ASC Report No. 18/2010

An Efficient Asymptotically Correct Error Estimator for Collocation Solutions to Singular Index-1 DAEs

Winfried Auzinger, Herbert Lehner, Ewa Weinmüller

Institute for Analysis and Scientific Computing
Vienna University of Technology — TU Wien
www.asc.tuwien.ac.at ISBN 978-3-902627-03-2

Most recent ASC Reports

17/2010 *Bertram Düring, Michel Fournié*
High-Order Compact Finite Difference Scheme for Option Pricing in Stochastic Volatility Models

16/2010 *Robert Hammerling, Othmar Koch, Christa Simon, Ewa B. Weinmüller*
Numerical Treatment od Singular ODE EVPs Using `bvpsuite`

15/2010 *Ansgar Jüngel, René Pinnau, Elisa Röhrig*
Analysis of a Bipolar Energy-Transport Model for a Metal-Oxide-Semiconductor Diode

14/2010 *Markus Aurada, Michael Ebner, Michael Feischl, Samuel Ferraz-Leite, Petra Goldenits, Michael Karkulik, Markus Mayr, Dirk Praetorius*
Hilbert (Release 2): A MATLAB Implementation of Adaptive BEM

13/2010 *Alexander Dick, Othmar Koch, Roswitha März, Ewa Weinmüller*
Convergence of Collocation Schemes for Nonlinear Index 1 DAEs with a Singular Point

12/2010 *Markus Aurada, Samuel Ferraz-Leite, Petra Goldenits, Michael Karkulik, Markus Mayr, Dirk Praetorius*
Convergence of Adaptive BEM for some Mixed Boundary Value Problem

11/2010 *Mario Bukal, Ansgar Jüngel, Daniel Matthes*
Entropies for Radially Symmetric Higher-Order Nonlinear Diffusion Equations

10/1010 *Karl Rupp, Ansgar Jüngel, Karl-Tibor Grasser*
Matrix Compression for Spherical Harmonics Expansions of the Boltzmann Transport Equation for Semiconductors

9/2010 *Markus Aurada, Samuel Ferraz-Leite, Dirk Praetorius*
Estimator Reduction and Convergence of Adaptive BEM

8/2010 *Robert Hammerling, Othmar Koch, Christa Simon, Ewa B. Weinmüller*
Numerical Solution of singular Eigenvalue Problems for ODEs with a Focus on Problems Posed on Semi-Infinite Intervals

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstraße 8–10
1040 Wien, Austria

E-Mail: `admin@asc.tuwien.ac.at`
WWW: `http://www.asc.tuwien.ac.at`
FAX: `+43-1-58801-10196`

ISBN 978-3-902627-03-2

ASC
TU WIEN

An efficient asymptotically correct error estimator for collocation solutions to singular index-1 DAEs

Winfried Auzinger, Herbert Lehner, Ewa Weinmüller

Abstract A computationally efficient a posteriori error estimator is introduced and analyzed for collocation solutions to linear index-1 differential-algebraic equations with properly stated leading term exhibiting a singularity of the first kind. The procedure is based on a modified defect correction principle, extending an established technique from the context of ordinary differential equations to the differential-algebraic case. Using recent convergence results for collocation methods, we prove that the resulting error estimate is asymptotically correct. Numerical examples demonstrate the performance of this approach. To keep the presentation reasonably self-contained, some arguments from the literature on differential-algebraic equations concerning the decoupling of the problem and its discretization, which is essential for our analysis, are also briefly reviewed. The appendix contains a remark about the interrelation between collocation and implicit Runge-Kutta methods for differential-algebraic equations.

Keywords Differential algebraic equations · singularity of the first kind · collocation · a posteriori error estimation · defect correction

PACS 02.60.Cb · 02.60.Lj · 02.70.Jn

Mathematics Subject Classification (2000) 65L80 · 65B05

1 Introduction

Differential-algebraic equations (DAEs) are a powerful tool in system modeling, analysis and simulation in various application fields as, for instance, electronic circuits, constrained mechanics, PDE semi-discretizations and control theory. Therefore, they have been intensively studied during the last 30 years. Most of the previous work is devoted to numerical integration methods for low index DAEs in Hessenberg form and to reduction techniques for smooth higher index DAEs via derivative arrays. Much effort is put into the development of numerical algorithms and solution of problems from industrial applications.

In the past, collocation methods have been successfully used by several authors to solve well-posed boundary value problems (BVPs) for index-1 DAEs without singularities. In [9] and [20] nonlinear systems of DAEs were studied and superconvergence results for Gaussian and Lobatto points were derived. First attempts to provide respective software go back to 1994, see [2]. Semi-explicit problems, where the algebraic constraints are separately specified as a set of nonlinear equations, are in the scope of the collocation code COLDAE. Collocation methods applied to solve linear and nonlinear BVPs for index-1 DAEs without singularities have recently been analyzed in [21] and [22]. Here, the system is assumed to be given in a form of separated sets of equations involving derivatives and derivative-free equations. Collocation at Gaussian or Lobatto points is used to treat these separated subsets of equations. In the approach of [21], an index reduction technique for higher index systems is used, aiming at the reduction of the DAEs system to the index-1 form, and then collocation is applied.

This brief overview indicates that much progress has been made concerning DAE theory and applications, but there are still many open questions.

A few years ago, an important new theoretical concept, DAEs *with properly stated leading term*, has been introduced and studied in [8,15,23]. This enables a proper and natural description of the solution derivatives involved. In particular, in the linear case one considers DAEs written in the form (1), (8) below, with continuous coefficient matrices $A(t) \in \mathbb{R}^{m\times n}$, $D(t) \in \mathbb{R}^{n\times m}$, $B(t) \in \mathbb{R}^{m\times m}$. The matrix functions A and D are supposed to have constant rank and to be well-matched in the sense that the intersection of the kernel of A and the range of D is empty, and their direct sum spans the entire $\mathbb{R}^n$. This setting rigorously indicates which derivatives of the unknown solution $x(t)$ are indeed involved, and it becomes fully natural to consider solutions $x(t)$ being continuous and having continuously differentiable components Dx, i.e.,, solutions from the linear function space $C^1_D := \{w \in C\colon Dw \in C^1\}$. One of the advantages of this precise description of the problem structure is that now there exists a uniquely determined (by the problem data) *inherent explicit ODE*, see [15–17].

The numerical treatment of critical points and singularities is rather in its beginning. In contrast to regular points of the DAE studied in [25,26], so-called *critical points* have been subdivided in [27,28] into A- and B-critical points. Further specification can be made in accordance to the level at which they arise in the matrix sequence. For DAEs with A-critical points, a decoupling procedure is given in [27]. By means of a modification of the projector construction described in [29], the projector-based decoupling is extended to DAEs with B-critical points in [28]. First results concerning the numerical integration of DAEs with somehow *harmless* critical points are reported in [12]. Roughly speaking, harmless critical points do not result in a singularity of the inherent ODE.

From the above discussion, it is clear that the numerical treatment of DAEs involving a singularity within its inherent ODE system is now on the agenda, and this paper shall constitute a contribution to the better understanding of this problem class. We use a *theoretical decoupling* of the DAE system to obtain the necessary information about the very sensitive dynamics related to the singularity, which will be necessary for the successful numerical analysis. We already have seen this in case of linear index-1 systems with singularities, cf. [19], where the information on the problem structure was crucial for better understanding of the numerical behavior in the present nonstandard situation.

This paper is concerned with a topic which has not received much attention so far, namely the design and analysis of an efficient, asymptotically correct a posteriori error estimator for collocation solutions to such a class of singular DAE systems. We make use of the convergence results from [5] and [19], and extend established defect-based techniques from the ODE case (see [4,5]) to the DAE setting. Our focus is on the index-1 case with a singularity of the first kind. The regular index-1 case can be seen as a simple special case, which is also considered in [7].

2 Problem setting

2.1 DAEs with properly stated leading term

We consider linear systems of DAEs of index 1 of the form

$$\boldsymbol{A}(t)(\boldsymbol{D}(t)\boldsymbol{x}(t))' + \boldsymbol{B}(t)\boldsymbol{x}(t) = \boldsymbol{g}(t), \quad t \in (0,1], \tag{1}$$

with appropriately smooth data $\boldsymbol{A}(t) \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$, $\boldsymbol{D}(t) \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}}$, $\boldsymbol{B}(t) \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$.

In our analysis, assume that (1) is well-posed as an initial value problem (IVP), admitting a certain singular behavior in the 'inherent ODE', see Sect. 2.2 for the detailed specifications. Furthermore we assume

$$\boldsymbol{m} > \boldsymbol{n}, \quad \text{and} \quad \ker \boldsymbol{A}(t) = \{\boldsymbol{0}\}, \quad \operatorname{im} \boldsymbol{D}(t) = \mathbb{R}^{\boldsymbol{n}}, \quad t \in (0,1]. \tag{2}$$

Remark. Properties (2) imply that $(\boldsymbol{AD})(t) \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$ is a singular matrix, with $\operatorname{rank}(\boldsymbol{AD})(t) \equiv \boldsymbol{n}$. The assumptions can be weakened in such a way that $\boldsymbol{A}(t)$ and $\boldsymbol{D}(t)$ themselves do not need to have full, but constant rank, replacing the requirement in (2) by $\ker \boldsymbol{A}(t) \oplus \operatorname{im} \boldsymbol{D}(t) = \mathbb{R}^{\boldsymbol{n}}$, cf., e.g., [30]. However, as in [19], we restrict our discussion to the standard case (2).

Our restriction to linear problems is reasonable because it enables the underlying ideas to be discussed precisely with moderate technical effort. In particular, the focus is on the effective design and analysis of an asymptotically correct a posteriori error estimator for collocation solutions to (1), with a uniform, 'black box' treatment of the differential and algebraic components and an appropriate handling of the case where $\boldsymbol{D}(t)$ is not constant. The generalization of the method and its analysis for nonlinear and/or higher index problems is out of the scope of this presentation.

The above assumptions mean that the system (1) is *properly stated*, cf. e.g. [8,14,15,21,23,30], which implies the existence of a unique, natural decoupling into an *inherent ODE* and associated algebraic components. In particular, we assume that the system has *tractability index 1*; see Sect. 2.2. First results for case where the inherent ODE is regular has been discussed in [7]. The aim of the present paper is to extend these results to the case where the inherent ODE has a singularity of the first kind at $t = 0$. □

Introducing $\boldsymbol{u}(t) = \boldsymbol{D}(t)\boldsymbol{x}(t)$ as a separate variable, we obtain the dilated system

$$\boldsymbol{A}(t)\boldsymbol{u}'(t) + \boldsymbol{B}(t)\boldsymbol{x}(t) = \boldsymbol{g}(t), \tag{3}$$

$$\boldsymbol{u}(t) - \boldsymbol{D}(t)\boldsymbol{x}(t) = \boldsymbol{0}, \tag{4}$$

which is equivalent to (1). In this formulation, it is obvious what initial or boundary conditions lead to a well-posed problem. For an initial value problem (IVP), only $\boldsymbol{x}(0)$ can be prescribed, while $\boldsymbol{u}(0)$ is uniquely fixed by the algebraic relation $\boldsymbol{u}(0) = \boldsymbol{D}(0)\boldsymbol{x}(0)$.

Note that the system (3),(4) can again be written in the original form (1),

$$\hat{\boldsymbol{A}}(t)(\hat{\boldsymbol{D}}(t)\hat{\boldsymbol{x}}(t))' + \hat{\boldsymbol{B}}(t)\hat{\boldsymbol{x}}(t) = \hat{\boldsymbol{g}}(t), \quad t \in (0,1], \tag{5}$$

with

$$\begin{aligned} \hat{\boldsymbol{A}}(t) &= \begin{pmatrix} \boldsymbol{A}(t) \\ \boldsymbol{0}_{n\times n} \end{pmatrix}, \quad \hat{\boldsymbol{D}}(t) = \begin{pmatrix} \boldsymbol{0}_{n\times m} \boldsymbol{I}_{n\times n} \end{pmatrix}, \\ \hat{\boldsymbol{B}}(t) &= \begin{pmatrix} \boldsymbol{B}(t) & \boldsymbol{0}_{m\times n} \\ \boldsymbol{D}(t) & -\boldsymbol{I}_{n\times n} \end{pmatrix}, \quad \hat{\boldsymbol{g}}(t) = \begin{pmatrix} \boldsymbol{g}(t) \\ \boldsymbol{0}_n \end{pmatrix}, \quad \hat{\boldsymbol{x}}(t) = \begin{pmatrix} \boldsymbol{x}(t) \\ \boldsymbol{u}(t) \end{pmatrix}, \end{aligned} \tag{6}$$

where, in particular, $\hat{\boldsymbol{D}} := \hat{\boldsymbol{D}}(t) \in \mathbb{R}^{\hat{\boldsymbol{n}}\times\hat{\boldsymbol{m}}}$ is now a constant matrix, $\hat{\boldsymbol{A}}(t) \in \mathbb{R}^{\hat{\boldsymbol{m}}\times\hat{\boldsymbol{n}}}, \hat{\boldsymbol{B}}(t) \in \mathbb{R}^{\hat{\boldsymbol{m}}\times\hat{\boldsymbol{m}}}, \hat{\boldsymbol{m}} = \boldsymbol{m} + \boldsymbol{n}$, $\hat{\boldsymbol{n}} = \boldsymbol{n}$. The analog of (2),

$$\hat{\boldsymbol{m}} > \hat{\boldsymbol{n}}, \quad \text{and} \quad \ker \hat{\boldsymbol{A}}(t) = \{\boldsymbol{0}\}, \quad \operatorname{im} \hat{\boldsymbol{D}}(t) = \mathbb{R}^{\hat{\boldsymbol{n}}} \tag{7}$$

is also satisfied.

In the sequel, we will use the symbols A, B, D, x, g, m, n as a generic denotation for either $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{x}, \boldsymbol{g}, \boldsymbol{m}, \boldsymbol{n}$ or $\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{g}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{n}}$ wherever the respective statements apply to both sets of variables, under the assumption that D is constant. The same convention applies to symbols introduced below like G, Q, M, p. Applying this convention, we may write (1) or (5), respectively, as

$$A(t)(Dx(t))' + B(t)x(t) = g(t), \quad t \in (0,1], \quad \text{with } D = \text{const.} \tag{8}$$

In other words: (8) represents the original system (1) if $\boldsymbol{D}(t) = \boldsymbol{D} = \text{const.}$ In general, (8) is to be identified with (5) where $\hat{\boldsymbol{D}}(t) = \hat{\boldsymbol{D}} = \text{const.}$ by construction. Actually, this latter property is the motivation for rewriting the system in the form (5), especially in view of favorable application of collocation methods, see Sect. 3 and Sect. 6 for a further discussion. For $\boldsymbol{D}(t) = \text{const.}$, the two interpretations of (8) are equivalent, also with respect to numerical approximation.

2.2 Decoupling of the index-1 DAE

Let $Q \in \mathbb{R}^{m\times m}$ be any projector onto $\ker D$ and define

$$G(t) := A(t)D + B(t)Q \in \mathbb{R}^{m\times m}. \tag{9}$$

We assume that the system (8), with $D \in \mathbb{R}^{n\times m}$, $D = \text{const.}$, satisfies assumptions specified in (2). Furthermore we are assuming that the system has tractability index 1, which, by definition means that $G(t)$ *is invertible*[1] for $t \in (0,1]$ (in contrast to $A(t)D$). Due to $G(t)Q = A(t)DQ + B(t)Q^2 = B(t)Q$ we have $G^{-1}(t)A(t)D = I_{m\times m} - G^{-1}(t)B(t)Q = I_{m\times m} - Q$, which is the key identity for the theoretical (semi-)decoupling of (9).

Let $D^- \in \mathbb{R}^{m\times n}$ be a generalized reflexive inverse of D (cf. [21,30,33]) such that $D^-D = I_{m\times m} - Q$ and $DD^- = I_{n\times n}$. Furthermore we denote

$$M(t) := -t\,G^{-1}(t)B(t)D^- \in \mathbb{R}^{m\times n}. \tag{10}$$

With $G^{-1}(t)B(t)Q = I_{m\times m} - D^-D = Q = Q^2$ and $DQ = 0$ we now obtain for $t \in (0,1]$:

$$\begin{pmatrix} DG^{-1} \\ QG^{-1} \end{pmatrix} \cdot \begin{pmatrix} AD & B & -I_{m\times m} \end{pmatrix} = \begin{pmatrix} D & -\frac{1}{t}DMD & -DG^{-1} \\ 0 & Q-\frac{1}{t}QMD & -QG^{-1} \end{pmatrix}. \tag{11}$$

Applying this identity decouples (8) into an ODE for $Dx(t)$ and an algebraic equation expressing $Qx(t)$ in terms of $Dx(t)$,

$$Dx'(t) - \frac{1}{t}DM(t)Dx(t) = DG^{-1}(t)g(t), \tag{12}$$

$$Qx(t) - \frac{1}{t}QM(t)Dx(t) = QG^{-1}(t)g(t), \tag{13}$$

for $t \in (0,1]$. Here, Dx and Qx can be considered as separate variables, as Q projects onto $\ker D$. The solution of (8) can then be represented as

$$x(t) = (I_{m\times m} - Q)x(t) + Qx(t) = D^-Dx(t) + Qx(t), \tag{14}$$

where $Dx(t)$ is the solution of (12).

2.3 Technical assumptions

With the above definitions we are able to specify the problem class under consideration precisely, following [19]. We assume that the inherent ODE system (12) is well-posed, featuring a *singularity of the first kind* at $t = 0$. This requires $DM(t) \in C[0,1]$, and we also will assume a sufficient degree of differentiability. For such a singular ODE, the solvability of an IVP ensuring a sufficiently smooth solution requires certain restrictions on the spectrum of $DM(0)$ (cf., e.g., [10]). In particular, for the IVP to be well posed, the eigenvalues of $DM(0)$ are not allowed to have positive real parts. Furthermore, the initial condition must satisfy $Dx(0) = a \in \ker DM(0)$. Of course, this initial condition must also be consistent with the algebraic condition (13) at $t = 0$.

To keep the presentation within a reasonable format, we also assume that $\lambda = 0$ can only be an semisimple eigenvalue of $DM(0)$. Otherwise, error bounds in Theorem 1 (Sect. 5) would in general be affected by an additional factor $|\ln(h)|^{n_0-1}$, where n_0 is the dimension of the largest Jordan block for $\lambda = 0$. This is a consequence of the convergence theory for collocation methods for singular problems, cf., e.g., [5,11], and the arguments given in Sect. 6 would have to be modified appropriately.

In the singular case, $G(t)$ from (9) is not invertible at $t = 0$, with $G^{-1}(t)$ unbounded at $t = 0$. As explicated in [19, Sect. 2.1], for the given problem to be well-posed a sufficient condition is that

$$t\,G^{-1}(t), \quad G^{-1}(t)g(t), \quad \text{and} \quad Q_{\text{can}}(t) := QG^{-1}(t)B(t) \tag{15}$$

[1] Under assumption (2), $\hat{\boldsymbol{G}}(t) = \hat{\boldsymbol{A}}(t)\hat{\boldsymbol{D}} + \hat{\boldsymbol{B}}(t)\hat{\boldsymbol{Q}}$ is invertible iff $\boldsymbol{G}(t) = \boldsymbol{A}(t)\boldsymbol{D} + \boldsymbol{B}(t)\boldsymbol{Q}$ is invertible, which can be seen as follows: Since the particular choice for $\hat{\boldsymbol{Q}}$ is irrelevant, we consider the natural orthogonal projector

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} \boldsymbol{I_{m\times m}} & \boldsymbol{0_{m\times n}} \\ \boldsymbol{0_{n\times m}} & \boldsymbol{0_{n\times n}} \end{pmatrix}, \quad \text{yielding} \quad \hat{\boldsymbol{G}} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{D} & \boldsymbol{0_{n\times n}} \end{pmatrix}.$$

Making use of the identities $\operatorname{im}\boldsymbol{Q} = \ker\boldsymbol{D}$, $\boldsymbol{D^-D} = \boldsymbol{I_{m\times m}} - \boldsymbol{Q}$ and $\boldsymbol{DD^-} = \boldsymbol{I_{n\times n}}$, we observe

$$\hat{\boldsymbol{G}} \cdot \underbrace{\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{D^-} \\ \boldsymbol{D} & \boldsymbol{0_{n\times n}} \end{pmatrix}}_{=:\,\hat{\boldsymbol{T}}} = \underbrace{\begin{pmatrix} \boldsymbol{G} & \boldsymbol{BD^-} \\ \boldsymbol{0_{n\times m}} & \boldsymbol{I_{n\times n}} \end{pmatrix}}_{=:\,\hat{\boldsymbol{H}}},$$

where $\hat{\boldsymbol{T}}$ is invertible (note that $\hat{\boldsymbol{T}}^2 = \hat{\boldsymbol{I}}$). This shows $\operatorname{rank}\hat{\boldsymbol{G}} = \operatorname{rank}\hat{\boldsymbol{H}}$; furthermore, $\hat{\boldsymbol{H}}$ is invertible iff $\boldsymbol{G}$ is invertible.

have continuous extensions to $[0,1]$ and are sufficiently differentiable on $[0,1]$. ($Q_{\rm can}(t)$ is the so-called canonical projector onto $\ker D$.) Together with (10), this also implies the analogous smoothness of

$$\frac{1}{t}\,QM(t) = -Q_{\rm can}(t)D^{-}. \tag{16}$$

These assumptions guarantee that for any continuous, smooth solution $Dx(t)$ of the inherent ODE (12) there exists a continuous and appropriately smooth solution of the entire DAE (8), which we denote by $x_*(t)$. Furthermore, higher-order convergence results apply for collocation methods considered below.

Conditions (15) are not strictly necessary. Cf. the discussion in [19]; see also Example 7.1.

A further assumption required in our analysis is that the matrix

$$G^{-1}(t+h)G(t) - I_{m\times m} \tag{17}$$

remains uniformly bounded for $h>0$ and $t \geq Ch$, where h plays the role of a stepsize in our numerical method.

3 Collocation methods

We consider collocation methods applied to the system (8), with $D = \text{const.}$. The collocation solution is a continuous piecewise polynomial function $p(t)$ of degree $\leq s$ which satisfies the given initial condition and obeys, in a pointwise sense, the DAE (8) at the collocation nodes

$$t_{ij} := \tau_i + c_j h_i, \tag{18}$$

where

$$\tau_0 = 0, \quad h_i := \tau_{i+1} - \tau_i > 0, \quad \tau_N = 1; \qquad 0 < c_1 < \ldots < c_s = 1, \tag{19}$$

for $i = 0,\ldots,N-1,\ j = 1,\ldots,s$. Note, in particular, that $c_s = 1$ is essential for our analysis. This means that the method is *stiffly accurate*, a property which in a natural way ensures stability of the scheme, cf. e.g. [15] for a more detailed discussion. We also denote $h := \max_{i=0,\ldots,N-1} h_i$.

The complete system of collocation equations reads

$$A(t_{ij})(Dp)'(t_{ij}) + B(t_{ij})p(t_{ij}) = g(t_{ij}), \quad i = 0,\ldots,N-1,\ j = 1,\ldots,s, \tag{20}$$

where $p(t)$ is represented by a polynomial $p_i(t)$ of degree $\leq s$ on each subinterval $[\tau_i, \tau_{i+1}]$, together with the continuity relations

$$p_{i-1}(\tau_i) = p_i(\tau_i), \quad i = 0,\ldots,N-1, \tag{21}$$

and $p_0(0)$ satisfying the initial conditions of the DAE (8). We also define

$$c_0 := 0, \quad t_{i0} := \tau_i, \quad i = 0,\ldots,N-1. \tag{22}$$

Remark. If we identify (8) with the DAE system rewritten in dilated form (3), the collocation conditions (20) are equivalent to

$$\boldsymbol{A}(t_{ij})\boldsymbol{q}'(t_{ij}) + \boldsymbol{B}(t_{ij})\boldsymbol{p}(t_{ij}) = \boldsymbol{g}(t_{ij}), \tag{23}$$

$$\boldsymbol{q}(t_{ij}) - \boldsymbol{D}(t_{ij})\boldsymbol{p}(t_{ij}) = \boldsymbol{0}, \tag{24}$$

together with the continuity conditions for $\boldsymbol{p}$ and $\boldsymbol{q}$ as in (21), where $p(t) = \hat{\boldsymbol{p}}(t) = (\boldsymbol{p}(t), \boldsymbol{q}(t))^T$, and $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ are piecewise polynomial approximations for $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$, respectively. Note that for $\boldsymbol{D}(t) \not\equiv \text{const.}$ we have $\boldsymbol{D}(t)\boldsymbol{p}_i(t) \not\equiv \boldsymbol{q}_i(t)$, and therefore the collocation scheme (23) cannot be interpreted as collocation applied directly to (1) because, in general,

$$\boldsymbol{A}(t_{ij})\boldsymbol{D}\boldsymbol{p}_i{}'(t_{ij}) + \boldsymbol{B}(t_{ij})\boldsymbol{p}(t_{ij}) \neq \boldsymbol{g}(t_{ij}). \tag{25}$$

On the other hand, (23) can be interpreted as as a well-established implicit Runge-Kutta (IRK) method for (1), see Appendix A. Therefore, the convergence results from [14] for IRK methods apply, at least for the regular case. These results are based on a natural decoupling of the problem based on its properly stated formulation and an analogous decoupling of the discrete scheme. The analysis of our error estimator in Sect. 6 will also be based on such a decoupling procedure, where we make use of the techniques and convergence results from [5, 19]. □

4 Defect-based error estimation

A posteriori error estimation in ODEs based on the defect correction principle is an old idea originally due to Zadunaisky [32], which was further developed by Stetter [31]. In the context of regular and singular ODEs, this approach was refined and analyzed in [4,5] and implemented in [3]. In particular, for a special realization of the defect, an efficient, asymptotically correct error estimator, the QDeC estimator, was designed in [4] for collocation solutions on arbitrary grids. These ideas will now be extended to the DAE context, permitting a singularity of the first kind in the inherent ODE. As will be seen below, this is not straightforward because of the coupling between differential and algebraic components. A detailed analysis of the proposed estimator is given for the linear index-1 case (1).

In abstract notation, the basic structure of a defect-based error estimator can be described as follows: Consider a numerical solution $x_\Delta \approx x_*$ for a problem

$$F(x(t)) = 0, \quad t \in [0,1], \tag{26}$$

on a grid Δ. Define the *defect* $d = d(t)$ by interpolating x_Δ by a continuous piecewise polynomial function $p(t)$ of degree $\leq s$ and substituting $p(t)$ into (26),

$$d(t) := F(p(t)), \quad t \in [0,1]. \tag{27}$$

Obviously, $p(t)$ is the exact solution to a 'neighboring problem'

$$F(x(t)) = d(t) \tag{28}$$

related to the original problem (26). Now use a procedure of low effort (typically a low order scheme), the so-called *auxiliary scheme* $\tilde{F}$, to obtain approximate discrete solutions $\tilde{x}_\Delta$ and $\tilde{x}_\Delta^d$ for both the original and neighboring problems on the grid Δ, i.e. $\tilde{F}(\tilde{x}_\Delta) = 0$ and $\tilde{F}(\tilde{x}_\Delta^d) = d_\Delta$, where d_Δ is an appropriate restriction of $d(t)$ to the grid Δ.

Since (26) and (28) differ only by the (presumably) small defect d, we expect that the global error

$$e := x_\Delta - x_*. \tag{29}$$

is well approximated by the computable a posteriori estimate

$$\varepsilon_\Delta := \tilde{x}_\Delta^d - \tilde{x}_\Delta. \tag{30}$$

This is exactly the procedure originally proposed in [31]. However, in concrete applications, the auxiliary scheme $\tilde{F}$ and a suitable representation for the defect d_Δ have to be carefully chosen. In [4] collocation for the ODE case was considered, where the defect $d(t)$ is directly defined in terms of the continuous collocation solution $p(t)$ (playing the role of x_Δ). If, for instance, $\tilde{F}$ is chosen as the backward Euler scheme, it is argued in [4] that use of the pointwise defect makes no sense; a modified, averaged version of the defect (27) has to be used in order to obtain an asymptotically correct estimate. In the following section this approach (the 'QDeC estimator') is described in more detail and is extended to the DAE case.

For linear problems, $F(x) = Lx + g$, the procedure can be realized in a simpler way, computing ε_Δ by a single application of the corresponding auxiliary scheme,

$$\tilde{L}\,\varepsilon_\Delta = d_\Delta. \tag{31}$$

See also [6] for a discussion and for further variants of this approach.

5 The QDeC estimator for DAE systems

Now we apply the procedure described in Sect. 4 to the linear DAE system (8). In addition to the collocation method introduced in Sect. 3, we use a scheme of backward Euler type over the complete collocation grid as an auxiliary method. Let $h_{ij} := t_{ij} - t_{i,j-1}$. Then the grid function ε_{ij} satisfying the auxiliary scheme with a defect term $\bar{d}_{ij}$ on the right hand side, and homogeneous initial condition $\varepsilon_{00} = 0$,

$$A(t_{ij})\,\frac{D\varepsilon_{ij} - D\varepsilon_{i,j-1}}{h_{ij}} + B(t_{ij})\varepsilon_{ij} = \bar{d}_{ij}, \tag{32}$$

defines the estimator $\varepsilon_{ij} \approx e(t_{ij})$. The scheme (32) is the analog of (31), the backward Euler scheme playing the role of $\tilde{L}$.

According to (27), the straightforward, classical way to define the defect $\bar{d}_{ij}$ would be to substitute $p(t)$ into (8) in the pointwise sense,

$$d(t) := A(t)Dp'(t) + B(t)p(t) - g(t), \quad t \in (0,1], \tag{33}$$

and using this pointwise defect $d(t_{ij})$ for $\bar{d}_{ij}$ in (32). However, as has been pointed out in [4] in the ODE context, this procedure does not lead to successful results. For collocation this is obvious: Since, by definition of the collocation solution (20), the defect $d(t_{ij})$ vanishes at each point t_{ij} ($i = 0,\ldots,N-1,\ j = 1,\ldots,s$) that enters the scheme (32), the error estimate ε would vanish.

In slight variation of the procedure introduced in [4] we now define a modified defect via a polynomial quadrature approximation of degree s for the integral mean of d,

$$\bar{d}_{ij} := \sum_{k=0}^{s} \alpha_{jk}\, d(t_{ik}) \approx \frac{1}{h_{ij}} \int_{t_{i,j-1}}^{t_{ij}} d(t)\,dt \tag{34}$$

for $i = 0,\dots,N-1,\ j = 1,\dots,s$. The quadrature coefficients α_{jk} are given by

$$\alpha_{jk} = \frac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} L_k(t)\,dt, \quad j = 1,\dots,s,\ \ k = 0,\dots,s, \tag{35}$$

with the Lagrange polynomials $L_k(t)$ of degree s satisfying $L_k(c_j) = \delta_{jk}$. Note that, in contrast to collocation at s nodes in each subinterval excluding the left endpoint $t_{i0} = \tau_i + c_0 h_i$, we now also include the leftmost node $c_0 = 0$ (see (22)) for the polynomial quadrature defining (34).

Remark. As explained in [4], the modified defect (34) is related to the residual of the p_{ij} w.r.t. a higher order Runge-Kutta scheme of collocation type including the nodes t_{i0}. Originally, the motivation for defining the $\bar{d}_{ij}$ in this way comes from the ODE context. Actually, due to (20) the sum in (34) reduces to one term,

$$\bar{d}_{ij} = \alpha_{j0}\, d(t_{i0}), \tag{36}$$

at least if we ignore numerical errors in the computation of $p(t)$.

In the DAE case, where the algebraic relations are satisfied exactly at the collocation nodes, see (23), such a weighting of pointwise defect values seems to be suspicious at first sight. However, in the following theorem, which extends the results from [4,5], we show that the outcome is an asymptotically correct error estimate also in the DAE case. The essential observation is that a separate handling of differential and algebraic system components is not necessary. This lets us expect that the procedure will also be successfully applicable to problems with nonlinear coupling. □

In the sequel, $\|\,u(t)\,\|_\infty$ denotes the the sup-norm for functions $u = u(t) \in C[0,1]$, $\|\,u(t)\,\|_\infty := \sup_{t\in[0,1]} |u(t)|_\infty$, and analogously for grid functions and matrix-valued functions. The collocation error is denoted by $p(t) - x_*(t) =: e(t)$. We occasionally use the subscript denotation $[\cdots]_{ij}$ for any scalar, vector or matrix valued function as a shortcut for evaluation at the grid point t_{ij}.

We are now in the position to state our main result; the proof is given in Sect. 6.

Theorem 1 *Let the collocation degree s be even. While the global error of the collocation method* (20) *is of order h^s (see [19]), i.e.,*

$$\|e(t)\|_\infty = \|p(t) - x_*(t)\|_\infty = \mathcal{O}(h^s), \tag{37}$$

the estimate for the global error $e(t)$ based on the modified defect (34) *and the auxiliary scheme* (32) *is asymptotically correct, i.e., its deviation satisfies*

$$\|\varepsilon_{ij} - e(t_{ij})\|_\infty = \mathcal{O}(h^{s+1}). \tag{38}$$

6 Analysis of the error estimator

6.1 Preparations. Relations to prior work

The analysis given below is based on a theoretical decoupling argument and a study of the associated inherent ODE, according to the ideas from [8,14,15], see Sect. 2. Note that the assumption $D = \text{const.}$ is important here: It guarantees the problem to be *numerically well-formulated*, i.e., the original problem and the discrete schemes decouple in a parallel way, which is essential to ensure stable integration (cf., e.g., [15]).

A convergence theory for collocation methods applied to the problem at hand is presented in [19]. The following proof does not directly rely on [19]. Rather, the respective decoupling arguments are described from scratch, and our analysis of the differential component is based on the results from [5], which contains an analysis of the QDeC estimator for an ODE system with a singularity of the first kind. In particular, in the second step of the proof below we shall make use of the convergence result [5, Theorem 4.3], which can be directly applied to the case of a singularity of the first kind considered here. We explicitly refer to the arguments from [5], without reproducing all of the technical details. The algebraic components are then estimated separately.

However, concerning the analysis of the error estimator there is a minor difference to [5]. Namely, in the version of the QDeC estimator considered in [5] the left endpoint of the collocation subintervals is not involved in the defect quadrature (rather the right endpoint, which is not assumed to be a collocation node). Reformulating the respective arguments from [5] for our purpose is straightforward, with the exception that we now have to take special care about evaluations at the singular point $t = 0$. The respective technical details are explicated in the proof below.

Remark. The pointwise defect (33) can also be written as

$$d(t) = A(t)De'(t) + B(t)e(t). \tag{39}$$

In contrast to the error $e(t)$, the functions $e'(t)$ and $d(t)$ are not continuous at the endpoints t_{i0} of the collocation intervals. In the sequel, evaluation at $t = t_{i0}$ is to be understood as

$$e'(t_{i0}) := \lim_{t\downarrow t_{i0}} e'(t), \quad d(t_{i0}) := \lim_{t\downarrow t_{i0}} d(t), \tag{40}$$

□

6.2 Proof of Theorem 1

We proceed in several steps.

- *Decoupling the collocation equations* (20)

  The decoupling transformation (11) which led us to (12) also applies to the collocation equations (20), yielding

$$Dp'(t_{ij}) - \frac{1}{t_{ij}} DM(t_{ij}) Dp(t_{ij}) = DG^{-1}(t_{ij}) g(t_{ij}), \tag{41}$$

$$Qp(t_{ij}) - \frac{1}{t_{ij}} QM(t_{ij}) Dp(t_{ij}) = QG^{-1}(t_{ij}) g(t_{ij}), \tag{42}$$

  for $i = 0, \ldots, N-1,\ j = 1, \ldots, s$. Together with (12), which is satisfied by $x = x_*$, we see that the global error $e(t) = p(t) - x_*(t)$ of the collocation solution satisfies

$$De'(t_{ij}) - \frac{1}{t_{ij}} DM(t_{ij}) De(t_{ij}) = 0, \tag{43}$$

$$Qe(t_{ij}) - \frac{1}{t_{ij}} QM(t_{ij}) De(t_{ij}) = 0, \tag{44}$$

  for $i = 0, \ldots, N-1,\ j = 1, \ldots, s$, and with $e(0) = 0$.

- *A priori error estimates for the error* $e(t) = p(t) - x_*(t)$

  First we invoke the convergence result [5, Theorem 4.3], which can be directly applied to the case of a singularity of the first kind considered here, with $Dp(t)$, $DM(t)$, $DG^{-1}(t)g(t)$ playing the role of $p(t)$, $M(t)$, $f(t)$ from [5]. This shows uniform convergence on $[0,1]$ for the error in the differential component and its first derivative,

$$\|De(t)\|_\infty = \|Dp(t) - Dx_*(t)\|_\infty = \mathcal{O}(h^s), \tag{45}$$

$$\|De'(t)\|_\infty = \|Dp'(t) - Dx'_*(t)\|_\infty = \mathcal{O}(h^s). \tag{46}$$

  Furthermore, the uniform estimate

$$Dp'(t) - \frac{1}{t} DM(t) Dp(t) - DG^{-1}(t) g(t) = \mathcal{O}(h^s) \quad \text{for } t \in [0,1] \tag{47}$$

  follows from [5, Theorem 4.3, (4.18c)]. Together with (12) (with $x = x_*$) this also shows

$$De'(t) - \frac{1}{t} DM(t) De(t) = \mathcal{O}(h^s) \quad \text{for } t \in [0,1]. \tag{48}$$

  The second collocation equation (42) says that for each $i = 0, \ldots, N-1$, the polynomial $Qp_i(t)$ of degree $\le s$ interpolates $QG^{-1}(t)g(t) + \frac{1}{t} QM(t) Dp_i(t)$ at $t = t_{ij}$ for $j = 1, \ldots, s$. The additional interpolation property at $t_{i0} = \tau_i$ follows from the continuity condition (21) for $i = 1, \ldots, N-1$ and the initial condition for $i = 0$. The standard estimate for the error of Lagrange interpolation gives

$$Qp(t) - \frac{1}{t} QM(t) Dp(t) - QG(t)^{-1} g(t) = \mathcal{O}(h^{s+1}) \quad \text{for } t \in [0,1]. \tag{49}$$

  The differentiability required for the function interpolated follows from our smoothness assumptions, see (15),(16). Together with (13), $x = x_*$, this also shows

$$Qe(t) - \frac{1}{t} QM(t) De(t) = \mathcal{O}(h^{s+1}) \quad \text{for } t \in [0,1]. \tag{50}$$

  By (16) and (45), this implies the uniform a priori estimate

$$\|Qe(t)\|_\infty = \|Qp(t) - Qx_*(t)\|_\infty = \mathcal{O}(h^s). \tag{51}$$

- *Decoupling the auxiliary scheme* (32) *and the defect. Defect estimates*

  The decoupling transformation (11) applies to the auxiliary scheme (32) as well, resulting in

$$\frac{D\varepsilon_{ij} - D\varepsilon_{i,j-1}}{h_{ij}} - \frac{1}{t_{ij}} DM(t_{ij}) D\varepsilon_{ij} = DG^{-1}(t_{ij}) \bar{d}_{ij}, \tag{52}$$

$$Q\varepsilon_{ij} - \frac{1}{t_{ij}} QM(t_{ij}) D\varepsilon_{ij} = QG^{-1}(t_{ij}) \bar{d}_{ij}, \tag{53}$$

  with $D\varepsilon_{00} = Q\varepsilon_{00} = 0$.

Decoupling the pointwise defect (33) according to (11) for $t \in (0,1]$ leads to

$$DG^{-1}(t)d(t) = Dp'(t) - \frac{1}{t}DM(t)Dp(t) - DG^{-1}(t)g(t), \tag{54}$$

$$QG^{-1}(t)d(t) = Qp(t) - \frac{1}{t}QM(t)Dp(t) - QG^{-1}(t)g(t). \tag{55}$$

Subtracting (12) with $x = x_*$ from (54) yields

$$DG^{-1}(t)d(t) = De'(t) - \frac{1}{t}DM(t)De(t), \tag{56}$$

$$QG^{-1}(t)d(t) = Qe(t) - \frac{1}{t}QM(t)De(t), \tag{57}$$

for $t \in (0,1]$, which can also be considered as a decoupled DAE system for the $e(t)$. Together with (16) and (48), this implies the uniform estimates for $t \in (0,1]$,

$$\|DG^{-1}(t)d(t)\|_\infty = \mathscr{O}(h^s), \tag{58}$$

$$\|QG^{-1}(t)d(t)\|_\infty = \mathscr{O}(h^s), \tag{59}$$

which also imply[2]

$$\begin{aligned} \|d(t)\|_\infty = \|G(t)G^{-1}(t)d(t)\|_\infty &\le C\,\|G^{-1}(t)d(t)\|_\infty \\ &\le C\,\|D^-DG^{-1}(t)d(t)\|_\infty + C\,\|QG^{-1}(t)d(t)\|_\infty = \mathscr{O}(h^s) \end{aligned} \tag{60}$$

with $C = \|G(t)\|_\infty$.

The limits of (56) and (57) for $t \downarrow 0$ are given by[3]

$$\begin{aligned} \lim_{t\downarrow 0} DG^{-1}(t)d(t) &= De'(0) - \lim_{t\downarrow 0}\frac{DM(t)De(t) - DM(0)De(0)}{t-0} \\ &= De'(0) - \frac{d}{dt}\big(DM(t)De(t)\big)\big|_{t=0} = (I_{n\times n} - DM(0))De'(0) \end{aligned} \tag{61}$$

due to $De(0) = 0$, and

$$\lim_{t\downarrow 0} QG^{-1}(t)d(t) = 0 \tag{62}$$

due to $e(0) = 0$ and by virtue of the smooth extension property (16).

The pointwise defect in the algebraic component satisfies

$$QG^{-1}(t_{ik})d(t_{ik}) = 0, \quad \text{for all } i = 0,\dots,N-1,\ k = 0,\dots,s. \tag{63}$$

(For $k > 0$ this follows simply from the collocation identities $d(t_{ik}) = 0$. For $i > 0$ and $k = 0$ it is true due to $t_{i0} = t_{i-1,s}$; this is the point in the proof where assumption $c_s = 1$ is essential. For $i = k = 0$, $t_{00} = 0$, (63) is nothing but (62).)

- *Difference scheme for the differential component of the deviation* $\delta_{ij} := \varepsilon_{ij} - e(t_{ij})$

  We express the difference quotients of De and its estimate $D\varepsilon$ in terms of quadratures (with the coefficients from (34),(35)). Integration of $De'(t)$ using (56) gives[4]

$$\begin{aligned} \frac{De_{ij} - De_{i,j-1}}{h_{ij}} &= \frac{1}{h_{ij}}\int_{t_{i,j-1}}^{t_{ij}} De'(t)\,dt = \sum_{k=0}^{s}\alpha_{jk}De'(t_{ik}) + \mathscr{O}(h^{s+1}) \\ &= \sum_{k=0}^{s}\alpha_{jk}\Big(\frac{1}{t_{ik}}DM(t_{ik})De(t_{ik}) + DG^{-1}(t_{ik})d(t_{ik})\Big) + \mathscr{O}(h^{s+1}). \end{aligned} \tag{64}$$

  Furthermore, substituting (34) into (52) gives

$$\frac{D\varepsilon_{ij} - D\varepsilon_{i,j-1}}{h_{ij}} = \frac{1}{t_{ij}}DM(t_{ij})D\varepsilon_{ij} + DG^{-1}(t_{ij})\sum_{k=0}^{s}\alpha_{jk}\,d(t_{ik}). \tag{65}$$

[2] $\|d(t)\|_\infty = \mathscr{O}(h^s)$ also follows directly from (39).

[3] The analogous relation as in (61) also holds true for $G^{-1}(t)d(t)$.

[4] The quadrature formula (34),(35) is of degree s and, since we are approximating an integral mean, an error estimate of the form Ch^{s+1} holds, where C depends on the $(s+1)$-th derivative of the integrand, i.e., on $De^{(s+2)} = Dp^{(s+2)} - Dx_*^{(s+2)} = -Dx_*^{(s+2)}$ since p is of degree s.

Subtracting (64) from (65) we obtain, after simple rearrangement making use of $\sum_{k=0}^{s}\alpha_{jk}=1$, a system of difference equations of backward Euler type (analogous to (52),(65)), for the deviation $D\delta_{ij}=D\varepsilon_{ij}-De(t_{ij})$:

$$\begin{aligned}&\frac{D\delta_{ij}-D\delta_{i,j-1}}{h_{ij}}-\frac{1}{t_{ij}}DM(t_{ij})D\delta_{ij}\\&\quad=\sum_{k=0}^{s}\alpha_{jk}\Big(\frac{1}{t_{ij}}DM(t_{ij})De(t_{ij})-\frac{1}{t_{ik}}DM(t_{ik})De(t_{ik})\Big)+\\&\qquad+\sum_{k=0}^{s}\alpha_{jk}\big(DG^{-1}(t_{ij})-DG^{-1}(t_{ik})\big)d(t_{ik})+\mathscr{O}(h^{s+1}),\end{aligned}\tag{66}$$

or equivalently, with $d(t_{ik})=0$ for $k>0$,

$$\begin{aligned}&\frac{D\delta_{ij}-D\delta_{i,j-1}}{h_{ij}}-\frac{1}{t_{ij}}DM(t_{ij})D\delta_{ij}\\&\quad=\sum_{k=0}^{s}\alpha_{jk}\Big(\frac{1}{t_{ij}}DM(t_{ij})De(t_{ij})-\frac{1}{t_{ik}}DM(t_{ik})De(t_{ik})\Big)+\\&\qquad+\alpha_{j0}\big(DG^{-1}(t_{ij})-DG^{-1}(t_{i0})\big)d(t_{i0})+\mathscr{O}(h^{s+1}).\end{aligned}\tag{67}$$

The denotation in (64)–(67) is somewhat informal. Namely, the nontrivial evaluations occurring for $i,k=0,0$ (i.e., at the singular point $t=t_{00}=0$) have to be understood in the sense of $\lim_{t\downarrow 0}\ldots$, and these limits must be well defined and have to be studied. This is done in the next step of the proof.

- *Estimation of the right hand side of* (67)

  Here our aim is to identify the right hand side of (67) as an object of the type

$$\frac{1}{t_{ij}}DM(0)\,\mathscr{O}(h^{s+1})+\frac{1}{t_{ij}}\,\mathscr{O}(h^{s+1})+\mathscr{O}(h^{s+1})\tag{68}$$

  uniformly for $i=0,\ldots,N-1,\ j=1,\ldots,s$. To this end we separately estimate the contributions to the right hand side of (67).

  – First we consider the terms under the sum on the right hand side of (67), depending on $De(t)$. For $i>0,\ t_{ij}>t_{i0}>0$ we can argue in a similar way as in [5, (5.14)]: With $\tau=t_{ij}+\sigma(t_{ik}-t_{ij})$ and writing $M(t)$ in the form $M(t)=M(0)+\bar{M}_1(t)\cdot t$, where $\bar{M}_1(t)$ is assumed to be appropriately smooth, we obtain [5]

$$\begin{aligned}&\frac{1}{t_{ij}}DM(t_{ij})De(t_{ij})-\frac{1}{t_{ik}}DM(t_{ik})De(t_{ik})\\&\quad=\int_0^1\frac{d}{dt}\Big(\frac{1}{t}DM(t)De(t)\Big)\big|_{t=\tau}d\sigma\cdot(t_{ij}-t_{ik})\\&\quad=\int_0^1\Big(-\frac{1}{\tau^2}DM(0)De(\tau)+\frac{1}{\tau}DM(0)De'(\tau)+\frac{d}{dt}\big(\bar{M}_1(t)De(t)\big)\Big)\big|_{t=\tau}d\sigma\cdot(t_{ij}-t_{ik})\\&\quad=\frac{1}{t_{ij}}DM(0)\,\mathscr{O}(h^{s+1})+\mathscr{O}(h^{s+1}),\end{aligned}\tag{69}$$

    as desired. Here we have used the a priori estimates (45) and the fact that $\frac{1}{\tau}\le\frac{C}{t_{ij}}$.

    For the special case $t_{00}=0$ ($i=k=0$) we make use of existence of the limit $\lim_{t\downarrow 0}\frac{1}{t}DM(t)De(t)=DM(0)De'(0)$ (cf. (61)). For $t=t_{0j},\ j=1,\ldots,s$, this also yields the form (68),

$$\begin{aligned}&\frac{1}{t_{0j}}DM(t_{0j})De(t_{0j})-\lim_{t\downarrow 0}\frac{1}{t}DM(t)De(t)\\&\quad=\Big(\frac{1}{t_{0j}}DM(0)+D\bar{M}_1(t_{0j})\Big)\Big(De(0)+t_{0j}D\bar{e}_1(t_{0j})\Big)-\frac{1}{t_{0j}}\underbrace{t_{0j}DM(0)De'(0)}_{=\mathscr{O}(h^{s+1})}\\&\quad=\frac{1}{t_{0j}}DM(0)\,\mathscr{O}(h^{s+1})+\frac{1}{t_{0j}}\,\mathscr{O}(h^{s+1})+\mathscr{O}(h^{s+1})\end{aligned}\tag{70}$$

    by virtue of (45) and with $De(0)=0$, $t_{0j}=\mathscr{O}(h)$.

[5] Here and also in (70), $\bar{M}_1(t)$ and $\bar{e}_1(t)$ denote the Taylor remainders of first order after expanding $M(t)$ and $e(t)$ about $t=0$.

– Concerning the second term on the right hand side of (67), depending on $d(t)$, we also first consider the case $i>0,\ t_{ij}>t_{i0}>0$. We have

$$\begin{aligned}\big(DG^{-1}(t_{ij})-DG^{-1}(t_{i0})\big)d(t_{i0}) &= D\left(G^{-1}(\tau_i+c_jh_i)-G^{-1}(\tau_i)\right)d(\tau_i)\\ &= DH_{ij}G^{-1}(\tau_i)d(\tau_i),\end{aligned} \tag{71}$$

with $H_{ij}:=G^{-1}(\tau_i+c_jh_i)G(\tau_i)-I_{m\times m}$ uniformly bounded by assumption (17), and where

$$\begin{aligned}G^{-1}(\tau_i)d(\tau_i) &= (D^-D+Q)G^{-1}(\tau_i)d(\tau_i)=D^-DG^{-1}(\tau_i)d(\tau_i)\\ &= D^-\big(DG^{-1}(\tau_i)d(\tau_i)-DG^{-1}(\tau_i+c_jh_i)d(\tau_i+c_jh_i)\big)\end{aligned} \tag{72}$$

holds due to (63) and $d(\tau_i+c_jh_i)=0$. Relations (71) and (72) result in

$$G^{-1}(\tau_i)d(\tau_i)=DH_{ij}D^-\big(DG^{-1}(\tau_i)d(\tau_i)-DG^{-1}(\tau_i+c_jh_i)d(\tau_i+c_jh_i)\big). \tag{73}$$

Using representation (56) we can estimate the right hand side of (72) in a similar way as in (69) above.

For the special case $t_{00}=0$ ($i=k=0$) we again make use of existence of the limit (61), which implies

$$\lim_{t\downarrow 0} tDG^{-1}(t)\,d(0)=\lim_{t\downarrow 0} tDG^{-1}(t)d(t)=0. \tag{74}$$

Together with Taylor expansion of $tDG^{-1}(t)$ about $t=0$ this gives

$$tDG^{-1}(t)d(0)=\big(R_1(0)t+R_2(0)\frac{t^2}{2}+\mathscr{O}(t^3)\big)d(0),\quad t\downarrow 0, \tag{75}$$

where $R_1(t):=\frac{d}{dt}\left(tDG^{-1}(t)\right)$, $R_2(t):=\frac{d^2}{dt^2}\left(tDG^{-1}(t)\right)$, yielding

$$\begin{aligned}&\big(DG^{-1}(t_{0j})-\lim_{t\downarrow 0}DG^{-1}(t)\big)d(0)\\ &\quad=\big(R_2(0)\frac{t_{0j}}{2}+\mathscr{O}({t_{0j}}^2)\big)d(0)=\mathscr{O}(h^{s+1}).\end{aligned} \tag{76}$$

- *Estimation of the deviation δ_{ij}*

With the estimate (68) for the right hand side of the difference scheme (67) at hand, its solution can be estimated via a stability argument for the backward Euler scheme, cf., e.g., the proof of [4, Theorem 5.1]. The argument in this proof is based on [4, Theorem 4.2] which is applicable to (67) due to the estimates for its right hand side indicated in the previous step of the proof. In this way we end up with

$$\|D\delta_{ij}\|_\infty=\mathscr{O}(h^{s+1}). \tag{77}$$

Finally, to estimate the algebraic component $Q\delta_{ij}$, we subtract (44) from (52), which results in

$$Q\delta_{ij}=\frac{1}{t_{ij}}QM(t_{ij})D\delta_{ij}+QG^{-1}(t_{ij})\bar{d}_{ij}. \tag{78}$$

For the term describing the influence of the algebraic defect component in (78) we now obtain the uniform estimate

$$\begin{aligned}QG^{-1}(t_{ij})\bar{d}_{ij} &= \sum_{k=0}^{s}\alpha_{jk}\,QG^{-1}(t_{ij})d(t_{ik})\\ &= \sum_{k=0}^{s}\alpha_{jk}\big(QG^{-1}(t_{ij})-QG^{-1}(t_{ik})\big)d(t_{ik})+\sum_{k=0}^{s}\alpha_{jk}QG^{-1}(t_{ik})d(t_{ik})\\ &= \alpha_{j0}\big(QG^{-1}(t_{ij})-QG^{-1}(t_{ik})\big)d(t_{i0})+\alpha_{j0}\,QG^{-1}(t_{ik})d(t_{i0}),\end{aligned} \tag{79}$$

where the first term on the right hand side can be estimated in the same way as above for the differential component, and the second one vanishes due to (63).

Together with (16) and (77) this shows

$$\|Q\delta_{ij}\|_\infty=\mathscr{O}(h^{s+1}),\quad\text{and}\quad \|\delta_{ij}\|_\infty\le\|D^-D\,\delta_{ij}\|_\infty+\|Q\delta_{ij}\|_\infty=\mathscr{O}(h^{s+1}), \tag{80}$$

which completes the proof of (38). □

7 Numerical examples

For a more detailed discussion of the analytical properties of the following examples, see [19].

7.1 Singular DAE system of dimension 2. Unbounded canonical projector

We consider the initial value problem with $x(0) = (0,-1)^T$ for

$$\begin{pmatrix} t \\ 1 \end{pmatrix} x'(t) + \begin{pmatrix} 1 & 0 \\ 0 & \cos t \end{pmatrix} x(t) = \begin{pmatrix} t(2\sin t + t\cos t) \\ -e^{2t} \end{pmatrix}, \tag{81}$$

with inherent ODE

$$x_1'(t) + \frac{1}{t} x_1(t) = 2\sin t + t\cos t. \tag{82}$$

This is even a 'more difficult' example which is not strictly covered by our theoretical investigations, since the canonical projector $Q_{\mathrm{can}}(t) = \begin{pmatrix} 0 & 0 \\ -\frac{1}{t\cos t} & 1 \end{pmatrix}$ is unbounded near $t = 0$ (entailing the limit (62) to be nonzero). However, the initial value problem is well posed (cf. [19]).

We use collocation at equidistant points with $s = 4$ on $N = 2, 4, 8, 16, 32$ intervals. In Table 1, columns 'e' and '$\delta = \varepsilon - e$', the maximum of $|e_{ij}|_\infty$ and $|\delta_{ij}|_\infty = |\varepsilon_{ij} - e_{ij}|_\infty$ over all grid points is displayed.

Table 1 Error and deviation of estimate for Example 7.1

N	e	ord_e	$\delta = \varepsilon - e$	ord_δ
4	2.886 e-06		9.495 e-07	
8	2.103 e-07	3.8	3.249 e-08	4.9
16	1.407 e-08	3.9	1.057 e-09	4.9
32	9.072 e-10	4.0	3.336 e-11	5.0

The asymptotical order $\delta = \varepsilon - e = \mathcal{O}(h^{s+1})$ is clearly observed also in this case.

7.2 Semi-explicit singular DAE system of dimension 4. Bounded canonical projector

Next we consider a semi-explicit initial value problem (with appropriate initial conditions) of the type (for more details see [19])

$$\begin{aligned} t\,x_1'(t) + B_{11}x_1(t) + B_{12}x_2(t) &= g_1(t), \\ B_{21}x_1(t) + B_{22}x_2(t) &= g_2(t), \end{aligned} \tag{83}$$

where $B_{ij} \in \mathbb{R}^{2\times 2}$, $g_k \in C[0,1]$, and B_{22} is invertible. The inherent ODE system takes the form

$$x_1'(t) - \frac{M}{t} x_1(t) = f(t), \quad M(t) = B_{12}B_{22}^{-1}B_{21} - B_{11}. \tag{84}$$

The eigenvalues of M are $\lambda_1 = 0$ and $\lambda_2 = -2$. We apply the same collocation method as for Example 7.1. In Table 2, columns 'e' and '$\delta = \varepsilon - e$', the maximum of $|e_{ij}|_\infty$ and $|\delta_{ij}|_\infty = |\varepsilon_{ij} - e_{ij}|_\infty$ over all grid points is displayed.

Table 2 Error and deviation of estimate for Example 7.2

N	e	ord_e	$\delta = \varepsilon - e$	ord_δ
4	3.059 e-05		7.578 e-06	
8	2.543 e-06	3.6	2.335 e-07	5.0
16	1.796 e-07	3.8	7.429 e-09	5.0
32	1.189 e-08	3.9	2.408 e-10	5.0

A On the interrelation between collocation and Runge-Kutta methods

For ODEs, collocation methods are a special case of implicit Runge-Kutta (IRK) methods, cf. e.g. [13]. Here we explicate the analogous interrelation for the DAE case.

For a linear DAE system (1), the following version of a Runge-Kutta method is described and analyzed in [14] (see also [15] for the nonlinear case): Assume that

$$\mathscr{A} = (a_{jk},\ j,k = 1,\dots,s) \tag{85}$$

is the coefficient matrix in the Butcher array for a given s-stage IRK scheme, with internal nodes $c = (c_1,\dots,c_s)$ and weights $b = (b_1,\dots,b_s)$. As in [14] we assume that the method is stiffly accurate, i.e. $c_s = 1$ (as in (19)), $a_{sk} \equiv b_k$ and $\mathscr{A}$ invertible, with

$$\mathscr{A}^{-1} =: \hat{\mathscr{A}} = (\hat{a}_{jk},\ \ j,k = 1,\dots,s)\,. \tag{86}$$

For $\tau_{i+1} = \tau_i + h_i$, let $t_{ij} := \tau_i + c_j h_i$ as in Sect. 3. As usual we denote $\boldsymbol{A}_{ij} := \boldsymbol{A}(t_{ij})$, etc. For notational convenience, let $h := h_i,\ c_0 := 0$ and $t_{i0} := \tau_i$. In [14], an IRK step for (1) relating $\boldsymbol{x}_{i+1}$ to $\boldsymbol{x}_i$ is defined via internal approximations stages $\boldsymbol{X}_{ij}$ satisfying the linear system

$$\boldsymbol{A}_{ij}\boldsymbol{U}'_{ij} + \boldsymbol{B}_{ij}\boldsymbol{X}_{ij} = \boldsymbol{g}_{ij}, \quad j = 1,\dots,s, \tag{87}$$

for the unknowns $\boldsymbol{X}_{ij},\ j = 1,\dots,s$. Here, $\boldsymbol{U}'_{ij}$ is the shortcut

$$\boldsymbol{U}'_{ij} := \frac{1}{h}\sum_{k=1}^{s} \hat{a}_{jk}(\boldsymbol{D}_{ik}\boldsymbol{X}_{ik} - \boldsymbol{D}_{i0}\boldsymbol{X}_{i0}), \quad j = 1,\dots,s. \tag{88}$$

Together with $\boldsymbol{X}_{i0} = \boldsymbol{x}_i$ and with $c_s = 1$, this defines $\boldsymbol{x}_{i+1} := \boldsymbol{X}_{is}$.

Now we additionally assume that the given IRK scheme is equivalent to a collocation scheme in the conventional ODE sense, and we shall demonstrate in which way (87), (88) can be interpreted as a collocation scheme applied to the given DAE. To this end, we observe that (88) can be written as

$$h(\mathscr{A}\otimes\boldsymbol{I})\cdot\begin{pmatrix}\boldsymbol{U}'_{i1}\\ \vdots\\ \boldsymbol{U}'_{is}\end{pmatrix} = \begin{pmatrix}\boldsymbol{D}_{i1}\boldsymbol{X}_{i1} - \boldsymbol{D}_{i0}\boldsymbol{X}_{i0}\\ \vdots\\ \boldsymbol{D}_{is}\boldsymbol{X}_{is} - \boldsymbol{D}_{i0}\boldsymbol{X}_{i0}\end{pmatrix}, \tag{89}$$

or equivalently,

$$\boldsymbol{D}_{ij}\boldsymbol{X}_{ij} = \boldsymbol{D}_{i0}\boldsymbol{X}_{i0} + h\sum_{k=1}^{s} a_{jk}\boldsymbol{U}'_{ik}, \quad j = 1,\dots,s, \tag{90}$$

which has a structure similar to a conventional IRK system.

For IRK schemes of collocation type, the coefficients a_{jk} satisfy the simplifying assumption $C(s)$ (cf. [13]),

$$\sum_{k=1}^{s} a_{jk}c_k^{\ell-1} = \frac{c_j^\ell}{\ell}, \quad j,\ell = 1,\dots,s, \tag{91}$$

which implies

$$h\sum_{k=1}^{s} a_{jk}\,q(t_{ik}) = \int_{t_{i0}}^{t_{ij}} q(t)\,dt, \quad j = 1,\dots,s, \tag{92}$$

for any polynomial $q(t)$ of degree $\le s-1$. Now, let $\boldsymbol{p}_i(t)$ and $\boldsymbol{q}_i(t)$ denote the unique polynomials of degree $\le s$ interpolating the $\boldsymbol{X}_{ij}$ and $\boldsymbol{D}_{ij}\boldsymbol{X}_{ij}$, respectively, i.e.,

$$\boldsymbol{p}_i(t_{ij}) = \boldsymbol{X}_{ij}, \quad \boldsymbol{q}_i(t_{ij}) = \boldsymbol{D}_{ij}\boldsymbol{X}_{ij}, \quad j = 0,\dots,s. \tag{93}$$

Then, due to (92) the following identities hold for $j = 1,\dots,s$:

$$\boldsymbol{D}_{ij}\boldsymbol{X}_{ij} = \boldsymbol{q}_i(t_{ij}) = \boldsymbol{q}_i(t_{i0}) + \int_{t_{i0}}^{t_{ij}} \boldsymbol{q}'_i(t)\,dt = \boldsymbol{q}_i(t_{i0}) + h\sum_{j=1}^{s} a_{jk}\,\boldsymbol{q}'_i(t_{ik}) = \boldsymbol{D}_{i0}\boldsymbol{X}_{i0} + h\sum_{k=1}^{s} a_{jk}\,\boldsymbol{q}'_i(t_{ik}), \tag{94}$$

or equivalently,

$$h(\mathscr{A}\otimes\boldsymbol{I})\cdot\begin{pmatrix}\boldsymbol{q}'_i(t_{i1})\\ \vdots\\ \boldsymbol{q}'_i(t_{is})\end{pmatrix} = \begin{pmatrix}\boldsymbol{D}_{i1}\boldsymbol{X}_{i1} - \boldsymbol{D}_{i0}\boldsymbol{X}_{i0}\\ \vdots\\ \boldsymbol{D}_{is}\boldsymbol{X}_{is} - \boldsymbol{D}_{i0}\boldsymbol{X}_{i0}\end{pmatrix}. \tag{95}$$

Since $\mathscr{A}$ has been assumed to be invertible, (89) and (95) have unique, identical solutions, i.e.,

$$\boldsymbol{q}'_i(t_{ij}) = \boldsymbol{U}'_{ij}, \quad j = 1,\dots,s. \tag{96}$$

Now, (93) and (96) imply that solving the IRK system (87),(88) is equivalent to determining polynomials $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ of degree $\le s$ which satisfy $\boldsymbol{p}_i(\tau_i) = \boldsymbol{x}_i$, $\boldsymbol{q}_i(\tau_i) = \boldsymbol{D}(\tau_i)\boldsymbol{x}_i$, and

$$\boldsymbol{A}(t_{ij})\boldsymbol{q}'(t_{ij}) + \boldsymbol{B}(t_{ij})\boldsymbol{p}(t_{ij}) = \boldsymbol{g}(t_{ij}), \tag{97}$$

$$\boldsymbol{q}(t_{ij}) - \boldsymbol{D}(t_{ij})\boldsymbol{p}(t_{ij}) = \boldsymbol{0}, \tag{98}$$

for $j = 1,\dots,s$, which is exactly (23). Starting from (97) we can also reverse the above argumentation, ending up with the IRK formulation (87),(88).

References

1. Ascher, U.M., Mattheij, R.M.M., Russell, R.D.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, Prentice Hall (1988)
2. Ascher, U.M., Spiteri, R.: Collocation software for boundary value differential-algebraic equations, SIAM J. Sci. Stat. Comp. **4**, 938–952 (1994)
3. Auzinger, W., Kneisl, G., Koch, O., Weinmüller, E.: SBVP 1.0 – A MATLAB solver for singular boundary value problems, Technical Report ANUM Preprint No. 2/02, Vienna University of Technology (2002)
4. Auzinger, W., Koch, O., Weinmüller, E.: Efficient collocation schemes for singular boundary value problems, Numer. Algorithms **31**, 5–25 (2002)
5. Auzinger, W., Koch, O., Weinmüller, E.: Analysis of a new error estimate for collocation methods applied to singular boundary value problems, SIAM J. Numer. Anal. **42**, 2366–2386 (2005)
6. Auzinger, W., Koch, O., Praetorius, D., Weinmüller, E.: New a posteriori error estimates for singular boundary value problems, Numer. Algorithms **40**, 79–100 (2005)
7. Auzinger, W., Lehner, H., Weinmüller, E.: Defect-based a posteriori error estimation for index-1 DAEs, ASC Report 20/2007, Institute for Analysis and Scientific Computing, Vienna University of Technology (2007)
8. Balla, K., März, R.: A unified approach to linear differential algebraic equations and their adjoints, J. Anal. Appl. **21/3**, 783–802 (2002)
9. Degenhardt, A.: Collocation for transferable differential-algebraic equations, Technical Report 1992-1, Humboldt University Berlin (1992)
10. de Hoog, F.R., Weiss, R.: Difference methods for boundary value problems with a singularity of the first kind, SIAM J. Numer. Anal. **13**, 775–813 (1976)
11. de Hoog, F.R., Weiss, R.: Collocation methods for singular boundary value problems, SIAM J. Numer. Anal. **15**, 198–217 (1978)
12. Dokchan, R.: Numerical integration of DAEs with harmless critical points, Humboldt University Berlin, working paper (2007)
13. Hairer, E., Wanner, G., Nørsett, S.P: Solving Ordinary Differential Equations I – Nonstiff Problems, Springer Series in Computational Mathematics 8 (1987)
14. Higueras, I., März, R.: Differential algebraic systems anew, Appl. Numer. Math. **42**, 315–335 (2002)
15. Higueras, I., März, R.: Differential algebraic equations with properly stated leading term, Comp. Math. Appl. **48**, 215–235 (2004)
16. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-1 DAEs, Appl. Numer. Math. **45**, 175–200 (2003)
17. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-2 DAEs, Appl. Numer. Math. **45** 201–229 (2003)
18. Koch, O., Kofler, P., Weinmüller, E.: Initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind, Analysis **21**, 373–389 (2001)
19. Koch, O., März, R., Praetorius, D., Weinmüller, E.: Collocation for solving DAEs with singularities, Math. Comp. **79**, 281–304 (2010)
20. Kopelmann, A.: Ein Kollokationsverfahren für überführbare Algebro-Differentialgleichungen, Preprint 1987-151, Humboldt University Berlin (1987)
21. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations – Analysis and Numerical Solution, EMS Publishing House (2006)
22. Kunkel, P., Stöver, R.: Symmetric collocation methods for linear differential-algebraic boundary value problems, Numer. Math. **91**, 475–501 (2002)
23. März, R.: Differential algebraic equations anew, Appl. Numer. Math. **42** 315–335 (2002)
24. März, R.: The index of linear differential algebraic equations with properly stated leading terms, Results Math. **42**, 308–338 (2002)
25. März, R., Riaza, R.: Linear index-1 DAEs: Regular and singular problems, Acta Appl. Math **84**, 24–53 (2004)
26. März, R., Riaza, R.: Linear differential-algebraic equations with properly stated leading term: Regular points, J. Math. Anal. Appl. **323**, 1279–1299 (2006)
27. März, R., Riaza, R.: Linear differential-algebraic equations with properly stated leading term: A-critical points, Math. Comp. Model. Dyn. Sys. **13**, 291–314 (2007)
28. März, R., Riaza, R.: Linear differential-algebraic equations with properly stated leading term: B-critical points, Preprint 2007-09, Humboldt University Berlin (2007)
29. Riaza, R., März, R.: A simpler construction of the matrix chain defining the tractability index of linear DAEs, Appl. Math. Letters **21/4**, 326–331 (2008)
30. Schulz, S.: Four Lectures on Differential-Algebraic Equations, Report Series 497, Dept. of Mathematics, Univ. of Auckland (2003)
31. Stetter, H.J.: The defect correction principle and discretization methods, Numer. Math. **29**, 425–443 (1978)
32. Zadunaisky, P.E.: On the estimation of errors propagated in the numerical integration of ODEs, Numer. Math. **27**, 21–39 (1976)
33. Zielke, G.: Motivation und Darstellung von verallgemeinerten Matrixinversen, Beiträge zur Numerischen Mathematik **7**, 177–218 (1979)