilde{A}(t) &= P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & \Sigma_A(t) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
\tilde{x}(t) &= Q(t)^T x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}, \\
\tilde{f}(t) &= P(t)f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{bmatrix},
\end{aligned}
\tag{2.2}
$$

with Σ_E and Σ_A point-wise nonsingular. All block sizes are allowed to be zero.

Proof. While in [28], the existence of (2.1) was implicitly assumed, we will show constructively that (2.1) always exists if the original DAE (1.1) is strangeness free.

By Theorem 10 there exist smooth matrix functions Z_1 of size $(m, \hat{d})$, Z_2 of size $(m, \hat{a})$ and Z_3 of size $(m, \hat{v})$, where $\hat{d} = \operatorname{rank}(E)$, $\hat{a} = \operatorname{rank}([Z_2 Z_3]^T A) = \operatorname{rank}(Z_2^T A)$ and $\hat{v} = m - \hat{d} - \hat{a}$, such that $[Z_1\ Z_2\ Z_3]$ is point-wise orthogonal and

$$
[Z_1\ Z_2\ Z_3]^T(E, A) = \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right).
$$

By Theorem 3 there exist orthogonal and smooth matrix functions U_1 and V_1 such that

$$
U_1^T \hat{E}_1 V_1 = [\Sigma_E \quad 0]
$$

with point-wise nonsingular Σ_E of size $(\hat{d}, \hat{d})$. Because $\hat{E}_1$ has full row rank, one can choose $U_1 = I$ and we get

$$
\begin{aligned}
\left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right) &\sim \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix} V_1, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} V_1 - \begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix} \dot{V}_1 \right) \\
&= \left(\begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \\ 0 & 0 \end{bmatrix} \right)
\end{aligned}
$$

where $[\hat{A}_{11}\ \hat{A}_{12}] = \hat{A}_1 V_1 - \hat{E}_1 \dot{V}_1$ and $[\hat{A}_{21}\ \hat{A}_{22}] = \hat{A}_2 V_1$ are partitioned according to the size of Σ_E. By Theorem 10 there exists a matrix function T_2 of size $(n, \hat{d})$ with point-wise orthogonal columns such that $\operatorname{rank}(E(t)T_2(t)) = \hat{d}$ in $[t_0, T]$ and $\hat{A}_2 T_2 = 0$. If we partition

$$
V_1^T T_2 = \begin{bmatrix} T_2' \\ T_2'' \end{bmatrix},
$$

such that T_2'' is of size $(\hat{d}, \hat{d})$, then

$$
\operatorname{rank}(E(t)T_2) = \operatorname{rank}\left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix} T_2 \right) = \operatorname{rank}\left(\begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} T_2' \\ T_2'' \end{bmatrix} \right) = \hat{d}
$$

implies that T_2' is nonsingular. From

$$\hat{A}_2 T_2 = \begin{bmatrix} \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \begin{bmatrix} T_2' \\ T_2'' \end{bmatrix} = \hat{A}_{21} T_2' + \hat{A}_{22} T_2''$$

it follows that $\hat{A}_{21} = -\hat{A}_{22} T_2'' T_2'^{-1}$ and this implies $\operatorname{rank}(\hat{A}_{22}) = \operatorname{rank}([\hat{A}_{21}\ \hat{A}_{22}]) = \hat{a}$.

Again, Theorem 3 shows the existence of smooth orthogonal matrix functions U_2 and V_2 such that

$$U_2^T \hat{A}_{22} V_2 = \begin{bmatrix} \Sigma_A & 0 \end{bmatrix},$$

and consequently we obtain

$$\begin{aligned} \left(\begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \\ 0 & 0 \end{bmatrix} \right) &\sim \left(\begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} V_2 \\ U_2^T \hat{A}_{21} & U_2^T \hat{A}_{22} V_2 \\ 0 & 0 \end{bmatrix} \right) \\ &= \left(\begin{bmatrix} \Sigma_E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & \Sigma_A & 0 \\ 0 & 0 & 0 \end{bmatrix} \right). \end{aligned}$$

□

Like the normal form (1.18), the orthogonal normal form allows to distinguish between the different components of the solution $\tilde{x}$, namely the differential part x_1, the algebraic part x_2 and the undetermined part x_3 of size $\hat{u} = n - \hat{d} - \hat{a}$. Suppose that $f_3(t) = 0$ for all t and thus the DAE is solvable, then x_3 can be chosen arbitrarily, just like an input variable in a control system. For any input x_3 the variable x_1 has to be the solution of the ordinary differential equation

$$\dot{x}_1(t) = -\Sigma_E^{-1}(t)(A_{11}(t)x_1(t) + A_{12}(t)x_2(t) + A_{13}(t)x_3(t) + f_1(t)),$$

where the component x_2 has to satisfy the algebraic condition

$$x_2(t) = -\Sigma_A^{-1}(t)(A_{21}(t)x_1(t) + f_2(t)).$$

It follows that any initial condition can be assigned to x_1, while the overall initial condition is consistent if and only if the equation

$$x_{20} = -\Sigma_A^{-1}(t_0)(A_{21}(t_0)x_{10} + f_2(t_0))$$

holds, where

$$\begin{bmatrix} x_{10} \\ x_{20} \\ x_{30} \end{bmatrix} = Q(t_0)^T x_0.$$

Note that it is in general difficult to compute the orthogonal standard form numerically due to the difficulties associated with realizing the necessary transformations smoothly, as already discussed at the end of Section 1.1.

2.2 The Moore-Penrose pseudoinverse

Given $A \in \mathbb{R}^{m,n}$ and $b \in \mathbb{R}^m$, a general system of linear equations

$$Ax = b,$$

which may be over- or underdetermined, can be "solved" uniquely by considering the minimization problem

$$\frac{1}{2}\|x\|_2^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Ax - b\|_2^2 = \min! \tag{2.3}$$

This problem always has a unique solution, which is called *least squares solution* and can be written in the form

$$x = A^+ b,$$

where $A^+ \in \mathbb{R}^{n,m}$ is the Moore-Penrose pseudoinverse of A [3]. The Moore-Penrose pseudoinverse can be computed by means of a singular value decomposition

$$A = U \begin{bmatrix} \Sigma_A & 0 \\ 0 & 0 \end{bmatrix} V^T,$$

where $U \in \mathbb{R}^{m,m}$ and $V \in \mathbb{R}^{n,n}$ are orthogonal and $\Sigma_A \in \mathbb{R}^{a,a}$, $a = \operatorname{rank}(A)$, is a diagonal matrix containing the nonzero singular values of A. Observe that

$$\begin{aligned} \|Ax - b\|_2 &= \|UAV^T Vx - U^T b\|_2 = \left\| \begin{bmatrix} \Sigma_A & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T x \\ V_2^T x \end{bmatrix} - \begin{bmatrix} U_1^T b \\ U_2^T b \end{bmatrix} \right\|_2 \\ &= \left\| \begin{bmatrix} \Sigma_A V_1^T x - U_1^T b \\ U_2^T b \end{bmatrix} \right\|_2 \end{aligned}$$

is minimized for $V_1^T x = \Sigma_A^{-1} U_1^T b$ and

$$\|x\|_2 = \|V^T x\|_2 = \left\| \begin{bmatrix} V_1^T x \\ V_2^T x \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} \Sigma_A^{-1} U_1^T b \\ V_2^T x \end{bmatrix} \right\|_2$$

is minimal for

$$x = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma_A^{-1} U_1^T b \\ 0 \end{bmatrix} = V_1 \Sigma_A^{-1} U_1^T b.$$

Therefore, we obtain

$$A^+ = V_1 \Sigma_A^{-1} U_1^T = V \begin{bmatrix} \Sigma_A^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T. \tag{2.4}$$

Here, $V = [U_1\, U_2]$ and $U = [V_1\, V_2]$ are partitioned according to the above block structure, i.e., $U_1 \in \mathbb{R}^{m,a}$ and $V_1 \in \mathbb{R}^{n,a}$.

The Moore-Penrose pseudoinverse satisfies the four Moore-Penrose axioms

$$\begin{aligned} &(1) \quad AA^+A = A, \\ &(2) \quad A^+AA^+ = A^+, \\ &(3) \quad (AA^+)^T = AA^+, \\ &(4) \quad (A^+A)^T = A^+A. \end{aligned} \tag{2.5}$$

On the other hand, for given $A \in \mathbb{R}^{m,n}$, the four axioms (2.5) fix a unique matrix $A^+ \in \mathbb{R}^{n,m}$, see [7], which can be computed via the formula (2.4).

If, for a given matrix $A \in \mathbb{R}^{m,n}$, a matrix $A^- \in \mathbb{R}^{n,m}$ satisfies only some of the Moore-Penrose axioms it is also called a generalized inverse of A. If, e.g., A^- satisfies the Moore-Penrose axioms (1), (2) and (3) it is called a (1,2,3)-inverse of A. Analogously, one can define (1,2,4)-inverses or (1,2)-inverses of a given matrix. Of course, in general, these generalized inverses are not uniquely defined [7].

Another way to interpret the definition of the Moore-Penrose pseudoinverse in the context of linear operators is to consider the homomorphism $A : \mathbb{R}^n \to \mathbb{R}^m$ induced by the matrix A. There exists a linear mapping which maps a vector $b \in \mathbb{R}^m$ onto the unique solution $x \in \mathbb{R}^n$ of the minimization problem (2.3). The matrix representation of this mapping is given by the matrix A^+.

2.2.1 The Moore-Penrose pseudoinverse for matrix functions

This well-known theory can be easily generalized to the case of matrix functions between spaces of smooth functions. To see this, let us consider an equation of the form

$$A(t)x(t) = f(t), \qquad t \in \mathbb{I} \tag{2.6}$$

with $A \in C^l(\mathbb{I}, \mathbb{R}^{m,n})$, $f \in C^l(\mathbb{I}, \mathbb{R}^m)$, $l \in \mathbb{N}_0$, on some interval $\mathbb{I}$, along with the minimization problem

$$\frac{1}{2}\|x\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Ax - f\|^2 = \min! \tag{2.7}$$

with respect to the norm

$$\|x\|_2 = \sqrt{(x,x)}, \qquad (x,y) = \int_{\mathbb{I}} x(t)^T y(t) dt.$$

Provided that the conditions of Theorem 3 hold, we will see in Lemma 16 below that the time-variant orthogonal decomposition (1.11) of A implies the existence of a unique solution of (2.7). In this case, a pseudoinverse operator can be computed similarly as for linear systems, see (2.4). We will now set up the appropriate spaces and reformulate the Moore-Penrose axioms for this problem, using basic tools from functional analysis [22].

Definition 12 *Let $\mathbb{X}$ be a vector space with an inner product $(\cdot,\cdot)$ and let $A : \mathbb{X} \to \mathbb{X}$ be an endomorphism. An endomorphism $A^* : \mathbb{X} \to \mathbb{X}$ is called a* conjugate *of A if and only if*

$$(Ax, x^*) = (x, A^* x^*)$$

for all $x, x^* \in \mathbb{X}$.

In the setting of Definition 12, a conjugate is always unique and there exists the following rule for the conjugate of a product of endomorphisms.

Lemma 13 *Let $\mathbb{X}$ be a vector space with an inner product $(\cdot,\cdot)$ and let $A : \mathbb{X} \to \mathbb{X}$ be an endomorphism. There is at most one endomorphism $A^* : \mathbb{X} \to \mathbb{X}$ conjugate to A.*

Let the endomorphisms $A^*, B^* : \mathbb{X} \to \mathbb{X}$ be conjugate to the endomorphisms $A, B : \mathbb{X} \to \mathbb{X}$. Then AB has the conjugate $(AB)^*$ given by

$$(AB)^* = B^* A^*.$$

Proof. See, e.g., [22]. □

Now we are able to define a Moore-Penrose pseudoinverse of a linear operator between vector spaces.

Definition 14 *Let $\mathbb{X}$ and $\mathbb{Y}$ be two vector spaces with an inner product $(\cdot,\cdot)$ and let $D : \mathbb{X} \to \mathbb{Y}$ be a homomorphism. A homomorphism $D^+ : \mathbb{Y} \to \mathbb{X}$ is called a* Moore-Penrose pseudoinverse *of D if DD^+ and D^+D possess conjugates $(DD^+)^*$ and $(D^+D)^*$, and the relations*

$$\begin{array}{ll} (1) & DD^+D = D, \\ (2) & D^+DD^+ = D^+, \\ (3) & (DD^+)^* = DD^+, \\ (4) & (D^+D)^* = D^+D \end{array} \tag{2.8}$$

hold.

As for linear systems, it will be shown in the following lemma that the four axioms of Definition 14 guarantee the uniqueness of the Moore-Penrose pseudoinverse. In general, the existence of such an inverse cannot be shown, but in the case of matrix functions $A \in C^l(\mathbb{I}, \mathbb{R}^{n,m})$ the function A^+ defined via the smooth orthogonal decomposition (1.11) satisfies these four axioms.

Lemma 15 *Let $\mathbb{X}$ and $\mathbb{Y}$ be two vector spaces with an inner product $(\cdot,\cdot)$ and $D : \mathbb{X} \to \mathbb{Y}$ be a homomorphism. Then D has at most one Moore-Penrose pseudoinverse $D^+ : \mathbb{Y} \to \mathbb{X}$.*

Proof. For completeness, let us recall the proof given in [28]. Let D^+, $\tilde{D}^+ : \mathbb{Y} \to \mathbb{X}$ be two Moore-Penrose pseudoinverses of D. Then we have

$$\begin{aligned} D^+ &= D^+DD^+ = D^+D\tilde{D}^+DD^+ = (D^+D)^*(\tilde{D}^+D)^*D^+ \\ &= (\tilde{D}^+DD^+D)^*D^+ = (\tilde{D}^+D)^*D^+ = \tilde{D}^+DD^+ = \tilde{D}^+(DD^+)^* \\ &= \tilde{D}^+(D\tilde{D}^+DD^+)^* = \tilde{D}^+(DD^+)^*(D\tilde{D}^+)^* = \tilde{D}^+DD^+D\tilde{D}^+ = \tilde{D}^+D\tilde{D}^+ = \tilde{D}^+. \end{aligned}$$

□

The following lemma shows that the Moore-Penrose pseudoinverse of a matrix function A can be defined and computed analogously to the case of systems of linear equations provided that the rank of $A(t)$ is constant for all t.

Lemma 16 *Let $A \in C^l(\mathbb{I}, \mathbb{R}^{m,n})$ be a matrix function with $\operatorname{rank}(A(t)) = a$ for all $t \in \mathbb{I}$. Then the minimization problem (2.7) possesses a unique solution $x \in C^l(\mathbb{I}, \mathbb{R}^n)$ for every inhomogeneity $f \in C^l(\mathbb{I}, \mathbb{R}^m)$. The matrix function $A^+ \in C^l(\mathbb{I}, \mathbb{R}^{n,m})$ that maps f by pointwise multiplication onto this solution, i.e. $x(t) = A^+(t)f(t)$, is the Moore-Penrose pseudoinverse of A.*

Proof. According to Theorem 3 there exist unitary matrix functions $U \in C^l(\mathbb{I}, \mathbb{R}^{m,m})$ and $V \in C^l(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$A = U \begin{bmatrix} \Sigma_A & 0 \\ 0 & 0 \end{bmatrix} V^T, \tag{2.9}$$

where $\Sigma_A \in C^l(\mathbb{I}, \mathbb{R}^{a,a})$ is point-wise nonsingular. We now define

$$\begin{aligned} \tilde{A} &= U^TAV = \begin{bmatrix} \Sigma_A & 0 \\ 0 & 0 \end{bmatrix}, \\ \tilde{f} &= U^Tf = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \\ \tilde{x} &= V^Tx = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

The minimization problem

$$\frac{1}{2}\|\tilde{x}\|^2 = \left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|^2 \text{min!} \quad \text{s.t.} \quad \frac{1}{2}\|\tilde{A}\tilde{x} - \tilde{f}\|^2 = \left\| \begin{bmatrix} \Sigma_A x_1 - f_1 \\ f_2 \end{bmatrix} \right\|^2 = \text{min!}$$

has the unique solution

$$\tilde{x} = \begin{bmatrix} \Sigma_A^{-1} f_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_A^{-1} & 0 \\ 0 & 0 \end{bmatrix} \tilde{f}.$$

Moreover, the matrix function

$$\tilde{A}^+ = \begin{bmatrix} \Sigma_A^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

satisfies, together with $\tilde{A}$, the four Moore-Penrose axioms and thus is the Moore-Penrose pseudoinverse of $\tilde{A}$. Because of

$$\|x\| = \|V^T x\| = \|\tilde{x}\|, \quad \|Ax - f\| = \|U^T AVV^T x - U^T f\| = \|\tilde{A}\tilde{x} - \tilde{f}\|,$$

the minimization problem (2.7) transforms covariantly with the transformations U and V and

$$A^+ = V\tilde{A}^+ U^T$$

is the Moore-Penrose pseudoinverse of A. This can be easily verified by inserting A and A^+ in the equations (2.8). The unique solution of the minimization problem (2.7) is given by $x = A^+ f$. □

We can now compute the least squares solution and the Moore-Penrose pseudoinverse for Example 1 following the lines of the proof of Lemma 16.

Example 9 ([28]) The strangeness free DAE (1.24) derived from system (1.4) can be written in the form $A(t)x(t) = f(t)$, where

$$A(t) = \begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) - tf_2(t) \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix}.$$

We have $A = \tilde{A}V^T$ with

$$\tilde{A} = \begin{bmatrix} \sqrt{1+t^2} & 0 \\ 0 & 0 \end{bmatrix}, \quad V = \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix}.$$

Then the Moore-Penrose pseudoinverse of A is given by

$$A^+ = V\tilde{A}^+ = V \begin{bmatrix} \frac{1}{\sqrt{1+t^2}} & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{1+t^2} \begin{bmatrix} 1 & 0 \\ -t & 0 \end{bmatrix}$$

and for the solution of the minimization problem (2.7) we get

$$x = A^+ f = \frac{1}{1+t^2} \begin{bmatrix} f_1(t) - tf_2(t) \\ -tf_1(t) + t^2 f_2(t) \end{bmatrix}.$$

◊

2.2.2 The Moore-Penrose pseudoinverse for differential-algebraic operators

The (algebraic) equation (2.6) can be interpreted as a special case of a differential-algebraic equation with $\hat{d} = 0$ (for the definition of $\hat{d}$, see Theorem 10). So the question arises naturally, if least squares or similarly generalized solutions can also be defined for general DAEs. For this purpose, we introduce linear *differential-algebraic operators* (DAOs) and appropriate spaces.

Let us first assume that the DAE is strangeness free and that it is given in the orthogonal standard form (2.1). This enables us to detect the differential components x_1 of the solution $\tilde{x} = Q^T x$. Only these components of the solution have to be differentiable and only for these components we can provide an initial condition. We must require that the DAE has the trivial solution for $\tilde{f} \equiv 0$ in the uniquely solvable case. Otherwise a mapping that maps $\tilde{f}$ onto the solution of (2.1) cannot be linear. Therefore, we only allow for initial values $\tilde{x}(t_0) = \tilde{x}_0 = 0$. This can always be obtained by shifting $\tilde{x}(t)$ to $\tilde{x}(t) - \tilde{x}_0$ and changing the inhomogeneity $\tilde{f}(t)$ to $\tilde{f}(t) + \tilde{A}(t)\tilde{x}_0$. (Of course, this can be done for any DAE, which is not necessarily in orthogonal standard form.)

In the following, we assume that the pair $(\tilde{E}, \tilde{A})$ is in orthogonal standard form. We set

$$\tilde{\mathbb{X}} = \{\tilde{x} \in C([t_0, T], \mathbb{R}^n) | x_1 \in C^1([t_0, T], \mathbb{R}), x_1(t_0) = 0\}, \tag{2.10}$$

$$\tilde{\mathbb{Y}} = C([t_0, T], \mathbb{R}^m), \tag{2.11}$$

where $\tilde{x}$ is partitioned as in (2.2). Moreover, let us define a *differential-algebraic operator* $\tilde{D} : \tilde{\mathbb{X}} \to \tilde{\mathbb{Y}}$ by

$$\tilde{D}\tilde{x}(t) = \tilde{E}(t)\dot{\tilde{x}}(t) - \tilde{A}(t)\tilde{x}(t). \tag{2.12}$$

This operator allows to rewrite the DAE (2.1) in the compact form

$$\tilde{D}\tilde{x} = \tilde{f}.$$

Having defined differential-algebraic operators for DAEs in orthogonal standard form, we can easily generalize this concept for general strangeness free DAEs by setting

$$D = P^T \tilde{D} Q^T, \tag{2.13}$$

where the product $P^T \tilde{D} Q^T$ should be understood point-wise and the operators P, Q represent the transformations to orthogonal standard form. Then we obtain

$$\begin{aligned}
Dx(t) &= P(t)^T \tilde{D} Q(t) x(t) \\
&= P(t)^T \tilde{D} \tilde{x}(t) \\
&= P(t)^T \tilde{E}(t)\dot{\tilde{x}}(t) - P(t)^T \tilde{A}(t)\tilde{x}(t) \\
&= P(t)^T P(t) E(t) Q(t) \left(\dot{Q}(t)^T x(t) + Q(t)^T \dot{x}(t)\right) \\
&\quad - P(t)^T \left(P(t)A(t)Q(t)Q(t)^T x(t) - P(t)E(t)\dot{Q}(t)Q^T(t)x(t)\right) \\
&= E(t)\dot{x}(t) - A(t)x(t) + E(t)\left(Q(t)\dot{Q}(t)^T + \dot{Q}(t)^T Q(t)\right) x(t) \\
&= E(t)\dot{x}(t) - A(t)x(t),
\end{aligned}$$

using the identity

$$Q(t)\dot{Q}(t)^T + \dot{Q}(t)Q(t)^T = \frac{d}{dt}\left(Q(t)Q(t)^T\right) = \frac{d}{dt} I = 0.$$

The spaces where the operator D acts upon can also be transformed, which gives $D : \mathbb{X} \to \mathbb{Y}$ with

$$\begin{aligned}\mathbb{X} &= \{x \in C([t_0, T], \mathbb{R}^n) | E^+Ex \in C^1([t_0, T], \mathbb{R}^n), E^+Ex(t_0) = 0\},\\ \mathbb{Y} &= C([t_0, T], \mathbb{R}^m).\end{aligned}$$

Here, the operator

$$E^+E = Q \begin{bmatrix} \Sigma_E^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} PP^T \begin{bmatrix} \Sigma_E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T = Q \begin{bmatrix} I_{\hat{d}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T$$

is a projector onto the differential component of $x = Q\tilde{x}$.

Summarizing the discussion, we are able to rewrite any strangeness free linear DAE (1.1) in the form

$$Dx = f.$$

Our aim is to show that the minimization problem

$$\frac{1}{2}\|x\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Dx - f\|^2 = \min! \tag{2.14}$$

possesses a unique solution and that this solution induces an operator $D^+ : \mathbb{Y} \to \mathbb{X}$, which satisfies the four Moore-Penrose axioms and is therefore a Moore-Penrose pseudoinverse of the differential-algebraic operator D.

First, we will show this result for the operator $\tilde{D}$ defined for DAEs in orthogonal standard form. The subsequent extension of this result to general linear DAEs will then be rather straightforward.

The minimization problem

$$\frac{1}{2}\|\tilde{x}\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|\tilde{D}\tilde{x} - \tilde{f}\|^2 = \min! \tag{2.15}$$

can be written in explicit form as

$$\begin{aligned}&\frac{1}{2}\int_{t_0}^{T} \tilde{x}(t)^T\tilde{x}(t)dt = \min!\\ &\quad \text{s.t.} \quad \frac{1}{2}\int_{t_0}^{T} \left(w_1(t)^T w_1(t) + w_2(t)^T w_2(t) + w_3(t)^T w_3(t)\right) dt = \min!,\end{aligned} \tag{2.16}$$

where

$$w_1(t) = \Sigma_E(t)\dot{x}_1(t) - A_{11}(t)x_1(t) - A_{12}(t)x_2(t) - A_{13}(t)x_3(t) - f_1(t), \tag{2.17}$$

$$w_2(t) = -A_{21}(t)x_1(t) - \Sigma_A(t)x_2(t) - f_2(t), \tag{2.18}$$

$$w_3(t) = -f_3(t). \tag{2.19}$$

The constraint can easily be satisfied, because we can solve the system $w_1 = 0$, $w_2 = 0$ in $\tilde{\mathbb{X}}$ for an arbitrary continuous function x_3. For this purpose we eliminate the function x_2 in (2.17) using

$$x_2(t) = -\Sigma_A(t)^{-1}\left(A_{21}(t)x_1(t) + f_2(t)\right). \tag{2.20}$$

Next, we solve the initial value problem

$$\begin{aligned}\dot{x}_1(t) = {} & \Sigma_E(t)^{-1}\left(A_{11}(t) - A_{12}(t)\Sigma_A(t)^{-1}A_{21}(t)\right)x_1(t)\\ & +\Sigma_E(t)^{-1}A_{13}(t)x_3(t) + \Sigma_E(t)^{-1}\left(f_1(t) - A_{12}(t)\Sigma_A(t)^{-1}f_2(t)\right), \quad x_1(t_0) = 0,\end{aligned}$$

and compute x_2 according to (2.20).

The problem (2.16) turns out to be a linear quadratic optimal control problem [21, 35]. If we set

$$\begin{aligned}A(t) &= \Sigma_E(t)^{-1}\left(A_{11}(t) - A_{12}(t)\Sigma_A(t)^{-1}A_{21}(t)\right),\\ B(t) &= \Sigma_E(t)^{-1}A_{13}(t),\\ C(t) &= -\Sigma_A(t)^{-1}A_{21}(t),\\ f(t) &= \Sigma_E(t)^{-1}\left(f_1(t) - A_{12}(t)\Sigma_A(t)^{-1}f_2(t)\right),\\ g(t) &= -\Sigma_A(t)^{-1}f_2(t),\end{aligned} \tag{2.21}$$

and rename the components of $\tilde{x}$ by setting $x = x_1$, $y = x_2$ and $u = x_3$, the problem (2.15) turns into

$$\begin{aligned}&\frac{1}{2}\int_{t_0}^{T}\left(x(t)^Tx(t) + y(t)^Ty(t) + u(t)^Tu(t)\right)dt = \min!\\ &\text{s.t.}\quad \begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t)u(t) + f(t), \quad x(t_0) = 0,\\ y(t) &= C(t)x(t) + g(t).\end{aligned}\end{aligned} \tag{2.22}$$

Hence the undetermined part x_3 of the variable $\tilde{x}$ can be interpreted as the input variable of this control problem. The problem (2.22) is a generalization of the standard linear quadratic control problems due to the inhomogeneities that appear in the constraints. As for standard control problems, it can be shown that (2.22) possesses a unique solution.

Theorem 17 *Let*

$$\begin{aligned}&A \in C([t_0,T],\mathbb{R}^{\hat{d},\hat{d}}), \quad B \in C([t_0,T],\mathbb{R}^{\hat{d},\hat{u}}), \quad C \in C([t_0,T],\mathbb{R}^{\hat{a},\hat{d}}),\\ &f \in C([t_0,T],\mathbb{R}^{\hat{d}}), \quad g \in C([t_0,T],\mathbb{R}^{\hat{a}}).\end{aligned}$$

Then the linear quadratic control problem (2.22) possesses a unique solution $x \in C^1([t_0,T],\mathbb{R}^{\hat{d}})$, $y \in C([t_0,T],\mathbb{R}^{\hat{a}})$, $u \in C([t_0,T],\mathbb{R}^{\hat{u}})$. This solution coincides with the corresponding part of the unique solution of the boundary value problem

$$\begin{aligned}\dot{\lambda}(t) &= (I + C(t)^TC(t))x(t) - A(t)^T\lambda(t) + C(t)^Tg(t), \quad \lambda(T) = 0,\\ \dot{x}(t) &= A(t)x(t) + B(t)u(t) + f(t), \quad x(t_0) = 0,\\ y(t) &= C(t)x(t) + g(t),\\ u(t) &= B(t)^T\lambda(t),\end{aligned} \tag{2.23}$$

which can be obtained by the successive solution of the initial value problems

$$\begin{aligned}\dot{P}(t) &= I + C(t)^TC(t) - P(t)A(t) - A(t)^TP(t) - P(t)B(t)B(t)^TP(t), \quad P(T) = 0,\\ \dot{v}(t) &= C(t)^Tg(t) - P(t)f(t) - A(t)^Tv(t) - P(t)B(t)B(t)^Tv(t), \quad v(T) = 0,\\ \dot{x}(t) &= A(t)x(t) + B(t)B(t)^T(P(t)x(t) + v(t)) + f(t), \quad x(t_0) = 0,\\ \lambda(t) &= P(t)x(t) + v(t),\\ y(t) &= C(t)x(t) + g(t),\\ u(t) &= B(t)^T\lambda(t).\end{aligned} \tag{2.24}$$

Proof. See [28]. □

The unique solvability of the problem (2.22) obviously implies that the minimization problem (2.15) has the unique solution $\tilde{x} = (x_1, x_2, x_3) = (x, y, u)$. This enables us to define an operator $\tilde{D}^+$ that maps the inhomogeneity $\tilde{f}$ onto this solution:

$$\tilde{D}^+ : \tilde{\mathbb{Y}} \to \tilde{\mathbb{X}}, \qquad \tilde{D}^+ : \tilde{f} \mapsto \tilde{x}, \tag{2.25}$$

see (2.10)–(2.11) for the definition of $\tilde{\mathbb{X}}$ and $\tilde{\mathbb{Y}}$. It is easy to see that this operator is linear because the Riccati differential equation in (2.24) does not depend on the inhomogeneities. The image of $\tilde{D}^+$ lies in $\tilde{\mathbb{X}}$ because the state variable x in (2.22) and therefore the differential component x_1 of $\tilde{x}$ is continuously differentiable.

Finally, it can be shown that the operator $\tilde{D}^+$, together with the differential-algebraic operator $\tilde{D}$, satisfies the four Moore-Penrose axioms and hence $\tilde{D}^+$ is the Moore-Penrose pseudoinverse of $\tilde{D}$.

Theorem 18 *The operator $\tilde{D}^+$, defined as in (2.25), is the Moore-Penrose pseudoinverse of the operator $\tilde{D}$ defined in (2.12), i.e., the endomorphisms $\tilde{D}\tilde{D}^+$ and $\tilde{D}^+\tilde{D}$ have conjugates such that (2.8) holds for $\tilde{D}$ and $\tilde{D}^+$.*

Proof. See [28]. □

Having shown the existence and uniqueness of the Moore-Penrose pseudoinverse of a differential-algebraic equation in orthogonal standard form, we can now generalize this result to strangeness free linear DAEs. Remember that the differential-algebraic operator D was defined indirectly via the standard form (2.1) by (2.13). Because of

$$\|x\| = \|Q^T x\| = \|\tilde{x}\|, \quad \|Dx - f\| = \|P(P^T \tilde{D} Q^T x - f)\| = \|\tilde{D}\tilde{x} - \tilde{f}\|,$$

the minimization problem (2.14) transforms covariantly with the application of the transformations P and Q. Thus for a general DAE we can first compute the orthogonal standard form along with the operators P and Q and solve the minimization problem (2.15). Having found the Moore-Penrose pseudoinverse $\tilde{D}^+$ of $\tilde{D}$, the Moore-Penrose pseudoinverse D^+ of D is given by

$$D^+ = Q\tilde{D}^+ P.$$

In this way we have found the Moore-Penrose pseudoinverse of a differential-algebraic operator that generalizes the Moore-Penrose pseudoinverses of matrices in a canonical way; it is defined via a similar minimization problem.

In the special case $E \equiv 0$, the problem (2.14) reduces to the algebraic minimization problem

$$\|x\| = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Ax + f\|^2 = \min!$$

and we get $D^+ = -A^+$, where A^+ is defined as in Lemma 15. Moreover, the transformation of a DAE to orthogonal standard form corresponds to the application of the smooth decomposition (2.9) that was necessary to compute the solution of the minimization problem (2.7) in the proof of Lemma 15.

2.3 (1,2,3)-inverses

There are other ways to generalize the theory of Moore-Penrose pseudoinverses to linear DAEs. The computation of the solution $x = D^+ f$ requires the solution of the boundary value problem

(2.23). Furthermore, no arbitrary initial values can be prescribed for the undetermined part of the solution (the component x_3 in the orthogonal normal form).

A possible approach to circumvent this problem is to relax the Moore-Penrose theory and fix a unique solution of (1.1). If the system can be transformed to the orthogonal standard form this can be done very easily by setting $x_3 \equiv 0$ and then solving the problem

$$\begin{bmatrix} \Sigma_E(t) & 0 \\ 0 & 0 \end{bmatrix} \dot{\tilde{x}}(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & \Sigma_A(t) \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad x_1(t_0) = 0, \tag{2.26}$$

which is uniquely solvable. For a general strangeness free linear DAE we can write down this approach in terms of a minimization problem using the matrix function

$$\Pi(t) = E(t)^+ E(t) + F(t)^+ F(t),$$

with

$$F(t) = (I - E(t)E(t)^+)A(t)(I - E(t)^+ E(t)).$$

The function Π has the following properties. For a system in orthogonal standard form we have

$$\tilde{E}(t)^+ = \begin{bmatrix} \Sigma_E(t)^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and thus

$$\begin{aligned} \tilde{F}(t) &= (I - \tilde{E}(t)\tilde{E}(t)^+)\tilde{A}(t)(I - \tilde{E}(t)^+\tilde{E}(t)) \\ &= \begin{bmatrix} 0 & & \\ & I & \\ & & I \end{bmatrix} \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & \Sigma_A(t) & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & & \\ & I & \\ & & I \end{bmatrix} \\ &= \begin{bmatrix} 0 & & \\ & \Sigma_A(t) & \\ & & 0 \end{bmatrix}. \end{aligned}$$

Therefore

$$\tilde{\Pi}(t) = \tilde{E}(t)^+\tilde{E}(t) + \tilde{F}(t)^+\tilde{F}(t) = \begin{bmatrix} I & & \\ & I & \\ & & 0 \end{bmatrix}$$

is an orthogonal projector onto the components x_1 and x_2 of $\tilde{x}$. In addition we get (omitting the argument t):

$$\begin{aligned} \tilde{F} &= (I - \tilde{E}\tilde{E}^+)\tilde{A}(I - \tilde{E}^+\tilde{E}) \\ &= (I - PEQQ^T E^+ P^T)(PAQ - PE\dot{Q})(I - Q^T E^+ P^T PEQ) \\ &= P(I - EE^+)(A - E\dot{Q}Q^T)(I - E^+E)Q \\ &= P(I - EE^+)A(I - E^+E)Q - P(E - EE^+E)\dot{Q}Q^T(I - E^+E)Q \\ &= PFQ. \end{aligned}$$

This shows that F is similarly transformed as E with respect to the transformations P and Q. Hence,

$$\begin{aligned} \tilde{\Pi}(t) &= \tilde{E}(t)^+\tilde{E}(t) + \tilde{F}(t)^+\tilde{F}(t) \\ &= Q(t)^T E(t)^+ P(t)^T P(t)E(t)Q(t) + Q(t)^T F(t)^+ P(t)^T P(t)F(t)Q(t) \\ &= Q(t)^T(E(t)^+E(t) + F(t)^+F(t))Q(t) \\ &= Q(t)^T \Pi Q(t). \end{aligned}$$

It follows that Π is also an orthogonal projector and the minimization problem

$$\frac{1}{2}\|(I-\Pi)x\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Dx-f\|^2 = \min! \tag{2.27}$$

transforms covariantly with the application of P and Q. Thus, the minimization problem (2.27) possesses a unique solution and this solution induces a (linear) operator $D^- : \mathbb{Y} \to \mathbb{X}$. The following theorem shows that D^- is a (1,2,3)-inverse of the differential-algebraic operator D.

Theorem 19 *The operator D^- defined as the solution operator of (2.27) is a (1,2,3)-inverse of D, i.e., the endomorphism DD^- has a conjugate such that (2.8 a,b,c) hold for D and D^-.*

Proof. For the sake of completeness, we provide a detailed proof of this result, which can also be found in [28].

First we show this result for a DAE in orthogonal standard form. The solution $\tilde{x} = \tilde{D}^-\tilde{f}$ of the minimization problem

$$\frac{1}{2}\|(I-\tilde{\Pi})x\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|\tilde{D}\tilde{x}-\tilde{f}\|^2 = \min! \tag{2.28}$$

satisfies (2.26) and we have $x_3 = 0$. If we set $\hat{f} = \tilde{D}\tilde{D}^-\tilde{f}$ and partition $\hat{f} = (\hat{f}_1, \hat{f}_2, \hat{f}_3)$ according to the block structure of $\tilde{f}$, we get

$$\begin{aligned}
\hat{f}_1(t) &= \Sigma_E(t)\dot{x}_1(t) - A_{11}(t)x_1(t) - A_{12}(t)x_2(t) - A_{13}(t) = \tilde{f}_1(t),\\
\hat{f}_2(t) &= -A_{21}(t)x_1(t) - \Sigma_A(t)x_2(t) = \tilde{f}_2(t),\\
\hat{f}_3(t) &= 0,
\end{aligned}$$

after inserting this solution. Thus the operator $\tilde{D}\tilde{D}^- : \mathbb{Y} \to \mathbb{Y}$ can be written as

$$\tilde{D}\tilde{D}^- = \begin{bmatrix} I & & \\ & I & \\ & & 0 \end{bmatrix}.$$

This operator is obviously self-conjugate. Furthermore, we get $\tilde{D}\tilde{D}^-\tilde{D} = \tilde{D}$ because $\tilde{D}$ has a vanishing third component, as well as $\tilde{D}^-\tilde{D}\tilde{D}^- = \tilde{D}^-$ because f_3 does not have any influence on the solution of (2.27).

Since the problem transforms covariantly if we apply the orthogonal transformations P and Q, the operator $D^- = Q\tilde{D}^-P$ maps onto the solution of (2.27). It then satisfies the axioms (2.8) (a)–(c). □

If we again consider the special case of a DAE with $E \equiv 0$, then the solution of (2.27) coincides with the solution of the minimization problem (2.7) for the algebraic system (2.6) and we get $D^- = D^+ = -A^+$. Thus the (1,2,3)-inverse D^- also generalizes the Moore-Penrose pseudoinverse of matrices.

The numerical computation of the generalized solutions that we have defined in this section can be carried out easily if the DAEs are given in orthogonal standard form or if the necessary transformations P and Q (and the derivative $\dot{Q}$) are known. For the computation of the solution $x = D^-f$ a uniquely solvable DAE has to be solved and the computation of $x = D^+f$ requires the solution of a boundary value problem.

But as already stated, the orthogonal standard form and in particular the necessary orthogonal transformations from the right are generally difficult to compute numerically. In the next chapters we will present possibilities to approximate these generalized solutions without any knowledge of these transformations.

Chapter 3

Numerical Determination of Generalized Solutions

In this chapter, we consider the numerical computation of generalized solutions of DAEs. Throughout this chapter, we will assume that the DAE is strangeness free; this choice is motivated by the results of the preceding chapters. Two methods will be presented, both based on time discretizations via BDF-methods [8, 20]. This class of methods has favourable stability properties [20] and has proved its robustness as well as reliability for solving DAEs in software packages such as DASSL [37], ODASSL [11, 12], and GELDA [33]. Moreover, the considerable simplicity of BDF-methods makes them particularly suitable for our purposes.

The two methods presented here allow to approximate numerically the solutions associated with the Moore-Penrose inverse and a (1,2,3)-inverse of differential-algebraic operators without involving any transformation of the solution space.

3.1 BDF-methods and discretization operators

BDF-methods are implicit k-step methods for the numerical solution of ordinary differential equations of the form

$$\dot{x}(t) = f(t, x), \quad t \in [t_0, T], \tag{3.1}$$

with an initial condition

$$x(t_0) = x_0. \tag{3.2}$$

They are defined by setting

$$\sum_{l=0}^{k} \alpha_l x_{i-k+l} = h f(t_i, x_i),$$

where $h = (T - t_0)/N$ with $N \in \mathbb{N}$ is a fixed step size, $t_i = (t_0 + ih)$ are the corresponding grid points in the time interval, and x_l, $l = i - k, \ldots, i - 1$, denote numerical approximations to the solution of (3.1,3.2) at these grid points. The coefficients α_l, $l = 0, \ldots, k$, are defined such that the method has the highest attainable order of convergence. Table 3.1 shows these coefficients for different values of k. BDF-methods are stable for $k \leq 6$ and unstable for $k > 6$ (see, e.g., [20]). They are consistent of order $p = k$ and thus convergent of order $p = k$ for $k \leq 6$.

In Chapter 1, we have seen that the implicit Euler method, i.e., the BDF-method for $k = 1$, can lead to misleading results if applied to general DAEs. However, if we discretize uniquely

α_l	$l=k$	$l=k-1$	$l=k-2$	$l=k-3$	$l=k-4$	$l=k-5$	$l=k-6$
$k=1$	1	-1					
$k=2$	$\frac{3}{2}$	-2	$\frac{1}{2}$				
$k=3$	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$			
$k=4$	$\frac{25}{12}$	-4	3	$-\frac{4}{3}$	$\frac{1}{4}$		
$k=5$	$\frac{137}{60}$	-5	5	$-\frac{10}{3}$	$\frac{5}{4}$	$-\frac{1}{5}$	
$k=6$	$\frac{147}{60}$	-6	$\frac{15}{2}$	$-\frac{20}{3}$	$\frac{15}{4}$	$-\frac{6}{5}$	$\frac{1}{6}$

Table 3.1: Coefficients for BDF-methods.

solvable strangeness free DAEs with BDF-methods, then we get the same convergence result as for ODEs, see Theorem 20 below. This discretization is computed as in the case of ODEs by replacing $\dot{x}$ with the sum $\frac{1}{h}\sum_{l=0}^{k}\alpha_l x_{i-k+l}$, such that, for given initial data $x_{i-k},\dots x_{i-1}$, we have to solve the systems

$$\frac{1}{h}E(t_i)\sum_{l=0}^{k}\alpha_l x_{i-k+l} = A(t_i)x_i + f(t_i) \tag{3.3}$$

successively for x_i, $i=k,\dots,N$.

Theorem 20 ([4]) *Let (1.1) be a strangeness free DAE with E, A and f sufficiently smooth. Let (1.1) together with (1.2) possess the unique solution $x\in C^1([t_0,T],\mathbb{R}^n)$. Furthermore, let $x_0,\dots,x_{k-1}$ be given with*

$$x(t_i)-x_i = O(h^p) \quad \textit{for} \quad h\to 0,$$

$i=0,\dots,k-1$. Define the sequence (x_i) for $i\geq k$ by solving the systems (3.3). Then we get

$$x(T;h)-x(T) = O(h^p) \quad \textit{for} \quad h\to 0.$$

Here, $x(T;h)$ denotes the approximation $x(T;h)=x_N$ to $x(T)$ computed with the step size $h=(T-t_0)/N$.

We now establish an alternative notation for the discretization of a linear DAE with BDF-methods using a fixed step size $h=(T-t_0)/N$. First, we introduce the restriction operator $R_{\mathbb{X}_h}$ as the restriction of a function z to its values at the grid points $t_i=t_0+ih$, $i=k,\dots,N$, in the finite dimensional space $\mathbb{X}_h=\mathbb{R}^{N_k n}$, $N_k=N-k+1$, i.e.,

$$R_{\mathbb{X}_h}z = \begin{bmatrix} z(t_k)\\ \vdots \\ z(t_N)\end{bmatrix}. \tag{3.4}$$

Note that we have used the subscript $\mathbb{X}_h$ to emphasize the facts that $R_{\mathbb{X}_h}$ will be used to discretize the solution space and that the action of the operator depends on the chosen step size (it also depends on the order of the BDF-method and t_0, which, however, will be assumed to be fixed).

We also define a discretization operator $R_{\mathbb{Y}_h}$, which maps a function f into the space $\mathbb{Y}_h = \mathbb{R}^{N_k m}$ and is defined by

$$R_{\mathbb{Y}_h} f = \begin{bmatrix} f(t_k) - \frac{1}{h}\sum_{l=0}^{k-1} \alpha_l E(t_k) x_l \\ \vdots \\ f(t_{2k-1}) - \frac{1}{h}\alpha_0 E(t_{2k-1}) x_{k-1} \\ f(t_{2k}) \\ \vdots \\ f(t_N) \end{bmatrix}. \tag{3.5}$$

Note that the operator $R_{\mathbb{Y}_h}$, in contrast to $R_{\mathbb{X}_h}$, includes additional information about the initial values $x_0, \ldots, x_{k-1}$. Apart from the step size h, we assume all other variables on which the definition of $R_{\mathbb{Y}_h}$ depends to be constant.

Setting $g_h := R_{\mathbb{Y}_h} f$, we can merge the systems (3.3) into one big linear system

$$D_h x_h = g_h, \tag{3.6}$$

where $x_h \in \mathbb{X}_h$ is given by

$$x_h = \begin{bmatrix} x_k \\ \vdots \\ x_N \end{bmatrix},$$

and the linear operator $D_h : \mathbb{X}_h \to \mathbb{Y}_h$ is defined as follows. Each of its blocks $[D_h]_{ij}$, $i, j = k, \ldots, N$, is given by

$$[D_h]_{ij} = \begin{cases} \frac{1}{h}\alpha_{k+j-i} E(t_i) & \text{for } i-k \le j < i, \\ \frac{1}{h}\alpha_k E(t_i) - A(t_i) & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

If the corresponding initial value problem (1.1) with (1.2) possesses a unique solution $x \in \mathbb{X}$, then the matrix D_h is nonsingular for sufficiently small step sizes h and the solution $x_h = D_h^{-1} g_h$ of (3.6) contains the approximations x_i, $i = k, \ldots, N$, to x at the grid points t_i.

With this notation we can now reformulate Theorem 20.

Remark 21 *Let the assumptions of Theorem 20 be satisfied. Then for the solution $x_h = D_h^{-1} g_h$ of (3.6) we get*

$$\|x_h - R_{\mathbb{X}_h} x\|_\infty \le C h^p \quad \textit{for} \quad h \to 0,$$

where the constant $C > 0$ does not *depend on h.*

3.2 Local Minimization

In Section 2.3, we have defined a (1,2,3)-inverse D^- of a differential-algebraic operator D. The corresponding generalized solution $x = D^- f$ of the strangeness free DAE (1.1) is uniquely defined by the minimization problem (2.27).

This solution can be easily computed analytically and numerically if the DAE is given in orthogonal standard form or if the necessary orthogonal transformations P and Q as defined in Theorem 11 are known. In this case, the system can be assumed to be in orthogonal standard

form and the undetermined component x_3 of $\tilde{x} = Q^T x$ can be set to zero. The solvability condition $f_3 = 0$ does not have any influence on the generalized solution and the remaining DAE

$$\begin{aligned} \Sigma_E(t)\dot{x}_1(t) &= A_{11}(t)x_1(t) + A_{12}(t)x_2(t) + f_1(t), \\ 0 &= A_{21}(t)x_1(t) + \Sigma_A(t)x_2(t) + f_2(t) \end{aligned} \tag{3.7}$$

possesses a unique solution. This system can be discretized directly with a BDF-method as shown in the previous section.

One can also solve (3.7) by first solving the ordinary differential equation

$$\begin{aligned} \dot{x}_1(t) = {} & \Sigma_E(t)^{-1}\left(A_{11}(t) - A_{12}(t)\Sigma_A(t)^{-1}A_{21}(t)\right)x_1(t) \\ & + \Sigma_E(t)^{-1}\left(f_1(t) - \Sigma_A(t)^{-1}f_2(t)\right) \end{aligned} \tag{3.8}$$

by an appropriate method and then compute the algebraic component of the solution by

$$x_2 = -\Sigma_A(t)^{-1}\left(A_{21}(t)x_1(t) + f_2(t)\right).$$

The solution of (3.7) is then given by

$$x = Q\tilde{x} = Q \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}.$$

Let us emphasize again that both approaches require the undetermined components of the solution and consequently the transformations P and Q to be known. However, since these quantities are usually not known in advance in realistic applications, we have to find other ways to approximate the generalized solution.

Discretizing the strangeness free DAE (1.1) directly with a k-step BDF method using a fixed step size $h = (T - t_0)/N$ leads to $N - k + 1$ systems of the form

$$\frac{1}{h}E_i \sum_{l=0}^{k} \alpha_l x_{i-k+l} = A_i x_i + f_i \tag{3.9}$$

for $i = k, \ldots, N$. Here E_i, A_i and f_i denote the values of E, A and f at the grid points $t_i = t_0 + ih$. It must be assumed that sufficiently good initial approximations $x_0, \ldots, x_{k-1}$ to the generalized solution $x = D^- f$ at the grid points $t_0, \ldots, t_{k-1}$ are provided. The systems in (3.9) can be written as

$$\left(\frac{\alpha_k}{h}E_i - A_i\right)x_i = -\frac{1}{h}\sum_{l=0}^{k-1} \alpha_l E_i x_{i-k+l} + f_i \tag{3.10}$$

and have to be solved with respect to x_i. Since our focus is on over- or underdetermined DAEs, the matrices $\left(\frac{\alpha_k}{h}E_i - A_i\right)$ are likely to be rank deficient. As in (3.6), the discretization (3.10) can be written in terms of the large linear system

$$D_h x_h = g_h \tag{3.11}$$

where the blocks $[D_h]_{ij} \in \mathbb{R}^{m,n}$, $i, j = k, \ldots, N$ of the matrix D_h are defined as

$$[D_h]_{ij} = \begin{cases} \dfrac{1}{h}\alpha_{k+j-i}E_i & \text{for } i - k \le j < i, \\ \dfrac{1}{h}\alpha_k E_i - A_i & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

and the right-hand side $g_h = R_{\mathbb{Y}_h} f$ with $R_{\mathbb{Y}_h}$ as in (3.5) contains values of the inhomogeneity f at the grid points as well as the initial values $x_0, \ldots, x_{k-1}$. The vector

$$x_h = \begin{bmatrix} x_k \\ \vdots \\ x_N \end{bmatrix}$$

contains the solutions x_i of (3.9), provided that these systems are solvable.

If the systems (3.9) are not solvable, then we can compute solutions in a least squares sense, i.e., we solve the $N - k + 1$ minimization problems

$$\frac{1}{2}\|x_i\|_2^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\left\| \frac{1}{h} E_i \sum_{l=0}^{k} \alpha_l x_{i-k+l} - A_i - f_i \right\|_2^2 = \min! \tag{3.13}$$

successively for $i = k, \ldots, N$. The corresponding solutions x_i can then be written in the form

$$x_i = \left(\frac{\alpha_k}{h} E_i - A_i \right)^+ \left(-\frac{1}{h} \sum_{l=0}^{k-1} \alpha_l E_i x_{i-k+l} + f_i \right), \tag{3.14}$$

where $\left(\frac{\alpha_k}{h} E_i - A_i\right)^+$ denotes the Moore-Penrose pseudoinverse of the matrix $\frac{\alpha_k}{h} E_i - A_i$.

The rest of this section is concerned with showing that the solutions x_i in (3.14) are approximations to the generalized solution $x = D^- f$ at the grid points t_i.

Synopsis of proof Before we go into the technical details to prove the main result, let us provide a brief synopsis of the proof.

Step 1: Starting from (3.14), we use the transformation matrices of the DAE to orthogonal standard form to get the reformulated equations (3.16). Multiplying by an appropriate regular factor from the left gives (3.23). Writing these reformulations in terms of the enlarged system as in (3.6) yields the system $\check{D}_h x_h = \check{g}_h$ in (3.32), which has the same solution as (3.11).

Step 2: On the other side, we consider the orthogonal standard form of the DAE and derive a slightly modified DAE, which has as its unique solution the $(1, 2, 3)$-solution $\tilde{D}^{-1}\tilde{f}$ as defined in (2.28). Then we map the transformed solution space back to the original solution space and obtain (3.27). This equation is discretized, see (3.28), and written as the enlarged system $\hat{D}_h \hat{x}_h = \hat{g}_h$ in (3.35).

Step 3: Lemma 25 shows that the coefficient matrices $\check{D}_h$ and $\hat{D}_h$ as well as the right-hand sides $\check{g}_h$ and $\hat{g}_h$ differ at most by $O(h)$. Combined with the boundedness of $\hat{D}_h^{-1}$, this fact is used in Theorem 26 to conclude the proof.

Remark: Let us emphasize that all the transformations are performed solely for theoretical purposes; our numerical method will directly employ the unmodified discretized equations (3.14).

$\Diamond$

Step 1

Since the DAE is strangeness free, we know that there exist orthogonal transformations P and Q such that $\tilde{E} = PEQ$ and $\tilde{A} = PAQ - PE\dot{Q}$ have the structure given in Theorem 11. Therefore, we have

$$\begin{aligned}\frac{\alpha_k}{h}E_i - A_i &= P_i^T(\frac{\alpha_k}{h}P_iE_iQ_i - P_iA_iQ_i)Q_i^T \\ &= P_i^T(\frac{\alpha_k}{h}\tilde{E}_i - (\tilde{A}_i + P_iE_i\dot{Q}_i))Q_i^T \\ &= P_i^T(\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i - \tilde{E}_iQ_i^T\dot{Q}_i)Q_i^T \\ &= P_i^T(\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i)Q_i^T.\end{aligned} \tag{3.15}$$

Again, the index i denotes the evaluation of the corresponding matrix function at $t = t_i$. The last identity follows from

$$0 = \dot{I} = \frac{d}{dt}(Q^TQ) = \dot{Q}^TQ + Q^T\dot{Q}.$$

Because of the term $\tilde{E}_i\dot{Q}_i^TQ_i$, Equation (3.15) shows that the discretization used here does not transform covariantly with the application of the transformation Q.

Using (3.15), we can rewrite (3.14) as

$$\begin{aligned}x_i &= \left(\frac{\alpha_k}{h}E_i - A_i\right)^+\left(-\frac{1}{h}\sum_{l=0}^{k-1}\alpha_lE_ix_{i-k+l} + f_i\right) \\ &= \left(P_i^T(\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i)Q_i^T\right)^+\left(-\frac{1}{h}\sum_{l=0}^{k-1}\alpha_lE_ix_{i-k+l} + f_i\right) \\ &= Q_i\left(\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i\right)^+P_i\left(-\frac{1}{h}\sum_{l=0}^{k-1}\alpha_lE_ix_{i-k+l} + f_i\right) \\ &= Q_i\left(\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i\right)^+\left(-\frac{1}{h}\sum_{l=0}^{k-1}\alpha_l\tilde{E}_iQ_i^Tx_{i-k+l} + \tilde{f}_i\right).\end{aligned} \tag{3.16}$$

The matrix $\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i$ has the structure

$$\frac{\alpha_k}{h}\tilde{E}_i - \tilde{A}_i + \tilde{E}_i\dot{Q}_i^TQ_i = \begin{bmatrix} R_i & S_i \\ 0 & 0\end{bmatrix}, \tag{3.17}$$

where

$$R_i = \begin{bmatrix} \frac{\alpha_k}{h}\Sigma_{E_i} - A_{11_i} + \Sigma_{E_i}\hat{Q}_{1_i} & -A_{12_i} + \Sigma_{E_i}\hat{Q}_{2_i} \\ -A_{21_i} & -\Sigma_{A_i}\end{bmatrix} \in \mathbb{R}^{\hat{d}+\hat{a},\hat{d}+\hat{a}} \tag{3.18}$$

is nonsingular for sufficiently small h and

$$S_i = \begin{bmatrix} -A_{13_i} + \Sigma_{E_i}\hat{Q}_{3_i} \\ 0\end{bmatrix} \in \mathbb{R}^{\hat{d}+\hat{a},\hat{u}}. \tag{3.19}$$

Here we use the notation

$$\hat{Q}_j := \dot{Q}_{11}^TQ_{1j} + \dot{Q}_{21}^TQ_{2j} + \dot{Q}_{31}^TQ_{3j}, \tag{3.20}$$

where $\dot{Q}_{ij}$ and Q_{ij} denote the block (i, j) of $\dot{Q}$ and Q, respectively, in accordance with the block structure of the orthogonal standard form. Hence, $\hat{Q}_j$ is the j-th block of $\dot{Q}^T Q$ in the first block row. Note that the perturbation induced by the derivative of Q only appears in the first block rows of R_i and S_i.

We will now use the following formula to compute the Moore-Penrose pseudoinverse.

Lemma 22 *For a matrix*

$$A = \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \tag{3.21}$$

with a nonsingular block R, we have

$$A^+ = \begin{bmatrix} \left(I - VW^{-1}V^T\right) R^{-1} & 0 \\ W^{-1}V^T R^{-1} & 0 \end{bmatrix}, \tag{3.22}$$

with

$$V := R^{-1}S, \quad W := I + V^T V.$$

Proof. For A and A^+ as defined in (3.21) and (3.22) it follows that

$$\begin{aligned} AA^+ &= \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \left(I - VW^{-1}V^T\right) R^{-1} & 0 \\ W^{-1}V^T R^{-1} & 0 \end{bmatrix} \\ &= \begin{bmatrix} R\left(I - VW^{-1}V^T\right) R^{-1} + SW^{-1}V^T R^{-1} & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} I - SW^{-1}V^T R^{-1} + SW^{-1}V^T R^{-1} & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

and this immediately shows that the first three Moore-Penrose axioms (2.5) (1)–(3) are satisfied. For the matrix A^+A we get

$$\begin{aligned} A^+A &= \begin{bmatrix} \left(I - VW^{-1}V^T\right) R^{-1} & 0 \\ W^{-1}V^T R^{-1} & 0 \end{bmatrix} \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} I - VW^{-1}V^T & \left(I - VW^{-1}V^T\right) V \\ W^{-1}V^T & W^{-1}V^T V \end{bmatrix}. \end{aligned}$$

The block $I - VW^{-1}V^T$ is symmetric because W and thus W^{-1} are symmetric. It follows that

$$\begin{aligned} \left(W^{-1}V^T V\right)^T &= V^T V W^{-T} = V^T V W^{-1} = \left(I + V^T V\right) W^{-1} - W^{-1} \\ &= W^{-1}\left(I + V^T V\right) - W^{-1} = W^{-1}V^T V. \end{aligned}$$

Furthermore,

$$\begin{aligned} \left(W^{-1}V^T\right)^T &= VW^{-1} = V\left(I - I + W^{-1}\right) = V\left(I - W^{-1}\left(I + V^T V\right) + W^{-1}\right) \\ &= V\left(I - W^{-1}V^T V\right) = \left(I - VW^{-1}V^T\right) V. \end{aligned}$$

This shows that A^+A is symmetric and the fourth Moore-Penrose axiom (2.5) (4) is satisfied. ∎

From Lemma 22, it follows that

$$\begin{bmatrix} R & S \\ 0 & I \end{bmatrix} \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix}^+ = \begin{bmatrix} I & 0 \\ W^{-1}V^T R^{-1} & 0 \end{bmatrix}.$$

Next, we turn Equation (3.16) for $i = k, \ldots, N$ into a linear system by multiplication from the left with the regular matrix

$$\begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} Q_i^T,$$

in order to obtain the system

$$\begin{aligned} \begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} Q_i^T x_i &= \begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix}^+ \left(-\frac{1}{h} \sum_{l=0}^{k-1} \alpha_l \tilde{E}_i Q_i^T x_{i-k+l} + \tilde{f}_i \right) \\ &= \begin{bmatrix} I_{\hat{d}} & 0 & 0 \\ 0 & I_{\hat{a}} & 0 \\ \mathcal{R}_{1_i} & \mathcal{R}_{2_i} & 0 \end{bmatrix} \left(-\frac{1}{h} \sum_{l=0}^{k-1} \alpha_l \begin{bmatrix} \Sigma_{E_i} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q_i^T x_{i-k+l} + \begin{bmatrix} f_{1_i} \\ f_{2_i} \\ f_{3_i} \end{bmatrix} \right) \\ &= -\frac{1}{h} \sum_{l=0}^{k-1} \alpha_l \begin{bmatrix} \Sigma_{E_i} & 0 & 0 \\ 0 & 0 & 0 \\ \mathcal{R}_{1_i}\Sigma_{E_i} & 0 & 0 \end{bmatrix} Q_i^T x_{i-k+l} + \begin{bmatrix} f_{1_i} \\ f_{2_i} \\ \mathcal{R}_{1_i} f_{1_i} + \mathcal{R}_{2_i} f_{2_i} \end{bmatrix}, \end{aligned} \tag{3.23}$$

where for

$$V_i = R_i^{-1} S_i, \quad W_i = I + V_i^T V_i, \tag{3.24}$$

the matrix

$$W_i^{-1} V_i^T R_i^{-1} =: [\mathcal{R}_{1_i} \quad \mathcal{R}_{2_i}] \tag{3.25}$$

is partitioned according to the block structure in the orthogonal standard form.

Step 2

Let us consider the system (3.7) together with the condition $x_3 \equiv 0$ and write this uniquely solvable DAE as

$$\bar{E}(t)\dot{\bar{x}}(t) = \bar{A}(t)\bar{x}(t) + \bar{f}(t), \tag{3.26}$$

with

$$\begin{aligned} \bar{E}(t) &= \begin{bmatrix} \Sigma_E(t) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in C(\mathbb{I}, \mathbb{R}^{n,n}), \\ \bar{A}(t) &= \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & \Sigma_A(t) & 0 \\ 0 & 0 & -I_{\hat{u}} \end{bmatrix} \in C(\mathbb{I}, \mathbb{R}^{n,n}), \\ \bar{f}(t) &= \begin{bmatrix} f_1(t) \\ f_2(t) \\ 0 \end{bmatrix} \in C(\mathbb{I}, \mathbb{R}^n). \end{aligned}$$

We then apply a change of basis with the matrix function Q to this system and obtain the new DAE

$$\hat{E}(t)\dot{\hat{x}}(t) = \hat{A}(t)\hat{x}(t) + \hat{f}(t), \tag{3.27}$$

where $\hat{E} = \bar{E}Q^T$, $\hat{A} = \bar{A}Q^T - \bar{E}\dot{Q}^T$ and $\hat{f} = \bar{f}$. Since the systems (3.26) and (3.27) are equivalent and uniquely solvable, we get the following equality:

$$\hat{x} = Q\bar{x} = Q\tilde{x} = x,$$

where $\tilde{x}$ is the solution of the DAE in orthogonal standard form with $x_3 \equiv 0$ and x is the (1,2,3)-solution of the original DAE (1.1), i.e., $x = D^- f$. Furthermore, a discretization of (3.27) with a k-step BDF-method is convergent of order $p = k$, see Theorem 20. Performing this discretization will give a reference discretization that can be compared with the systems (3.23) in order to show that the solutions x_i of (3.23) actually approximate the solution $\hat{x}$ of (3.27).

BDF-methods applied to (3.27) lead to systems of the form

$$\left(\frac{\alpha_k}{h}\hat{E}_i - \hat{A}_i\right)\hat{x}_i = -\sum_{l=0}^{k-1} \alpha_l \hat{E}_i \hat{x}_{i-k+l} + \hat{f}_i \tag{3.28}$$

for $i = k, \ldots, N$, which are uniquely solvable for sufficiently small step sizes h. Analogously to (3.15) we get

$$\begin{aligned} \frac{\alpha_k}{h}\hat{E}_i - \hat{A}_i &= \left(\frac{\alpha_k}{h}\bar{E}_i - \bar{A}_i + \bar{E}_i\dot{Q}_i^T Q_i\right) Q_i^T \\ &= \begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} Q_i^T, \end{aligned}$$

where R_i and S_i are defined as in (3.18) and (3.19) and thus we can rewrite the systems (3.28) in the form

$$\begin{aligned} \begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} Q_i^T \hat{x}_i &= -\frac{1}{h}\sum_{l=0}^{k-1} \alpha_l \bar{E}_i Q_i^T \hat{x}_{i-k+l} + \bar{f}_i \\ &= -\frac{1}{h}\sum_{l=0}^{k-1} \alpha_l \begin{bmatrix} \Sigma_{E_i} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q_i^T \hat{x}_{i-k+l} + \begin{bmatrix} f_{1_i} \\ f_{2_i} \\ 0 \end{bmatrix}. \end{aligned} \tag{3.29}$$

This system is the same as the final system (3.23) obtained in Step 1, apart from the third block row, which does not vanish in (3.23).

Step 3

In the following we show that the systems (3.23) approximate the systems (3.29) by investigating the third block row of (3.23) in more detail. For this purpose, we prove the following lemma, which shows that the blocks $\mathcal{R}_{1_i}$ and $\mathcal{R}_{2_i}$ in (3.23) are sufficiently small.

Before, let us introduce the following definition for notational convenience.

Definition 23 *Let $A \in C([t_0, T], \mathbb{R}^{m,n})$ and $2 \leq p \leq \infty$. Then the constant $C_{\|A\|_p}$ is defined as*

$$C_{\|A\|_p} = \max_{t\in[t_0,T]} \{\|A(t)\|_p, \|A(t)^T\|_p\}.$$

Lemma 24 *Let $\mathcal{R}_{1_i}$ and $\mathcal{R}_{2_i}$ be defined as in (3.24) and (3.25). Then there exist constants C_{01}, $C_{02} \in \mathbb{R}^+$ such that*

$$\|\mathcal{R}_{1_i}\|_\infty \leq C_{01} h^2,$$

and

$$\|\mathcal{R}_{2_i}\|_\infty \leq C_{02} h,$$

for sufficiently small h.

Proof. We introduce the constants

$$\begin{aligned} C_{11} &= C_{\|-A_{11}+\Sigma_E \hat{Q}_1\|_\infty}, \\ C_{12} &= C_{\|-A_{12}+\Sigma_E \hat{Q}_2\|_\infty}, \\ C_{13} &= C_{\|-A_{13}+\Sigma_E \hat{Q}_3\|_\infty}, \end{aligned}$$

with $\hat{Q}_j$, $j = 1, 2, 3$, as defined in (3.20). We will first estimate the block columns of the matrix R_i^{-1} according to the given block structure. If we denote the blocks in the upper block row of the matrix R_i (omitting the index i) as

$$\begin{aligned} R_{11} &= \frac{\alpha_k}{h}\Sigma_E - A_{11} + \Sigma_E \hat{Q}_1, \\ R_{12} &= -A_{12} + \Sigma_E \hat{Q}_2, \end{aligned}$$

we see that the matrix R_{11} is nonsingular for sufficiently small h and we can compute the inverse of R using the block LU decomposition

$$R = \begin{bmatrix} R_{11} & R_{12} \\ -A_{21} & -\Sigma_A \end{bmatrix} = \begin{bmatrix} I & 0 \\ -A_{21}R_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} R_{11} & 0 \\ 0 & A_{21}R_{11}^{-1}R_{12} - \Sigma_A \end{bmatrix} \begin{bmatrix} I & R_{11}^{-1}R_{12} \\ 0 & I \end{bmatrix}.$$

Then, if $T = A_{21}R_{11}^{-1}R_{12} - \Sigma_A$ is nonsingular, the inverse of R is given by

$$\begin{aligned} R^{-1} &= \begin{bmatrix} I & -R_{11}^{-1}R_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ A_{21}R_{11}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} R_{11}^{-1}\left(I + R_{12}T^{-1}A_{21}R_{11}^{-1}\right) & -R_{11}^{-1}R_{12}T^{-1} \\ -T^{-1}A_{21}R_{11}^{-1} & T^{-1} \end{bmatrix}. \end{aligned}$$

The individual block columns of R^{-1} will be denoted by

$$[R^{-1}]_1 = \begin{bmatrix} R_{11}^{-1}\left(I + R_{12}T^{-1}R_{21}R_{11}^{-1}\right) \\ -T^{-1}A_{21}R_{11}^{-1} \end{bmatrix}$$

and

$$[R^{-1}]_2 = \begin{bmatrix} -R_{11}^{-1}R_{12}T^{-1} \\ T^{-1} \end{bmatrix}.$$

Because of

$$\begin{aligned} R_{11}^{-1} &= \left(\frac{\alpha_k}{h}\Sigma_E - A_{11} + \Sigma_E \hat{Q}_1\right)^{-1} \\ &= \left(\frac{\alpha_k}{h}\Sigma_E \left(I - \frac{h}{\alpha_k}\left(\Sigma_E^{-1}A_{11} - \hat{Q}_1\right)\right)\right)^{-1} \\ &= \frac{h}{\alpha_k}\left(I - \frac{h}{\alpha_k}\left(\Sigma_E^{-1}A_{11} - \hat{Q}_1\right)\right)^{-1}\Sigma_E^{-1}, \end{aligned}$$

we can compute R_{11}^{-1} using the Neumann series if

$$\frac{h}{\alpha_k}\left\|\Sigma_E^{-1}A_{11}-\hat{Q}_1\right\|_\infty < 1.$$

This inequality is fulfilled if

$$h < \frac{\alpha_k}{C_{\|\Sigma_E^{-1}\|_\infty}C_{11}} \leq \frac{\alpha_k}{\left\|\Sigma_E^{-1}\right\|_\infty\left\|A_{11}-\Sigma_E\hat{Q}_1\right\|_\infty} \leq \frac{\alpha_k}{\left\|\Sigma_E^{-1}A_{11}-\hat{Q}_1\right\|_\infty},$$

with $C_{\|\Sigma_E^{-1}\|_\infty}$ defined as in Definition 23. We then get

$$\left(I-\frac{h}{\alpha_k}\left(\Sigma_E^{-1}A_{11}-\hat{Q}_1\right)\right)^{-1} = \sum_{l=0}^{\infty}\left(\frac{h}{\alpha_k}\left(\Sigma_E^{-1}A_{11}-\hat{Q}_1\right)\right)^l$$

and therefore

$$\begin{aligned}
\left\|R_{11}^{-1}\right\|_\infty &= \left\|\frac{h}{\alpha_k}\sum_{l=0}^{\infty}\left(\frac{h}{\alpha_k}\left(\Sigma_E^{-1}A_{11}-\hat{Q}_1\right)\right)^l\Sigma_E^{-1}\right\|_\infty \\
&\leq \frac{h}{\alpha_k}\sum_{l=0}^{\infty}\left\|\frac{h}{\alpha_k}(\Sigma_E^{-1}A_{11}-\hat{Q}_1)\right\|_\infty^l\left\|\Sigma_E^{-1}\right\|_\infty \\
&= \frac{h}{\alpha_k}\frac{1}{1-\frac{h}{\alpha_k}\left\|\Sigma_E^{-1}A_{11}-\hat{Q}_1\right\|_\infty}\left\|\Sigma_E^{-1}\right\|_\infty \\
&\leq \frac{h}{\alpha_k}\frac{1}{1-\frac{h}{\alpha_k}\left\|\Sigma_E^{-1}\right\|_\infty\left\|A_{11}-\Sigma_E\hat{Q}_1\right\|_\infty}\left\|\Sigma_E^{-1}\right\|_\infty \\
&\leq \frac{h}{\alpha_k}\frac{1}{1-\frac{h}{\alpha_k}C_{\|\Sigma_E^{-1}\|_\infty}C_{11}}C_{\|\Sigma_E^{-1}\|_\infty} \\
&\leq \frac{2h}{\alpha_k}C_{\|\Sigma_E^{-1}\|_\infty} \\
&= hC_1 \quad \text{for} \quad h \leq \frac{\alpha_k}{2C_{\|\Sigma_E^{-1}\|_\infty}C_{11}}, \quad C_1 = \frac{2}{\alpha_k}C_{\|\Sigma_E^{-1}\|_\infty}.
\end{aligned}$$

We estimate the norm of T^{-1} using

$$\begin{aligned}
\left\|T^{-1}\right\|_\infty &= \left\|(A_{21}R_{11}^{-1}R_{12}-\Sigma_A)^{-1}\right\|_\infty \\
&= \left\|(I-\Sigma_A^{-1}A_{21}R_{11}^{-1}R_{12})^{-1}\Sigma_A^{-1}\right\|_\infty \\
&\leq \sum_{l=0}^{\infty}\left\|\Sigma_A^{-1}A_{21}R_{11}^{-1}R_{12}\right\|_\infty^l\left\|\Sigma_A^{-1}\right\|_\infty \\
&\leq \sum_{l=0}^{\infty}\left(hC_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1C_{12}\right)^l C_{\|\Sigma_A^{-1}\|_\infty} \\
&= \frac{C_{\|\Sigma_A^{-1}\|_\infty}}{1-hC_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1C_{12}} \\
&\leq 2C_{\|\Sigma_A^{-1}\|_\infty} \quad \text{for} \quad h \leq \frac{1}{2C_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1C_{12}}.
\end{aligned}$$

Again, the constants $C_{\|\Sigma_A^{-1}\|_\infty}$ and $C_{\|A_{21}\|_\infty}$ are defined according to Definition 23.

The remaining estimates are straightforward. For the block entries of $[R^{-1}]_1$ and $[R^{-1}]_2$ we get

$$\begin{aligned}
\left\|R_{11}^{-1}\left(I + R_{12}T^{-1}A_{21}R_{11}^{-1}\right)\right\|_\infty &\leq hC_1\left(1 + 2hC_{12}C_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1\right)\\
&\leq 2hC_1 \quad \text{for} \quad h \leq \frac{\alpha_k}{4C_{12}C_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1},\\
\| - T^{-1}A_{21}R_{11}^{-1}\|_\infty &\leq 2hC_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1\\
&= hC_2, \quad C_2 = 2C_{\|\Sigma_A^{-1}\|_\infty}C_{\|A_{21}\|_\infty}C_1,
\end{aligned}$$

and

$$\begin{aligned}
\left\|-R_{11}^{-1}R_{12}T^{-1}\right\|_\infty &\leq 2hC_1C_{12}C_{\|\Sigma_A^{-1}\|_\infty}\\
&= hC_3, \quad C_3 = 2C_1C_{12}C_{\|\Sigma_A^{-1}\|_\infty}.
\end{aligned}$$

Analogously it can be proved for the transposed block entries of $[R^{-1}]_1$ that

$$\left\|\left(R_{11}^{-1}\left(I + R_{12}T^{-1}A_{21}R_{11}^{-1}\right)\right)^T\right\|_\infty \leq 2hC_1$$

and

$$\left\|\left(-T^{-1}A_{21}R_{11}^{-1}\right)^T\right\|_\infty \leq hC_2.$$

Thus, we obtain for the two block columns of R^{-1} the estimates

$$\begin{aligned}
\left\|[R^{-1}]_1\right\|_\infty &\leq \max\{hC_1, hC_2\}\\
&= h\max\{C_1, C_2\}\\
&= hC_4, \quad C_4 = \max\left(C_1, C_2\right),\\
\left\|[R^{-1}]_1^T\right\|_\infty &\leq h\left(C_1 + C_2\right)\\
&= hC_5, \quad C_5 = C_1 + C_2,
\end{aligned}$$

and

$$\begin{aligned}
\|[R^{-1}]_2\|_\infty &\leq \max\{hC_3, 2C_{\|\Sigma_A^{-1}\|_\infty}\}\\
&\leq 2C_{\|\Sigma_A^{-1}\|_\infty} \quad \text{for} \quad h \leq \frac{2C_{\|\Sigma_A^{-1}\|_\infty}}{C_3} = \frac{1}{C_1C_{12}}.
\end{aligned}$$

For $V = R^{-1}S$ we get, owing to the block structure of S,

$$\begin{aligned}
\|V\|_\infty &= \left\|R^{-1}S\right\|_\infty\\
&= \left\|[R^{-1}]_1\left(-A_{13} + \Sigma_E\hat{Q}_3\right)\right\|_\infty\\
&\leq \left\|[R^{-1}]_1\right\|_\infty\left\|-A_{13} + \Sigma_E\hat{Q}_3\right\|_\infty\\
&\leq hC_4C_{13},
\end{aligned}$$

and

$$\begin{aligned}
\|V^T\|_\infty &= \left\|\left(-A_{13} + \Sigma_E\hat{Q}_3\right)^T[R^{-1}]_1^T\right\|_\infty\\
&\leq hC_5C_{13}.
\end{aligned}$$

Hence, $\|V^TV\|_\infty \le \|V^T\|_\infty \|V\|_\infty \le h^2 C_4 C_5 C_{13}^2$. In particular, we have

$$\left\|V^TV\right\|_\infty \le \frac{1}{2} \quad \text{for} \quad h \le \frac{1}{\sqrt{2C_4C_5}C_{13}}.$$

Thus for sufficiently small h we can again employ the Neumann series to estimate the norm of the inverse of $W = I + V^TV$, which gives

$$\begin{aligned}\left\|W^{-1}\right\|_\infty &= \left\|\left(I + V^TV\right)^{-1}\right\|_\infty = \left\|\sum_{l=0}^{\infty}\left(-V^TV\right)^l\right\|_\infty \\ &\le \sum_{l=0}^{\infty}\left\|\left(V^TV\right)^l\right\|_\infty \le \sum_{l=0}^{\infty}\frac{1}{2^l} = 2.\end{aligned}$$

Finally, we can combine all these results and obtain the assertion, because

$$\begin{aligned}\|\mathcal{R}_{1_i}\|_\infty &= \left\|W_i^{-1}V_i^T[R_i^{-1}]_1\right\|_\infty \\ &\le 2h^2C_4C_5C_{13} \\ &= h^2C_{01}, \quad C_{01} = 2C_4C_5C_{13},\end{aligned}$$

and

$$\begin{aligned}\|\mathcal{R}_{2_i}\|_\infty &= \left\|W_i^{-1}V_i^T[R_i^{-1}]_2\right\|_\infty \\ &\le 4hC_5C_{13}C_{\|\Sigma_A^{-1}\|_\infty} \\ &= hC_{02}, \quad C_{02} = 4C_5C_{13}C_{\|\Sigma_A^{-1}\|_\infty}.\end{aligned}$$

Note that none of the occurring constants depends on h. □

We now consider the systems (3.23) as well as (3.29), and represent both discretizations in terms of two large linear systems. Let us define the matrix $\check{D}_h = [\check{D}_h]_{i,j=k,\dots,N} \in \mathbb{R}^{N_k n, N_k n}$, $N_k = N - k + 1$, blockwise by

$$[\check{D}_h]_{ij} = \begin{cases} \dfrac{1}{h}\alpha_{k+j-i}\begin{bmatrix} \Sigma_{E_i} & 0\ 0 \\ 0 & 0\ 0 \\ \mathcal{R}_{1_i}\Sigma_{E_i} & 0\ 0 \end{bmatrix} Q_i^T & \text{for } i-k \le j < i, \\ \begin{bmatrix} R_i & S_i \\ 0 & I_{\hat{u}} \end{bmatrix} Q_i^T & \text{for } j = i, \\ 0 & \text{otherwise,} \end{cases} \tag{3.30}$$

and the right-hand side

$$[\check{g}_h]_i = [\check{f}_h]_i - \frac{1}{h}\sum_{l=0}^{2k-i-1}\alpha_l \begin{bmatrix} \Sigma_{E_i} & 0\ 0 \\ 0 & 0\ 0 \\ \mathcal{R}_{1_i}\Sigma_{E_i} & 0\ 0 \end{bmatrix} Q_i^T x_{i-k+l} \tag{3.31}$$

with

$$[\check{f}_h]_i = \begin{bmatrix} f_{1_i} \\ f_{2_i} \\ \mathcal{R}_{1_i}f_{1_i} + \mathcal{R}_{2_i}f_{2_i} \end{bmatrix}.$$

Then we can write the $N-k+1$ discretization steps (3.23) as

$$\check{D}_h x_h = \check{g}_h. \tag{3.32}$$

Note that the (unique) solution x_h of the system (3.32) is the solution of the system (3.11), where the components x_i of x_h are computed according to (3.16).

Analogously, we can define a matrix $\hat{D}_h$ via

$$[\hat{D}_h]_{ij} = \begin{cases} \dfrac{1}{h}\alpha_{k+j-i}\bar{E}_i Q_i^T & \text{for } i-k \le j < i, \\ \begin{bmatrix} R_i & S_i \\ 0 & I_u \end{bmatrix} Q_i^T & \text{for } j = i, \\ 0 & \text{otherwise,} \end{cases} \tag{3.33}$$

and the corresponding right-hand side by

$$[\hat{g}_h]_i = \bar{f}_i - \frac{1}{h}\sum_{l=0}^{2k-i-1} \alpha_l \bar{E}_i Q_i^T x_{i-k+l}. \tag{3.34}$$

Then we can represent the systems (3.29) by the large system

$$\hat{D}_h \hat{x}_h = \hat{g}_h. \tag{3.35}$$

The following lemma shows that transferring Lemma 24 to the large systems yields $\check{D}_h - \hat{D}_h = O(h)$ and $\check{g}_h - \hat{g}_h = O(h)$.

Lemma 25 *Let $\check{D}_h$, $\hat{D}_h$, $\check{g}_h$ and $\hat{g}_h$ be defined as in (3.30), (3.31), (3.33) and (3.34). Then there exist positive constants C_D and C_g, which do not depend on h, such that*

$$\left\| \check{D}_h - \hat{D}_h \right\|_\infty \le h C_D$$

and

$$\left\| \check{g}_h - \hat{g}_h \right\|_\infty \le h C_g$$

for sufficiently small h.

Proof. From Lemma 24 it follows that

$$\begin{aligned} \left\| \check{D}_h - \hat{D}_h \right\|_\infty &\le \max_i \sum_{j=k}^{N} \left\| [\check{D}_h]_{ij} - [\hat{D}_h]_{ij} \right\|_\infty \\ &\le \max_i \sum_{j=i-k}^{i-1} \left\| \frac{1}{h}\alpha_{k+j-i} \left(\begin{bmatrix} \Sigma_{E_i} & 0\ 0 \\ 0 & 0\ 0 \\ \mathcal{R}_{1_i}\Sigma_{E_i} & 0\ 0 \end{bmatrix} - \bar{E}_i \right) Q_i^T \right\|_\infty \\ &\le \max_i \frac{1}{h} \sum_{j=i-k}^{i-1} |\alpha_{k+j-i}| \left\| \begin{bmatrix} 0 & 0\ 0 \\ 0 & 0\ 0 \\ \mathcal{R}_{1_i}\Sigma_{E_i} & 0\ 0 \end{bmatrix} Q_i^T \right\|_\infty \\ &\le \frac{1}{h} \sum_{l=0}^{k-1} |\alpha_l| h^2 C_{01} C_{\|\Sigma_E\|_\infty} C_{\|Q\|_\infty} \\ &= h C_D. \end{aligned} \tag{3.36}$$

Similarly, we get for the right-hand sides that

$$\begin{aligned}\left\|[\check{f}_h]_i - \bar{f}_i\right\|_\infty &= \left\|W_i^{-1}V_i^T[R_i^{-1}]_1 f_{1_i} + W_i^{-1}V_i^T[R_i^{-1}]_2 f_{2_i}\right\|_\infty \\ &\le \left\|W_i^{-1}V_i^T[R_i^{-1}]_1\right\|_\infty \|f_{1_i}\|_\infty + \left\|W_i^{-1}V_i^T[R_i^{-1}]_2\right\|_\infty \|f_{2_i}\|_\infty \\ &\le h^2 C_{01} \max_{t\in[t_0,T]}\{\|f_1(t)\|_\infty\} + hC_{02} \max_{t\in[t_0,T]}\{\|f_2(t)\|_\infty\} \\ &\le 2hC_{02} \max_{t\in[t_0,T]}\{\|f_1(t)\|_\infty, \|f_2(t)\|_\infty\}\end{aligned}$$

for sufficiently small h. Together with (3.36), we consequently obtain

$$\begin{aligned}\|\check{g}_h - \hat{g}_h\|_\infty &\le 2hC_{02} \max_{t\in[t_0,T]}\{\|f_1(t)\|_\infty, \|f_2(t)\|_\infty\} + hC_D \max_{i=0,\dots,k-1}\|x_i\|_\infty \\ &\le hC_g.\end{aligned}$$

□

From these observations it is clear that the system (3.32) is nothing else than the system (3.35) plus a perturbation of $O(h)$. This fact allows us to prove the main result of this section.

Theorem 26 *Let (1.1) be a strangeness free linear DAE. Let $x = D^- f$ be the solution of the minimization problem (2.27), i.e., D^- is the (1,2,3)-inverse of the differential-algebraic operator D associated with (1.1).*

Suppose that (1.1) is discretized with a k-step BDF-method, $k \le 6$, using a fixed step size $h = (T - t_0)/N$ and k initial values $x_0, \dots, x_{k-1}$ that satisfy

$$x_l - x(t_l) = O(h),$$

where $t_l = t_0 + lh$. If the systems (3.10) are solved in the least squares sense, i.e., the minimization problems (3.13) are solved successively for $i = k, \dots, N$, then we have for the solutions x_i of these problems the estimate

$$\|x_i - x(t_i)\|_\infty \le hC_{local},$$

with a positive constant C_{local} not depending on the step size h.

Proof. Let $\hat{D}_h\hat{x}_h = \hat{g}_h$, defined as in (3.35), represent the BDF-discretization of the uniquely solvable strangeness free DAE (3.27) and consider the restriction operator $R_{\mathbb{X}_h}$ from (3.4). Then the fact that BDF-methods applied to uniquely solvable strangeness free DAEs are convergent (see Theorem 20) implies

$$\|\hat{x}_h - R_{\mathbb{X}_h}x\|_\infty \le C_1 h \quad \text{for } h \to 0,$$

with some constant $C_1 > 0$. (Note that they would actually be convergent of order $p = k$ if we assumed sufficient accuracy in the initial values.)

Since BDF-methods are stable for $k \le 6$, we also have

$$\|\hat{D}_h^{-1}\|_\infty \le C_2$$

for some constant $C_2 > 0$, see [4] and also Section 3.3.2.

We now consider the linear system (3.32), whose solution represents the output of the BDF method applied to the DAE (1.1), where the systems (3.10) are solved in the least squares sense.

Applying Lemma 25 and using Neumann series, we can estimate the difference of the inverses of the discretization matrices $\check{D}_h$ and $\hat{D}_h$ for $h < \frac{1}{2}C_2C_D$ as follows:

$$\begin{aligned}
\left\|\check{D}_h^{-1} - \hat{D}_h^{-1}\right\|_\infty &= \left\|\left(\hat{D}_h - \left(\hat{D}_h - \check{D}_h\right)\right)^{-1} - \hat{D}_h^{-1}\right\|_\infty \\
&= \left\|\left(I - \hat{D}_h^{-1}\left(\hat{D}_h - \check{D}_h\right)\right)^{-1}\hat{D}_h^{-1} - \hat{D}_h^{-1}\right\|_\infty \\
&= \left\|\sum_{l=0}^{\infty}\left(\hat{D}_h^{-1}\left(\hat{D}_h - \check{D}_h\right)\right)^{l}\hat{D}_h^{-1} - \hat{D}_h^{-1}\right\|_\infty \\
&= \left\|\sum_{l=1}^{\infty}\left(\hat{D}_h^{-1}\left(\hat{D}_h - \check{D}_h\right)\right)^{l}\hat{D}_h^{-1}\right\|_\infty \\
&\leq \left\|\hat{D}_h^{-1}\right\|_\infty \sum_{l=1}^{\infty}\left(\left\|\hat{D}_h^{-1}\right\|_\infty \left\|\hat{D}_h - \check{D}_h\right\|\right)^{l} \\
&\leq C_2\sum_{l=1}^{\infty}\left(hC_2C_D\right)^{l} = hC_2^2C_D\sum_{l=0}^{\infty}\left(hC_2C_D\right)^{l} \\
&= \frac{hC_2^2C_D}{1 - hC_2C_D} \leq 2hC_2^2C_D.
\end{aligned}$$

This yields for the solution $x_h = \check{D}_h^{-1}\check{g}_h$,

$$\begin{aligned}
\|x_h - R_{\mathbb{X}_h}x\|_\infty &= \left\|\check{D}_h^{-1}\check{g}_h - R_{\mathbb{X}_h}x\right\|_\infty \\
&\leq \left\|\check{D}_h^{-1}\check{g}_h - \hat{D}_h^{-1}\check{g}_h + \hat{D}_h^{-1}\check{g}_h - \hat{D}_h^{-1}\hat{g}_h + \hat{D}_h^{-1}\hat{g}_h - R_{\mathbb{X}_h}x\right\|_\infty \\
&\leq \left\|\check{D}_h^{-1} - \hat{D}_h^{-1}\right\|_\infty \|\check{g}_h\|_\infty + \left\|\hat{D}_h^{-1}\right\|_\infty \|\check{g}_h - \hat{g}_h\|_\infty + \left\|\hat{D}_h^{-1}\hat{g}_h - R_{\mathbb{X}_h}x\right\|_\infty \\
&\leq hC_D \max_{t\in[t_0,T]}\|g(t)\|_\infty + hC_2C_g + hC_1 \\
&\leq hC_{local},
\end{aligned}$$

which implies the assertion. □

Remark 27 *It is important to remark that, no matter how high the order of the BDF method is chosen, the order of the bounds in Lemma 25 is always $O(h)$. This is due to presence of the error terms $\mathcal{R}_{1_i}$ (which is always $O(h^2)$) and $\mathcal{R}_{2_i}$ (which is always $O(h)$) in (3.23). Our numerical methods will compute the solution of (3.23), which represents the direct discretization of the DAE combined with local least squares solutions. Thus there is little justification for employing higher order BDF methods in our setting. This observation will be confirmed by the numerical experiments in Section 4.2.1.*

A (1,2,3)-inverse of D_h

Let us take a different look at the linear system (3.11). Solving the systems (3.10) according to the minimization problem (3.13), i.e., computing the solutions x_i via (3.14), defines a matrix D_h^- such that the solution x_h of (3.11) can be written as

$$x_h = D_h^- g_h.$$

Despite this notation, it is not clear yet that D_h^- is actually a (1,2,3)-inverse of D_h. This relationship will be proved in the following.

Theorem 28 *Let $x_h = [x_1^T, \ldots, x_N^T]^T$ denote the solution of the system $D_h x_h = g_h$ as in (3.11), obtained by computing x_i via (3.14) for $i = 1, \ldots, N$. Let $D_h^- \in \mathbb{R}^{mN,nN}$ be given such that $x_h = D_h^- g_h$.*

Then D_h^- is a (1,2,3)-inverse of D_h.

Proof. The i-th block row of the vector $D_h x_h$ can be written as

$$\begin{aligned}[D_h x_h]_i &= \left(\frac{1}{h}E_i - A_i\right) x_i + \frac{1}{h}\sum_{l=l_0}^{k-1} \alpha_l E_i x_{i-k+l} \\ &= P_i^T \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix} Q_i^T x_i + \frac{1}{h}\sum_{l=l_i}^{k-1} \alpha_l P_i^T \tilde{E}_i Q_i^T x_{i-k+l},\end{aligned} \tag{3.37}$$

for $i = k, \ldots, N$ with $l_i = \max(0, 2k - i)$. Here we have used (3.15) and (3.17). The solution $x_i = [x_h]_i = [D_h^- g_h]_i$ of the minimization problem (3.13) is given by (3.16). Inserting this solution into (3.37) yields for every $g_h = R_{\mathbb{Y}_h} f$, with $R_{\mathbb{Y}_h}$ defined in (3.5),

$$\begin{aligned}[D_h x_h]_i &= [D_h D_h^- g_h]_i \\ &= P_i^T \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix}^+ \left(-\frac{1}{h}\sum_{l=0}^{k-1} \alpha_l \tilde{E}_i Q_i^T x_{i-k+l} + \tilde{f}_i\right) + \frac{1}{h}\sum_{l=l_i}^{k-1} \alpha_l P_i^T \tilde{E}_i Q_i^T x_{i-k+l} \\ &= P_i^T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \left(-\frac{1}{h}\sum_{l=0}^{k-1} \alpha_l \tilde{E}_i Q_i^T x_{i-k+l} + \tilde{f}_i\right) + \frac{1}{h}\sum_{l=l_i}^{k-1} \alpha_l P_i^T \tilde{E}_i Q_i^T x_{i-k+l} \\ &= P_i^T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P_i \left(f_i - \frac{1}{h}\sum_{l=0}^{l_i-1} \alpha_l E_i x_{i-k+l}\right) \\ &= P_i^T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P_i [g_h]_i.\end{aligned}$$

Therefore we have

$$D_h D_h^- = \begin{bmatrix} P_k^T \hat{I} P_k & & \\ & \ddots & \\ & & P_N^T \hat{I} P_N \end{bmatrix}$$

with

$$\hat{I} = \begin{bmatrix} I_{\hat{d}+\hat{a}} & 0 \\ 0 & 0 \end{bmatrix}.$$

From

$$\hat{I} \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix}$$

and $\hat{I}\tilde{E} = \tilde{E}$ it immediately follows that

$$[D_h D_h^- D_h x_h]_i = [D_h x_h]_i$$

for $k = 1, \ldots, N$ and for all vectors x_h. From

$$\begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix}^+ \hat{I} = \begin{bmatrix} R_i & S_i \\ 0 & 0 \end{bmatrix}^+$$

we get

$$[D_h^- D_h D_h^- g_h]_i = [D_h^- g_h]_i$$

for $i = k, \ldots, N$ and for all right hand sides g_h. This shows that D_h and D_h^- satisfy the first three Moore-Penrose axioms (2.5) (1)–(3), thus D_h^- is a $(1,2,3)$-inverse of D_h. □

3.3 Global Minimization

In Section 2.2, a Moore-Penrose solution of a linear strangeness free DAE was defined as the solution of the minimization problem (2.14). For the analytical solution of (2.14) it is necessary to transform the system to the orthogonal standard form (2.1) and to solve a linear boundary value problem of the form (2.23). This is a viable approach for compute the Moore-Penrose solution numerically, provided that the orthogonal standard form of the DAE can be computed. However, as we have pointed on several occasions, this is not always possible for general DAEs.

This requires to develop a different approach for computing numerical approximations to the Moore-Penrose solution. Let us again discretize the DAE with a BDF-method. As defined in (3.11), we write down the whole discretization in terms of the large linear system

$$D_h x_h = f_h. \tag{3.38}$$

In the previous chapter we have examined the continuous solution $x = D^- f$ of a strangeness free differential-algebraic equation (1.1) with homogeneous initial conditions, where the $(1,2,3)$-inverse D^- of the differential-algebraic operator D was defined by the solution of the minimization problem (2.27). We have shown that this solution can be approximated by solving the decoupled minimization problems (3.13) after discretizing the DAE. We have shown that this solution of the discretization can be written as

$$x_h = D_h^- f_h,$$

where D_h is a $(1,2,3)$-inverse of D_h. This result motivates much of the following approach to compute an approximation of the Moore-Penrose solution $x = D^+ f$ of the DAE (1.1).

Instead of computing least squares solutions in every single step of the BDF discretization, as it was done in Section 3.2, we solve the complete system (3.38) in a least squares sense by solving the minimization problem

$$\frac{1}{2}\|x_h\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|D_h x_h - f_h\|^2 = \min!. \tag{3.39}$$

The solution can be written as

$$x_h = D_h^+ f_h, \tag{3.40}$$

where D_h^+ denotes the Moore-Penrose pseudoinverse of the matrix D_h. We will prove that this rather intuitive approach leads in fact to a numerical approximation to the Moore-Penrose solution $x = D^+ f$ of the DAE (1.1). Note that the orthogonal standard form of the DAE does not have to be computed to achieve this solution.

Being considerably long and technical, the proof of this assertion will be taken in two major steps. In a first step, we assume that the system is given in orthogonal standard form. In this case, solving the least squares problem (3.39) corresponds to the solution of a discrete boundary value problem, which can be interpreted as a convergent discretization of the continuous boundary value problem (2.23). In a second step, this result will be extended to general strangeness

free DAEs. We will restrict ourselves here to the implicit Euler method instead of considering general k-step BDF-methods. Using higher order BDF-methods requires highly accurate approximations to $k-1$ additional initial values and it is not clear how to obtain these approximations in an efficient manner, as we implicitly solve a boundary value problem. In view of Remark 27, it is moreover questionable whether this would result in higher order of convergence.

3.3.1 Systems in orthogonal standard form

Let

$$\tilde{E}(t)\dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{f}(t), \tag{3.41}$$

be a strangeness free DAE in orthogonal standard form, i.e. $\tilde{E}$, $\tilde{A}$ and $\tilde{f}$ are given as in (2.2). A discretization of (3.41) with the implicit Euler method using a fixed step size $h = (T - t_0)/N$ leads to equations of the form

$$\frac{1}{h}\tilde{E}_i(\tilde{x}_i - \tilde{x}_{i-1}) = \tilde{A}_i\tilde{x}_i + \tilde{f}_i \tag{3.42}$$

for $i = 1, \ldots, N$. This discretization can be written as

$$\tilde{D}_h\tilde{x}_h = \tilde{f}_h \tag{3.43}$$

with $\tilde{f}_h = [\tilde{f}_1^T, \ldots, \tilde{f}_N^T]^T$ and the *discretization matrix* $\tilde{D}_h$ is defined blockwise by

$$[\tilde{D}_h]_{ij} = \begin{cases} -\frac{1}{h}\tilde{E}_i & \text{for } j = i-1, \\ \frac{1}{h}\tilde{E}_i - \tilde{A}_i & \text{for } j = i, \\ 0 & \text{otherwise,} \end{cases} \tag{3.44}$$

where $i, j = 1, \ldots, N$. The goal is to show that the solution $\tilde{x}_h = [\tilde{x}_1^T, \ldots, \tilde{x}_N^T]^T$ of the minimization problem

$$\frac{1}{2}\left\|\tilde{x}_h\right\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\left\|\tilde{D}_h\tilde{x}_h - \tilde{f}_h\right\|^2 = \min!. \tag{3.45}$$

approximates the least squares solution of (3.41). Using the structure of the orthogonal standard form, the systems (3.42) can be written component-wise as

$$\begin{aligned} \frac{1}{h}\Sigma_{E_i}\left(x_{1_i} - x_{1_{i-1}}\right) &= A_{11_i}x_{1_i} + A_{12_i}x_{2_i} + A_{13_i}x_{3_i} + f_{1_i}, \\ 0 &= A_{21_i}x_{1_i} + \Sigma_{A_i}x_{2_i} + f_{2_i}, \\ 0 &= f_{3_i}, \end{aligned}$$

for $i = 1, \ldots, N$. The third equation is independent of $\tilde{x}_h$ and will be omitted in the following considerations. The algebraic components x_{2_i} of the unknown $\tilde{x}_i$ can be eliminated from the first equation by multiplying from the left with the regular matrices

$$\begin{bmatrix} \Sigma_{E_i}^{-1} & -\Sigma_{E_i}^{-1}A_{12_i}\Sigma_{A_i}^{-1} \\ 0 & \Sigma_{A_i}^{-1} \end{bmatrix}. \tag{3.46}$$

Here we assume that h is sufficiently small such that $\tilde{D}_h$, with $0 = f_{3_i}$ discarded, has full row rank.

After applying the substitutions

$$\begin{aligned} A_i &= \Sigma_{E_i}^{-1}\left(A_{11_i} - A_{12_i}\Sigma_{A_i}^{-1}A_{21_i}\right), \\ B_i &= \Sigma_{E_i}^{-1}A_{13_i}, \\ C_i &= -\Sigma_{A_i}^{-1}A_{21_i}, \\ f_i &= \Sigma_{E_i}^{-1}\left(f_{1_i} - A_{12_i}\Sigma_{A_i}^{-1}f_{2_i}\right), \\ g_i &= -\Sigma_{A_i}^{-1}f_{2_i}, \end{aligned} \tag{3.47}$$

these systems can be written in the form

$$\begin{aligned} \frac{1}{h}(x_{1_i} - x_{1_{i-1}}) &= A_i x_{1_i} + B_i x_{3_i} + f_i, \\ x_{2_i} &= C_i x_{1_i} + g_i. \end{aligned} \tag{3.48}$$

Finally, we rearrange the first equation and introduce the matrices

$$\begin{aligned} \bar{A}_i &= \frac{1}{h}\left(\frac{1}{h}I - A_i\right)^{-1}, \\ \bar{B}_i &= \left(\frac{1}{h}I - A_i\right)^{-1}B_i, \\ \bar{f}_i &= \left(\frac{1}{h}I - A_i\right)^{-1}f_i, \end{aligned} \tag{3.49}$$

where h is assumed to be sufficiently small, such that $\frac{1}{h}I - A_i$ is invertible, and rename the components of the solutions $\tilde{x}_i$ as $x_i = x_{1_i}$, $y_i = x_{2_i}$ and $u_i = x_{3_i}$. Then we can write (3.48) as the linear discrete-time system

$$\begin{aligned} x_i &= \bar{A}_i x_{i-1} + \bar{B}_i u_i + \bar{f}_i \\ y_i &= C_i x_i + g_i. \end{aligned}$$

All the transformations applied to obtain this system are regular and do not have any influence on the solution of the equations (3.42). Thus, the minimization problem (3.39) can be written in the form

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{N}\left(x_i^T x_i + y_i^T y_i + u_i^T u_i\right) = \min! \\ s.t. \quad &x_i = \bar{A}_i x_{i-1} + \bar{B}_i u_i + \bar{f}_i, \quad i = 1, ..., N, \quad x_0 = 0 \\ &y_i = C_i x_i + g_i, \end{aligned} \tag{3.50}$$

which turns out to be a discrete-time linear quadratic optimal control problem (see [35]), where the undetermined components x_{3_i} of the solution can be interpreted as an input u_i and the algebraic components x_{2_i} play the role of the output y_i. In our case, it is already known that the solution of the control problem (3.50) is unique, as it is represented by the unique solution $\tilde{x}_h = \tilde{D}_h^+ \tilde{f}_h$ of the minimization problem (3.43). The following theorem shows that this solution corresponds to the solution of a discrete linear boundary value problem.

Theorem 29 *For $i = 1, ..., N$, let*

$$A_i \in \mathbb{R}^{d,d}, \quad B_i \in \mathbb{R}^{d,k}, \quad C_i \in \mathbb{R}^{l,d}, \quad f_i \in \mathbb{R}^d, \quad g \in \mathbb{R}^l.$$

Let $u_* = [u_{*1}^T, ..., u_{*N}^T]^T$ solve the linear quadratic optimal control problem

$$\begin{aligned} & \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T x_i + y_i^T y_i + u_i^T u_i\right) = \min! \\ s.t. \quad x_i &= A_i x_{i-1} + B_i u_i + f_i, \quad i = 1, ..., N, \quad x_0 = 0, \\ y_i &= C_i x_i + g_i, \end{aligned} \tag{3.51}$$

and let $x_* = [x_{*0}^T, ..., x_{*N}^T]^T$ denote the associated trajectory. Furthermore, we set $y_* = [y_{*1}^T, ..., y_{*N}^T]^T$ with $y_{*i} = C_i x_{*i} + g_i$ for $i = 1, ...N$.

Then there exists $\lambda = [\lambda_0^T, ..., \lambda_N^T]^T$, $\lambda_i \in \mathbb{R}^d$, such that (x_*, λ, u_*) solves the discrete linear boundary value problem

$$\begin{aligned} A_{i+1}^T \lambda_i &= (I + C_i^T C_i)x_i + \lambda_{i-1} + C_i^T g_i, \quad \lambda_N = 0, \\ x_i &= A_i x_{i-1} + B_i u_i + f_i, \quad x_0 = 0, \\ y_i &= C_i x_i + g_i, \\ u_i &= B_i^T \lambda_{i-1} \end{aligned} \tag{3.52}$$

for $i = 1, ..., N$.

Proof. By inserting the y_i into the objective functional we obtain

$$\begin{aligned} \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T x_i + y_i^T y_i + u_i^T u_i\right) &= \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T x_i + (C_i x_i + g_i)^T (C_i x_i + g_i) + u_i^T u_i\right) \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T (I + C_i^T C_i) x_i + 2 g_i^T C_i x_i + u_i^T u_i + g_i^T g_i\right). \end{aligned}$$

The summands $g_i^T g_i$ can be neglected and thus, the minimization problem

$$\begin{aligned} & \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T (I + C_i^T C_i) x_i + 2 g_i^T C_i x_i + u_i^T u_i + g_i^T g_i\right) = \min! \\ s.t. \quad x_i &= A_i x_{i-1} + B_i u_i + f_i, \quad i = 1, ..., N, \quad x_0 = 0 \end{aligned} \tag{3.53}$$

is equivalent to the discrete linear quadratic control optimization problem

$$\frac{1}{2}\left(x^T M x + 2 g^T x + u^T u\right) = \min! \quad s.t. \quad Ax + Bu = f \tag{3.54}$$

with $x = [x_1^T, ..., x_N^T]^T$, $u = [u_1^T, ..., u_N^T]^T$, $M = \mathrm{diag}(I + C_1^T C_1, ..., I + C_N^T C_N)$, $g = [g_1^T C_1, ..., g_N^T C_N]^T$, $B = \mathrm{diag}(-B_1, ..., -B_N)$, $f = [f_1^T, ..., f_N^T]^T$ and

$$A = \begin{bmatrix} I & & \\ -A_2 & \ddots & \\ & \ddots & I \\ & & -A_N \; I \end{bmatrix}.$$

It is well-known (see, e.g., [14]), that due to the fact that the blockdiagonal matrix M is positive definite, the problem (3.54) possesses a unique solution. This solution is given by the solution of the unconstrained minimization problem

$$S(x,u,\lambda) = \frac{1}{2}\left(x^T M x + 2g^T x + u^T u\right) + \lambda^T(Ax + Bu - f) = \min! \tag{3.55}$$

with the Lagrange multiplier $\lambda = [\lambda_0^T, ..., \lambda_N^T]^T$. A necessary and sufficient condition for the solution of (3.55) is that the partial derivatives of the functional S with respect to x, u and λ are equal to zero, which yields the system

$$\begin{aligned} Mx + g + A^T\lambda &= 0, \\ u + B^T\lambda &= 0, \\ Ax + Bu - f &= 0. \end{aligned} \tag{3.56}$$

Together with the output equations $y_i = C_i x_i + g_i$, $i = 1, ..., N$, the system (3.56) is equivalent to the discrete boundary value problem (3.52). □

Theorem 29 can be directly applied to the control problem (3.50). It shows the existence of a vector $\bar\lambda_h = [\bar\lambda_0^T, \ldots, \bar\lambda_N^T]^T$ such that $\bar\lambda_h$ together with the components $(x_i, y_i, u_i) = (x_{1_i}, x_{2_i}, x_{3_i}) = \tilde x_i$ of $\tilde x_h = \tilde D_h^+ \tilde f_h$ satisfies the discrete boundary value problem

$$\begin{aligned} \bar A_{i+1}^T \bar\lambda_i &= (I + C_i^T C_i)x_i + \bar\lambda_{i-1} + C_i^T g_i, \quad \bar\lambda_N = 0, \\ x_i &= \bar A_i x_{i-1} + \bar B_i u_i + \bar f_i, \quad x_0 = 0, \\ y_i &= C_i x_i + g_i, \\ u_i &= \bar B_i^T \bar\lambda_{i-1}, \end{aligned} \tag{3.57}$$

$i = 1, ..., N$. Setting $\lambda_i = (\frac{1}{h} I - A_{i+1})^{-T} \bar\lambda_i$ and applying the transformations (3.49), this system can be written as

$$\begin{aligned} \frac{1}{h}(\lambda_i - \lambda_{i-1}) &= (I + C_i^T C_i)x_i - A_i^T \lambda_{i-1} + C_i^T g_i, \quad \lambda_N = 0, \\ \frac{1}{h}(x_i - x_{i-1}) &= A_i x_i + B_i u_i + f_i, \quad x_0 = 0, \\ y_i &= C_i x_i + g_i, \\ u_i &= B_i^T \lambda_{i-1}, \end{aligned} \tag{3.58}$$

$i = 1, ..., N$. This system coincides with a discretization by the implicit Euler method, forward in x and backward in λ, of the continuous boundary value problem (2.23) yielding the Moore-Penrose solution $(x, y, u) = \tilde x = \tilde D^+ \tilde f$. The solutions (x_i, y_i, u_i) are approximations to (x, y, u) at the grid point $t_i = t_0 + ih$ and consequently $\tilde x_h$ contains approximations to $\tilde x$. This fact will be proved by the following theorem, which shows that the systems (3.58) lead to a convergent discretization of the boundary value problem (2.23).

Theorem 30 *Let the boundary value problem*

$$\dot x(t) = A(t)x(t) + f(t), \quad R_{t_0} x(t_0) + R_T x(T) = r, \quad t \in [t_0, T], \tag{3.59}$$

with $A \in C([t_0, T], \mathbb{R}^{n,n})$, $f \in C([t_0, T], \mathbb{R}^n)$ sufficiently smooth, be uniquely solvable.

Then for any $d \in \mathbb{N}$ with $d < n$, the single step method defined by the recursion

$$x_{k+1} = x_k + hA(t_{k+1})\left(\hat I x_{k+1} + \left(I - \hat I\right) x_k\right) + hf(t_{k+1}), \tag{3.60}$$

with

$$\hat{I} = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n,n},$$

applied to (3.59) is convergent of order 1.

Proof. First, we have to show that the discretization defined by (3.60) is consistent. For the solution x of (3.59), using the Taylor expansion, we get for $t_1 = t_0 + h$

$$\begin{aligned} x(t_1) &= x(t_0) + h\dot{x}(t_0) + O(h^2) \\ &= x(t_0) + h\left(A(t_0)x(t_0) + f(t_0)\right) + O(h^2). \end{aligned}$$

For $x_0 = x(t_0)$, the recursion (3.60) yields

$$\begin{aligned} x(t_1) - x_1 &= h\left(A(t_0)x(t_0) + f(t_0) - A(t_1)\left(\hat{I}x_1 + x(t_0) - \hat{I}x(t_0)\right) - f(t_1)\right) + O(h^2) \\ &= h\left(\left(A(t_0) - A(t_1)\right)x(t_0) + A(t_1)\hat{I}(x(t_0) - x_1) + f(t_0) - f(t_1)\right) + O(h^2) \\ &= O(h^2) \end{aligned}$$

for sufficiently smooth functions A and f. This implies that the discretization defined by (3.60) is consistent of order 1. To show that it is convergent if applied to (3.59), we have to show that it is stable. The discretization can be written in terms of the linear system

$$L_h x_h = g_h,$$

with the discretization matrix

$$L_h = \left[\begin{array}{c|cccc} R_{t_0} & 0 & \dots & 0 & R_T \\ \hline L_{1,s} & L_{1,d} & & & \\ 0 & L_{2,s} & L_{2,d} & & \\ \vdots & & \ddots & \ddots & \\ 0 & & & L_{N,s} & L_{N,d} \end{array}\right]$$

and the right hand side

$$g_h = \begin{bmatrix} r \\ f_1 \\ \vdots \\ f_N \end{bmatrix}.$$

Here, the subdiagonal blocks $L_{i,s}$ and the diagonal blocks $L_{i,d}$ are defined as

$$\begin{aligned} L_{i,s} &= -\left(\frac{1}{h}I + A(t_i)\left(I - \hat{I}\right)\right), \\ L_{i,d} &= \left(\frac{1}{h}I - A(t_i)\hat{I}\right), \end{aligned}$$

and $f_i = f(t_i)$, $t_i = a + ih$, for $i = 1, \dots, N$.

The matrix

$$\bar{L}_h = \begin{bmatrix} L_{1,d} & & & \\ L_{2,s} & L_{2,d} & & \\ & \ddots & \ddots & \\ & & L_{N,s} & L_{N,d} \end{bmatrix}$$

can be interpreted as the discretization matrix belonging to the discretization (3.60) of the initial value problem

$$\dot{x} = Ax + f, \quad x(t_0) = 0, \quad t \in [t_0, T].$$

Since single step methods applied to initial value problems are always stable, we can conclude that $\bar{L}_h^{-1}$ is uniformly bounded, i.e.,

$$\left\|\bar{L}_h^{-1}\right\|_\infty \leq C_{\bar{L}}$$

for some positive constant $C_{\bar{L}}$ that does not depend on the step size h. The inverse of L_h can be computed using Schur complements via

$$L_h = \begin{bmatrix} R_{t_0} & \bar{R}_T \\ \bar{L}_{1,s} & \bar{L}_h \end{bmatrix} = \begin{bmatrix} I & \bar{R}_T \bar{L}_h^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} R_{t_0} - \bar{R}_T \bar{L}_h^{-1} \bar{L}_{1,s} & 0 \\ 0 & \bar{L}_h \end{bmatrix} \begin{bmatrix} I & o \\ \bar{L}_h^{-1} \bar{L}_{1,s} & I \end{bmatrix}$$

with

$$\bar{R}_T = \begin{bmatrix} 0 \cdots 0 \; R_T \end{bmatrix}, \qquad \bar{L}_{1,s} = \begin{bmatrix} L_{1,s} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This leads to

$$L_h^{-1} = \begin{bmatrix} S_h^{-1} & -S_h^{-1} \bar{R}_T \bar{L}_h^{-1} \\ -\bar{L}_h^{-1} \bar{L}_{1,s} S_h^{-1} & L_h^{-1} + L_h^{-1} \bar{L}_{1,s} S_h^{-1} \bar{R}_T \bar{L}_h^{-1} \end{bmatrix},$$

provided that the Schur complement

$$S_h = R_{t_0} - \bar{R}_T \bar{L}_h^{-1} \bar{L}_{1,s}$$

is nonsingular. The matrix S_h can be interpreted as follows.

A boundary value problem of the form (3.59) is uniquely solvable if and only if the matrix

$$R_{t_0} + R_T W(T, t_0)$$

is nonsingular, where the transfer function $W(t, a) \in C^1([t_0, T], \mathbb{R}^{n,n})$ is the solution of the initial value problem

$$\frac{d}{dt} W(t, t_0) = A(t) W(t, t_0), \quad W(t_0, t_0) = I, \quad t \in [t_0, T], \tag{3.61}$$

see, e.g., [2]. This initial value problem can be solved numerically by applying the discretization defined by (3.60) to (3.61). This discretization can be written as

$$\bar{L}_h W_h = \begin{bmatrix} 0 - L_{1,s} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = -\bar{L}_{1,s},$$

where the solution

$$W_h = \begin{bmatrix} W_1 \\ \vdots \\ W_N \end{bmatrix} = -\bar{L}_h^{-1} \bar{L}_{1,s}$$

contains approximations W_i to $W(t_0 + ih, t_0)$, i.e.,

$$W_i - W(t_0 + ih, t_0) = O(h) \quad \text{for} \quad i = 1, \ldots, N.$$

In particular, we have

$$\begin{aligned} S_h &= R_{t_0} - \bar{R}_T \bar{L}_h^{-1} \bar{L}_{1,s} \\ &= R_{t_0} + \bar{R}_T W_h \\ &= R_{t_0} + R_T W_N \\ &= R_{t_0} + R_T W(T, t_0) + O(h). \end{aligned}$$

Thus, for sufficiently small h the matrix S_h is nonsingular and its inverse S_h^{-1} is uniformly bounded. Therefore the matrix L_h^{-1} is also uniformly bounded and the consistent discretization defined by the recursion (3.60) applied to (3.59) is stable and thus convergent. □

Theorem 30 shows that the solution of the systems (3.58) is unique and that it converges to the solution of the boundary value problem (2.23). The following theorem summarizes the results obtained so far.

Theorem 31 *Let $\tilde{x} = \tilde{D}^+ \tilde{f}$ be the (unique) solution of the minimization problem (2.15) and let $\tilde{x}_h = \tilde{D}_h^+ \tilde{f}_h$ be the solution of the minimization problem (3.45). Let $R_{\mathbb{X}_h}$ be as in (3.4).*

Then there exists a positive constant $\tilde{C}$ such that

$$\|\tilde{x}_h - R_{\mathbb{X}_h} \tilde{x}\|_\infty \leq \tilde{C} h,$$

provided that $h > 0$ is sufficiently small.

Proof. According to Theorem 17, the solution (x, y, u) of the minimization problem (2.22) is given by the corresponding parts of the (unique) solution of the boundary value problem (2.23). This implies that the function

$$\tilde{x} = \begin{bmatrix} x \\ y \\ u \end{bmatrix}$$

solves the minimization problem (2.15).

The systems (3.58) can be written in the form

$$\begin{aligned} \frac{1}{h} \left(\begin{bmatrix} \lambda_i \\ x_i \end{bmatrix} - \begin{bmatrix} \lambda_{i-1} \\ x_{i-1} \end{bmatrix} \right) &= \begin{bmatrix} -A_i^T & I + C_i^T C_i \\ B_i B_i^T & A_i \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \lambda_i \\ x_i \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{i-1} \\ x_{i-1} \end{bmatrix} \right) + \begin{bmatrix} C_i^T g_i \\ f_i \end{bmatrix}, \\ y_i &= C_i x_i + g_i, \\ u_i &= B_i^T \lambda_{i-1}, \end{aligned}$$

$i = 1, \ldots, N$, with the initial values satisfying

$$\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \lambda_0 \\ x_0 \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_N \\ x_N \end{bmatrix} = 0.$$

By Theorem 30, this system is uniquely solvable for sufficiently small h and a convergent discretization of the boundary value problem (2.23). Therefore, by setting

$$\bar{x}_h = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_N \end{bmatrix} \quad \text{with} \quad \bar{x}_i = \begin{bmatrix} x_i \\ y_i \\ u_i \end{bmatrix}, \quad i = 1, \ldots, N,$$

we get

$$\|\bar{x}_h - R_{\mathbb{X}_h}\tilde{x}\| \leq \tilde{C}h$$

for some positive constant $\tilde{C}$.

The discrete boundary value problem (3.57) can be computed from (3.58) by applying the substitutions (3.49). Moreover, the unique solution $(\bar{\lambda}_i, x_i, y_i, u_i)$, $i = 0, \ldots, N$, of (3.57) is given by the solution $(\lambda_i, x_i, y_i, u_i)$ of (3.58) with $\bar{\lambda}_i = (\frac{1}{h}I - A_{i+1})^T\lambda_i$ for $i = 0, \ldots, N$.

By Theorem 29, the part (x_i, y_i, u_i), $i = 0, \ldots, N$, of this solution solves the minimization problem (3.50). By the substitutions (3.49), (3.47), and the regular transformation (3.46), the problem (3.50) can be written in the form

$$\begin{aligned} \frac{1}{2}\|\tilde{x}_h\| &= \min! \qquad \text{s.t.} \\ \frac{1}{h}\Sigma_{E_i}(x_{1_i} - x_{1_{i-1}}) &= A_{11_i}x_{1_i} + A_{2_i}x_{2_i} + A_{3_i}x_{3_i} + f_{1_i}, \quad x_{1_0} = 0, \\ 0 &= A_{21_i}x_{1_i} + \Sigma_{A_i}x_{2_i} + f_{2_i}, \end{aligned} \tag{3.62}$$

$i = 1, \ldots, N$, with

$$\tilde{x}_h = \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_N \end{bmatrix}, \quad \tilde{x}_i = \begin{bmatrix} x_{1_i} \\ x_{2_i} \\ x_{3_i} \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \\ u_i \end{bmatrix}, \quad i = 1, \ldots, N.$$

Hence, we get $\tilde{x}_h = \bar{x}_h$ for the solution of (3.62). The assertion follows because the minimization problem (3.62) is equivalent to the minimization problem (3.45). □

3.3.2 General strangeness free systems

To generalize the result of Theorem 31 to differential-algebraic systems that are not given in orthogonal standard form, we use the following observation. Let the large linear system $D_h x_h = f_h$ represent the discretization of the general strangeness free DAE (1.1) with the implicit Euler method. D_h can be defined analogously to (3.44) and the right-hand side is given by $f_h = (f_1, \ldots, f_N)$. Furthermore, consider the orthogonal matrices

$$\begin{aligned} P_h &= \operatorname{diag}(P_1, \ldots, P_N), \\ Q_h &= \operatorname{diag}(Q_1, \ldots, Q_N), \end{aligned} \tag{3.63}$$

which contain the orthogonal transformations P and Q evaluated at the grid points $t_1, \ldots, t_N$. If $\tilde{f}_h$ denotes the right-hand side corresponding to the discretization of the system transformed to orthogonal standard form, we have $\tilde{f}_h = P_h f_h$. It will be seen that the relation $\tilde{D}_h = P_h D_h Q_h$ unfortunately only holds if Q is constant over the interval $[t_0, T]$. This particularly implies that the minimization problems (3.39) and (3.43) do not transform covariantly with the application of these transformations if Q is time-dependent. Our major goal is to show that x_h nevertheless approximates $Q_h\tilde{x}_h$, where x_h and $\tilde{x}_h$ are the least squares solution of the discretizations $D_h x_h = f_h$ and $\tilde{D}_h\tilde{x}_h = \tilde{f}_h$ of the original DAE and the transformed DAE in orthogonal standard form, respectively.

The blocks of the matrix $P_h D_h Q_h$ are given by

$$[P_h D_h Q_h]_{ij} = \begin{cases} -\dfrac{1}{h}P_i E_i Q_{i-1} & \text{for } j = i-1, \\ P_i\left(\dfrac{1}{h}E_i - A_i\right)Q_i & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.64}$$

From (2.2), we know that

$$\tilde{A}_i = P_i A_i Q_i - P_i E_i \dot{Q}_i$$

and

$$\tilde{E}_i = P_i E_i Q_i.$$

This implies that

$$\begin{aligned} P_i \left(\frac{1}{h} E_i - A_i \right) Q_i &= \frac{1}{h} \tilde{E}_i - \tilde{A}_i - P_i E_i \dot{Q}_i \\ &= \frac{1}{h} \tilde{E}_i - \tilde{A}_i - \tilde{E}_i Q_i^T \dot{Q}_i, \end{aligned} \tag{3.65}$$

and for the subdiagonal blocks in (3.64), we get

$$\begin{aligned} -\frac{1}{h} P_i E_i Q_{i-1} &= -\frac{1}{h} \tilde{E}_i + \frac{1}{h} \tilde{E}_i Q_i^T Q_i - \frac{1}{h} \tilde{E}_i Q_i^T Q_{i-1} \\ &= -\frac{1}{h} \tilde{E}_i + \tilde{E}_i Q_i^T \frac{Q_i - Q_{i-1}}{h} \\ &= -\frac{1}{h} \tilde{E}_i + \tilde{E}_i Q_i^T \dot{Q}_i + \tilde{E}_i Q_i^T \left(\frac{Q_i - Q_{i-1}}{h} - \dot{Q}_i \right). \end{aligned} \tag{3.66}$$

For the large system, this implies the relationship

$$P_h D_h Q_h = \tilde{D}_h + \Delta_h, \tag{3.67}$$

with a perturbation matrix Δ_h that has the structure

$$\begin{bmatrix} \Delta_1 & & & & \\ -\Delta_2 - \delta_2 & \Delta_2 & & & \\ & -\Delta_3 - \delta_3 & \Delta_3 & & \\ & & \ddots & \ddots & \\ & & & -\Delta_N - \delta_N & \Delta_N \end{bmatrix}, \tag{3.68}$$

where $\Delta_i = -\tilde{E}_i Q_i^T \dot{Q}_i$ and $\delta_i = -\tilde{E}_i Q_i^T \left(\frac{Q_i - Q_{i-1}}{h} - \dot{Q}_i \right)$. Note that for sufficiently smooth orthogonal functions Q and sufficiently small h we have

$$\left\| \frac{Q_i - Q_{i-1}}{h} - \dot{Q}_i \right\|_\infty \leq C_{\dot{Q}} h$$

for some positive constant $C_{\dot{Q}}$. We will show in the following that for the perturbation Δ_h we get

$$(\tilde{D}_h + \Delta_h)^+ - \tilde{D}_h^+ = (P_h D_h Q_h)^+ - \tilde{D}_h^+ = O(h), \tag{3.69}$$

which then implies

$$\begin{aligned} \|x_h - Q_h \tilde{x}_h\|_\infty &= \left\| D_h^+ f_h - Q_h \tilde{D}_h^+ \tilde{f}_h \right\|_\infty \\ &= \left\| Q_h \left(Q_h^T D_h^+ P_h^T P_h f_h - \tilde{D}_h^+ \tilde{f}_h \right) \right\|_\infty \\ &\leq \|Q_h\|_\infty \left\| \left(Q_h^T D_h^+ P_h^T - \tilde{D}_h^+ \right) \tilde{f}_h \right\|_\infty \\ &= C_{\|Q\|_\infty} \left\| \left((P_h D_h Q_h)^+ - \tilde{D}_h^+ \right) \tilde{f}_h \right\|_\infty \\ &= O(h). \end{aligned}$$

The assertion (3.69) will be proved in three steps. First, it will be shown for the special case of ordinary differential equations, then this result will be extended to the case of uniquely solvable DAEs and finally to the case of general underdetermined DAEs.

The ODE case

Let us assume that $\hat{a} = \hat{u} = \hat{v} = 0$, with $\hat{a}, \hat{u}, \hat{v}$ as in Theorem 10, such that the DAE (2.1) in orthogonal standard form reduces to the implicit ordinary differential equation

$$\Sigma_E(t)\dot{\tilde{x}}(t) = A_{11}(t)\tilde{x}(t) + \tilde{f}_1(t), \quad t \in \mathbb{I} = [t_0, T], \tag{3.70}$$

with matrix functions $\Sigma_E \in C(\mathbb{I}, \mathbb{R}^{n,n})$, $A_{11} \in C(\mathbb{I}, \mathbb{R}^{n,n})$, $f_1 \in C(\mathbb{I}, \mathbb{R}^n)$, Σ_E pointwise nonsingular and $n = \hat{d}$. Moreover, we consider the initial condition

$$\tilde{x}(t_0) = 0.$$

Let the step size $h = (T - t_0)/N$ be constant. A discretization of (3.70) with the implicit Euler method and the step size h can be written in terms of the linear system

$$\tilde{D}_h^o \tilde{x}_h = \tilde{f}_h, \tag{3.71}$$

with

$$\tilde{D}_h^o = \begin{bmatrix} \frac{1}{h}\Sigma_{E_1} - A_{11_1} & & & \\ -\frac{1}{h}\Sigma_{E_2} & \frac{1}{h}\Sigma_{E_2} - A_{11_2} & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}\Sigma_{E_N} & \frac{1}{h}\Sigma_{E_N} - A_{11_N} \end{bmatrix} \tag{3.72}$$

and

$$\tilde{f}_h = \begin{bmatrix} \tilde{f}_1 \\ \vdots \\ \tilde{f}_N \end{bmatrix}.$$

On the other hand, let

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad t \in \mathbb{I} = [t_0, T] \tag{3.73}$$

denote an ODE, which is equivalent to (3.70), i.e., there exist orthogonal matrix functions $P \in C([t_0, T], \mathbb{R}^{n,n})$ and $Q \in C^1([t_0, T], \mathbb{R}^{n,n})$ such that $\Sigma_E = PEQ$, $A_{11} = PAQ - PE\dot{Q}$ and $f_1 = Pf$. The discretization of (3.73) with the implicit Euler method using a constant step size h can be represented by a linear system

$$D_h^o x_h = f_h \tag{3.74}$$

with

$$D_h^o = \begin{bmatrix} \frac{1}{h}E_1 - A_1 & & & \\ -\frac{1}{h}E_2 & \frac{1}{h}E_2 - A_2 & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}E_N & \frac{1}{h}E_N - A_N \end{bmatrix} \tag{3.75}$$

and

$$f_h = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix}.$$

If we define block diagonal matrices P_h and Q_h according to (3.63) and set

$$\hat{D}_h^o = P_h D_h^o Q_h, \tag{3.76}$$

then we have

$$\hat{D}_h^o = \tilde{D}_h^o + \Delta_h, \tag{3.77}$$

with a perturbation matrix Δ_h having the structure

$$\begin{bmatrix} \Delta_{11_1} & & & & & \\ -\Delta_{11_2} - \delta_{11_2} & \Delta_{11_2} & & & & \\ & -\Delta_{11_3} - \delta_{11_3} & \Delta_{11_3} & & & \\ & & \ddots & \ddots & & \\ & & & -\Delta_{11_N} - \delta_{11_N} & \Delta_{11_N} \end{bmatrix}.$$

In particular, we have $\|\Delta_{11_i}\|_\infty = \|\Sigma_{E_i} Q_i^T \dot{Q}_i\|_\infty \leq C_{\Delta_{11}}$ and $\|\delta_{11_i}\|_\infty \leq C_{\delta_{11}} h$ for sufficiently small h. Since the ODE systems (3.70) and (3.73) are globally equivalent, we have

$$\tilde{x} = Q^T x$$

for the (unique) solutions $\tilde{x}$ and x of (3.70) and (3.73). The implicit Euler method applied to these systems is convergent of order one; for sufficiently small step sizes h we therefore get

$$\|\tilde{x}_h - R_{\mathbb{X}_h} \tilde{x}\|_\infty = O(h)$$

and

$$\|x_h - R_{\mathbb{X}_h} x\|_\infty = O(h)$$

for the solutions $\tilde{x}_h$ and x_h of (3.71) and (3.74), respectively. If we let $\hat{x}_h$ denote the solution of the linear system

$$\hat{D}_h^o \hat{x}_h = \hat{f}_h,$$

with $\hat{f}_h = P_h f_h = \tilde{f}_h$, it follows that

$$\begin{aligned} \|\hat{x}_h - \tilde{x}_h\|_\infty &= \left\| \hat{D}_h^{o^{-1}} \hat{f}_h - \tilde{x}_h \right\|_\infty \\ &= \left\| Q_h^T D_h^{o^{-1}} P_h^T P_h f_h - \tilde{x}_h \right\|_\infty \\ &= \left\| Q_h^T x_h - \tilde{x}_h \right\|_\infty \\ &= \left\| Q_h^T x_h - Q_h^T R_{\mathbb{X}_h} x + Q_h^T R_{\mathbb{X}_h} x - \tilde{x}_h \right\|_\infty \\ &\leq \left\| Q_h^T \right\|_\infty \|x_h - R_{\mathbb{X}_h} x\|_\infty + \left\| R_{\mathbb{X}_h} Q^T x - \tilde{x}_h \right\|_\infty \\ &= C_{\|Q\|_\infty} \|x_h - R_{\mathbb{X}_h} x\|_\infty + \|R_{\mathbb{X}_h} \tilde{x} - \tilde{x}_h\|_\infty \\ &= O(h). \end{aligned}$$

This suggests that the inverses of the coefficient matrices $\hat{D}_h^o$ and $\tilde{D}_h^o$ satisfy

$$\left\| \hat{D}_h^{o^{-1}} - \tilde{D}_h^{o^{-1}} \right\|_\infty = O(h). \tag{3.78}$$

We will show that this is indeed true. In addition, we will show that for the blocks $[\hat{D}_h^{o^{-1}}]_{ij}$ and $[\tilde{D}_h^{o^{-1}}]_{ij}$, $i,j = 1,\dots,N$, of $\hat{D}_h^{o^{-1}}$ and $\tilde{D}_h^{o^{-1}}$, we have

$$\left\| [\hat{D}_h^{o^{-1}}]_{ij} - [\tilde{D}_h^{o^{-1}}]_{ij} \right\|_\infty = O(h^2).$$

To show this property we first define a third discretization of the ODE (3.70) that allows us to compute the inverse of its discretization matrix in a simple way. This explicit discretization is defined by

$$\begin{aligned} \frac{1}{h}\Sigma_{E_i}(x_{1_i} - x_{1_{i-1}}) &= A_{11_i}x_{1_{i-1}} + f_{1_i}, \quad i = 1,...,N, \\ x_{1_0} &= 0, \end{aligned} \tag{3.79}$$

and leads to a discretization matrix of the form

$$\bar{D}_h^o = \begin{bmatrix} \frac{1}{h}\Sigma_{E_1} & & & \\ -\frac{1}{h}\Sigma_{E_2} - A_{11_2} & \frac{1}{h}\Sigma_{E_2} & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}\Sigma_{E_N} - A_{11_N} & \frac{1}{h}\Sigma_{E_N} \end{bmatrix}. \tag{3.80}$$

If we denote the diagonal and (negative) subdiagonal entries of $\bar{D}_h^o$ by $\bar{D}_{h_{d,i}}^o$ and $\bar{D}_{h_{s,i}}^o$, respectively, i.e.,

$$\begin{aligned} [\bar{D}_h^o]_{d,i} &= \frac{1}{h}\Sigma_{E_i}, \\ [\bar{D}_h^o]_{s,i} &= \frac{1}{h}\Sigma_{E_i} + A_{11_i}, \end{aligned} \tag{3.81}$$

then we can write down the inverse blockwise according to the block structure of $\bar{D}_h^o$ by

$$\begin{aligned} [\bar{D}_h^{o^{-1}}]_{ij} &= \begin{cases} \left[\coprod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1}[\bar{D}_h^o]_{s,k}\right] [\bar{D}_h^o]_{d,j}^{-1} & \text{for } i \geq j, \\ 0 & \text{otherwise}, \end{cases} \\ &= \begin{cases} h\left[\coprod_{k=i}^{j+1} \left(I + h\Sigma_{E_k}^{-1}A_{11_k}\right)\right] \Sigma_{E_j}^{-1} & \text{for } i \geq j, \\ 0 & \text{otherwise}. \end{cases} \end{aligned} \tag{3.82}$$

Here, we employ the following non-standard notation: for matrices $A_{k_1}, A_{k_1+1}, \dots, A_{k_2}$, $k_2 \geq k_1$, we define the reversed product

$$\coprod_{k=k_2}^{k_1} A_k := A_{k_2}A_{k_2-1}...A_{k_1+1}A_{k_1}. \tag{3.83}$$

If $k_2 < k_1$, we set

$$\coprod_{k=k_2}^{k_1} A_k := I.$$

Since one-step formulas are always stable, we know that $\|\bar{D}_h^{o^{-1}}\|_\infty$ is uniformly bounded. The following lemma can be used to compute this bound explicitly.

Lemma 32 *Let $\bar{D}_h^o$ be defined as in (3.80). Then for*

$$C_1^o = C_{\|\Sigma_E^{-1}\|_\infty} \exp\left(C_{\|\Sigma_E^{-1}A_{11}\|_\infty}\right)$$

we have

$$\left\|[\bar{D}_h^{o^{-1}}]_{ij}\right\|_\infty \leq C_1^o h,$$

which implies

$$\left\|\bar{D}_h^{o^{-1}}\right\|_\infty \leq C_1^o (T - t_0).$$

Here, the constants $C_{\|\Sigma_E^{-1}\|_\infty}$ and $C_{\|\Sigma_E^{-1}A_{11}\|_\infty}$ are defined according to Definition 23.

Proof. For $[\bar{D}_h^o]_{d,k}$ and $[\bar{D}_h^o]_{s,k}$ as defined in (3.81) we have, using the notation (3.83),

$$\begin{aligned}
\left\|\prod_{k=i}^{j+1}[\bar{D}_h^o]_{d,k}^{-1}[\bar{D}_h^o]_{s,k}\right\|_\infty &= \left\|\prod_{k=i}^{j+1}(I + h\Sigma_{E_k}^{-1}A_{11_k})\right\|_\infty \\
&\leq \prod_{k=j+1}^{i}\left\|I + h\Sigma_{E_k}^{-1}A_{11_k}\right\|_\infty \\
&\leq \prod_{k=1}^{N}\left\|I + h\Sigma_{E_k}^{-1}A_{11_k}\right\|_\infty \\
&\leq \left(1 + hC_{\|\Sigma_E^{-1}A_{11}\|_\infty}\right)^N \\
&= \left(1 + \frac{1}{N}C_{\|\Sigma_E^{-1}A_{11}\|_\infty}(T - t_0)\right)^N \\
&\leq \exp\left(C_{\|\Sigma_E^{-1}A_{11}\|_\infty}(T - t_0)\right),
\end{aligned}$$

and thus for $1 \leq j \leq i \leq N$, it follows from (3.82) that

$$\begin{aligned}
\left\|[\bar{D}_h^{o^{-1}}]_{ij}\right\|_\infty &= \left\|h\left[\prod_{k=i}^{j+1}\left(I + h\Sigma_{E_k}^{-1}A_{11_k}\right)\right]\Sigma_{E_j}^{-1}\right\|_\infty \\
&\leq C_{\|\Sigma_E^{-1}\|_\infty}\exp\left(C_{\|\Sigma_E^{-1}A_{11}\|_\infty}(T - t_0)\right)h = C_1^o h.
\end{aligned}$$

Hence,

$$\left\|\bar{D}_h^{o^{-1}}\right\|_\infty \leq \max_{i=1,\ldots,N}\sum_{j=1}^{N}\left\|[\bar{D}_h^{o^{-1}}]_{ij}\right\|_\infty \leq NC_1^o h = C_1^o(T - t_0),$$

which concludes the proof. □

Additionally, we can use $\bar{D}_h^{o^{-1}}$ as an approximation to $\hat{D}_h^{o^{-1}}$ as well as to $D_h^{o^{-1}}$, which will help proving (3.78). The following lemma shows that $\bar{D}_h^{o^{-1}}$ has the desired properties.

Lemma 33 *Let $\bar{D}_h^o$ and $\hat{D}_h^o$ be defined as in (3.80) and (3.76). Then there exists a constant $C_2^o \in \mathbb{R}^+$, which does not depend on the step size h, such that for $i, j = 1, ..., N$,*

$$\left\|[\bar{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty \leq C_2^o h^2.$$

Proof. First of all, similar to (3.82), we write $\hat{D}_h^{o^{-1}}$ blockwise in the form

$$[\hat{D}_h^{o^{-1}}]_{ij} = \begin{cases} \left[\coprod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k}\right] [\hat{D}_h^o]_{d,j}^{-1} & \text{for } i \geq j, \\ 0 & \text{otherwise,} \end{cases} \tag{3.84}$$

using the notation

$$\begin{aligned} [\hat{D}_h^o]_{d,i} &:= \frac{1}{h}\Sigma_{E_i} - A_{11_i} + \Delta_{11_i}, \\ [\hat{D}_h^o]_{s,i} &:= \frac{1}{h}\Sigma_{E_i} + \Delta_{11_i} + \delta_{11_i}, \end{aligned}$$

for the diagonal and (negative) subdiagonal blocks of $\hat{D}_h^o$, respectively. The following observation already proves the desired result for the diagonal blocks of the inverted matrices. We have

$$\begin{aligned} [\hat{D}_h^o]_{d,i}^{-1} &= \left(\frac{1}{h}\Sigma_{E_i} - A_{11_i} + \Delta_{11_i}\right)^{-1} \\ &= h\left(\Sigma_{E_i}\left(I - h\Sigma_{E_i}^{-1}\left(A_{11_i} - \Delta_{11_i}\right)\right)\right)^{-1} \\ &= h\left(I - h\Sigma_{E_i}^{-1}\left(A_{11_i} - \Delta_{11_i}\right)\right)^{-1}\Sigma_{E_i}^{-1}. \end{aligned}$$

Furthermore, for

$$h < \min\left(1, C^{-1}_{\|\Sigma_E^{-1}(A_{11}-\Delta_{11})\|_\infty}\right),$$

where $C_{\|\Sigma_E^{-1}(A_{11}-\Delta_{11})\|_\infty}$ is defined according to Definition 23, we have

$$\left\|h\Sigma_{E_i}^{-1}\left(A_{11_i} - \Delta_{11_i}\right)\right\|_\infty < 1$$

for $i = 1, \ldots, N$. Hence, we can use the Neumann series to compute the inverse by

$$\begin{aligned} [\hat{D}_h^o]_{d,i}^{-1} &= h\left(\sum_{k=0}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i} - \Delta_{11_i})\right)^k\right)\Sigma_{E_i}^{-1} \\ &= h\left(I + \sum_{k=1}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i} - \Delta_{11_i})\right)^k\right)\Sigma_{E_i}^{-1}. \end{aligned}$$

Because of

$$\begin{aligned} \left\|\sum_{k=0}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i} - \Delta_{11_i})\right)^k\right\|_\infty &\leq \sum_{k=0}^{\infty} h^k \left\|\Sigma_{E_i}^{-1}(A_{11_i} - \Delta_{11_i})\right\|_\infty^k \\ &= \frac{1}{1 - h\left\|\Sigma_{E_i}^{-1}(A_{11_i} - \Delta_{11_i})\right\|_\infty} \\ &\leq 2 \end{aligned}$$

and

$$\begin{aligned}\left\|\sum_{k=1}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right)^k\right\|_\infty &\le h\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty \sum_{k=1}^{\infty} h^k\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty^{k-1}\\ &= h\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty \sum_{k=0}^{\infty} h^k\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty^{k}\\ &= h\frac{\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty}{1-h\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty}\\ &\le 2h\left\|\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right\|_\infty\\ &\le 2hC_{\|\Sigma_E^{-1}(A_{11}-\Delta_{11})\|_\infty},\end{aligned}$$

we obtain for $h \le \frac{1}{2}\min\left(1, C^{-1}_{\|\Sigma_E^{-1}(A_{11}-\Delta_{11})\|_\infty}\right)$ that

$$\left\|[\hat{D}_h^o]_{d,i}^{-1}\right\|_\infty \le 2C_{\|\Sigma_E^{-1}\|_\infty}h \quad \text{for} \quad h\to 0. \tag{3.85}$$

Moreover,

$$\begin{aligned}\left\|[\hat{D}_h^o]_{d,i}^{-1}-[\bar{D}_h^o]_{d,i}^{-1}\right\|_\infty &= \left\|h\left(I+\sum_{k=1}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right)^k\right)\Sigma_{E_i}^{-1}-h\Sigma_{E_i}^{-1}\right\|_\infty\\ &= h\left\|\left(\sum_{k=1}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right)^k\right)\Sigma_{E_i}^{-1}\right\|_\infty\\ &\le h\left\|\sum_{k=1}^{\infty}\left(h\Sigma_{E_i}^{-1}(A_{11_i}-\Delta_{11_i})\right)^k\right\|_\infty\left\|\Sigma_{E_i}^{-1}\right\|_\infty\\ &\le C_3^o h^2 \quad \text{for} \quad h\to 0,\end{aligned}$$

where

$$C_3^o = 2C_{\|\Sigma_E^{-1}(A_{11}-\Delta_{11})\|_\infty}C_{\|\Sigma_E^{-1}\|_\infty}.$$

To get a similar estimate for the remaining blocks, we first define

$$M_k := [\hat{D}_h^o]_{d,k}^{-1}[\hat{D}_h^o]_{s,k} - [\bar{D}_h^o]_{d,k}^{-1}[\bar{D}_h^o]_{s,k}$$

and observe that

$$\begin{aligned}M_k &= \left(\frac{1}{h}\Sigma_{E_k}-A_{11_k}+\Delta_{11_k}\right)^{-1}\left(\frac{1}{h}\Sigma_{E_k}+\Delta_{11_k}+\delta_{11_k}\right)-\left(I+h\Sigma_{E_k}^{-1}A_{11_k}\right)\\ &= I+\left(\frac{1}{h}\Sigma_{E_k}-A_{11_k}+\Delta_{11_k}\right)^{-1}(A_{11_k}+\delta_{11_k})-\left(I+h\Sigma_{E_k}^{-1}A_{11_k}\right)\\ &= \left(\frac{1}{h}\Sigma_{E_k}-A_{11_k}+\Delta_{11_k}\right)^{-1}A_{11_k}-h\Sigma_{E_k}^{-1}A_{11_k}+\left(\frac{1}{h}\Sigma_{E_k}-A_{11_k}+\Delta_{11_k}\right)^{-1}\delta_{11_k}\\ &= \left([\hat{D}_h^o]_{d,k}^{-1}-[\bar{D}_h^o]_{d,k}^{-1}\right)A_{11_k}+[\hat{D}_h^o]_{d,k}^{-1}\delta_{11_k},\end{aligned}$$

and thus

$$\|M_k\|_\infty \le h^2\left(C_3^o C_{\|A_{11}\|_\infty}+2C_{\|\Sigma_E^{-1}\|_\infty}C_\delta\right) =: h^2 C_4^o$$

for $k = 1, \ldots, N$. Applying Lemma 32, we get

$$
\begin{aligned}
&\left\| \coprod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k} - \coprod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k} \right\|_\infty \\
&\quad = \left\| \coprod_{k=i}^{j+1} \left(I + h\Sigma_{E_k}^{-1} A_{11_k} + M_k \right) - \coprod_{k=i}^{j+1} \left(I + h\Sigma_{E_k}^{-1} A_k \right) \right\|_\infty \\
&\quad \leq \sum_{l=1}^{i-j} \binom{i-j}{l} \left(\max_{k \in \{j+1,\ldots,i\}} \left\| I + h\Sigma_{E_k}^{-1} A_{11_k} \right\|_\infty \right)^{i-j-l} \left(\max_{k \in \{j+1,\ldots,i\}} \|M_k\|_\infty \right)^l \\
&\quad \leq \sum_{l=1}^{N} N^l \left(1 + h \max_{t \in [t_0,T]} \left\| \Sigma_E^{-1}(t) A_{11}(t) \right\|_\infty \right)^N (h^2 C_4^o)^l \\
&\quad \leq \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right) \sum_{l=1}^{\infty} (h C_4^o (T - t_0))^l \\
&\quad = h \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right) C_4^o (T - t_0) \sum_{l=0}^{\infty} (h C_4^o (T - t_0))^l \\
&\quad = h \frac{C_4^o (T - t_0) \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right)}{1 - h C_4^o (T - t_0)} \\
&\quad \leq 2h C_4^o (T - t_0) \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right)
\end{aligned}
$$

where the last equality holds for

$$
h \leq \frac{1}{2C_4^o (T - t_0)}. \tag{3.86}
$$

This implies that the product appearing in the representation (3.84) of $\hat{D}_h^{o^{-1}}$ is also bounded, i.e.,

$$
\begin{aligned}
\left\| \coprod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k} \right\|_\infty &\leq \left\| \coprod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k} - \coprod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k} \right\|_\infty + \left\| \coprod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k} \right\|_\infty \\
&\leq 2h C_4^o (T - t_0) \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right) + \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right) \\
&\leq 2 \exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty} (T - t_0) \right)
\end{aligned}
$$

for h satisfying (3.86). Thus, it follows that

$$\begin{aligned}
\left\|[\bar{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty &= \left\| \left[\prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k}\right] [\hat{D}_h^o]_{d,j}^{-1} - \left[\prod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k}\right] [\bar{D}_h^o]_{d,j}^{-1} \right\|_\infty \\
&\leq \left\| \left[\prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k}\right] [\hat{D}_h^o]_{d,j}^{-1} - \left[\prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k}\right] [\bar{D}_h^o]_{d,j}^{-1} \right\|_\infty \\
&\quad + \left\| \left[\prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k}\right] [\bar{D}_h^o]_{d,j}^{-1} - \left[\prod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k}\right] [\bar{D}_h^o]_{d,j}^{-1} \right\|_\infty \\
&\leq \left\| \prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k} \right\|_\infty \left\| [\hat{D}_h^o]_{d,j}^{-1} - [\bar{D}_h^o]_{d,j}^{-1} \right\|_\infty \\
&\quad + \left\| \prod_{k=i}^{j+1} [\hat{D}_h^o]_{d,k}^{-1} [\hat{D}_h^o]_{s,k} - \prod_{k=i}^{j+1} [\bar{D}_h^o]_{d,k}^{-1} [\bar{D}_h^o]_{s,k} \right\|_\infty \left\| [\bar{D}_h^o]_{d,j}^{-1} \right\|_\infty \\
&\leq 2h^2 \exp(C_{\|\Sigma_E^{-1} A_{11}\|_\infty}(T - t_0)) C_3^o \\
&\quad + 2h^2 C_4^o \exp(C_{\|\Sigma_E^{-1} A_{11}\|_\infty}(T - t_0)) C_{\|\Sigma_E^{-1}\|_\infty} \\
&= C_2^o h^2,
\end{aligned}$$

where

$$C_2^o = 2\exp(C_{\|\Sigma_E^{-1} A_{11}\|_\infty}(T - t_0))(C_3^o + C_4^o C_{\|\Sigma_E^{-1}\|_\infty}).$$

□

Note that the result of Lemma 33 also applies if $\hat{D}_h^{o^{-1}}$ is replaced by $\tilde{D}_h^{o^{-1}}$, since $\hat{D}_h^o$ reduces to $\tilde{D}_h^o$ for $\Delta_h = 0$.

Theorem 34 *Let $\tilde{D}_h^o$ and $\hat{D}_h^o$ be defined as in (3.75) and (3.80). Then there exists a constant $C^o \in \mathbb{R}^+$, which does not depend on the step size h, such that for $i, j = 1, ..., N$,*

$$\left\|[\tilde{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty \leq C^o h^2.$$

Moreover,

$$\left\|\tilde{D}_h^{o^{-1}} - \hat{D}_h^{o^{-1}}\right\|_\infty \leq C^o (T - t_0) h.$$

Proof. Using Lemma 33 and the triangular inequality, we get

$$\begin{aligned}
\left\|[\tilde{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty &\leq \left\|[\tilde{D}_h^{o^{-1}}]_{ij} - [\bar{D}_h^{o^{-1}}]_{ij}\right\|_\infty + \left\|[\bar{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty \\
&\leq \left(C_2^{o(1)} + C_2^{o(2)}\right) h^2,
\end{aligned}$$

where both constants, $C_2^{o(1)}$ and $C_2^{o(2)}$, can be computed as shown in the proof of Lemma 33. In particular, we obtain $C_2^{o(1)}$ by setting $\Delta_{11} = 0$ and $\delta_{11} = 0$, i.e.,

$$\begin{aligned}
C_2^{o(1)} &= 4\exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty}(T - t_0)\right) C_{\|\Sigma_E^{-1}\|_\infty} C_{\|\Sigma_E^{-1} A_{11}\|_\infty} \left(1 + C_{\|A_{11}\|_\infty}\right), \\
C_2^{o(2)} &= 4\exp\left(C_{\|\Sigma_E^{-1} A_{11}\|_\infty}(T - t_0)\right) C_{\|\Sigma_E^{-1}\|_\infty} \left(C_{\|\Sigma_E^{-1}(A_{11} - \Delta_{11})\|_\infty} \left(1 + C_{\|A_{11}\|_\infty}\right) + C_\delta\right).
\end{aligned}$$

Let us define

$$C^o = C_2^{o(1)} + C_2^{o(2)}.$$

Then it follows that

$$\begin{aligned}\left\|\tilde{D}_h^{o^{-1}} - \hat{D}_h^{o^{-1}}\right\|_\infty &= \max_{i=1,\ldots,N} \sum_{j=1}^{N} \left\|[\tilde{D}_h^{o^{-1}}]_{ij} - [\hat{D}_h^{o^{-1}}]_{ij}\right\|_\infty \\ &\leq NC^o h^2 = C^o(T-t_0)h.\end{aligned}$$

□

The uniquely solvable DAE case

The next step is to extend the results of the previous step for ODEs to the case of a uniquely solvable strangeness free DAE

$$\tilde{E}(t)\dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{f}(t), \quad t \in \mathbb{I} = [t_0, T], \quad \tilde{x}(t_0) = 0, \tag{3.87}$$

with the coefficient functions

$$\tilde{E} = \begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \end{bmatrix},$$
$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & \Sigma_A \end{bmatrix},$$

where Σ_E and Σ_A are are both square, pointwise nonsingular matrix functions of size $\hat{d}$ and $\hat{a}$, respectively. The matrix functions A_{11}, A_{12} and A_{21} are assumed to be of matching size. We proceed as in the previous chapter by discretizing (3.87) with the implicit Euler method and writing the discretization as a linear system

$$\tilde{D}_h^u \tilde{x}_h = \tilde{f}_h. \tag{3.88}$$

Next, we consider a linear DAE

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad t \in \mathbb{I} = [t_0, T], \quad x(t_0) = 0, \tag{3.89}$$

which is equivalent to the system (3.87), i.e., there exist orthogonal matrix functions $P \in C([t_0, T], \mathbb{R}^{n,n})$ and $Q \in C^1([t_0, T], \mathbb{R}^{n,n})$, $n = \hat{d} + \hat{a}$, such that $\tilde{E} = PEQ$, $\tilde{A} = PAQ - PE\dot{Q}$ and $\tilde{f} = Pf$. A discretization of (3.89) with the implicit Euler method leads to a linear system of the form

$$D_h^u x_h = f_h.$$

We will show that the solution $\hat{x}_h$ of the system

$$\hat{D}_h^u \hat{x}_h = \hat{f}_h$$

with $\hat{D}_h^u = P_h D_h Q_h$ and $\hat{f}_h = P_h f_h = \tilde{f}_h$ approximates the solution $\tilde{x}_h$ of (3.88); an assertion that is motivated by the fact that the solutions $\tilde{x}$ and x of (3.87) and (3.89) satisfy $\tilde{x} = Q^T x$. In particular, we will show that $\|\hat{D}_h^{u^{-1}} - \tilde{D}_h^{u^{-1}}\|_\infty = O(h)$ and estimate the norm of the differences between certain blocks of these inverses.

Similarly as for the ODE case in (3.67), we get the following relation for the matrices $\tilde{D}_h^u$ and $\hat{D}_h^u$:

$$\hat{D}_h^u = \tilde{D}_h^u + \Delta_h,$$

where the perturbation matrix Δ_h has the block bidiagonal structure given in (3.68), with the blocks

$$\Delta_i = -\tilde{E}_i Q_i^T \dot{Q}_i = \begin{bmatrix} \Delta_{11_i} & \Delta_{12_i} \\ 0 & 0 \end{bmatrix}$$

and

$$\delta_i = -\tilde{E}_i Q_i^T \left(\frac{Q_i - Q_{i-1}}{h} - \dot{Q}_i \right) = \begin{bmatrix} \delta_{11_i} & \delta_{12_i} \\ 0 & 0 \end{bmatrix}.$$

In particular, for sufficiently small h and for a sufficiently smooth matrix function Q there exist positive constants $C_{\Delta_{11}}$, $C_{\Delta_{12}}$, $C_{\delta_{11}}$ and $C_{\delta_{12}}$ such that

$$\begin{aligned} \|\Delta_{11_i}\|_\infty &\le C_{\Delta_{11}}, \\ \|\Delta_{12_i}\|_\infty &\le C_{\Delta_{12}}, \\ \|\delta_{11_i}\|_\infty &\le C_{\delta_{11}} h, \\ \|\delta_{12_i}\|_\infty &\le C_{\delta_{12}} h, \end{aligned}$$

for $i = 1, \ldots, N$. To compute the inverses $\tilde{D}_h^{u^{-1}}$ and $\hat{D}_h^{u^{-1}}$, we introduce permutations $P_{l,h}$ and $P_{r,h}$ such that

$$P_{h,l} \tilde{D}_h^u P_{h,r} = \begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix} \tag{3.90}$$

and

$$P_{h,l} \hat{D}_h^u P_{h,r} = \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}, \tag{3.91}$$

where

$$\hat{A}_{12_h} = A_{12_h} + \Delta_{12_h} \tag{3.92}$$

and

$$\begin{aligned} A_{12_h} &= \begin{bmatrix} -A_{12_1} & & \\ & \ddots & \\ & & -A_{12_N} \end{bmatrix}, \\ \Delta_{12_h} &= \begin{bmatrix} \Delta_{12_1} & & & \\ -\Delta_{12_2} - \delta_{12_2} & \Delta_{12_2} & & \\ & \ddots & \ddots & \\ & & -\Delta_{12_N} - \delta_{12_N} & \Delta_{12_N} \end{bmatrix}, \\ A_{21_h} &= \begin{bmatrix} -A_{21_1} & & \\ & \ddots & \\ & & -A_{21_N} \end{bmatrix}, \\ \Sigma_{A_h} &= \begin{bmatrix} -\Sigma_{A_1} & & \\ & \ddots & \\ & & -\Sigma_{A_N} \end{bmatrix}. \end{aligned} \tag{3.93}$$

Furthermore, $\tilde{D}_h^o$ as well as $\hat{D}_h^o = \tilde{D}_h^o + \Delta_{11_h}$ are defined as in (3.72) and (3.77), respectively. The structure of the transformed discretization matrices allows us to compute the inverse matrices with the help of the block LU decomposition

$$\begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix} = \begin{bmatrix} I & A_{12_h}\Sigma_{A_h}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{D}_h^o - A_{12_h}\Sigma_{A_h}^{-1}A_{21_h} & 0 \\ 0 & \Sigma_{A_h} \end{bmatrix} \begin{bmatrix} I & 0 \\ \Sigma_{A_h}^{-1}A_{21_h} & I \end{bmatrix}.$$

This immediately yields

$$\begin{aligned} \begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1} &= \begin{bmatrix} I & 0 \\ -\Sigma_{A_h}^{-1}A_{21_h} & I \end{bmatrix} \begin{bmatrix} \left(\tilde{D}_h^o - A_{12_h}\Sigma_{A_h}^{-1}A_{21_h}\right)^{-1} & 0 \\ 0 & \Sigma_{A_h}^{-1} \end{bmatrix} \begin{bmatrix} I & -A_{12_h}\Sigma_{A_h}^{-1} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} S_h^{-1} & -S_h^{-1}A_{12_h}\Sigma_{A_h}^{-1} \\ -\Sigma_{A_h}^{-1}A_{21_h}S_h^{-1} & \Sigma_{A_h}^{-1} + \Sigma_{A_h}^{-1}A_{21_h}S_h^{-1}A_{12_h}\Sigma_{A_h}^{-1} \end{bmatrix}, \end{aligned} \tag{3.94}$$

with

$$\begin{aligned} \tilde{S}_h &= \tilde{D}_h^o - A_{12_h}\Sigma_{A_h}^{-1}A_{21_h} \\ &= \begin{bmatrix} \frac{1}{h}\Sigma_{E_1} - \bar{A}_1 & & & \\ -\frac{1}{h}\Sigma_{E_2} & \frac{1}{h}\Sigma_{E_2} - \bar{A}_2 & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}\Sigma_{E_N} & \frac{1}{h}\Sigma_{E_N} - \bar{A}_N \end{bmatrix}, \\ \bar{A}_i &= A_{11_i} + A_{12_i}\Sigma_{A_i}^{-1}A_{21_i}. \end{aligned} \tag{3.95}$$

Analogously, we get

$$\begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1} = \begin{bmatrix} \hat{S}_h^{-1} & -\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma_{A_h}^{-1} \\ -\Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1} & \Sigma_{A_h}^{-1} + \Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma_{A_h}^{-1} \end{bmatrix},$$

where

$$\begin{aligned} \hat{S}_h &= \hat{D}_h^o - \hat{A}_{12_h}\Sigma_{A_h}^{-1}A_{21_h} \\ &= \hat{D}_h^o - \begin{bmatrix} -A_{12_1} + \Delta_{12_1} & & & \\ -\Delta_{12_2} - \delta_{12_2} & -A_{12_2} + \Delta_{12_2} & & \\ & \ddots & \ddots & \\ & & -\Delta_{12_N} - \delta_{12_N} & -A_{12_N} + \Delta_{12_N} \end{bmatrix} \Sigma_{A_h}^{-1}A_{21_h} \end{aligned}$$

can be written in the form

$$\hat{S}_h = \tilde{S}_h + \hat{\Delta}_h \tag{3.96}$$

with

$$\begin{aligned} \hat{\Delta}_h &= \begin{bmatrix} \hat{\Delta}_1 & & & \\ -\hat{\Delta}_2 - \hat{\delta}_2 & \hat{\Delta}_2 & & \\ & \ddots & \ddots & \\ & & -\hat{\Delta}_N - \hat{\delta}_N & \hat{\Delta}_N \end{bmatrix}, \\ \hat{\Delta}_i &= \Delta_{11_i} - \Delta_{12_i}\Sigma_{A_i}^{-1}A_{21_i}, \end{aligned}$$

and

$$\hat{\delta}_i = \delta_{11_i} - \Delta_{12_i}\left(\Sigma_{A_{i-1}}^{-1}A_{21_{i-1}} - \Sigma_{A_i}^{-1}A_{21_i}\right) - \delta_{12_i}\Sigma_{A_i}^{-1}A_{21_i}.$$

Before proceeding, we need the following definition.

Definition 35 *Let $A \in C^1([t_0,T],\mathbb{R}^{m,n})$ and $2 \le p \le \infty$. Then the constant $L_{\|A\|_p}$ is defined as*

$$L_{\|A\|_p} = \max_{t\in[t_0,T]} \{\|\dot{A}(t)\|_p, \|\dot{A}(t)^T\|_p\}.$$

Because of

$$\begin{aligned}\left\|\hat{\delta}_i\right\|_\infty &= \left\|\delta_{11_i} - \Delta_{12_i}\left(\Sigma_{A_{i-1}}^{-1}A_{21_{i-1}} - \Sigma_{A_i}^{-1}A_{21_i}\right) - \delta_{12_i}\Sigma_{A_i}^{-1}A_{21_i}\right\|_\infty \\ &\le \left(C_{\delta_{11}} + C_{\Delta_{12}}L_{\|\Sigma_A^{-1}A_{21}\|_\infty} + C_{\delta_{12}}C_{\|\Sigma_A^{-1}A_{21}\|_\infty}\right)h,\end{aligned}$$

it follows that we can apply Lemma 32 and Lemma 33 directly to the matrices $\tilde{S}_h$ and $\hat{S}_h$. Here and in the following, by writing $L_{\|\cdot\|_\infty}$ we implicitly assume that the argument is differentiable, which, by Theorem 3, can be guaranteed if we assume the matrix pair (E,A) to be sufficiently smooth. Then the above inequality follows from the mean value theorem.

The matrix $\tilde{S}_h$ can be interpreted as the discretization matrix of the ordinary differential equation

$$\Sigma_E(t)\dot{x}(t) = \bar{A}(t)x(t) + f(t), \tag{3.97}$$

with

$$\bar{A} = A_{11} + A_{12}\Sigma_A A_{21},$$

according to a discretization with the implicit Euler method. As in the previous section, we consider an explicit discretization of (3.97) similar to (3.79) with the discretization matrix

$$\bar{S}_h = \begin{bmatrix} \frac{1}{h}\Sigma_{E_1} & & & \\ -\frac{1}{h}\Sigma_{E_2} - \bar{A}_2 & \frac{1}{h}\Sigma_{E_2} & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}\Sigma_{E_N} - \bar{A}_N & \frac{1}{h}\Sigma_{E_N} \end{bmatrix}, \tag{3.98}$$

which possesses the inverse $\bar{S}_h^{-1}$ with the blocks

$$[\bar{S}_h^{-1}]_{ij} = \begin{cases} h\left[\prod_{k=i}^{j+1}(I + h\Sigma_{E_k}^{-1}\bar{A}_k)\right]\Sigma_{E_j}^{-1} & \text{for } i \ge j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.99}$$

Corollary 36 *Let $\bar{S}_h$ and $\hat{S}_h$ be defined as in (3.98) and (3.96). Then there exists a constant $C_1^u \in \mathbb{R}$, which does not depend on the step size h, such that for $i,j = 1,...,N$,*

$$\left\|[\bar{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{ij}\right\|_\infty \le C_1^u h^2 \tag{3.100}$$

for sufficiently small h. Furthermore,

$$\left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty \le C_2^u h \tag{3.101}$$

and

$$\left\|\hat{S}_h^{-1}\right\|_\infty \le C_2^u(T - t_0) \tag{3.102}$$

with

$$C_2^u = 2C_{\|\Sigma_E^{-1}\|_\infty}\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right).$$

Proof. The assertion (3.100) follows immediately from Lemma 33. In particular, we get

$$C_1^u = 4\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right) C_{\|\Sigma_E^{-1}\|_\infty}\left(C_{\|\Sigma_E^{-1}(\bar{A}-\hat{\Delta})\|_\infty}\left(1+C_{\|\bar{A}\|_\infty}\right)+C_{\hat{\delta}}\right).$$

From Lemma 32, it follows that $\|[\bar{S}_h^{-1}]_{ij}\|_\infty \leq C_1^o h$ for $i,j=1,\ldots,N$ with

$$C_1^o = C_{\|\Sigma_E^{-1}\|_\infty}\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right),$$

and thus

$$\begin{aligned}\left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty &\leq \left\|[\hat{S}_h^{-1}]_{ij}-[\bar{S}_h^{-1}]_{ij}\right\|_\infty + \left\|[\bar{S}_h^{-1}]_{ij}\right\|_\infty \\ &\leq C_1^u h^2 + C_1^o h \leq 2C_1^o h \leq C_2^u h\end{aligned}$$

holds for sufficiently small h. Then the estimate (3.102) follows from

$$\begin{aligned}\left\|\hat{S}_h^{-1}\right\|_\infty &\leq \max_i \sum_{j=1}^N \left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty \\ &\leq 2NC_1^o h = C_2^u(T-t_0).\end{aligned}$$

□

The results of Corollary 36 help us to show two further properties of $\hat{S}_h^{-1}$ and $\tilde{S}_h^{-1}$.

Lemma 37 *Let $\hat{S}_h$ and $\tilde{S}_h$ be defined as in (3.96) and (3.95). Then there exist constants C_3^u, $C_4^u \in \mathbb{R}^+$, which do not depend on the step size h, such that*

$$\left\|[\tilde{S}_h^{-1}]_{ij}-[\hat{S}_h^{-1}]_{ij}\right\|_\infty \leq C_3^u h^2 \tag{3.103}$$

for $i,j=1,...,N$, and

$$\left\|[\hat{S}_h^{-1}]_{ij}-[\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty \leq C_4^u h^2 \tag{3.104}$$

for $i=1,...,N$, $j=1,...,N-1$.

Proof. The first estimate follows directly from Corollary 36, because

$$\begin{aligned}\left\|[\tilde{S}_h^{-1}]_{ij}-[\hat{S}_h^{-1}]_{ij}\right\|_\infty &\leq \left\|[\tilde{S}_h^{-1}]_{ij}-[\bar{S}_h^{-1}]_{ij}\right\|_\infty + \left\|[\bar{S}_h^{-1}]_{ij}-[\hat{S}_h^{-1}]_{ij}\right\|_\infty \\ &\leq (C_1^{u(1)}+C_1^{u(2)})h^2,\end{aligned}$$

where both $C_1^{u(1)}$ and $C_1^{u(2)}$ can be computed as in Corollary 36, i.e.,

$$\begin{aligned}C_1^{u(1)} &= 4\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right) C_{\|\Sigma_E^{-1}\|_\infty} C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}\left(1+C_{\|\bar{A}\|_\infty}\right), \\ C_1^{u(2)} &= 4\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right) C_{\|\Sigma_E^{-1}\|_\infty}\left(C_{\|\Sigma_E^{-1}(\bar{A}-\hat{\Delta})\|_\infty}\left(1+C_{\|\bar{A}\|_\infty}\right)+C_{\hat{\delta}}\right).\end{aligned}$$

Thus, the inequality (3.103) follows from setting

$$C_3^u = C_1^{u(1)} + C_1^{u(2)}.$$

To prove the second estimate, we first show that the matrix $\bar{S}_h^{-1}$ satisfies an inequality of the form (3.104):

$$\begin{aligned}
\left\|[\bar{S}_h^{-1}]_{ij} - [\bar{S}_h^{-1}]_{i,j+1}\right\|_\infty &= \left\| h \left[\prod_{k=i}^{j+1}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right)\right]\Sigma_{E_j}^{-1} + h\left[\prod_{k=i}^{j+2}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right)\right]\Sigma_{E_{j+1}}^{-1}\right\|_\infty \\
&\le h\left(\left\|\prod_{k=i}^{j+1}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right)\right\|_\infty \left\|\Sigma_{E_j}^{-1} - \Sigma_{E_{j+1}}^{-1}\right\|_\infty \right. \\
&\quad \left. + \left\|\prod_{k=i}^{j+1}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right) - \prod_{k=i}^{j+2}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right)\right\|_\infty \left\|\Sigma_{E_{j+1}}^{-1}\right\|_\infty\right) \\
&\le h\left(h\exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right)L_{\|\Sigma_E^{-1}\|_\infty}\right. \\
&\quad \left. + \left\|\left[\prod_{k=i}^{j+2}\left(I + h\Sigma_{E_k}^{-1}\bar{A}_k\right)\right]\left(\left(I + h\Sigma_{E_{j+1}}^{-1}\bar{A}_{j+1}\right) - I\right)\right\|_\infty \left\|\Sigma_{E_{j+1}}^{-1}\right\|_\infty\right) \\
&\le \exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right)\left(L_{\|\Sigma_E^{-1}\|_\infty} + C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}C_{\|\Sigma_E^{-1}\|_\infty}\right)h^2.
\end{aligned}$$

Thus, by applying Corollary 36, we get

$$\begin{aligned}
\left\|[\hat{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty &\le \left\|[\hat{S}_h^{-1}]_{ij} - [\bar{S}_h^{-1}]_{i,j}\right\|_\infty + \left\|[\bar{S}_h^{-1}]_{ij} - [\bar{S}_h^{-1}]_{i,j+1}\right\|_\infty \\
&\quad + \left\|[\bar{S}_h^{-1}]_{i,j+1} - [\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty \\
&\le \underbrace{\left(2C_1^u + \exp\left(C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}(T-t_0)\right)\left(L_{\|\Sigma_E^{-1}\|_\infty} + C_{\|\Sigma_E^{-1}\bar{A}\|_\infty}C_{\|\Sigma_E^{-1}\|_\infty}\right)\right)}_{=:\,C_4^u} h^2.
\end{aligned}$$

□

Next, we take further steps to prove the main result for uniquely solvable DAEs. First, we extend the result of Corollary 36 to the remaining blocks of the inverted transformed discretization matrices. We denote these blocks as follows:

$$\begin{bmatrix} \hat{D}_{h11}^{u^{inv}} & \hat{D}_{h12}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} & \hat{D}_{h22}^{u^{inv}} \end{bmatrix} := \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1}, \tag{3.105}$$

i.e.,

$$\begin{aligned}
\hat{D}_{h11}^{u^{inv}} &= \hat{S}_h^{-1}, \\
\hat{D}_{h12}^{u^{inv}} &= -\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma_{A_h}^{-1}, \\
\hat{D}_{h21}^{u^{inv}} &= -\Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1}, \\
\hat{D}_{h22}^{u^{inv}} &= \Sigma_{A_h}^{-1} + \Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma_{A_h}^{-1}.
\end{aligned} \tag{3.106}$$

Analogously, we define the blocks $\tilde{D}_{h11}^{u^{inv}}$, $\tilde{D}_{h12}^{u^{inv}}$, $\tilde{D}_{h21}^{u^{inv}}$ and $\tilde{D}_{h22}^{u^{inv}}$ for the case $\Delta_h = 0$.

Lemma 38 *Let*

$$\begin{bmatrix} \hat{D}_{h11}^{u^{inv}} & \hat{D}_{h12}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} & \hat{D}_{h22}^{u^{inv}} \end{bmatrix}$$

be defined as in (3.105). Then there exist positive constants $C^u_{11}, C^u_{12}, C^u_{21}, C^u_{22}$ such that

$$\left\|\hat{D}^{u^{inv}}_{h11}\right\|_\infty \leq C^u_{11}, \tag{3.107}$$

$$\left\|\hat{D}^{u^{inv}}_{h12}\right\|_\infty \leq C^u_{12}, \tag{3.108}$$

$$\left\|\hat{D}^{u^{inv}}_{h21}\right\|_\infty \leq C^u_{21}, \tag{3.109}$$

$$\left\|\hat{D}^{u^{inv}}_{h22}\right\|_\infty \leq C^u_{22}. \tag{3.110}$$

Proof. The assertion (3.107) follows directly from Corollary 36. From the equality $\hat{D}^{u^{inv}}_{h11} = \hat{S}_h^{-1}$ it follows that

$$\left\|\hat{D}^{u^{inv}}_{h11}\right\|_\infty = \left\|\hat{S}_h^{-1}\right\|_\infty \leq C^u_2(T-t_0)$$

and we can set $C^u_{11} = C^u_2(T-t_0)$. For the remaining assertions we note that the matrix $\hat{S}_h^{-1}\hat{A}_{12_h}$ can be written blockwise as

$$[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ij} = \begin{cases} [\hat{S}_h^{-1}]_{ii}\left(-A_{12_i} + \Delta_{12_i}\right) & \text{for } i = j, \\ [\hat{S}_h^{-1}]_{ij}\left(-A_{12_j} + \Delta_{12_j}\right) - [\hat{S}_h^{-1}]_{i,j+1}\left(\Delta_{12_{j+1}} + \delta_{12_{j+1}}\right) & \text{for } i > j, \\ 0 & \text{otherwise.} \end{cases}$$

From Corollary 36, we get

$$\begin{aligned} \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ii}\right\|_\infty &\leq \left\|[\hat{S}_h^{-1}]_{ii}\right\|_\infty \left(\|A_{12_i}\|_\infty + \|\Delta_{12_i}\|_\infty\right) \\ &\leq C^u_2\left(C_{\|A_{12}\|_\infty} + C_{\Delta_{12}}\right)h \end{aligned}$$

and for $i > j$, using Corollary 36 and Lemma 37, we obtain

$$\begin{aligned} \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ij}\right\|_\infty &\leq \left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty \left\|A_{12_j}\right\|_\infty + \left\|[\hat{S}_h^{-1}]_{ij}\Delta_{12_j} - [\hat{S}_h^{-1}]_{i,j+1}\Delta_{12_{j+1}}\right\|_\infty \\ &\quad + \left\|[\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty \left\|\delta_{12_{j+1}}\right\|_\infty \\ &\leq C^u_2\left(C_{\|A_{12}\|_\infty} + C_{\delta_{12}}h\right)h \\ &\quad + \left\|[\hat{S}_h^{-1}]_{ij}\left(\Delta_{12_j} - \Delta_{12_{j+1}}\right)\right\|_\infty + \left\|\left([\hat{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{i,j+1}\right)\Delta_{12_{j+1}}\right\|_\infty \\ &\leq C^u_2 C_{\|A_{12}\|_\infty}h + \left(C^u_2\left(L_{\|\Delta_{12}\|_\infty} + C_{\delta_{12}}\right) + C^u_4 C_{\Delta_{12}}\right)h^2 \end{aligned}$$

for sufficiently small h. It follows that

$$\begin{aligned} \left\|\hat{S}_h^{-1}\hat{A}_{12_h}\right\|_\infty &\leq \max_i \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ii}\right\|_\infty + \sum_{j=1}^{i-1}\left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ij}\right\|_\infty \\ &\leq \max_i \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ii}\right\|_\infty + (N-1)\left\|[\hat{S}_h^{-1}\hat{A}_{12_h}]_{ij}\right\|_\infty \\ &\leq C^u_2(T-t_0)\left(C_{\|A_{12}\|_\infty} + (T-t_0)C_{\Delta_{12}}h\right) + \left(C^u_2\left(L_{\|\Delta_{12}\|_\infty} + C_{\delta_{12}}\right) + C^u_4 C_{\Delta_{12}}\right)h \\ &\leq 2C^u_{11}C_{\|A_{12}\|_\infty} \end{aligned}$$

for sufficiently small h.

The assertions (3.108), (3.109) and (3.110) follow from (3.106) by observing that $\Sigma_{A_h}^{-1}$ and A_{21_h} are block diagonal matrices, and thus $\|\Sigma_{A_h}^{-1}\|_\infty \leq C_{\|\Sigma_A^{-1}\|_\infty}$, $\|A_{21_h}\|_\infty \leq C_{\|A_{21}\|_\infty}$. Hence, we can choose

$$\begin{aligned} C_{12}^u &= 2C_{11}^u C_{\|A_{12}\|_\infty} C_{\|\Sigma_A^{-1}\|_\infty}, \\ C_{21}^u &= C_{\|\Sigma_A^{-1}\|_\infty} C_{\|A_{21}\|_\infty} C_{11}^u, \\ C_{22}^u &= C_{\|\Sigma_A^{-1}\|_\infty} \left(1 + 2C_{\|A_{21}\|_\infty} C_{11}^u C_{\|A_{12}\|_\infty} C_{\|\Sigma_A^{-1}\|_\infty}\right). \end{aligned}$$

□

Lemma 39 *There exist constants $\bar{C}_{11}$, $\bar{C}_{12d}$, $\bar{C}_{12s}$, $\bar{C}_{21}$, $\bar{C}_{22d}$ and $\bar{C}_{22s} \in \mathbb{R}^+$, which do not depend on the step size h, such that the following inequalities hold:*

1. $\left\|[\hat{D}_{h11}^{u^{inv}}]_{ij} - [\tilde{D}_{h11}^{u^{inv}}]_{ij}\right\|_\infty \leq \bar{C}_{11}h^2$ *for* $i,j = 1,...,N$;

2. $\left\|[\hat{D}_{h12}^{u^{inv}}]_{ii} - [\tilde{D}_{h12}^{u^{inv}}]_{ii}\right\|_\infty \leq \bar{C}_{12d}h$ *for* $i = 1,...,N$,
   $\left\|[\hat{D}_{h12}^{u^{inv}}]_{ij} - [\tilde{D}_{h12}^{u^{inv}}]_{ij}\right\|_\infty \leq \bar{C}_{12s}h^2$ *for* $i,j = 1,...,N$, $j < i$;

3. $\left\|[\hat{D}_{h21}^{u^{inv}}]_{ij} - [\tilde{D}_{h21}^{u^{inv}}]_{ij}\right\|_\infty \leq \bar{C}_{21}h^2$ *for* $i,j = 1,...,N$;

4. $\left\|[\hat{D}_{h22}^{u^{inv}}]_{ii} - [\tilde{D}_{h22}^{u^{inv}}]_{ii}\right\|_\infty \leq \bar{C}_{22d}h$ *for* $i = 1,...,N$,
   $\left\|[\hat{D}_{h22}^{u^{inv}}]_{ij} - [\tilde{D}_{h22}^{u^{inv}}]_{ij}\right\|_\infty \leq \bar{C}_{22s}h^2$ *for* $i,j = 1,...,N$, $j < i$.

Proof.

1. Lemma 37 implies

$$\begin{aligned} \left\|[\hat{D}_{h11}^{u^{inv}}]_{ij} - [D_{h11}^{u^{inv}}]_{ij}\right\|_\infty &= \left\|[\hat{S}_h^{-1}]_{ij} - [\tilde{S}_h^{-1}]_{ij}\right\|_\infty \\ &\leq \bar{C}_{11}h^2, \end{aligned}$$

   where $\bar{C}_{11} := C_3^u$.

2. Corollary 36 and Lemma 37 imply

$$\begin{aligned} \left\|[\hat{D}_{h12}^{u^{inv}}]_{ii} - [\tilde{D}_{h12}^{u^{inv}}]_{ii}\right\|_\infty &= \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma_{A_h}^{-1}]_{ii} - [\tilde{S}_h^{-1}A_{12_h}\Sigma_{A_h}^{-1}]_{ii}\right\|_\infty \\ &\leq \left\|\left([\hat{S}_h^{-1}]_{ii}[\hat{A}_{12_h}]_{ii} - [\tilde{S}_h^{-1}]_{ii}[A_{12_h}]_{ii}\right)[\Sigma_{A_h}^{-1}]_{ii}\right\|_\infty \\ &\leq \left\|[\hat{S}_h^{-1}]_{ii}\left(-A_{12_i} + \Delta_{12_i}\right) + [\tilde{S}_h^{-1}]_{ii}A_{12_i}\right\|_\infty \left\|\Sigma_{A_i}^{-1}\right\|_\infty \\ &\leq \left(\left\|\left([\hat{S}_h^{-1}]_{ii} - [\tilde{S}_h^{-1}]_{ii}\right)A_{12_i}\right\|_\infty + \left\|[\hat{S}_h^{-1}]_{ii}\Delta_{12_i}\right\|_\infty\right) C_{\|\Sigma_A^{-1}\|_\infty} \\ &\leq \left(C_3^u C_{\|A_{12}\|_\infty}h^2 + 2C_{\|\Sigma_E^{-1}\|_\infty}C_{\Delta_{12}}h\right) C_{\|\Sigma_A^{-1}\|_\infty} \\ &\leq \underbrace{3C_{\|\Sigma_E^{-1}\|_\infty}C_{\Delta_{12}}C_{\|\Sigma_A^{-1}\|_\infty}}_{=:\,\bar{C}_{12d}}h \end{aligned}$$

for sufficiently small h. Moreover,

$$\begin{aligned}
\left\|[\hat{D}^{u^{inv}}_{h12}]_{ij} - [\tilde{D}^{u^{inv}}_{h12}]_{ij}\right\|_\infty &= \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma^{-1}_{A_h}]_{ij} - [\tilde{S}_h^{-1}A_{12_h}\Sigma^{-1}_{A_h}]_{ij}\right\|_\infty \\
&\le \left\|\left([\hat{S}_h^{-1}]_{ij}[\hat{A}_{12_h}]_{jj} + [\hat{S}_h^{-1}]_{i,j+1}[\hat{A}_{12_h}]_{j+1,j} - [\tilde{S}_h^{-1}]_{ij}[A_{12_h}]_{jj}\right)[\Sigma^{-1}_{A_h}]_{jj}\right\|_\infty \\
&\le \Big\|[\hat{S}_h^{-1}]_{ij}\left(-A_{12_j} + \Delta_{12_j}\right) - [\hat{S}_h^{-1}]_{i,j+1}\left(\Delta_{12_{j+1}} + \delta_{12_{j+1}}\right) \\
&\qquad +[\tilde{S}_h^{-1}]_{ij}A_{12_j}\Big\|_\infty \left\|\Sigma^{-1}_{A_i}\right\|_\infty \\
&\le \Big(\left\|\left([\tilde{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{ij}\right)A_{12_j}\right\|_\infty + \left\|[\hat{S}_h^{-1}]_{i,j+1}\delta_{12_{j+1}}\right\|_\infty \\
&\qquad + \left\|[\hat{S}_h^{-1}]_{ij}\Delta_{12_j} - [\hat{S}_h^{-1}]_{i,j+1}\Delta_{12_{j+1}}\right\|_\infty\Big) C_{\|\Sigma^{-1}_A\|_\infty} \\
&\le \Big(\left(C_3^u C_{\|A_{12}\|_\infty} + C_2^u C_{\delta_{12}}\right)h^2 + \left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty \left\|\Delta_{12_j} - \Delta_{12_{j+1}}\right\|_\infty \\
&\qquad + \left\|[\hat{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty \left\|\Delta_{12_{j+1}}\right\|_\infty\Big) C_{\|\Sigma^{-1}_A\|_\infty} \\
&\le \underbrace{\left(C_3^u C_{\|A_{12}\|_\infty} + C_2^u C_{\delta_{12}} + C_2^u L_{\|\Delta_{12}\|_\infty} + C_4^u C_{\|\Delta_{12}\|_\infty}\right) C_{\|\Sigma^{-1}_A\|_\infty}}_{=:\,\bar{C}_{12s}} h^2
\end{aligned}$$

for sufficiently small h.

3. Once again, Lemma 37 implies

$$\begin{aligned}
\left\|[\hat{D}^{u^{inv}}_{h21}]_{ij} - [\tilde{D}^{u^{inv}}_{h21}]_{ij}\right\|_\infty &= \left\|[\Sigma^{-1}_{A_h}A_{21_h}\hat{S}_h^{-1}]_{ij} - [\Sigma^{-1}_{A_h}A_{21_h}\tilde{S}_h^{-1}]_{ij}\right\|_\infty \\
&= \left\|-\Sigma^{-1}_{A_i}A_{21_i}[\hat{S}_h^{-1}]_{ij} + \Sigma^{-1}_{A_i}A_{21_i}[S_h^{-1}]_{ij}\right\|_\infty \\
&\le \left\|\Sigma^{-1}_{A_i}A_{21_i}\right\|_\infty \left\|[S_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{ij}\right\|_\infty \\
&\le \underbrace{C_{\|\Sigma^{-1}_A\|_\infty}C_{\|A_{21}\|_\infty}C_3^u}_{=:\,\bar{C}_{21}} h^2
\end{aligned}$$

for sufficiently small h.

4. Similarly,

$$\begin{aligned}
\left\|[\hat{D}^{u^{inv}}_{h22}]_{ii} - [\tilde{D}^{u^{inv}}_{h22}]_{ii}\right\|_\infty &= \Big\|[\Sigma^{-1}_{A_h} + \Sigma^{-1}_{A_h}A_{21_h}\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma^{-1}_{A_h}]_{ii} \\
&\qquad -[\Sigma^{-1}_{A_h} + \Sigma^{-1}_{A_h}A_{21_h}\tilde{S}_h^{-1}A_{12_h}\Sigma^{-1}_{A_h}]_{ii}\Big\|_\infty \\
&= \left\|\Sigma^{-1}_{A_i}A_{21_i}\right\|_\infty \left\|[\hat{S}_h^{-1}\hat{A}_{12_h}\Sigma^{-1}_{A_h}]_{ii} - [\tilde{S}_h^{-1}A_{12_h}\Sigma^{-1}_{A_h}]_{ii}\right\|_\infty \\
&\le \underbrace{C_{\|\Sigma^{-1}_A\|_\infty}C_{\|A_{12}\|_\infty}\bar{C}_{12d}}_{=:\,\bar{C}_{22d}} h,
\end{aligned}$$

$$\begin{aligned}\left\|[\hat{D}_{h22}^{u^{inv}}]_{ij} - [\tilde{D}_{h22}^{u^{inv}}]_{ij}\right\|_\infty &= \Big\|[\Sigma_{A_h}^{-1} + \Sigma_{A_h}^{-1} A_{21_h} \hat{S}_h^{-1} \hat{A}_{12_h} \Sigma_{A_h}^{-1}]_{ij} \\ &\quad -[\Sigma_{A_h}^{-1} + \Sigma_{A_h}^{-1} A_{21_h} \tilde{S}_h^{-1} A_{12_h} \Sigma_{A_h}^{-1}]_{ij}\Big\|_\infty \\ &= \left\|\Sigma_{A_i}^{-1} A_{21_i}\right\|_\infty \left\|[\hat{S}_h^{-1} \hat{A}_{12_h} \Sigma_{A_h}^{-1}]_{ij} - [\tilde{S}_h^{-1} A_{12_h} \Sigma_{A_h}^{-1}]_{ij}\right\|_\infty \\ &\leq \underbrace{C_{\|\Sigma_A^{-1}\|_\infty} C_{\|A_{12}\|_\infty} \bar{C}_{12s}}_{=: \bar{C}_{22s}} h^2\end{aligned}$$

for sufficiently small h.

□

Now we are prepared to prove the main result for uniquely solvable DAEs.

Theorem 40 *Let $\tilde{D}_h^u$ and D_h^u be the discretization matrices of the globally equivalent differential-algebraic equations (3.87) and (3.89) and let P and Q be the corresponding transformation functions. Let $\hat{D}_h^u = P_h D_h Q_h$ with P_h and Q_h defined as in (3.63). Then there exists a positive constant C^u, which does not depend on the step size h, such that*

$$\left\|\hat{D}_h^{u^{-1}} - \tilde{D}_h^{u^{-1}}\right\|_\infty \leq C^u h.$$

Proof. Using the permutations $P_{l,h}$ and $P_{r,h}$ as defined in (3.90) and (3.91), we get

$$\begin{aligned}\left\|\hat{D}_h^{u^{-1}} - \tilde{D}_h^{u^{-1}}\right\|_\infty &= \left\|P_{h,r}^{u^T} \hat{D}_h^{u^{-1}} P_{h,l}^{u^T} - P_{h,r}^{u^T} \tilde{D}_h^{u^{-1}} P_{h,l}^{u^T}\right\|_\infty \\ &= \left\|\left(P_{h,l}^u \hat{D}_h^u P_{h,r}^u\right)^{-1} - \left(P_{h,l}^u \tilde{D}_h^u P_{h,r}^u\right)^{-1}\right\|_\infty \\ &= \left\|\begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1} - \begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1}\right\|_\infty \\ &= \left\|\begin{bmatrix} \hat{D}_{h11}^{u^{inv}} & \hat{D}_{h12}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} & \hat{D}_{h22}^{u^{inv}} \end{bmatrix} - \begin{bmatrix} \tilde{D}_{h11}^{u^{inv}} & \tilde{D}_{h12}^{u^{inv}} \\ \tilde{D}_{h21}^{u^{inv}} & \tilde{D}_{h22}^{u^{inv}} \end{bmatrix}\right\|_\infty \\ &\leq \max\Big(\left\|\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right\|_\infty + \left\|\hat{D}_{h12}^{u^{inv}} - \tilde{D}_{h12}^{u^{inv}}\right\|_\infty, \\ &\qquad \left\|\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}}\right\|_\infty + \left\|\hat{D}_{h22}^{u^{inv}} - \tilde{D}_{h22}^{u^{inv}}\right\|_\infty\Big).\end{aligned}$$

By Lemma 39, we have

$$\begin{aligned}\left\|\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right\|_\infty &\leq \max_{i=1,\ldots,N} \sum_{j=1}^{N} \left\|[\hat{D}_{h11}^{u^{inv}}]_{ij} - [\tilde{D}_{h11}^{u^{inv}}]_{ij}\right\|_\infty \\ &\leq N\bar{C}_{11} h^2 = \bar{C}_{11} h(T - t_0).\end{aligned} \tag{3.111}$$

Analogously, we get

$$\left\|\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}}\right\|_\infty \leq \bar{C}_{21} h(T - t_0). \tag{3.112}$$

For the remaining blocks we achieve the same result by observing that

$$\begin{aligned}\left\|\hat{D}_{h12}^{u^{inv}} - \tilde{D}_{h12}^{u^{inv}}\right\|_\infty &\leq \max_{i=1,\dots,N} \sum_{j=1}^{N} \left\|[\hat{D}_{h12}^{u^{inv}}]_{ij} - [\tilde{D}_{h12}^{u^{inv}}]_{ij}\right\|_\infty \\ &= \max_{i=1,\dots,N} \left(\left\|[\hat{D}_{h12}^{u^{inv}}]_{ii} - [\tilde{D}_{h12}^{u^{inv}}]_{ii}\right\|_\infty + \sum_{j=1}^{i-1} \left\|[\hat{D}_{h12}^{u^{inv}}]_{ij} - [\tilde{D}_{h12}^{u^{inv}}]_{ij}\right\|_\infty \right) \\ &\leq \left(\bar{C}_{12d} h + N \bar{C}_{12s} h^2\right) = \left(\bar{C}_{12d} + (T - t_0)\bar{C}_{12s}\right) h\end{aligned}$$

and, analogously,

$$\left\|\hat{D}_{h22}^{u^{inv}} - D_{h22}^{u^{inv}}\right\|_\infty \leq (\bar{C}_{22d} + (T - t_0)\bar{C}_{22s})h.$$

In summary, we conclude

$$\begin{aligned}\left\|\hat{D}_h^{u^{-1}} - D_h^{u^{-1}}\right\|_\infty &\leq \max\left(\bar{C}_{12d} + (T - t_0)(\bar{C}_{11} + \bar{C}_{12s}), \bar{C}_{22d} + (T - t_0)(\bar{C}_{21} + \bar{C}_{22s})\right) h \\ &= C^u h.\end{aligned}$$

□

The general DAE case

The final step to prove the main result of the whole section is to show that the results we have obtained for regular DAEs can be extended to the least squares solution of over- and underdetermined DAEs.

For this purpose, we consider a general strangeness free differential-algebraic equation (1.1) with a homogeneous initial condition. The discretization of this system with the implicit Euler method and a fixed step size $h = (T - t_0)/N$ is written in terms of the linear system

$$D_h x_h = f_h,$$

with D_h defined as in (3.12) and

$$f_h = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix}.$$

Furthermore, consider a globally equivalent DAE in orthogonal standard form (2.1) and let P and Q be the corresponding transformation functions. Applying the same discretization to the orthogonal standard form system leads to a linear system

$$\tilde{D}_h \tilde{x}_h = \tilde{f}_h,$$

with $\tilde{D}_h$ defined as in (3.44) and $\tilde{f}_h$ defined analogously to f_h. From (3.65) and (3.66), we know that for

$$\hat{D}_h = P_h D_h Q_h, \tag{3.113}$$

where the orthogonal transformations P_h and Q_h are defined as in (3.63), we get

$$\hat{D}_h = \tilde{D}_h + \Delta_h.$$

The perturbation matrix Δ_h has the structure given in (3.68), i.e.,

$$\Delta_h = \begin{bmatrix} \Delta_1 & & & & \\ -\Delta_2 - \delta_2 & \Delta_2 & & & \\ & -\Delta_3 - \delta_3 & \Delta_3 & & \\ & & \ddots & \ddots & \\ & & & -\Delta_N - \delta_N & \Delta_N \end{bmatrix}$$

with

$$\Delta_i = -\tilde{E}_i Q_i^T \dot{Q}_i = \begin{bmatrix} \Delta_{11_i} & \Delta_{12_i} & \Delta_{13_i} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\delta_i = -\tilde{E}_i Q_i^T \left(\frac{Q_i - Q_{i-1}}{h} - \dot{Q}_i \right) = \begin{bmatrix} \delta_{11_i} & \delta_{12_i} & \delta_{13_i} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

for $i = 1, \ldots, N$. Thus, for a sufficiently smooth transformation function Q we obtain the inequalities

$$\|\Delta_{1j_i}\|_\infty \leq C_{\Delta_{1j}}$$

and

$$\|\delta_{1j_i}\|_\infty \leq C_{\delta_{1j}} h$$

with positive constants $C_{\Delta_{1j}}$ and $C_{\delta_{1j}}$ for $j = 1, 2, 3$. The corresponding right-hand sides satisfy $\hat{f}_h = P_h f_h = \tilde{f}_h$.

Our aim is to show that the difference between the Moore-Penrose pseudoinverses of $\hat{D}_h$ and $\tilde{D}_h$ can be bounded as

$$\left\| \hat{D}_h^+ - \tilde{D}_h^+ \right\|_\infty \leq Ch \tag{3.114}$$

for some positive constant C which does not depend on h.

We compute the pseudoinverses of $\tilde{D}_h$ and $\hat{D}_h$ using the formula (3.22) given in Lemma 22. We apply permutations $P_{h,l}$ and $P_{h,r}$ from the left and from the right, respectively, to the discretization matrices, such that we obtain the structures

$$P_{h,l} \hat{D}_h P_{h,r} = \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} & \hat{A}_{13_h} \\ A_{21_h} & \Sigma_{A_h} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$P_{h,l} \tilde{D}_h P_{h,r} = \begin{bmatrix} \tilde{D}_h^o & A_{12_h} & A_{13_h} \\ A_{21_h} & \Sigma_{A_h} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where the matrices $\tilde{D}_h^o$ and $\hat{D}_h^o$ are defined as in (3.72) and (3.77). The blocks A_{12_h}, $\hat{A}_{12_h}$, A_{21_h} and Σ_{A_h} are defined as in (3.92) and (3.93), while the matrices A_{13_h} and $\hat{A}_{13_h}$ are given by

$$A_{13_h} = \begin{bmatrix} -A_{13_1} & & \\ & \ddots & \\ & & -A_{13_N} \end{bmatrix} \tag{3.115}$$

and

$$\hat{A}_{13_h} = A_{13_h} + \Delta_{13_h},$$

with

$$\Delta_{13_h} = \begin{bmatrix} \Delta_{13_1} & & & \\ -\Delta_{13_2} - \delta_{13_2} & \Delta_{13_2} & & \\ & \ddots & \ddots & \\ & & -\Delta_{13_N} - \delta_{13_N} & \Delta_{13_N} \end{bmatrix}.$$

According to Lemma 22, the pseudoinverses of $\hat{D}_h$ and $\tilde{D}_h$ can be computed as follows. From the previous chapters we know that the matrices

$$R_h = \begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix} \tag{3.116}$$

as well as

$$\hat{R}_h = \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}$$

are nonsingular for sufficiently small step sizes h, and that the inverses of both matrices are uniformly bounded.

If we set

$$B_h = \begin{bmatrix} A_{13_h} \\ 0 \end{bmatrix}, \tag{3.117}$$

then the Moore-Penrose pseudoinverse of $P_{h,l}\tilde{D}_h P_{h,r}$ is given by

$$\begin{aligned} \left(P_{h,l}\tilde{D}_h P_{h,r}\right)^+ &= \begin{bmatrix} D_h^o & A_{12_h} & A_{13_h} \\ A_{21_h} & \Sigma_{A_h} & 0 \\ 0 & 0 & 0 \end{bmatrix}^+ \\ &= \begin{bmatrix} R_h & B_h \\ 0 & 0 \end{bmatrix}^+ \\ &= \begin{bmatrix} \left(I - V_h W_h^{-1} V_h^T\right) R_h^{-1} & 0 \\ W_h^{-1} V_h^T R_h^{-1} & 0 \end{bmatrix}, \end{aligned}$$

where

$$V_h = R_h^{-1} B_h \tag{3.118}$$

and

$$W_h = I + V_h^T V_h. \tag{3.119}$$

Analogously, we can compute the pseudoinverse of $P_{h,l}\hat{D}_h P_{h,r}$ as

$$\left(P_{h,l}\hat{D}_h P_{h,r}\right)^+ = \begin{bmatrix} \left(I - \hat{V}_h \hat{W}_h^{-1} \hat{V}_h^T\right) \hat{R}_h^{-1} & 0 \\ \hat{W}_h^{-1} \hat{V}_h^T \hat{R}_h^{-1} & 0 \end{bmatrix}$$

with

$$\hat{B}_h = \begin{bmatrix} \hat{A}_{13_h} \\ 0 \end{bmatrix}, \quad \hat{V}_h = \hat{R}_h^{-1} \hat{B}_h, \quad \hat{W}_h = I + \hat{V}_h^T \hat{V}_h. \tag{3.120}$$

These representations of the pseudoinverses will be convenient for proving the main result (3.114).

Before we can achieve this proof, we have to show the following property.

Lemma 41 *Let $W_h = I + V_h^T V_h$ with*

$$V_h = R_h^{-1} B_h,$$

where R_h and B_h are defined as in (3.116) and (3.117). Then the inverse of W_h is uniformly bounded, i.e., there exists a constant $C_{W^{-1}} \in \mathbb{R}^+$ such that

$$\left\| W_h^{-1} \right\|_\infty \leq C_{W^{-1}}.$$

Proof. See Appendix A. □

Theorem 42 *Let $\hat{D}_h$ and $\tilde{D}_h$ be defined as in (3.44) and (3.113). Then there exists a positive constant $\hat{C}$, which does not depend on the step size h, such that*

$$\left\| \hat{D}_h^+ - \tilde{D}_h^+ \right\|_\infty \leq \hat{C} h \tag{3.121}$$

for sufficiently small h.

Proof. First of all we can transform the matrices using the permutations $P_{h,l}$ and $P_{h,r}$, which leads to

$$\begin{aligned}
\left\| \hat{D}_h^+ - \tilde{D}_h^+ \right\|_\infty &= \left\| P_{h,r}^T \hat{D}_h^+ P_{h,l}^T - P_{h,r}^T \tilde{D}_h^+ P_{h,l}^T \right\|_\infty \\
&= \left\| \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} & \hat{A}_{13_h} \\ A_{21_h} & \Sigma_{A_h} & 0 \\ 0 & 0 & 0 \end{bmatrix}^+ - \begin{bmatrix} \tilde{D}_h^o & A_{12_h} & A_{13_h} \\ A_{21_h} & \Sigma_{A_h} & 0 \\ 0 & 0 & 0 \end{bmatrix}^+ \right\|_\infty \\
&= \left\| \begin{bmatrix} \left(I - \hat{V}_h \hat{W}_h^{-1} \hat{V}_h^T \right) \hat{R}_h^{-1} & 0 \\ \hat{W}_h^{-1} \hat{V}_h^T \hat{R}_h^{-1} & 0 \end{bmatrix} - \begin{bmatrix} \left(I - V_h W_h^{-1} V_h^T \right) R_h^{-1} & 0 \\ W_h^{-1} V_h^T R_h^{-1} & 0 \end{bmatrix} \right\|_\infty \\
&= \left\| \begin{bmatrix} \left(I - \hat{V}_h \hat{W}_h^{-1} \hat{V}_h^T \right) \hat{R}_h^{-1} - \left(I - V_h W_h^{-1} V_h^T \right) R_h^{-1} & 0 \\ \hat{W}_h^{-1} \hat{V}_h^T \hat{R}_h^{-1} - W_h^{-1} V_h^T R_h^{-1} & 0 \end{bmatrix} \right\|_\infty .
\end{aligned}$$

Here, the matrices R_h, V_h, W_h, $\hat{R}_h$, $\hat{V}_h$ and $\hat{W}_h$ are defined as in (3.116), (3.118), (3.119) and (3.120). The rest of this proof is concerned with estimating the norms of the nonzero blocks in $\hat{D}_h^+ - D_h^+$. By basic algebraic manipulations,

$$\begin{aligned}
&\left(I - \hat{V}_h \hat{W}_h^{-1} \hat{V}_h^T \right) \hat{R}_h^{-1} - \left(I - V_h W_h^{-1} V_h^T \right) R_h^{-1} \\
&\qquad = \left(I - V_h W_h^{-1} V_h^T \right) \left(\hat{R}_h^{-1} - R_h^{-1} \right) + V_h W_h^{-1} \left(V_h^T - \hat{V}_h^T \right) \hat{R}_h^{-1} \\
&\qquad\quad + \hat{V}_h \left(W_h^{-1} - \hat{W}_h^{-1} \right) \hat{V}_h^T \hat{R}_h^{-1} + \left(V_h - \hat{V}_h \right) W_h^{-1} \hat{V}_h^T \hat{R}_h^{-1}
\end{aligned} \tag{3.122}$$

and

$$\begin{aligned}
\hat{W}_h^{-1} \hat{V}_h^T \hat{R}_h^{-1} - W_h^{-1} V_h^T R_h^{-1} &= W_h^{-1} V_h^T \left(\hat{R}_h^{-1} - R_h^{-1} \right) + W_h^{-1} \left(\hat{V}_h^T - V_h^T \right) \hat{R}_h^{-1} \\
&\quad + \left(\hat{W}_h^{-1} - W_h^{-1} \right) \hat{V}_h^T \hat{R}_h^{-1}.
\end{aligned} \tag{3.123}$$

In order to show the bound (3.121), we have to examine the different parts of these sums. For $\hat{R}_h^{-1}$, we get from Lemma 38 that

$$\begin{aligned}
\left\|\hat{R}_h^{-1}\right\|_\infty &= \left\| \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1} \right\|_\infty \\
&= \left\| \begin{bmatrix} \hat{D}_{h11}^{u^{inv}} & \hat{D}_{h12}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} & \hat{D}_{h22}^{u^{inv}} \end{bmatrix} \right\|_\infty \\
&\leq \max\left(\left\|\hat{D}_{h11}^{u^{inv}}\right\|_\infty + \left\|\hat{D}_{h12}^{u^{inv}}\right\|_\infty, \left\|\hat{D}_{h21}^{u^{inv}}\right\|_\infty + \left\|\hat{D}_{h22}^{u^{inv}}\right\|_\infty \right) \\
&\leq \max\left(C_{11}^u + C_{12}^u, C_{21}^u + C_{22}^u\right) =: C_{\hat{R}^{-1}}.
\end{aligned}$$

The difference $\hat{R}_h^{-1} - R_h^{-1}$ can be estimated using Theorem 40 and (3.91), (3.90):

$$\begin{aligned}
\left\|\hat{R}_h^{-1} - R_h^{-1}\right\|_\infty &\leq \left\| P_{h,r}^T \hat{D}_h^{u^{-1}} P_{h,l}^T - P_{h,r}^T D_h^{u^{-1}} P_{h,l}^T \right\|_\infty \\
&\leq \left\| \hat{D}_h^{u^{-1}} - \tilde{D}_h^{u^{-1}} \right\|_\infty \\
&\leq C^u h.
\end{aligned}$$

The matrices $\hat{V}_h$ and V_h are given by

$$\begin{aligned}
\hat{V}_h &= \hat{R}_h^{-1} \hat{B}_h \\
&= \begin{bmatrix} \hat{D}_h^o & \hat{A}_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}^{-1} \begin{bmatrix} \hat{A}_{13_h} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \hat{D}_{h11}^{u^{inv}} & \hat{D}_{h12}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} & \hat{D}_{h22}^{u^{inv}} \end{bmatrix} \begin{bmatrix} \hat{A}_{13_h} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \hat{D}_{h11}^{u^{inv}} \\ \hat{D}_{h21}^{u^{inv}} \end{bmatrix} \hat{A}_{13_h},
\end{aligned}$$

and, analogously,

$$V_h = R_h^{-1} B_h = \begin{bmatrix} \tilde{D}_{h11}^{u^{inv}} & \tilde{D}_{h21}^{u^{inv}} \end{bmatrix} A_{13_h}.$$

Lemma 38 implies

$$\begin{aligned}
\left\|\hat{V}_h\right\|_\infty &\leq \max\left(\left\|\hat{D}_{h11}^{u^{inv}}\right\|_\infty, \left\|\hat{D}_{h21}^{u^{inv}}\right\|_\infty \right) \left\|\hat{A}_{13_h}\right\|_\infty \\
&\leq \max\left(C_{11}^u, C_{21}^u\right) \left(C_{\|A_{13}\|_\infty} + 2C_{\|\Delta_{13}\|_\infty} + C_{\|\delta_{13}\|_\infty} h \right) \\
&\leq \max\left(C_{11}^u, C_{21}^u\right) \left(C_{\|A_{13}\|_\infty} + 3C_{\|\Delta_{13}\|_\infty} \right) \\
&=: C_{\hat{V}}
\end{aligned}$$

for sufficiently small h, and analogously

$$\begin{aligned}
\|V_h\|_\infty &\leq \max\left(\left\|\tilde{D}_{h11}^{u^{inv}}\right\|_\infty, \left\|\tilde{D}_{h21}^{u^{inv}}\right\|_\infty \right) \|A_{13_h}\|_\infty \\
&\leq \max\left(C_{11}^u, C_{21}^u\right) C_{\|A_{13}\|_\infty} =: C_V.
\end{aligned}$$

For the transposed matrices $\hat{V}_h^T$ and V_h^T we get

$$\begin{aligned}
\left\|\hat{V}_h^T\right\|_\infty &\leq \left(\left\| \left(\hat{D}_{h11}^{u^{inv}}\right)^T \right\|_\infty + \left\| \left(\hat{D}_{h21}^{u^{inv}}\right)^T \right\|_\infty \right) \left\|\hat{A}_{13_h}^T\right\|_\infty \\
&\leq \left(C_{11}^u + C_{21}^u\right) \left(C_{\|A_{13}\|_\infty} + 3C_{\|\Delta_{13}\|_\infty} \right) =: C_{\hat{V}^T}
\end{aligned}$$

and

$$\begin{aligned}\left\|V_h^T\right\|_\infty &\leq \left(\left\|\left(\tilde{D}_{h11}^{u^{inv}}\right)^T\right\|_\infty + \left\|\left(\tilde{D}_{h21}^{u^{inv}}\right)^T\right\|_\infty\right)\left\|A_{13_h}^T\right\|_\infty \\ &\leq \left(C_{11}^u + C_{21}^u\right) C_{\|A_{13}\|_\infty} =: C_{V^T}\end{aligned}$$

for sufficiently small h. Here, we used the fact that the definition of $C_{\|\cdot\|_\infty}$, see Definition 23, is invariant under transposition, which implies that the constants C_{11}^u and C_{21}^u defined in Lemma 38 share the same property.

Considering the difference

$$\hat{V}_h - V_h = \begin{bmatrix} \hat{D}_{h11}^{u^{inv}}\hat{A}_{13_h} - D_{h11}^{u^{inv}}A_{13_h} \\ \hat{D}_{h21}^{u^{inv}}\hat{A}_{13_h} - D_{h21}^{u^{inv}}A_{13_h}\end{bmatrix},$$

we get for the first block row, using $\hat{D}_{h11}^{u^{inv}} = \hat{S}_h^{-1}$ and (3.111), that

$$\begin{aligned}\left\|\hat{D}_{h11}^{u^{inv}}\hat{A}_{13_h} - D_{h11}^{u^{inv}}A_{13_h}\right\|_\infty &= \left\|\left(\hat{D}_{h11}^{u^{inv}} - D_{h11}^{u^{inv}}\right)A_{13_h} + \hat{S}_h^{-1}\Delta_{13_h}\right\|_\infty \\ &\leq \bar{C}_{11}C_{\|A_{13}\|_\infty}h + \left\|\hat{S}_h^{-1}\Delta_{13_h}\right\|_\infty.\end{aligned}$$

To show that $\|\hat{S}_h^{-1}\Delta_{13_h}\|_\infty$ is small, we compute the matrix $\hat{S}_h^{-1}\Delta_{13_h}$ blockwise, using the structure of Δ_{13_h} and the fact that $\hat{S}_h^{-1}$ is a lower block triangular matrix:

$$[\hat{S}_h^{-1}\Delta_{13_h}]_{ij} = \begin{cases}[\hat{S}_h^{-1}]_{ii}\Delta_{13_i} & \text{for } i = j, \\ [\hat{S}_h^{-1}]_{ij}\Delta_{13_j} - [\hat{S}_h^{-1}]_{i,j+1}\left(\Delta_{13_{j+1}} + \delta_{13_{j+1}}\right) & \text{for } i > j, \\ 0 & \text{otherwise.}\end{cases}$$

The estimate $\|[\hat{S}_h]_{d,i}^{-1}\|_\infty \leq 2hC_{\Sigma_E^{-1}}$ (see (3.85)) implies the following bound for the diagonal blocks of $\hat{S}_h^{-1}\Delta_{13_h}$,

$$\begin{aligned}\left\|[\hat{S}_h^{-1}\Delta_{13_h}]_{ii}\right\|_\infty &= \left\|[\hat{S}_h^{-1}]_{ii}\Delta_{13_i}\right\|_\infty \\ &\leq \left\|[\hat{S}_h^{-1}]_{ii}\right\|_\infty C_{\|\Delta_{13}\|_\infty} \\ &\leq \left\|[\hat{S}_h]_{d,i}^{-1}\right\|_\infty C_{\|\Delta_{13}\|_\infty} \\ &\leq 2C_{\|\Sigma_E^{-1}\|_\infty}C_{\|\Delta_{13}\|_\infty}h.\end{aligned} \tag{3.124}$$

Applying Lemma 36 and Lemma 37 yields for $i > j$,

$$\begin{aligned}\left\|[\hat{S}_h^{-1}\Delta_{13_h}]_{ij}\right\|_\infty &= \left\|[\hat{S}_h^{-1}]_{ij}\Delta_{13_j} - [\hat{S}_h^{-1}]_{i,j+1}\left(\Delta_{13_{j+1}} + \delta_{13_{j+1}}\right)\right\|_\infty \\ &\leq \left\|[\hat{S}_h^{-1}]_{ij}\right\|_\infty \left\|\Delta_{13_j} - \Delta_{13_{j+1}}\right\|_\infty \\ &\quad + \left\|[\hat{S}_h^{-1}]_{ij} - [\hat{S}_h^{-1}]_{i,j+1}\right\|_\infty \left\|\Delta_{13_{j+1}}\right\|_\infty + \left\|[\hat{S}_h^{-1}]_{i,j+1}\delta_{13_{j+1}}\right\|_\infty \\ &\leq \left(C_2^u L_{\|\Delta_{13}\|_\infty} + C_5^u C_{\|\Delta_{13}\|_\infty} + C_2^u C_{\|\delta_{13}\|_\infty}\right)h^2.\end{aligned} \tag{3.125}$$

This leads to

$$\begin{aligned}\left\|\hat{S}_h^{-1}\Delta_{13_h}\right\|_\infty &\leq \max_{i=1,\dots,N}\sum_{j=1}^{N}\left\|[\hat{S}_h^{-1}\Delta_{13_h}]_{ij}\right\|_\infty \\ &\leq 2C_{\|\Sigma_E^{-1}\|_\infty}C_{\|\Delta_{13}\|_\infty}h + N\left(C_2^u L_{\|\Delta_{13}\|_\infty} + C_5^u C_{\|\Delta_{13}\|_\infty} + C_2^u C_{\|\delta_{13}\|_\infty}\right)h^2 \\ &= \left(2C_{\|\Sigma_E^{-1}\|_\infty}C_{\|\Delta_{13}\|_\infty} + (T-t_0)(C_2^u L_{\|\Delta_{13}\|_\infty} + C_5^u C_{\|\Delta_{13}\|_\infty} + C_2^u C_{\|\delta_{13}\|_\infty})\right)h \\ &=: C_{S^{-1}\Delta_{13}}h,\end{aligned}$$

and thus,

$$\left\|\hat{D}_{h11}^{u^{inv}}\hat{A}_{13_h} - \tilde{D}_{h11}^{u^{inv}}A_{13_h}\right\|_\infty \leq \left(\bar{C}_{11}C_{\|A_{13}\|_\infty} + C_{S^{-1}\Delta_{13}}\right)h.$$

Because of the identity $\hat{D}_{h21}^{u^{inv}} = -\Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1}$, we can estimate the second block of $\hat{V}_h - V_h$ analogously. Using (3.112), we achieve

$$\begin{aligned}\left\|\hat{D}_{h21}^{u^{inv}}\hat{A}_{13_h} - \tilde{D}_{h21}^{u^{inv}}A_{13_h}\right\|_\infty &= \left\|\left(\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}}\right)A_{13_h} - \Sigma_{A_h}^{-1}A_{21_h}\hat{S}_h^{-1}\Delta_{13_h}\right\|_\infty \\ &\leq \left(\bar{C}_{21}C_{\|A_{13}\|_\infty} + C_{\|\Sigma_A^{-1}A_{21}\|_\infty}C_{S^{-1}\Delta_{13}}\right)h.\end{aligned}$$

Combining both inequalities yields

$$\begin{aligned}\|\hat{V}_h - V_h\|_\infty &= \max\left(\left\|\hat{D}_{h11}^{u^{inv}}\hat{A}_{13_h} - \tilde{D}_{h11}^{u^{inv}}A_{13_h}\right\|_\infty, \left\|\hat{D}_{h21}^{u^{inv}}\hat{A}_{13_h} - \tilde{D}_{h21}^{u^{inv}}A_{13_h}\right\|_\infty\right) \\ &\leq \left(\bar{C}_{11}C_{\|A_{13}\|_\infty} + \max\left(1, C_{\|\Sigma_A^{-1}A_{21}\|_\infty}\right)C_{S^{-1}\Delta_{13}}\right)h \\ &=: C_{\Delta V}h.\end{aligned}$$

The difference

$$\begin{aligned}\hat{V}_h^T - V_h^T &= \left[\hat{A}_{13_h}^T\left(\hat{D}_{h11}^{u^{inv}}\right)^T - A_{13_h}^T\left(\tilde{D}_{h11}^{u^{inv}}\right)^T \quad \hat{A}_{13_h}^T\left(\hat{D}_{h21}^{u^{inv}}\right)^T - A_{13_h}^T\left(\tilde{D}_{h21}^{u^{inv}}\right)^T\right] \\ &= \left[A_{13_h}^T\left(\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right)^T + \Delta_{13_h}^T\left(\hat{D}_{h11}^{u^{inv}}\right)^T \quad A_{13_h}^T\left(\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}}\right)^T + \Delta_{13_h}^T\left(\hat{D}_{h21}^{u^{inv}}\right)^T\right]\end{aligned}$$

between the transposed matrices $\hat{V}_h^T$ and V_h^T can be estimated as follows. In Lemma 39 we have proved that

$$\left\|[\hat{D}_{h11}^{u^{inv}}]_{ij} - [\tilde{D}_{h11}^{u^{inv}}]_{ij}\right\|_\infty \leq \bar{C}_{11}h^2$$

for $i,j = 1,\dots,N$. From $\|A^T\|_\infty \leq m\|A\|_\infty$ for all $A \in \mathbb{R}^{m,n}$, it follows that

$$\begin{aligned}\left\|A_{13_h}^T\left(\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right)^T\right\|_\infty &= \max_i\sum_{j=1}^{N}\left\|\left[\left(\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right)A_{13_h}\right]_{ij}^T\right\|_\infty \\ &\leq \hat{d}\max_i\sum_{j=1}^{N}\left\|\left[\left(\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}}\right)A_{13_h}\right]_{ij}\right\|_\infty \\ &\leq \hat{d}\max_i\sum_{j=1}^{N}\left(\left\|[\hat{D}_{h11}^{u^{inv}}]_{ij} - [\tilde{D}_{h11}^{u^{inv}}]_{ij}\right\|_\infty \|A_{13_i}\|_\infty\right) \\ &\leq \hat{d}C_{\|A_{13}\|_\infty}\bar{C}_{11}h.\end{aligned}$$

Analogously, we get

$$\left\| A_{13_h}^T \left(\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}} \right)^T \right\|_\infty \leq \hat{d} C_{\|A_{13}\|_\infty} \bar{C}_{21} h.$$

From (3.124) and (3.125), we obtain for

$$\Delta_{13_h}^T \left(\hat{D}_{h11}^{u^{inv}} \right)^T = \Delta_{13_h}^T \hat{S}_h^{-T} = \left(\hat{S}_h^{-1} \Delta_{13_h} \right)^T$$

that

$$\left\| \left[\Delta_{13_h}^T \left(\hat{D}_{h11}^{u^{inv}} \right)^T \right]_{ij} \right\|_\infty = \left\| \left[\left(\hat{S}_h^{-1} \Delta_{13_h} \right)^T \right]_{ij} \right\|_\infty = \left\| \left[\hat{S}_h^{-1} \Delta_{13_h} \right]_{ji}^T \right\|_\infty \leq \hat{d} \left\| \left[\hat{S}_h^{-1} \Delta_{13_h} \right]_{ji} \right\|_\infty$$

for $i, j = 1, \ldots, N$. Consequently,

$$\left\| \Delta_{13_h}^T \left(\hat{D}_{h11}^{u^{inv}} \right)^T \right\|_\infty \leq \hat{d} \left\| \hat{S}_h^{-1} \Delta_{13_h} \right\|_\infty \leq \hat{d} C_{S^{-1}\Delta_{13}} h.$$

The identity

$$\Delta_{13_h}^T \left(\hat{D}_{h21}^{u^{inv}} \right)^T = \left(\Sigma_{A_h}^{-1} A_{21_h} \hat{S}_h^{-1} \Delta_{13_h} \right)^T$$

yields

$$\left\| \Delta_{13_h}^T \left(\hat{D}_{h21}^{u^{inv}} \right)^T \right\|_\infty \leq \hat{d} C_{\|\Sigma_A^{-1} A_{21}\|_\infty} C_{S^{-1}\Delta_{13}} h,$$

which implies

$$\begin{aligned} \left\| \hat{V}_h^T - V_h^T \right\|_\infty \leq & \left\| A_{13_h}^T \left(\hat{D}_{h11}^{u^{inv}} - \tilde{D}_{h11}^{u^{inv}} \right)^T \right\|_\infty + \left\| \Delta_{13_h}^T \left(\hat{D}_{h11}^{u^{inv}} \right)^T \right\|_\infty \\ & + \left\| A_{13_h}^T \left(\hat{D}_{h21}^{u^{inv}} - \tilde{D}_{h21}^{u^{inv}} \right)^T \right\|_\infty + \left\| \Delta_{13_h}^T \left(\hat{D}_{h21}^{u^{inv}} \right)^T \right\|_\infty \\ \leq & \hat{d} \left(C_{\|A_{13}\|_\infty} \left(\bar{C}_{11} + \bar{C}_{21} \right) + \left(1 + C_{\|\Sigma_A^{-1} A_{21}\|_\infty} \right) C_{S^{-1}\Delta_{13}} \right) h \\ =: & C_{\Delta V^T} h. \end{aligned}$$

Now we consider the matrices

$$\hat{W}_h = I + \hat{V}_h^T \hat{V}_h, \qquad W_h = I + V_h^T V_h.$$

From the observation

$$\begin{aligned} \left\| \hat{W}_h - W_h \right\|_\infty &= \left\| V_h^T V_h - \hat{V}_h^T \hat{V}_h \right\|_\infty \\ &\leq \left\| V_h^T - \hat{V}_h^T \right\|_\infty \|V_h\|_\infty + \left\| \hat{V}_h^T \right\|_\infty \left\| V_h - \hat{V}_h \right\|_\infty, \end{aligned}$$

we conclude

$$\|\hat{W}_h - W_h\|_\infty \leq \left(C_{\Delta V^T} C_V + C_{\Delta V} C_{V^T} \right) h =: C_{\Delta W} h.$$

Using Neumann series, we obtain for $h < (2C_{W^{-1}}C_{\Delta W})^{-1}$ that

$$\begin{aligned}
\left\|\hat{W}_h^{-1} - W_h^{-1}\right\|_\infty &= \left\|\left(\hat{W}_h - W_h + W_h\right)^{-1} - W_h^{-1}\right\|_\infty \\
&= \left\|\left(W_h\left(I - W_h^{-1}\left(W_h - \hat{W}_h\right)\right)\right)^{-1} - W_h^{-1}\right\|_\infty \\
&= \left\|\left(I + \sum_{i=1}^{\infty}\left(W_h^{-1}\left(W_h - \hat{W}_h\right)\right)^i\right)W_h^{-1} - W_h^{-1}\right\|_\infty \\
&\le \left\|\sum_{i=1}^{\infty}\left(W_h^{-1}\left(W_h - \hat{W}_h\right)\right)^i\right\|_\infty \|W_h^{-1}\|_\infty \\
&\le \|W_h^{-1}\|_\infty \sum_{i=1}^{\infty}\left(\|W_h^{-1}\|_\infty \left\|W_h - \hat{W}_h\right\|_\infty\right)^i \\
&\le C_{W^{-1}}\sum_{i=1}^{\infty}(C_{W^{-1}}C_{\Delta W}h)^i \\
&\le C_{W^{-1}}^2 C_{\Delta W}h\sum_{i=0}^{\infty}(C_{W^{-1}}C_{\Delta W}h)^i \\
&\le \frac{C_{W^{-1}}^2 C_{\Delta W}h}{1 - C_{W^{-1}}C_{\Delta W}h} \\
&\le 2C_{W^{-1}}^2 C_{\Delta W}h =: C_{\Delta W^{-1}}h.
\end{aligned}$$

The above considerations are combined to show the assertion (3.121). For the upper nonzero block of $\hat{D}_h^+ - D_h^+$ we get, according to (3.122), that

$$\begin{aligned}
&\left\|\left(I - \hat{V}_h\hat{W}_h^{-1}\hat{V}_h^T\right)\hat{R}_h^{-1} - \left(I - V_hW_h^{-1}V_h^T\right)R_h^{-1}\right\|_\infty \\
&\quad\le \left\|\left(I - V_hW_h^{-1}V_h^T\right)\left(\hat{R}_h^{-1} - R_h^{-1}\right)\right\|_\infty + \left\|V_hW_h^{-1}\left(V_h^T - \hat{V}_h^T\right)\hat{R}_h^{-1}\right\|_\infty \\
&\qquad + \left\|\hat{V}_h\left(W_h^{-1} - \hat{W}_h^{-1}\right)\hat{V}_h^T\hat{R}_h^{-1}\right\|_\infty + \left\|\left(V_h - \hat{V}_h\right)W_h^{-1}\hat{V}_h^T\hat{R}_h^{-1}\right\|_\infty \\
&\quad\le \left(1 + C_VC_{W^{-1}}C_{V^T}\right)C_{\Delta R}h + C_VC_{W^{-1}}C_{\Delta V}C_{\hat{R}^{-1}}h \\
&\qquad + C_{\hat{V}}C_{\Delta W^{-1}}C_{\hat{V}^T}C_{\hat{R}^{-1}}h + C_{\Delta V}C_{W^{-1}}C_{\hat{V}^T}C_{\hat{R}^{-1}}h \\
&\quad=: C_1h
\end{aligned}$$

and for the lower block according to (3.123)

$$\begin{aligned}
\left\|\hat{W}_h^{-1}\hat{V}_h^T\hat{R}_h^{-1} - W_h^{-1}V_h^TR_h^{-1}\right\|_\infty &\le \left\|W_h^{-1}V_h^T\left(\hat{R}_h^{-1} - R_h^{-1}\right)\right\|_\infty + \left\|W_h^{-1}\left(\hat{V}_h^T - V_h^T\right)\hat{R}_h^{-1}\right\|_\infty \\
&\quad + \left\|\left(\hat{W}_h^{-1} - W_h^{-1}\right)\hat{V}_h^T\hat{R}_h^{-1}\right\|_\infty \\
&\le C_{W^{-1}}C_{V^T}C_{\Delta R}h + C_{W^{-1}}C_{\Delta V}C_{\hat{R}^{-1}}h + C_{\Delta W^{-1}}C_{\hat{V}^T}C_{\hat{R}^{-1}}h \\
&=: C_2h
\end{aligned}$$

and therefore, finally, (3.121) holds with $\hat{C} := \max(C_1, C_2)$. □

3.3.3 The main theorem

Let us recall the objective of this section. Consider a strangeness free linear differential-algebraic equation of the form (1.1) with a homogeneous initial condition $x(t_0) = 0$, and a linear system

$$D_hx_h = f_h,$$

which represents the discretization of (1.1) with the implicit Euler method. Then, our aim is to show that the vector

$$x_h = D_h^+ f_h,$$

where D_h^+ is the Moore-Penrose pseudoinverse of D_h, contains approximations to the Moore-Penrose solution

$$x = D^+ f$$

of (1.1). Here, D^+ denotes the Moore-Penrose pseudoinverse of the differential-algebraic operator D, as defined in Section 2.2.2.

In Theorem 31, it has been shown that this approximation property is indeed true for systems that are given in orthogonal standard form. Theorem 42 can be used to generalize this result to general linear strangeness free differential algebraic equations.

Theorem 43 *Consider the linear strangeness free differential algebraic equation (1.1), where E and A are sufficiently smooth matrix functions. Let $x = D^+ f$ be the (unique) solution of the minimization problem (2.14) and let $x_h = D_h^+ f_h$ be the solution of the minimization problem (3.39), where the system $D_h x_h = f_h$ represents the discretization of (1.1) with the implicit Euler method using a fixed step size $h = (T - t_0)/N$.*

Then there exists a positive constant C, which does not depend on the step size h, such that

$$\|x_h - R_{\mathbb{X}_h} x\|_\infty \leq Ch \tag{3.126}$$

holds for sufficiently small h, where $R_{\mathbb{X}_h}$ is defined as in (3.4).

Proof. By Theorem 11, there exist pointwise orthogonal matrix functions P and Q such that we can transform (1.1) to the orthogonal standard form (2.1). Let the system $\tilde{D}_h \tilde{x}_h = \tilde{f}_h$ represent the discretization of (2.1) with the implicit Euler method. Furthermore, let P_h and Q_h be defined as in (3.63). Then by Theorem 31 and Theorem 42, setting $\hat{D}_h = P_h D_h Q_h$, we get

$$\begin{aligned}
\|x_h - R_{\mathbb{X}_h} x\|_\infty &\leq \|Q_h\|_\infty \left\|Q_h^T x_h - Q_h^T R_{\mathbb{X}_h} x\right\|_\infty \\
&\leq C_{\|Q\|_\infty} \left\|Q_h^T D_h^+ f_h - R_{\mathbb{X}_h} Q^T x\right\|_\infty \\
&= C_{\|Q\|_\infty} \left\|Q_h^T D_h^+ P_h^T P_h f_h - R_{\mathbb{X}_h} \tilde{x}\right\|_\infty \\
&= C_{\|Q\|_\infty} \left\|\hat{D}_h^+ \tilde{f}_h - R_{\mathbb{X}_h} \tilde{x}\right\|_\infty \\
&\leq C_{\|Q\|_\infty} \left(\left\|\left(\hat{D}_h^+ - \tilde{D}_h^+\right) \tilde{f}_h\right\|_\infty + \left\|\tilde{D}_h^+ \tilde{f}_h - R_{\mathbb{X}_h} \tilde{x}\right\|_\infty\right) \\
&\leq C_{\|Q\|_\infty} \left(\left\|\hat{D}_h^+ - \tilde{D}_h^+\right\|_\infty \left\|\tilde{f}_h\right\|_\infty + \left\|\tilde{D}_h^+ \tilde{f}_h - R_{\mathbb{X}_h} \tilde{x}\right\|_\infty\right) \\
&\leq C_{\|Q\|_\infty} \left(\hat{C} C_{\|\tilde{f}\|_\infty} + \tilde{C}\right) h.
\end{aligned}$$

This concludes the proof by setting $C = C_{\|Q\|_\infty}(\hat{C} C_{\|\tilde{f}\|_\infty} + \tilde{C})$. □

With this theorem, we have finally shown that global minimization yields a viable approach for computing $O(h)$ approximations to the least squares solution of a general linear, strangeness free DAE.

Chapter 4

Numerical Computations

The theoretical results of Chapter 3 have been turned into practicable algorithms and software. The purpose of this chapter is to describe these developments and to provide numerous numerical experiments substantiating the theoretical results.

4.1 Algorithms

The numerical methods considered in Chapter 3 require the DAE (1.1) to be strangeness free. To satisfy this requirement, a general DAE with a well-defined strangeness index is transformed into an equivalent strangeness free system having the same solution set, by the methods described in Section 1.2. Reliable algorithms have been derived in [27, 30, 33] and are part of the software package GELDA [33]. Our software is based on slightly modified versions of the corresponding routines in GELDA, see Chapter B.

In the following, we may therefore assume w.l.o.g. that the original DAE (1.1) is already strangeness free. By neglecting the trivial third block row in the strangeness free form of (1.1), we may furthermore assume that $\frac{1}{h}E(t) - A(t)$ is of full row rank for all sufficiently small $h > 0$ and all $t \in [t_0, T]$.

4.1.1 Local minimization

Let E_i, A_i and f_i denote the values of the coefficients E, A and f of the DAE (1.1) at the grid points $t_0 + ih$ for $i = 0, \ldots, N$ with $h = (T - t_0)/N$. Then local minimization amounts to the following algorithm, based on formula (3.14).

Algorithm 1 (local minimization)

Input: *Matrices $E_1, \ldots, E_N \in \mathbb{R}^{m\times n}$ and $A_1, \ldots, A_N \in \mathbb{R}^{m\times n}$ with $m \geq n$, vectors $f_1, \ldots, f_N \in \mathbb{R}^m$, a scalar $h > 0$. Starting values $x_0, \ldots, x_{k-1} \in \mathbb{R}^n$ and the parameters $\alpha_1, \ldots, \alpha_k$ of a k-step BDF-method.*

Output: *Vectors $x_k, x_{k+1}, \ldots, x_N$ approximating a generalized (1,2,3)-solution of the DAE (1.1).*

`FOR` $i = k, k+1, \ldots, N$
  Set $r = -\frac{1}{h}E_i\left(\alpha_0 x_{i-k} + \cdots \alpha_{k-1}x_{i-1}\right) + f_i$.
  Compute an LQ decomposition $(\frac{1}{h}E_i - A_i) = [L, 0]Q$.
  Compute $x_i = Q^T \begin{bmatrix} L^{-1}r \\ 0 \end{bmatrix}$.
`END`

Note that $B = [L, 0]Q$ is an LQ decomposition of B if and only if $B^T = Q^T[L, 0]^T$ is a QR decomposition of B^T. Algorithm 1 has favourably low computational requirements. The dominating computational part is the computation of the LQ decompositions, which require an overall amount of $O(Nm^2n)$ floating point operations (flops) [15]. The minimal memory requirement is $O(mn)$, provided that intermediate data is not longer than necessary stored.

Remark 44 *Motivated by the discussion in Remark 27, the final implementation of Algorithm 1 only covers the case $k = 1$ (implicit Euler). In this case, only the starting value x_0 needs to be provided; we always set $x_0 = 0$. Even if this starting value is not a consistent initial value for the DAE, it can be seen from (3.23) that only the "differential part" of x_0 is referenced and thus the first step of Algorithm 1 automatically forces x_1 to be (approximately) consistent. There is no need to compute consistent initial values explicitly.*

4.1.2 Global minimization

Using the same notation as in the previous section, global optimization amounts to the following algorithm, based on formula (3.40).

Algorithm 2 (global minimization)

Input: *Matrices $E_1, \ldots, E_N \in \mathbb{R}^{m\times n}$ and $A_1, \ldots, A_N \in \mathbb{R}^{m\times n}$ with $m \geq n$, vectors $f_1, \ldots, f_N \in \mathbb{R}^m$, a scalar $h > 0$.*

Output: *Vectors $x_1, x_2, \ldots, x_N$ approximating the least squares solution of the DAE (1.1) with homogeneous initial conditions.*

`FOR` $i = k, k+1, \ldots, N$
  Set $f_h = [f_1^T, f_2^T, \ldots, f_N^T]^T$.
  Compute an LQ decomposition $D_h = [L, 0]Q$ with D_h defined as in (3.12).
  Set $[x_1^T, x_2^T, \ldots, x_N^T]^T = Q^T \begin{bmatrix} L^{-1}f_h \\ 0 \end{bmatrix}$.
`END`

The computational cost of Algorithm 2 is dominated by determining the LQ decomposition of the $Nm \times Nn$ matrix D_h, which requires $O(N^3m^2n)$ flops using standard algorithms. Also the memory requirement of $O(N^2mn)$ is rather high.

Both expenses can be considerably reduced by using an algorithm that takes the special structure of D_h into account. For this purpose, let us recall the structure of the matrix D_h:

$$D_h = \frac{1}{h}\begin{bmatrix} E_1 - hA_1 & & & & \\ -E_2 & E_2 - hA_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -E_N & E_N - hA_N \end{bmatrix}, \tag{4.1}$$

where $E_i, A_i \in \mathbb{R}^{m\times n}$, $m \leq n$. In the following, the proposed algorithm is illustrated for $N = 3$:

$$D_h = \frac{1}{h}\begin{bmatrix} E_1 - hA_1 & 0 & 0 \\ -E_2 & E_2 - hA_2 & 0 \\ 0 & -E_3 & E_3 - hA_3 \end{bmatrix}.$$

First, we compute LQ decompositions $(E_3 - hA_3) = [L_3, 0]Q_{11}$ and $E_3 = [F_3, 0]Q_{12}$, where Q_{11}, Q_{12} are orthogonal and L_3, F are lower triangular matrices. Next, we compute an RQ decomposition $[F_3, L_3] = [0, R_3]Q_{13}$, where Q_{13} is orthogonal and R_3 is an upper triangular

matrix. Applying the orthogonal transformations to the corresponding block columns of D_h yields the updated matrix

$$D_h \leftarrow \frac{1}{h}\begin{bmatrix} E_1 - hA_1 & 0 & 0 & 0 \\ -E_2 & T_3 & U_3 & 0 \\ 0 & 0 & R_3 & 0 \end{bmatrix}.$$

An analogous step is applied to $-E_2$ and T_3, resulting in

$$D_h \leftarrow \frac{1}{h}\begin{bmatrix} T_1 & U_2 & 0 & 0 & 0 \\ 0 & R_2 & 0 & U_3 & 0 \\ 0 & 0 & 0 & R_3 & 0 \end{bmatrix}.$$

Finally, an RQ decomposition is used to reduce T_1 to upper triangular form:

$$D_h \leftarrow \frac{1}{h}\begin{bmatrix} R_1 & 0 & U_2 & 0 & 0 & 0 \\ 0 & 0 & R_2 & 0 & U_3 & 0 \\ 0 & 0 & 0 & 0 & R_3 & 0 \end{bmatrix}.$$

This is essentially an upper triangular matrix (leaving out zero block columns) and the solution of $\min\{\|x_h\|_2 : \ D_h x_h = f_h\}$ can be obtained by backward substitution. Note that this backward substitution can be combined with the reduction process; there is no need for saving all matrices R_i and U_i. The transformation matrices Q_{ij}, however, must be saved and applied afterwards to update the obtained solution x.

For general N, the algorithm reads as follows.

Algorithm 3

Input and Output: *See Algorithm 2.*

Set $T = E_N - hA_N$.
FOR $i = N, N-1, \ldots, 2$
  Compute an LQ decomposition $T = [L, 0]Q_{i1}$.
  Compute an LQ decomposition $-E_i = [F, 0]Q_{i2}$.
  Update $T = (E_{i-1} - hA_{i-1})Q_{i2}^T$.
  Compute an RQ decomposition $[F, L] = [0, R]Q_{i3}$.
  Update $[T(:, 1:m), U] = [T(:, 1:m), 0_{m\times m}]Q_{i3}^T$.
  Compute $x_i = \begin{bmatrix} R^{-1}f_i \\ 0 \end{bmatrix}$.
  Update $f_{i-1} = f_{i-1} - UR^{-1}f_i$.
END FOR
Compute RQ decomposition $T = [R, 0]Q_{11}$.
Compute $x_1 = Q_{11}^T \begin{bmatrix} R^{-1}f_1 \\ 0 \end{bmatrix}$.
FOR $i = 2, 3, \ldots, N$
  Update $\begin{bmatrix} x_{i-1}(1{:}m) \\ x_i(1{:}m)) \end{bmatrix} = Q_{i3}^T \begin{bmatrix} x_{i-1}(1{:}m) \\ x_i(1{:}m)) \end{bmatrix}$.
  Update $x_{i-1} \leftarrow Q_{i2}^T x_{i-1}$.
  Update $x_i \leftarrow Q_{i1}^T x_i$.
END FOR
$x \leftarrow hx$

Note that we used the colon notation $x(i:j)$ to denote the elements $i, i+1, \ldots, j$ of a vector x. Using compact LR and RQ decompositions, as implemented in LAPACK [1], each of the individual steps of Algorithm 3 requires at most $O(mn^2)$ flops, which yields an overall cost of $O(Nm^2n)$. This compares favourably with the cost of Algorithm 2 ($O(N^3m^2n)$ flops). There is

$O(Nmn)$ extra memory necessary to store information about the transformation matrices in the course of Algorithm 3. This is still favourable compared with Algorithm 2 ($O(N^2m^2n)$ memory) but significantly higher than the minimal memory requirements of Algorithm 1 ($O(m^2n)$ memory). The fact that Algorithm 3 implicitly solves a boundary value problem makes it questionable whether its memory requirements can be further reduced.

Let us emphasize that Algorithm 3 represents a (structure-exploiting) RQ decomposition of D_h combined with backward substitution using the upper triangular factor and matrix-vector multiplication using the orthogonal factor. The main differences to Algorithm 2 are that an RQ instead of an LQ decomposition is computed and that the backward substitution is carried out as soon as the corresponding blocks in the upper triangular matrix become available. None of these changes affects the numerical stability properties of Algorithm 3. In particular, as a consequence of the fact that the "big" RQ decomposition is computed from numerically backward stable LQ and RQ decompositions, the computed factors $\hat{R}$ and $\hat{Q}$ of the RQ decomposition of D_h satisfy

$$D_h + \triangle D_h = [\hat{R}, 0]\hat{Q}, \quad \|\triangle D_h\|_2 \leq c_D \mathbf{u} \|D_h\|_2, \quad \|\hat{Q}\hat{Q}^T - I\|_2 \leq c_Q \mathbf{u},$$

where c_D, c_Q are constants only depending on the dimension, and $\mathbf{u}$ denotes the unit roundoff, see [23]. Provided that the subsystems $R^{-1}f_i$ in Algorithm 3 are solved in a backward stable manner, the whole process of backward substitution is also numerically backward stable. Thus, Algorithm 3 can be expected to have the same numerical behaviour as Algorithm 2.

Remark 45 *Note that both Algorithm 2 and Algorithm 3 need not be initialized with a (consistent) starting value. This follows from the fact that both algorithms implicitly solve the boundary value problem (3.52), which forces the "differential part" of x_0 to be zero and the "algebraic part" of the solution to be (approximately) consistent.*

4.2 Numerical experiments

If not otherwise stated, the numerical experiments described in this section were performed using the Fortran routines listed in Appendix B. We used the Compaq Visual Fortran environment (along with the included precompiled BLAS and LAPACK libraries) on a Pentium IV 2.4 GHz processor with 512 MByte RAM to compile and execute these routines.

The following academical test example has been used to perform some of the numerical tests presented here.

The DAE

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{\tilde{x}}(t) = \begin{bmatrix} 1 - \frac{t}{2} & \frac{t}{2} & 1 \\ -1 & 1 & 0 \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} t\left(\frac{t}{2} + e^t\right) \\ t - 2(1 - e^t) \end{bmatrix}, \quad t \in [0, 1]. \tag{4.2}$$

is a strangeness free system and in orthogonal normal form (2.1), where the entries of the matrix functions $\tilde{E}$ and $\tilde{A}$, according to the block structure in (2.2), are given by

$$\begin{aligned} \Sigma_E(t) = 1, \;\; A_{11}(t) = 1 - \tfrac{t}{2}, \;\; A_{12}(t) = \tfrac{t}{2}, \;\; A_{13}(t) = 1, \\ A_{21}(t) = -1, \qquad \Sigma_A(t) = 1, \end{aligned} \tag{4.3}$$

and the inhomogeneity $\tilde{f}$ has the components

$$f_1(t) = t\left(\frac{t}{2} + e^t\right), \quad f_2(t) = t - 2\left(1 - e^t\right). \tag{4.4}$$

The (1,2,3)-generalized solution of (4.2) as defined in Section 2.3 can be obtained by setting the undetermined part x_3 of $\tilde{x}$ to zero and computing the remaining solution components by solving the reduced system

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 1 - \frac{t}{2} & \frac{t}{2} \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} t\left(\frac{t}{2} + e^t\right) \\ t - 2(1 - e^t) \end{bmatrix}, \quad t \in [0, 1]. \tag{4.5}$$

The (1,2,3)-generalized solution of (4.2) is therefore given by

$$\tilde{x}_{\mathrm{ge}}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} e^t - t - 1 \\ 1 - 2t - e^t \\ 0 \end{bmatrix}. \tag{4.6}$$

To compute the least squares solution of (4.2) as defined in Section 2.2, we consider the boundary value problem (2.23), where the coefficient functions can be computed from (4.3) and (4.4) by the substitutions (2.21), i.e.,

$$\begin{aligned} \dot{\lambda}(t) &= 2x(t) - \lambda(t) + 2(1 - e^t), & \lambda(1) &= 0, \\ \dot{x}(t) &= x(t) + u(t) + t, & x(0) &= 0, \\ y(t) &= x(t) + 2(1 - e^t) - t, & & \\ u(t) &= \lambda(t). & & \end{aligned} \tag{4.7}$$

This system has the unique solution

$$\lambda(t) = u(t) = 1 - t, \quad x(t) = e^t - 1, \quad y(t) = 1 - t - e^t, \tag{4.8}$$

and thus, the least square solution of (4.2) is given by

$$\tilde{x}_{\mathrm{ls}}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ y(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} e^t - 1 \\ 1 - t - e^t \\ 1 - t \end{bmatrix}. \tag{4.9}$$

In order to obtain a DAE system which is *not* in orthogonal standard form we have defined a time-variant, smooth orthogonal transformation $Q \in C^1(\mathbb{I}, \mathbb{R}^{3,3})$, based on the idea of Householder transformations (see, e.g., [15]). Given a vector function $v \in C^1(\mathbb{I}, \mathbb{R}^3)$, $v(t) \neq 0 \in \mathbb{I}$, and its derivative $\dot{v}$, the matrix function $Q = I - 2\frac{vv^T}{v^T v}$ is orthogonal and its derivative can be computed by

$$\dot{Q} = 2\left(\frac{vv^T \frac{d}{dt}(v^T v)}{(v^T v)^2} - \frac{\frac{d}{dt}(vv^T)}{v^T v}\right), \text{with} \quad \frac{d}{dt}(v^T v) = 2\dot{v}^T v, \quad \frac{d}{dt}(vv^T) = \dot{v}v^T + (\dot{v}v^T)^T.$$

Throughout the tests presented here, we have chosen

$$v(t) = \begin{bmatrix} t + 2 \\ t^2 + t + 1 \\ 1 \end{bmatrix}.$$

Except otherwise stated, the following numerical tests have been carried out with the DAE system

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \tag{4.10}$$

where

$$E(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T,$$
$$A(t) = \begin{bmatrix} 1 - \frac{t}{2} & \frac{t}{2} & 1 \\ -1 & 1 & 0 \end{bmatrix} Q^T - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{Q}^T,$$
$$f(t) = \begin{bmatrix} t\left(\frac{t}{2} + e^t\right) \\ t - 2(1 - e^t) \end{bmatrix}.$$

The (1,2,3)-generalized solution of 4.10 can be obtained by applying the transformation Q to the corresponding solution of the DAE 4.2, i. e.,

$$x_{\mathrm{ge}}(t) = Q\tilde{x}_{\mathrm{ge}}(t) = Q \begin{bmatrix} e^t - t - 1 \\ 1 - 2t - e^t \\ 0 \end{bmatrix}. \tag{4.11}$$

Analogously we can compute the least square solution of (4.10) as

$$x_{\mathrm{ls}}(t) = Q\tilde{x}_{\mathrm{ls}}(t) = Q \begin{bmatrix} e^t - 1 \\ 1 - t - e^t \\ 1 - t \end{bmatrix}. \tag{4.12}$$

The L_2 norms of $x_{\mathrm{ge}}(t)$ and $x_{\mathrm{ls}}(t)$ are given by

$$\|x_{\mathrm{ge}}\| = \|\tilde{x}_{\mathrm{ge}}\| = \sqrt{e^2 - 4e + \frac{23}{3}}, \quad \|x_{\mathrm{ls}}\| = \|\tilde{x}_{\mathrm{ls}}\| = \sqrt{e^2 - 4e + \frac{20}{3}},$$

where e denotes exp(1).

4.2.1 Higher order BDF-methods for local minimization

A separate implementation of Algorithm 1 using higher order BDF-methods has been used to compare the order of convergence of k-step BDF-methods for $k = 1, 2, 3, 4$. The algorithm has been applied to the test example 4.10 where the initial values $x_0, ..., x_{k-1}$ were given by evaluating the exact (1,2,3)-generalized solution as given in 4.11.

The error curves are shown in Figure 4.1. The maximum norm of the error is contained in the following table:

h	$k = 1$	$k = 2$	$k = 3$	$k = 4$
0.1	2.87×10^{-1}	1.61×10^{-1}	1.34×10^{-1}	1.09×10^{-1}
0.01	2.76×10^{-2}	1.62×10^{-2}	1.33×10^{-2}	1.17×10^{-2}
0.001	2.73×10^{-3}	1.62×10^{-3}	1.33×10^{-3}	1.17×10^{-3}
0.0001	2.73×10^{-4}	1.62×10^{-4}	1.33×10^{-4}	1.17×10^{-4}

It can be seen that the implicit Euler method ($k = 1$) displays approximate linear convergence. Using higher order BDF-methods leads to a slightly decreased error, apparently because of the *exact* initial values, but does not lead to higher order of convergence.

4.2.2 Convergence of global minimization

Algorithm 3 has been applied to (4.10) and the results have been compared with the analytic solution (4.12). The error curves are shown in Figure 4.2 and the maximum norm of the error is shown in the following table:

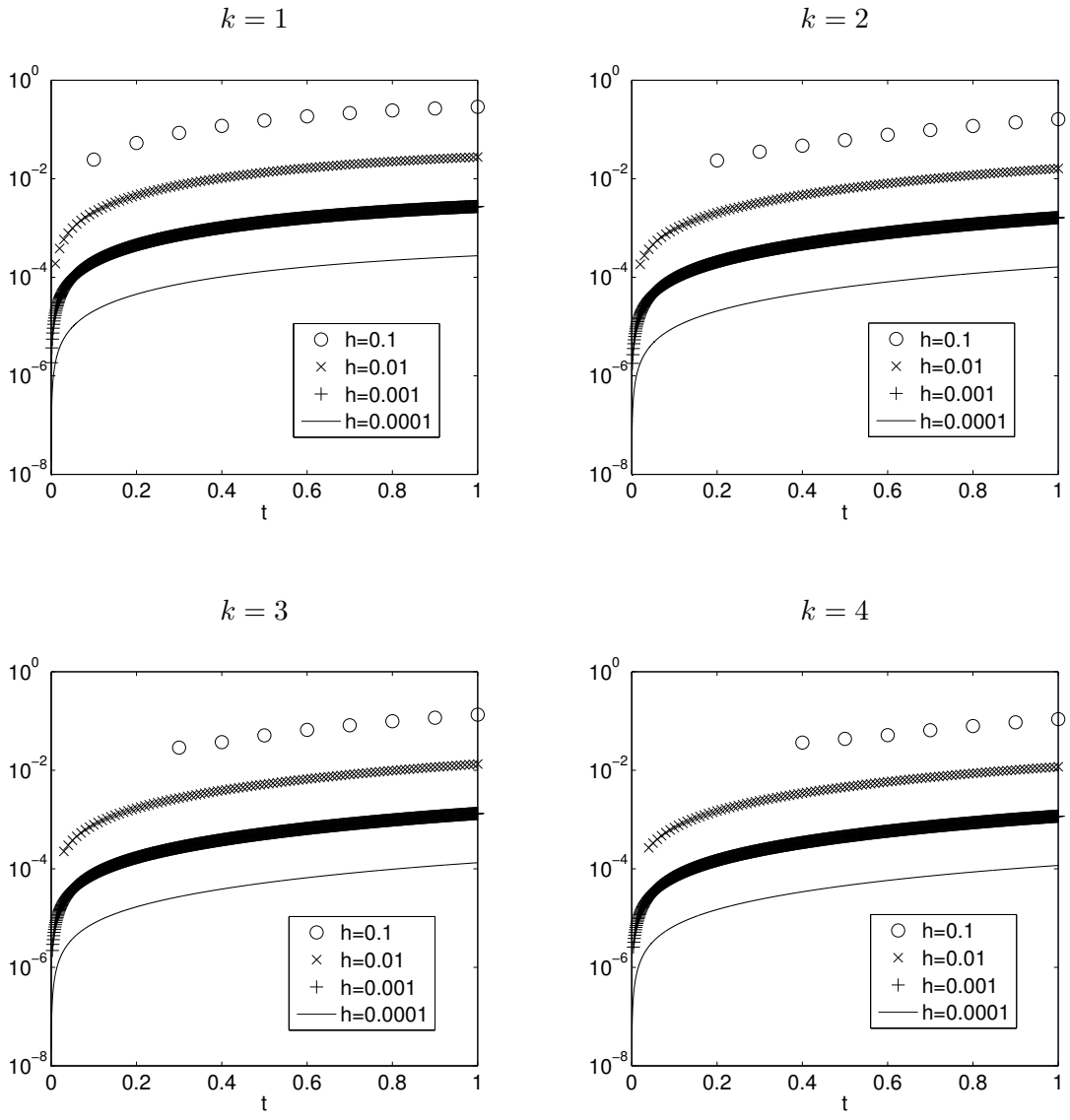

Figure 4.1: Error curves between exact (1,2,3)-generalized solution and approximated solution by local minimization with $h \in \{0.1, 0.01, 0.001, 0.0001\}$ and $k \in \{1, \ldots, 4\}$ for Example (4.10).

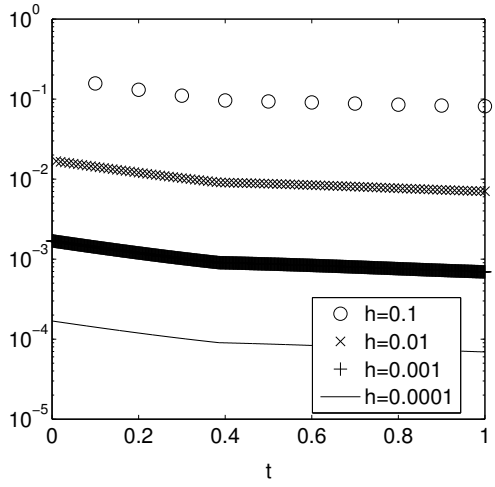

Figure 4.2: Error curves between exact least squares solution and approximated solution by global minimization with $h \in \{0.1, 0.01, 0.001, 0.0001\}$ for Example (4.10).

h	error
0.1	1.57×10^{-1}
0.01	1.67×10^{-2}
0.001	1.69×10^{-3}
0.0001	1.69×10^{-4}

4.2.3 Comparison with GELDA

The software package GELDA [33], which solves linear DAEs, also allows for the application to underdetermined DAEs. As in Algorithm 1, the linear systems arising during the discretization are solved (locally) in the least squares sense. In addition, an order and stepsize control is implemented in GELDA. We have applied Algorithm 1 with a step size $h = 0.001$ to the test example in orthogonal standard form (4.2) and compared this result to the result obtained by GELDA, where the absolute and relative error tolerances have been set to $ATOL = RTOL = 0.0001$. The corresponding results for all three solution components are shown in Figure 4.3.

One can see that the approximation computed by Algorithm 1 is closer to the analytical (1,2,3) generalized solution (4.6) in the sense that the undetermined solution component computed by Algorithm 1 is smaller and, hence, closer to the exact value $x_3 \equiv 0$, despite the relatively tight error tolerances used in GELDA. The maximum value of the undetermined solution component amounts to 2.996×10^{-3} using Algorithm 1 and to 2.735×10^{-2} using GELDA. This is due to the fact that the stepsize control implemented in GELDA does not consider *which* solution is to be approximated and thus, the stepsize is increased up to $h = 1.4 \times 10^{-2}$ by GELDA.

GELDA allows the user to limit the maximimum stepsize. If the maximum stepsize is set to $h = 0.001$ in GELDA, the maximum value of the approximation to the undetermined solution component reduces to a value of 1.60×10^{-3}. The corresponding solution is computed with 1006 discretization steps.

Hence, the applicability of software packages like GELDA seems to be limited if one wants

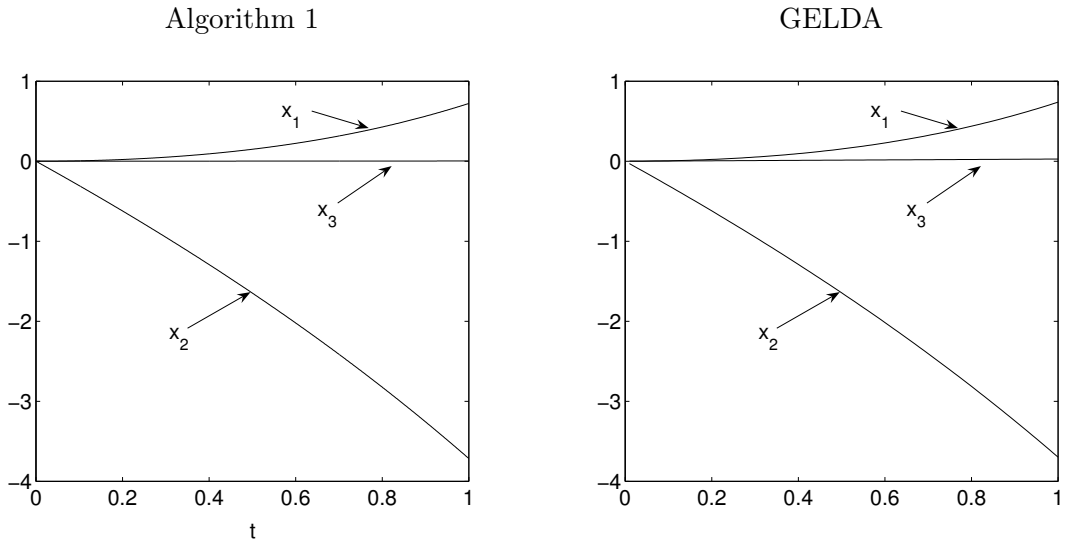

Figure 4.3: Approximate solutions of the DAE (4.2) computed with Algorithm 1 (with $h = 0.001$) and GELDA (with absolute and relative tolerances set to 0.0001).

to compute approximations to the (1,2,3) generalized solutions as presented here. In addition, it seems to be difficult to improve Algorithm 1 considerably by using stepsize control techniques.

The L_2 norm of the solution computed with GELDA is approximately given by $\sqrt{\sum_{i=1}^{1000} x_i^2/1000} \approx 2.040$, where x_i, $i = 1, ..., 1000$, denote the values of this solution evaluated at $t_i = i/1000$. Applying Algorithm 3 to the DAE (4.2) yields the solution which is displayed in Figure 4.4. The approximate norm of this solution computed as above using all intermediate grid points x_i, $i = 1, ..., 1000$, yields the value 1.785. This is a good approximation to the L_2 norm of the least squares solution x_{ls} of (4.2), which equals $\|x_{\text{ls}}\| = \sqrt{e^2 - 4e + \frac{20}{3}} \approx 1.7840$, and it is considerably smaller than the approximate norm of the solution computed with GELDA.

This confirms that GELDA cannot be used to compute approximations to the least squares solution (4.9) of the test example 4.2.

4.2.4 Performance

In order to compare the performance of the three Algorithms presented in Section 4.1 we applied them to a test example of size 30×60 with random coefficients.

The following table shows the execution time in seconds for the Algorithms using $N \in \{10, 20, 40\}$ discretization steps.

N	Algorithm 1	Algorithm 2	Algorithm 3
10	1.8×10^{-2}	5.2×10^{-1}	6.9×10^{-2}
20	3.5×10^{-2}	3.6×10^{0}	1.4×10^{-1}
40	7.1×10^{-2}	2.6×10^{1}	3.0×10^{-1}

As expected, the execution times for Algorithm 1 and Algorithm 3 are approximately $O(N)$, while Algorithm 2 needs approximately $O(N^3)$ seconds to compute the result obtained with Algorithm 3.

Algorithm 3

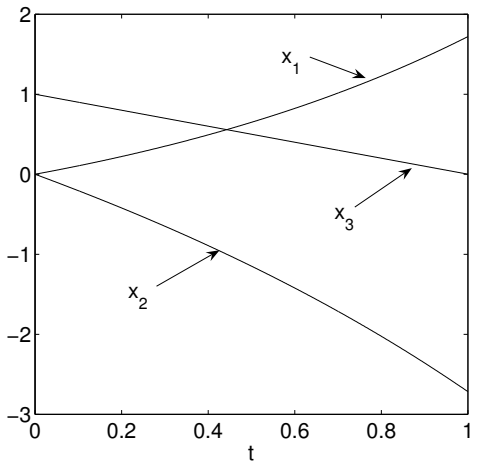

Figure 4.4: Approximate solution of the DAE (4.2) computed with Algorithm 3 (with $h = 0.001$).

4.2.5 A purely algebraic example

We have applied Algorithm 1 and Algorithm 3 to the DAE given in Example 1. This DAE has strangeness index one but the equivalent strangeness free system, computed with the index reduction techniques presented in Section 1.2, turns out to be purely algebraic. In this case, the (1,2,3)-generalized solution and the least squares solution coincide. Using both algorithms, we obtained the exact solution given in Example 9 within the realm of roundoff errors.

The results obtained by Algorithm 1 and Algorithm 3 applied to purely algebraic DAEs will always coincide. This can be explained by the fact that, in this case, the subdiagonal block entries of the discretization matrix D_h, as defined in (3.12), vanish. Thus, the linear systems which have to be solved in every discretization step are decoupled, and solving them independently (as it is done by Algorithm 1) yields the same result as solving the complete discretization at once (as it is done by Algorithm 3).

Chapter 5

Conclusions and Outlook

In this thesis, we have investigated the numerical computation of generalized solutions of linear time-variant differential-algebraic equations (DAEs). The major contributions are as follows.

- Development of a local minimization algorithm, which is shown in Theorem 26 to yield an $O(h)$ approximation to a particularly fixed (1,2,3)-generalized solution of the DAE.
- Development of a global minimization algorithm, which is shown in Theorem 43 to yield an $O(h)$ approximation to the least squares solution of the DAE. As a by-product, Lemma 41 shows that the inverse of a matrix with a certain structure related to BDF-methods is uniformly bounded.
- Both algorithms are based on a rather intuitive approach, but the investigation of their approximation properties turned out to be much more involved and has been addressed for the first time in this thesis.
- The straightforward implementation of global minimization requires $O(N^3m^2n)$ flops; this figure has been reduced to $O(Nm^2n)$ by exploiting the block bidiagonal structure of the discretization matrix.
- The developed algorithms have been implemented in Fortran routines.
- Various numerical experiments verify the obtained theoretical results.

Although the obtained results cover a wide range of tasks associated with the numerical computation of generalized solutions of DAEs, several extensions of these results remain to be studied, for instance:

- use of other minimization criteria for defining the least squares solution;
- application to control-related problems;
- extension of the developed algorithms to large and possibly sparse DAEs;
- investigation of adaptive time discretization schemes;
- extension to nonlinear DAEs.

Appendix A

Proof of Lemma 41

Before proving Lemma 41, we need the following preparatory results.

Theorem 46 ([23]) *Consider $A, \Delta A \in \mathbb{R}^{n,n}$ with $\operatorname{rank}(A) = \operatorname{rank}(A + \Delta A) = n$. Let*

$$A = QR$$

and

$$A + \Delta A = (Q + \Delta Q)(R + \Delta R)$$

be QR decompositions of A and $A + \Delta A$, normalized such that R and $R + \Delta R$ have positive diagonal elements. Then for sufficiently small ΔA we get

$$\frac{\|\Delta R\|_F}{\|R\|_F} \leq C_n \kappa_F(A) \frac{\|\Delta A\|_F}{\|A\|_F}, \tag{A.1}$$

$$\|\Delta Q\|_F \leq C_n \kappa_F(A) \frac{\|\Delta A\|_F}{\|A\|_F}, \tag{A.2}$$

where the constant $C_n \in \mathbb{R}^+$ depends only on the size n of A and $\kappa_F(A)$ denotes the condition number of A with respect to the Frobenius norm $\|\cdot\|_F$.

Lemma 47 *For $i \in \mathbb{N}$, let a matrix V be given with the block structure*

$$V = \begin{bmatrix} 0 & \cdots & 0 & I \\ M_1 & \cdots & M_{i-1} & 0 \\ N_1 & \cdots & N_{i-1} & N_i \end{bmatrix} \in \mathbb{R}^{3n,in},$$

where $M_j, N_j \in \mathbb{R}^{n,n}$ and $\|N_j\|_2$ is sufficiently small for $j = 1, \ldots, i$. It is assumed that there exists a matrix $B \in \mathbb{R}^{n,n}$ such that

$$N_j = BM_j \tag{A.3}$$

for $j = 1, \ldots, i-1$.

Then there exists an orthogonal matrix $Q \in \mathbb{R}^{3n,3n}$ such that

$$\hat{V} = Q^T V = \begin{bmatrix} \tilde{N}_1 & \cdots & \tilde{N}_{i-1} & \tilde{N}_i \\ \hat{M}_1 & \cdots & \hat{M}_{i-1} & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{A.4}$$

where for $j = 1, \ldots, i-1$,

$$\|\tilde{N}_j\|_2 \leq \sqrt{2}\, nC_n \|N_i\|_2 \|N_j\|_2, \tag{A.5}$$

with a constant $C_n \in \mathbb{R}^+$ that depends only on the blocksize n,

$$\|\hat{M}_j\|_2 \leq \sqrt{\|M_j\|_2^2 + \|N_j\|_2^2} \tag{A.6}$$

and

$$\|\tilde{N}_i^{-1}\|_2 \leq 1. \tag{A.7}$$

Furthermore, there exists a matrix $\hat{B} \in \mathbb{R}^{n,n}$ such that $\hat{M}_j = \hat{B}M_j$ for $j = 1, \cdots, i-1$.

Proof. The idea of the proof is as follows. The third block row of V can be eliminated by means of two QR decompositions. In a first step we can eliminate the block entry N_i by applying a QR decomposition to the first and the third block row of V. Then, by a QR decomposition of the second and the modified third block row, the third block row can be eliminated completely due to the assumption (A.3).

We first consider the QR decomposition

$$\begin{bmatrix} I & 0 \\ N_i & I \end{bmatrix} = \begin{bmatrix} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{bmatrix} \begin{bmatrix} \tilde{N}_i & R_{12} \\ 0 & R_{22} \end{bmatrix}, \tag{A.8}$$

where $\tilde{N}_i$ and R_{22} are upper triangular matrices. The property (A.7) follows from

$$\|\tilde{N}_i^{-1}\|_2 = \sigma_{\min}^{-1}(\tilde{N}_i) = \sigma_{\min}^{-1}\left(\begin{bmatrix} \tilde{N}_i \\ 0 \end{bmatrix}\right) = \sigma_{\min}^{-1}\left(\begin{bmatrix} I \\ N_i \end{bmatrix}\right) \leq 1,$$

where the latter inequality holds because

$$\sigma_{\min}\left(\begin{bmatrix} I \\ N_i \end{bmatrix}\right) = \min_{x \in \mathbb{R}^n} \frac{\left\|\begin{bmatrix} I \\ N_i \end{bmatrix} x\right\|_2}{\|x\|_2} = \min_{x \in \mathbb{R}^n} \frac{\sqrt{\|x\|_2^2 + \|N_i x\|_2^2}}{\|x\|_2} \geq 1. \tag{A.9}$$

The QR decomposition (A.8) can be interpreted as a perturbation of the QR decomposition $A = QR$ with $A = Q = R = I_{2n}$, where A is perturbed by

$$\Delta A = \begin{bmatrix} 0 & 0 \\ N_i & 0 \end{bmatrix}.$$

From Theorem 46, it follows for sufficiently small N_i that

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ N_i & 0 \end{bmatrix} = \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & \Delta Q_{22} \end{bmatrix}\right)(I_{2n} + \Delta R),$$

where

$$\left\|\begin{bmatrix} \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & \Delta Q_{22} \end{bmatrix}\right\|_F \leq C_n \kappa_F(I_{2n}) \frac{\left\|\begin{bmatrix} 0 & 0 \\ N_i & 0 \end{bmatrix}\right\|_F}{\|I_{2n}\|_F} = \sqrt{2n}\, C_n \|N_i\|_F.$$

Because of $\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\, \|A\|_2$ for all $A \in \mathbb{R}^{n,n}$, we get

$$\left\|\begin{bmatrix} \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & \Delta Q_{22} \end{bmatrix}\right\|_2 \leq \left\|\begin{bmatrix} \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & \Delta Q_{22} \end{bmatrix}\right\|_F \leq \sqrt{2n}\, C_n \|N_i\|_F \leq \sqrt{2}\, n C_n \|N_i\|_2.$$

From

$$\begin{bmatrix} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{bmatrix} = \begin{bmatrix} I + \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & I + \Delta Q_{22} \end{bmatrix}$$

we can estimate the norm of $\tilde{Q}_{21} = \Delta Q_{21}$ by

$$\|\tilde{Q}_{21}\|_2 \le \left\| \begin{bmatrix} \Delta Q_{11} & \Delta Q_{12} \\ \Delta Q_{21} & \Delta Q_{22} \end{bmatrix} \right\|_2 \le \sqrt{2}\, nC_n \|N_i\|_2.$$

By setting

$$\tilde{Q} = \begin{bmatrix} \tilde{Q}_{11} & 0 & \tilde{Q}_{12} \\ 0 & I & 0 \\ \tilde{Q}_{21} & 0 & \tilde{Q}_{22} \end{bmatrix}$$

we get

$$\tilde{Q}^T V = \begin{bmatrix} \tilde{N}_1 & \cdots & \tilde{N}_{i-1} & \tilde{N}_i \\ M_1 & \dots & M_{i-1} & 0 \\ \tilde{Q}_{22}^T N_1 & \cdots & \tilde{Q}_{22}^T N_{i-1} & 0 \end{bmatrix},$$

with $\tilde{N}_j = \tilde{Q}_{21}^T N_j$, where

$$\|\tilde{N}_j\|_2 \le \|\tilde{Q}_{21}^T\|_2 \|N_j\|_2 = \|\tilde{Q}_{21}\|_2 \|N_j\|_2 \le \sqrt{2}\, nC_n \|N_i\|_2 \|N_j\|_2,$$

and thus (A.5).

To eliminate the third block row of $\tilde{Q}^T V$, we consider a QR decomposition

$$\begin{bmatrix} I \\ \tilde{Q}_{22}^T B \end{bmatrix} = \begin{bmatrix} \hat{Q}_{11} & \hat{Q}_{12} \\ \hat{Q}_{21} & \hat{Q}_{22} \end{bmatrix} \begin{bmatrix} \hat{B} \\ 0 \end{bmatrix}, \tag{A.10}$$

where C_i is upper triangular. From (A.10) it follows that

$$\begin{bmatrix} \hat{Q}_{11}^T & \hat{Q}_{21}^T \\ \hat{Q}_{12}^T & \hat{Q}_{22}^T \end{bmatrix} \begin{bmatrix} I \\ \tilde{Q}_{22}^T B \end{bmatrix} = \begin{bmatrix} \hat{Q}_{11}^T + \hat{Q}_{21}^T \tilde{Q}_{22}^T B \\ \hat{Q}_{12}^T + \hat{Q}_{22}^T \tilde{Q}_{22}^T B \end{bmatrix} = \begin{bmatrix} \hat{B} \\ 0 \end{bmatrix},$$

and thus

$$\begin{aligned} \begin{bmatrix} \hat{Q}_{11}^T & \hat{Q}_{21}^T \\ \hat{Q}_{12}^T & \hat{Q}_{22}^T \end{bmatrix} \begin{bmatrix} M_j \\ \tilde{Q}_{22}^T N_j \end{bmatrix} &= \begin{bmatrix} \hat{Q}_{11}^T & \hat{Q}_{21}^T \\ \hat{Q}_{12}^T & \hat{Q}_{22}^T \end{bmatrix} \begin{bmatrix} M_j \\ \tilde{Q}_{22}^T B M_j \end{bmatrix} \\ &= \begin{bmatrix} (\hat{Q}_{11}^T + \hat{Q}_{21}^T \tilde{Q}_{22}^T B) M_j \\ (\hat{Q}_{12}^T + \hat{Q}_{22}^T \tilde{Q}_{22}^T B) M_j \end{bmatrix} \\ &= \begin{bmatrix} \hat{B} M_j \\ 0 \end{bmatrix}. \end{aligned}$$

We define

$$\hat{Q} = \begin{bmatrix} I & 0 & 0 \\ 0 & \hat{Q}_{11} & \hat{Q}_{12} \\ 0 & \hat{Q}_{21} & \hat{Q}_{22} \end{bmatrix}$$

and get

$$\hat{Q}^T \tilde{Q}^T V_i = \begin{bmatrix} \tilde{N}_1 & \cdots & \tilde{N}_{i-1} & \tilde{N}_i \\ \hat{M}_1 & \cdots & \hat{M}_{i-1} & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix},$$

which consequently implies (A.4) with $Q = \tilde{Q}\hat{Q}$ and $\hat{M}_j = \hat{B} M_j$. The blocks in the second block row can be estimated by

$$\begin{aligned} \|\hat{M}_j\|_2 = \left\| \begin{bmatrix} \hat{M}_j \\ 0 \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} \hat{Q}_{11}^T & \hat{Q}_{21}^T \\ \hat{Q}_{12}^T & \hat{Q}_{22}^T \end{bmatrix} \begin{bmatrix} M_j \\ \tilde{Q}_{22}^T N_j \end{bmatrix} \right\|_2 &= \left\| \begin{bmatrix} M_j \\ \tilde{Q}_{22}^T N_j \end{bmatrix} \right\|_2 \\ \le \sqrt{\|M_j\|_2^2 + \|\tilde{Q}_{22}^T N_j\|_2^2} \le \sqrt{\|M_j\|_2^2 + \|N_j\|_2^2}, \end{aligned}$$

which shows the assertion (A.6). □

Lemma 48 *For $N \in \mathbb{N}$ and $h \geq 0$, let*

$$A = \begin{bmatrix} I & & & \\ A_{21} & I & & \\ \vdots & \ddots & \ddots & \\ A_{N1} & \cdots & A_{N,N-1} & I \end{bmatrix}$$

with $A_{ij} \in \mathbb{R}^{n,n}$, $\|A_{ij}\| \leq C_A h$, for $j = 1, \ldots, N-1$, $i > j$. Then

$$A^{-1} = \begin{bmatrix} I & & & \\ \tilde{A}_{21} & I & & \\ \vdots & \ddots & \ddots & \\ \tilde{A}_{N1} & \cdots & \tilde{A}_{N,N-1} & I \end{bmatrix}$$

with $\|\tilde{A}_{ij}\| \leq C_A(1+C_A h)^{i-j-1}h$. Here, $\|\cdot\|$ denotes an arbitrary submultiplicative matrix norm.

Proof. We can write down A in the form

$$A = F_1 \cdot \ldots \cdot F_{N-1},$$

with the Frobenius matrices

$$F_k = \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ & & & I & & \\ & & & A_{k+1,k} & I & \\ & & & \vdots & & \ddots \\ & & & A_{Nk} & & I \end{bmatrix}.$$

The inverse of A can then be computed by

$$A^{-1} = F_{N-1}^{-1} \cdot \ldots \cdot F_1^{-1}$$

with

$$F_k^{-1} = \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ & & & I & & \\ & & & -A_{k+1,k} & I & \\ & & & \vdots & & \ddots \\ & & & -A_{Nk} & & I \end{bmatrix}.$$

We will show inductively that for $k = N-1, \ldots, 1$,

$$F_{N-1}^{-1} \cdot \ldots \cdot F_k^{-1} = \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ & & & I & & \\ & & & \tilde{A}_{k+1,k} & I & \\ & & & \vdots & \ddots & \ddots \\ & & & \tilde{A}_{Nk} & \cdots & \tilde{A}_{N,N-1} \; I \end{bmatrix}$$

with

$$\|\tilde{A}_{ij}\| \leq C_A(1 + C_A h)^{i-j-1} h. \tag{A.11}$$

For $k = N - 1$, this is a direct consequence of

$$F_{N-1}^{-1} = \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & \tilde{A}_{N,N-1} & I \end{bmatrix}$$

with $\tilde{A}_{N,N-1} = -A_{N,N-1}$ and thus $\|\tilde{A}_{N,N-1}\| = \|A_{N,N-1}\| \leq C_A h$.

For $k < N - 1$, we get

$$\begin{aligned} F_{N-1}^{-1} \cdot \ldots \cdot F_{k-1}^{-1} &= F_{N-1}^{-1} \cdot \ldots \cdot F_k^{-1} F_{k-1}^{-1} \\ &= \begin{bmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ & & & I & & & \\ & & & & I & & \\ & & & & \tilde{A}_{k+1,k} & I & \\ & & & & \vdots & \ddots & \ddots \\ & & & & \tilde{A}_{Nk} & \cdots & \tilde{A}_{N,N-1} \; I \end{bmatrix} \begin{bmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ & & & I & & & \\ & & & & A_{k,k-1} & I & \\ & & & & A_{k+1,k-1} & & I \\ & & & & \vdots & & \ddots \\ & & & & A_{N,k-1} & & I \end{bmatrix} \\ &= \begin{bmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ & & & I & & & \\ & & & \tilde{A}_{k,k-1} & I & & \\ & & & \tilde{A}_{k+1,k-1} & \tilde{A}_{k+1,k} & I & \\ & & & \vdots & \vdots & \ddots & \ddots \\ & & & \tilde{A}_{N,k-1} & \tilde{A}_{Nk} & \cdots & \tilde{A}_{N,N-1} \; I \end{bmatrix}, \end{aligned}$$

with

$$\tilde{A}_{i,k-1} = A_{i,k-1} + \sum_{l=0}^{i-k-1} \tilde{A}_{i,k+l} A_{k+l,k-1}$$

and thus

$$\begin{aligned}
\|\tilde{A}_{i,k-1}\| &\leq \|A_{i,k-1}\| + \sum_{l=0}^{i-k-1} \|\tilde{A}_{i,k+l}\| \|A_{k+l,k-1}\| \\
&\leq C_A h + \sum_{l=0}^{i-k-1} C_A^2 h^2 (1 + C_A h)^{i-k-l-1} \\
&= C_A h \left(1 + C_A h \sum_{l=0}^{i-k-1} (1 + C_A h)^{i-k-l-1} \right) \\
&= C_A h \left(1 + C_A h \sum_{l=0}^{i-k-1} (1 + C_A h)^{l} \right) \\
&= C_A h \left(1 + C_A h \frac{(1 + C_A h)^{i-k} - 1}{(1 + C_A h) - 1} \right) \\
&= C_A h (1 + C_A h)^{i-k} \\
&= C_A h (1 + C_A h)^{i-(k-1)-1}.
\end{aligned}$$

This shows the assertion for $A^{-1} = F_{N-1}^{-1} \cdot \ldots \cdot F_1^{-1}$. □

We are now prepared to prove Lemma 41. For convenience, let us restate this lemma in a more detailed form.

Lemma 41 *Consider the matrix* $W_h = I + V_h^T V_h$ *with*

$$V_h = R_h^{-1} B_h, \tag{A.12}$$

where

$$R_h = \begin{bmatrix} \tilde{D}_h^o & A_{12_h} \\ A_{21_h} & \Sigma_{A_h} \end{bmatrix}, \quad B_h = \begin{bmatrix} A_{13_h} \\ 0 \end{bmatrix}, \tag{A.13}$$

with D_h^o, A_{12_h}, A_{21_h}, Σ_{A_h} *and* A_{13_h} *defined as in (3.72), (3.93) and (3.115).*

Then the inverse of W_h *is uniformely bounded, i.e., there exists a positive constant* $C_{W^{-1}} \in \mathbb{R}$ *such that*

$$\|W_h^{-1}\|_\infty \leq C_{W^{-1}} \tag{A.14}$$

and $C_{W^{-1}}$ *does* not *depend on the step size* h*.*

Proof. In a first step, we show that $W_h = I + V_h^T V_h = \tilde{V}_h^T \tilde{V}_h$ with a lower block triangular matrix $\tilde{V}_h$, where the inverses of the diagonal blocks of $\tilde{V}_h$ have spectral norm less than one, while the subdiagonal blocks of $\tilde{V}_h$ have spectral norm $O(h)$. In a second step, we use this result to show that the inverse of $\tilde{V}_h$ is uniformly bounded.

From (3.94) and (3.95) it follows that the inverse of the matrix R_h takes the form

$$R_h^{-1} = \begin{bmatrix} S_h^{-1} & * \\ -\Sigma_{A_h}^{-1} A_{21_h} S_h^{-1} & * \end{bmatrix},$$

with

$$\begin{aligned}
S_h &= \begin{bmatrix} \frac{1}{h}\Sigma_{E_1} - \bar{A}_1 & & & \\ -\frac{1}{h}\Sigma_{E_2} & \frac{1}{h}\Sigma_{E_2} - \bar{A}_2 & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h}\Sigma_{E_N} & \frac{1}{h}\Sigma_{E_N} - \bar{A}_N \end{bmatrix}, \\
\bar{A}_i &= A_{11_i} + A_{12_i} \Sigma_{A_i}^{-1} A_{21_i}.
\end{aligned} \tag{A.15}$$

Thus, we have

$$R_h^{-1}B_h = \begin{bmatrix} S_h^{-1}A_{13_h} \\ -\Sigma_{A_h}^{-1}A_{21_h}S_h^{-1}A_{13_h} \end{bmatrix} = \begin{bmatrix} I \\ -\Sigma_{A_h}^{-1}A_{21_h} \end{bmatrix} S_h^{-1}A_{13_h}$$

and

$$\begin{aligned} V_h^T V_h &= A_{13_h}^T S_h^{-T} \begin{bmatrix} I & -A_{21_h}^T \Sigma_{A_h}^{-T} \end{bmatrix} \begin{bmatrix} I \\ -\Sigma_{A_h}^{-1}A_{21_h} \end{bmatrix} S_h^{-1}A_{13_h} \\ &= A_{13_h}^T S_h^{-T} \left(I + A_{21_h}^T \Sigma_{A_h}^{-T}\Sigma_{A_h}^{-1}A_{21_h} \right) S_h^{-1}A_{13_h}. \end{aligned}$$

Since Σ_{A_h} and A_{21_h} are block diagonal matrices, we get

$$I + A_{21_h}^T \Sigma_{A_h}^{-T}\Sigma_{A_h}^{-1}A_{21_h} = \operatorname{diag}(I + A_{21_1}^T \Sigma_{A_1}^{-T}\Sigma_{A_1}^{-1}A_{21_1}, \ldots, I + A_{21_N}^T \Sigma_{A_N}^{-T}\Sigma_{A_N}^{-1}A_{21_N}).$$

By a QR decomposition

$$\begin{bmatrix} I \\ -\Sigma_{A_i}^{-1}A_{21_i} \end{bmatrix} = Q \begin{bmatrix} \mathcal{A}_i \\ 0 \end{bmatrix},$$

the diagonal blocks can be written as

$$\begin{aligned} I + A_{21_i}^T \Sigma_{A_i}^{-T}\Sigma_{A_i}^{-1}A_{21_i} &= \begin{bmatrix} I & -A_{21_i}^T \Sigma_{A_i}^{-T} \end{bmatrix} \begin{bmatrix} I \\ -\Sigma_{A_i}^{-1}A_{21_i} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A}_i^T & 0 \end{bmatrix} Q^T Q \begin{bmatrix} \mathcal{A}_i \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A}_i^T & 0 \end{bmatrix} \begin{bmatrix} \mathcal{A}_i \\ 0 \end{bmatrix} = \mathcal{A}_i^T \mathcal{A}_i. \end{aligned}$$

Obviously $\mathcal{A}_i$ is nonsingular for $i = 1, \ldots, N$. The norm of $\mathcal{A}_i$ can be estimated by

$$\begin{aligned} \|\mathcal{A}_i\|_2 = \left\| \begin{bmatrix} I \\ -\Sigma_{A_i}^{-1}A_{21_i} \end{bmatrix} \right\|_2 &\le \sqrt{1 + \|\Sigma_{A_i}^{-1}A_{21_i}\|_2^2} \\ &\le 1 + \|\Sigma_{A_i}^{-1}A_{21_i}\|_2 \le 1 + C_{\|\Sigma_A^{-1}A_{21}\|_2}, \end{aligned}$$

where $C_{\|\Sigma_A^{-1}A_{21}\|_2}$ is defined as in Definition 23.

Setting $\mathcal{A}_h = \operatorname{diag}(\mathcal{A}_1, \ldots, \mathcal{A}_N)$ and $\mathcal{V}_h = \mathcal{A}_h S_h^{-1}A_{13_h}$, we have

$$W_h = I + V_h^T V_h = I + \mathcal{V}_h^T \mathcal{V}_h.$$

The inverse of S_h is given blockwise by

$$\begin{aligned} [S_h^{-1}]_{ij} &= \begin{cases} \left[\coprod_{k=i}^{j+1} \frac{1}{h} \left(\frac{1}{h}\Sigma_{E_k} - \bar{A}_k \right)^{-1} \Sigma_{E_k} \right] \left(\frac{1}{h}\Sigma_{E_j} - \bar{A}_j \right)^{-1} & \text{for } i \ge j, \\ 0 & \text{otherwise,} \end{cases} \\ &= \begin{cases} \left[\coprod_{k=i}^{j+1} \left(I - h\Sigma_{E_k}^{-1}\bar{A}_k \right)^{-1} \right] \left(\frac{1}{h}\Sigma_{E_j} - \bar{A}_j \right)^{-1} & \text{for } i \ge j, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus, it follows for all $j = 1, \ldots, N-1$ and $i = j, \ldots, N-1$ that

$$\begin{aligned}
[S_h^{-1}]_{i+1,j} &= \left[\coprod_{k=i+1}^{j+1} \left(I - h\Sigma_{E_k}^{-1}\bar{A}_k\right)^{-1}\right] \left(\frac{1}{h}\Sigma_{E_j} - \bar{A}_j\right)^{-1} \\
&= (I - h\Sigma_{E_{i+1}}^{-1}\bar{A}_{i+1})^{-1} \left[\coprod_{k=i}^{j+1} \left(I - h\Sigma_{E_k}^{-1}\bar{A}_k\right)^{-1}\right] \left(\frac{1}{h}\Sigma_{E_j} - \bar{A}_j\right)^{-1} \\
&= \left(I - h\Sigma_{E_{i+1}}^{-1}\bar{A}_{i+1}\right)^{-1} [S_h^{-1}]_{ij}.
\end{aligned}$$

For the blocks $\mathcal{V}_{ij}$ of $\mathcal{V}_h$ from

$$\mathcal{V}_{ij} = \mathcal{A}_i [S_h^{-1}]_{ij} A_{13_j},$$

we obtain

$$\begin{aligned}
\mathcal{V}_{i+1,j} &= \mathcal{A}_{i+1}[S_h^{-1}]_{i+1,j} A_{13_j} \\
&= \mathcal{A}_{i+1}\left(I - h\Sigma_{E_{i+1}}^{-1}\bar{A}_{i+1}\right)^{-1} \mathcal{A}_i^{-1}\mathcal{A}_i [S_h^{-1}]_{ij} A_{13_j} \\
&= \mathcal{B}_i \mathcal{V}_{ij},
\end{aligned} \tag{A.16}$$

with

$$\mathcal{B}_i = \mathcal{A}_{i+1}\left(I - h\Sigma_{E_{i+1}}^{-1}\bar{A}_{i+1}\right)^{-1} \mathcal{A}_i^{-1}.$$

It has been shown in Corollary 36 that $\|[S_h^{-1}]_{ij}\|_\infty \leq C_2^u h$. Hence, we can conclude for the blocks of $\mathcal{V}_h$ that

$$\begin{aligned}
\|\mathcal{V}_{ij}\|_2 &\leq \|\mathcal{A}_i\|_2 \|[S_h^{-1}]_{ij}\|_2 \|A_{13_h}\|_2 \\
&\leq (1 + C_{\|\Sigma_A^{-1}A_{21}\|_\infty})\sqrt{n}\, \|[S_h^{-1}]_{ij}\|_\infty C_{\|A_{13}\|_\infty} \\
&\leq (1 + C_{\|\Sigma_A^{-1}A_{21}\|_\infty})\sqrt{n}\, C_2^u C_{\|A_{13}\|_\infty} h \\
&= C_{\mathcal{V}} h.
\end{aligned}$$

Note that the matrix W_h can be factored as

$$W_h = \begin{bmatrix} I & \mathcal{V}_h^T \end{bmatrix} \begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix}.$$

We now apply Lemma 47 inductively to the factor

$$\begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} = \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ \mathcal{V}_{11} & & & \\ \vdots & \ddots & & \\ \mathcal{V}_{N1} & \cdots & \mathcal{V}_{NN} & \end{bmatrix}.$$

From (A.16), it follows that $\mathcal{V}_{Nj} = \mathcal{B}_{N-1}\mathcal{V}_{N-1,j}$, $j = 1, \ldots, N-1$. For sufficiently small h, Lemma 47 implies the existence of an orthogonal matrix Q_N such that

$$Q_N \begin{bmatrix} 0 & \cdots & 0 & I \\ \mathcal{V}_{N-1,1} & \cdots & \mathcal{V}_{N-1,N-1} & 0 \\ \mathcal{V}_{N1} & \cdots & \mathcal{V}_{N,N-1} & \mathcal{V}_{NN} \end{bmatrix} = \begin{bmatrix} \tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{N,N-1} & \tilde{\mathcal{V}}_{NN} \\ \hat{\mathcal{V}}_{N-1,1} & \cdots & \hat{\mathcal{V}}_{N-1,N-1} & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix},$$

and consequently the existence of an orthogonal matrix Q_{N_h} such that

$$\begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} = Q_{N_h}^T \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ & & & I & & \\ \tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{N,N-2} & \tilde{\mathcal{V}}_{N,N-1} & \tilde{\mathcal{V}}_{NN} \\ \mathcal{V}_{11} & & & & \\ \vdots & \ddots & & & \\ \mathcal{V}_{N-2,1} & \cdots & \mathcal{V}_{N-2,N-2} & & \\ \hat{\mathcal{V}}_{N-1,1} & \cdots & \hat{\mathcal{V}}_{N-1,N-2} & \hat{\mathcal{V}}_{N-1,N-1} & \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix},$$

where for a positive constant C_n,

$$\begin{aligned} \|\tilde{\mathcal{V}}_{Nj}\|_2 &\leq \sqrt{2}\, nC_n \|\mathcal{V}_{NN}\|_2 \|\mathcal{V}_{Nj}\|_2 \leq \sqrt{2}\, nC_n C_{\mathcal{V}}^2 h^2, \\ \|\hat{\mathcal{V}}_{N-1,j}\|_2 &\leq \sqrt{\|\mathcal{V}_{N-1,j}\|_2^2 + \|\mathcal{V}_{N,j}\|_2^2} \leq \sqrt{2C_{\mathcal{V}}^2 h^2} = \sqrt{2}\, C_{\mathcal{V}} h, \end{aligned}$$

for $j = 1, \ldots, N-1$, and

$$\|\tilde{\mathcal{V}}_{NN}^{-1}\|_2 \leq 1.$$

Furthermore, there exists a matrix $\hat{B}_{N-1}$ such that $\hat{\mathcal{V}}_{N-1,j} = \hat{B}_{N-1}\mathcal{V}_{N-1,j}$ and thus $\hat{\mathcal{V}}_{N-1,j} = \hat{B}_{N-1}\mathcal{B}_{N-2}\mathcal{V}_{N-2,j}$ according to (A.16). This implies for sufficiently small h that the matrix

$$\begin{bmatrix} 0 & \cdots & 0 & I \\ \mathcal{V}_{N-2,1} & \cdots & \mathcal{V}_{N-2,N-2} & 0 \\ \hat{\mathcal{V}}_{N-1,1} & \cdots & \hat{\mathcal{V}}_{N-1,N-2} & \hat{\mathcal{V}}_{N-1,N-1} \end{bmatrix}$$

satisfies the assumptions of Lemma 47. Thus, there exists an orthogonal matrix Q_{N-1_h} such that

$$\begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} = Q_{N_h}^T Q_{N-1_h}^T \begin{bmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ & & & I & & & \\ \tilde{\mathcal{V}}_{N-1,1} & \cdots & \tilde{\mathcal{V}}_{N-1,N-3} & \tilde{\mathcal{V}}_{N-1,N-2} & \tilde{\mathcal{V}}_{N-1,N-1} & & \\ \tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{N,N-3} & \tilde{\mathcal{V}}_{N,N-2} & \tilde{\mathcal{V}}_{N,N-1} & \tilde{\mathcal{V}}_{NN} \\ \mathcal{V}_{11} & & & & & \\ \vdots & \ddots & & & & \\ \mathcal{V}_{N-3,1} & \cdots & \mathcal{V}_{N-3,N-3} & & & \\ \hat{\mathcal{V}}_{N-2,1} & \cdots & \hat{\mathcal{V}}_{N-2,N-3} & \hat{\mathcal{V}}_{N-2,N-2} & & \\ 0 & \cdots & 0 & 0 & 0 & \\ 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}$$

with

$$\begin{aligned} \|\tilde{\mathcal{V}}_{N-1,j}\|_2 &\leq \sqrt{2}\, nC_n \|\hat{\mathcal{V}}_{N-1,N-1}\|_2 \|\hat{\mathcal{V}}_{N-1,j}\|_2 \leq 2\sqrt{2}\, nC_n C_{\mathcal{V}}^2 h^2, \\ \|\hat{\mathcal{V}}_{N-2,j}\|_2 &\leq \sqrt{\|\mathcal{V}_{N-2,j}\|_2^2 + \|\hat{\mathcal{V}}_{N-1,j}\|_2^2} \leq \sqrt{C_{\mathcal{V}}^2 h^2 + 2C_{\mathcal{V}}^2 h^2} = \sqrt{3}\, C_{\mathcal{V}} h \end{aligned}$$

for $j = 1, \ldots, N-2$,

$$\|\tilde{\mathcal{V}}_{N-1,N-1}^{-1}\|_2 \leq 1.$$

Furthermore, there exists a matrix $\hat{B}_{N-2}$ such that $\hat{\mathcal{V}}_{N-2,j} = \hat{B}_{N-2}\mathcal{V}_{N-2,j} = \hat{B}_{N-2}\mathcal{B}_{N-3}\mathcal{V}_{N-3,j}$.

For $i = N-2, \ldots, 2$, we can apply Lemma 47 to the matrix

$$\begin{bmatrix} 0 & \cdots & 0 & I \\ \mathcal{V}_{i-1,1} & \cdots & \mathcal{V}_{i-1,i-1} & 0 \\ \hat{\mathcal{V}}_{i,1} & \cdots & \hat{\mathcal{V}}_{i,i-1} & \hat{\mathcal{V}}_{ii} \end{bmatrix},$$

where $\hat{\mathcal{V}}_{i-1,j} = \hat{B}_i\mathcal{B}_{i-1}\mathcal{V}_{i-1,j}$ and

$$\|\hat{\mathcal{V}}_{ij}\|_2 \leq \sqrt{N-i+1}\, C_{\mathcal{V}} h$$

for $j = 1, \ldots, i-1$. Then an orthogonal matrix Q_{i_h} is computed, such that

$$\begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} = \left(\coprod_{k=N}^{i} Q_{i_h}^T \right) \begin{bmatrix} I & & & & & & & \\ & \ddots & & & & & & \\ & & I & & & & & \\ & & & I & & & & \\ \tilde{\mathcal{V}}_{i,1} & \cdots & \tilde{\mathcal{V}}_{i,i-2} & \tilde{\mathcal{V}}_{i,i-1} & \tilde{\mathcal{V}}_{ii} & & & \\ \vdots & & \vdots & \vdots & \vdots & \ddots & & \\ \tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{N,i-2} & \tilde{\mathcal{V}}_{N,i-1} & \tilde{\mathcal{V}}_{Ni} & \ldots & \tilde{\mathcal{V}}_{NN} & \\ \mathcal{V}_{11} & & & & & & & \\ \vdots & \ddots & & & & & & \\ \mathcal{V}_{i-2,1} & \cdots & \mathcal{V}_{i-2,i-2} & & & & & \\ \hat{\mathcal{V}}_{i-1,1} & \cdots & \hat{\mathcal{V}}_{i-1,i-2} & \hat{\mathcal{V}}_{i-1,i-1} & & & & \\ 0 & \cdots & 0 & 0 & 0 & & & \\ \vdots & & \vdots & \vdots & \vdots & \ddots & & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \end{bmatrix},$$

with

$$\begin{aligned} \|\tilde{\mathcal{V}}_{ij}\|_2 &\leq \sqrt{2}\, nC_n \|\hat{\mathcal{V}}_{ii}\|_2 \|\hat{\mathcal{V}}_{ij}\|_2 \leq (N-i+1)\sqrt{2}\, nC_n C_{\mathcal{V}}^2 h^2, \\ \|\hat{\mathcal{V}}_{i-1,j}\|_2 &\leq \sqrt{\|\mathcal{V}_{i-1,j}\|_2^2 + \|\hat{\mathcal{V}}_{ij}\|_2^2} \leq \sqrt{C_{\mathcal{V}}^2 h^2 + (N-i+1)C_{\mathcal{V}}^2 h^2} = \sqrt{N-i+2}\, C_{\mathcal{V}} h, \end{aligned}$$

for $j = 1, \ldots, i-1$, and

$$\|\tilde{\mathcal{V}}_{N-1,N-1}^{-1}\|_2 \leq 1.$$

Furthermore, there exists a matrix $\hat{B}_{i-1}$ such that $\hat{\mathcal{V}}_{i-1,j} = \hat{B}_{i-1}\mathcal{V}_{i-1,j} = \hat{B}_{i-1}\mathcal{B}_{i-2}\mathcal{V}_{i-2,j}$.

When this elimination procedure is completed, we have computed a sequence $Q_{N_h}, \ldots, Q_{2_h}$ of orthogonal matrix functions such that, using the notation (3.83),

$$\begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} = \left(\coprod_{k=N}^{2} Q_{i_h}^T \right) \begin{bmatrix} \tilde{V}_h \\ 0 \end{bmatrix},$$

with

$$\tilde{V}_h = \begin{bmatrix} \tilde{\mathcal{V}}_{11} & & \\ \vdots & \ddots & \\ \tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{NN} \end{bmatrix},$$

where

$$\|\tilde{\mathcal{V}}_{ii}^{-1}\|_2 \leq 1 \tag{A.17}$$

and

$$\|\tilde{\mathcal{V}}_{ij}\|_2 \leq \sqrt{2}\, nC_n(N-i+1)C_{\mathcal{V}}^2 h^2 \leq \sqrt{2}\, nC_n C_{\mathcal{V}}^2 h =: C_{\tilde{\mathcal{V}}} h \tag{A.18}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, i-1$. In addition, we have

$$\begin{aligned} W_h &= \begin{bmatrix} I & \mathcal{V}_h^T \end{bmatrix} \begin{bmatrix} I \\ \mathcal{V}_h \end{bmatrix} \\ &= \begin{bmatrix} \tilde{V}_h^T & 0 \end{bmatrix} \left(\prod_{k=2}^{N} Q_{i_h} \right) \left(\coprod_{k=N}^{2} Q_{i_h}^T \right) \begin{bmatrix} \tilde{V}_h \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \tilde{V}_h^T & 0 \end{bmatrix} \begin{bmatrix} \tilde{V}_h \\ 0 \end{bmatrix} = \tilde{V}_h^T \tilde{V}_h. \end{aligned}$$

The norm of the inverse of $\tilde{V}_h$ can be estimated blockwise according to Lemma 48, because

$$\tilde{V}_h = \operatorname{diag}(\tilde{\mathcal{V}}_{11}, \ldots, \tilde{\mathcal{V}}_{NN}) \hat{V}_h,$$

with

$$\hat{V}_h = \begin{bmatrix} I & & & \\ \tilde{\mathcal{V}}_{22}^{-1}\tilde{\mathcal{V}}_{21} & I & & \\ \vdots & \ddots & \ddots & \\ \tilde{\mathcal{V}}_{NN}^{-1}\tilde{\mathcal{V}}_{N1} & \cdots & \tilde{\mathcal{V}}_{NN}^{-1}\tilde{\mathcal{V}}_{N,N-1} & I \end{bmatrix}.$$

From (A.17) and (A.18) we know that

$$\|\tilde{\mathcal{V}}_{ii}^{-1}\tilde{\mathcal{V}}_{ij}\|_2 \leq \|\tilde{\mathcal{V}}_{ii}^{-1}\|_2 \|\tilde{\mathcal{V}}_{ij}\|_2 \leq \|\tilde{\mathcal{V}}_{ij}\|_2 \leq C_{\tilde{\mathcal{V}}} h.$$

Therefore, Lemma 48 implies

$$\hat{V}_h^{-1} = \begin{bmatrix} I & & & \\ \left[\hat{V}^{-1}\right]_{21} & I & & \\ \vdots & \ddots & \ddots & \\ \left[\hat{V}^{-1}\right]_{N1} & \cdots & \left[\hat{V}^{-1}\right]_{N,N-1} & I \end{bmatrix},$$

where

$$\begin{aligned} \|[\hat{V}^{-1}]_{ij}\|_2 &\leq C_{\tilde{\mathcal{V}}}(1 + C_{\tilde{\mathcal{V}}} h)^{i-j-1} h \leq C_{\tilde{\mathcal{V}}}(1 + C_{\tilde{\mathcal{V}}} h)^{N} h \\ &\leq C_{\tilde{\mathcal{V}}} \exp(C_{\tilde{\mathcal{V}}}(T - t_0)) h =: C_{\hat{V}^{-1}} h. \end{aligned}$$

For the blocks $[\tilde{V}^{-1}]_{ij}$ of $\tilde{V}_h^{-1}$ we get, due to the relation

$$\tilde{V}_h^{-1} = \hat{V}_h^{-1} \operatorname{diag}(\tilde{\mathcal{V}}_{11}^{-1}, \ldots, \tilde{\mathcal{V}}_{NN}^{-1}),$$

the estimates

$$\|[\tilde{V}^{-1}]_{ii}\|_2 \leq \|\tilde{\mathcal{V}}_{ii}^{-1}\|_2 \leq 1$$

and

$$\|[\tilde{V}^{-1}]_{ij}\|_2 \leq \|\tilde{\mathcal{V}}_{jj}^{-1}\|_2 \|[\hat{V}^{-1}]_{ij}\|_2 \leq \|[\hat{V}^{-1}]_{ij}\|_2 \leq C_{\hat{V}^{-1}} h.$$

The blocks $[W_h^{-1}]_{ij}$ of $W_h^{-1} = \tilde{V}_h^{-1}\tilde{V}_h^{-T}$ can be computed as

$$[W_h^{-1}]_{ij} = \sum_{k=1}^{N} [\tilde{V}^{-1}]_{ik} [\tilde{V}^{-T}]_{kj} = \sum_{k=1}^{\min(i,j)} [\tilde{V}^{-1}]_{ik} [\tilde{V}^{-1}]_{jk}^T.$$

In particular, the diagonal elements can be written as

$$[W_h^{-1}]_{ii} = [\tilde{V}^{-1}]_{ii}[\tilde{V}^{-1}]_{ii}^T + \sum_{k=1}^{i-1}[\tilde{V}^{-1}]_{ik}[\tilde{V}^{-1}]_{ik}^T,$$

which implies

$$\|[W_h^{-1}]_{ii}\|_2 \leq \|[\tilde{V}^{-1}]_{ii}\|_2^2 + (i-1)\|[\tilde{V}^{-1}]_{ik}\|_2^2 \leq 1 + (i-1)C_{\hat{V}^{-1}}^2 h^2.$$

For the upper offdiagonal elements ($j > i$), we get

$$[W_h^{-1}]_{ij} = \sum_{k=1}^{i}[\tilde{V}^{-1}]_{ik}[\tilde{V}^{-1}]_{jk}^T = [\tilde{V}^{-1}]_{ii}[\tilde{V}^{-1}]_{ji}^T + \sum_{k=1}^{i-1}[\tilde{V}^{-1}]_{ik}[\tilde{V}^{-1}]_{jk}^T,$$

and thus

$$\begin{aligned}\|[W_h^{-1}]_{ij}\|_2 &\leq \|[\tilde{V}^{-1}]_{ii}\|_2\|[\tilde{V}^{-1}]_{ji}^T\|_2 + \sum_{k=1}^{i-1}\|[\tilde{V}^{-1}]_{ik}\|_2\|[\tilde{V}^{-1}]_{jk}\|_2 \\ &\leq C_{\hat{V}^{-1}}h + (i-1)C_{\hat{V}^{-1}}^2 h^2.\end{aligned}$$

From the symmetry of W_h^{-1} it follows that for $i > j$,

$$\|[W_h^{-1}]_{ij}\|_2 = \|[W_h^{-1}]_{ji}\|_2 \leq C_{\hat{V}^{-1}}h + (j-1)C_{\hat{V}^{-1}}^2 h^2.$$

Combining these estimates and using the fact that $\|A\|_\infty \leq \sqrt{n}\ \|A\|_2$ for all $A \in \mathbb{R}^{n,n}$ finally leads to

$$\begin{aligned}\|W_h^{-1}\|_\infty &\leq \max_{i=1,\ldots,N}\sum_{j=1}^{N}\|[W_h^{-1}]_{ij}\|_\infty \\ &\leq \sqrt{n}\ \max_{i=1,\ldots,N}\sum_{j=1}^{N}\|[W_h^{-1}]_{ij}\|_2 \\ &\leq \sqrt{n}\ \left(1 + (N-1)C_{\hat{V}^{-1}}h + \max_i\sum_{j=1}^{N}(\min(i,j)-1)C_{\hat{V}^{-1}}^2 h^2\right) \\ &\leq \sqrt{n}\ \left(1 + C_{\hat{V}^{-1}}(T-t_0) + \sum_{j=1}^{N}(j-1)C_{\hat{V}^{-1}}^2 h^2\right) \\ &\leq \sqrt{n}\ \left(1 + C_{\hat{V}^{-1}}(T-t_0) + \frac{(N-1)(N-2)}{2}C_{\hat{V}^{-1}}^2 h^2\right) \\ &\leq \sqrt{n}\ \left(1 + C_{\hat{V}^{-1}}(T-t_0) + \frac{1}{2}(T-t_0)^2 C_{\hat{V}^{-1}}^2\right),\end{aligned}$$

which implies (A.14) with $C_{W^{-1}} = \sqrt{n}\ (1 + C_{\hat{V}^{-1}}(T-t_0) + \frac{1}{2}(T-t_0)^2 C_{\hat{V}^{-1}}^2)$. □

Appendix B

Software

Algorithms 1 and 3 have been implemented in Fortran routines, according to the Fortran 77 standards. The user interface has a similar design as the interface of the GELDA software package [33]. All linear algebra operations, such as computing QR and LQ decompositions, are performed by calls to BLAS [9] and LAPACK [1]. In the following, we list and briefly explain the individual routines of our implementation. For further details, we refer to the inline documentation.

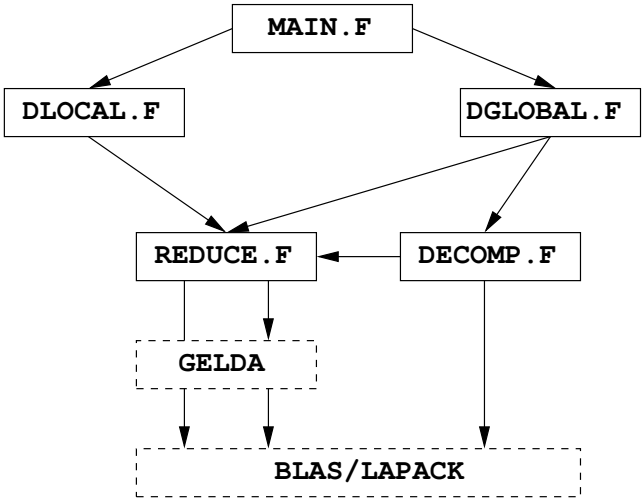

Figure B.1: Graph of dependencies between the implemented Fortran routines.

`MAIN.F` This is a driver routine, which allows to conveniently call the corresponding routines for performing the local and global minimization algorithms. The user must provide three routines which evaluate the coefficient functions $E(\cdot)$, $A(\cdot)$ and $f(\cdot)$ at an arbitrary time point, as well as the corresponding derivatives of order $1, \ldots, s$, where s is an upper bound on the strangeness index of the underlying DAE. Optionally, the computed results can be compared with a reference solution.

`DLOCAL.F` This is an implementation of the local minimization algorithm, Algorithm 1. Optionally, `REDUCE.F` is called, either to compute an equivalent strangeness free system or to reduce the size of the system by removing redundant equations.

`DGLOBAL.F` This routine initializes the global minimization algorithm and calls `DECOMP.F`.

`DECOMP.F` This routine performs parts of the global minimization algorithm, Algorithm 3. Optionally, `REDUCE.F` is called, either to compute an equivalent strangeness free system or to reduce the size of the system by removing redundant equations.

`REDUCE.F` Performs calls to slightly modified routines of GELDA to compute the reduced form (1.22) and to remove solvability conditions of the form $0 = f_i(t)$.

All routines along with some example programs are contained on the enclosed CD or available on request from the author.

Bibliography

[1] E. Anderson, Z. Bai, C. H. Bischof, S. Blackford, J. W. Demmel, J. J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. C. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999.

[2] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, volume 13 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1995.

[3] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[4] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, volume 14 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1996.

[5] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numer. Math.*, 60:1–40, 1991.

[6] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18(4):1101–1115, 1987.

[7] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Pitman, San Francisco, 1979.

[8] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Springer, 2002.

[9] J. J. Dongarra, J. Du Croz, I. S. Duff, and S. Hammarling. A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Software*, 16:1–17, 1990.

[10] D. Estévez-Schwarz and C. Tischendorf. Structural analysis for electrical circuits and consequences for MNA. *Int. J. Circ. Theor. Appl.*, 28:131–162, 2000.

[11] C. Führer. *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen*. Dissertationsschrift, Mathematisches Institut, Universität München, 1988.

[12] C. Führer and B. J. Leimkuhler. Numerical solution of differential-algebraic equations for constrained mechanical motion. *Numer. Math.*, 59:55–69, 1991.

[13] C. W. Gear and L. R. Petzold. Differential/algebraic systems and matrix pencils. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, pages 75–89. Springer-Verlag, 1983.

[14] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, 2002.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.

[16] E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment*. Teubner Texte zur Mathematik. Teubner-Verlag, Leipzig, 1986.

[17] A. Griewank. *Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, 2000.

[18] M. Günther, M. Hoschek, and P. Rentrop. Differential-algebraic equations in electrical circuit simulation. *Int. J. Electron. Commun.*, 54:101–107, 2000.

[19] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Lecture Notes in Mathematics No. 1409. Springer-Verlag, Berlin, 1989.

[20] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd ed.* Springer Verlag, Berlin, 1996.

[21] M. R. Hestenes. *Calculus of Variations and Optimal Control Theory*. John Wiley & Sons, New York, 1966.

[22] H. Heuser. *Funktionalanalysis*. B. G. Teubner, 1992.

[23] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, second edition, 2002.

[24] I. Higueras, R. März, and C. Tischendorf. Stability preserving integration of index-2 DAEs. *Appl. Numer. Math.*, 45:201–229, 2003.

[25] A. Ilchmann and V. Mehrmann. A behavioural approach to time-varying linear systems, part 1: General theory, 2005. To appear in SIAM J. Cont. Optim.

[26] A. Ilchmann and V. Mehrmann. A behavioural approach to time-varying linear systems, part 2: Descriptor systems, 2005. To appear in SIAM J. Cont. Optim.

[27] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comput. Appl. Math.*, 56:225–259, 1994.

[28] P. Kunkel and V. Mehrmann. Generalized inverses of differential-algebraic operators. *SIAM J. Matrix Anal. Appl.*, 17:426–442, 1996.

[29] P. Kunkel and V. Mehrmann. Local and global invariants of linear differential-algebraic equations and their relation. *Electr. Trans. Num. Anal.*, 4:138–157, 1996.

[30] P. Kunkel and V. Mehrmann. A new class of discretization methods for the solution of linear differential algebraic equations with variable coefficients. *SIAM J. Numer. Anal.*, 33(5):1941–1961, 1996.

[31] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations. Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006. To appear.

[32] P. Kunkel, V. Mehrmann, and W. Rath. Analysis and numerical solution of control problems in descriptor form. *Math. Control, Signals, Sys.*, 14:29–61, 2001.

[33] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert. GELDA: A software package for the solution of general linear differential algebraic equations. *SIAM J. Sci. Comput.*, 18:115 – 138, 1997.

[34] R. März. A matrix chain for analyzing differential-algebraic equations. Preprint 162, Sektion Mathematik, Humboldt-Universität zu Berlin, 1987.

[35] V. Mehrmann. *The Autonomous Linear Quadratic Control Problem.* Springer-Verlag, Berlin, 1991.

[36] V. Mehrmann and W. Rath. Numerical methods for the computation of analytic singular value decompositions. *Electr. Trans. Num. Anal.*, 1:72–88, 1993.

[37] L. R. Petzold. A description of DASSL: A differential/algebraic system solver. In R. S. Stepleman et al., editors, *IMACS Trans. Scientific Computing Vol. 1*, pages 65–68. North-Holland, Amsterdam, 1983.

[38] J. W. Polderman and J. .C. Willems. *Introduction to Mathematical Systems Theory: A Behavioural Approach.* Springer Verlag, New York, 1998.

[39] P. J. Rabier and W. C. Rheinboldt. Classical and generalized solutions of time-dependent linear differential-algebraic equations. *Linear Algebra Appl.*, 245:259–293, 1996.

[40] P. J. Rabier and W. C. Rheinboldt. *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint.* SIAM, Philadelphia, PA 19104-2688, USA, 2000.

[41] C. Tischendorf. *Solution of index-2-DAEs and its application in circuit simulation.* Dissertation, Humboldt-Univ. zu Berlin, 1996.