Revista Colombiana de Matemáticas
Volumen **28** *(1994), páginas 43–82*

The numerical solution of linear time-varying DAEs with index 2 by IRK methods

EBROUL IZQUIERDO

Heinrich Hertz Institut, Berlin

ABSTRACT. Differential-algebraic equations (DAEs) with a higher index can be approximated by implicit Runge-Kutta methods (IRK). Until now, a number of initial value problems have been approximated by Runge-Kutta methods, but all these problems have a special semi-explicit or Hessenberg form. In the present paper we consider IRK methods applied to general linear time-varying (nonautonomous) DAEs tractable with index 2.

For some stiffly accurate IRK formulas we show that the order of accuracy in the differential component is the same nonstiff order, if the DAE has constant nullspace. We prove that IRK methods cannot be feasible or become exponentially unstable when applied to linear DAEs with variable nullspace. In order to overcome these difficulties we propose a new approach for this case. Feasibility, weak instability and convergence are proved. Order results are given in terms of the Butcher identities.

Keywords and phrases. Ordinary differential equations, differential-algebraic equations, initial value problems, implicit Runge-Kutta methods.

1991 Mathematics Subject Classification. 65L06, 34D20.

1. Introduction

In this paper we consider linear differential algebraic equations (DAEs) of the form

$$A(t)x'(t) + B(t)x(t) = q(t). \tag{1.1 a}$$

We assume that the coefficients A, B in (1.1 a) are continuous matrix functions $A, B : [t_0, T] \longrightarrow L(\mathbb{R}^m)$ and the matrix $A(t)$ has a smooth nullspace $N(t)$

for $t \in [t_0, T]$, *i.e.*, there are continuously differentiable projection functions Q, $P : [t_0, T] \longrightarrow L(\mathbb{R}^m)$ so that $Q(t)$ is a projection onto the nullspace $N(t)$ and $P(t) \equiv I - Q(t)$. Note that a smooth nullspace $N(t)$ has always constant dimension on $[t_0, T]$ and $A(t)$ has constant rank. Such equations arise in a variety of applications, *e.g.*, electrical network theory, dynamical systems subjects to constraints, optimal control of lumped-parameter systems and reduced equations in singulary perturbed systems.

In the present paper, we are interested in index 2 equations (more precisely, DAEs being tractable with index 2). Besides the BDF implicit Runge-Kutta methods are popular for approximating initial value problems in DAEs. Since higher index DAEs are known to be ill-posed, some kind of instability should be expected. Nevertheless, some IRK methods are reported to be adequate for some special types of higher index equations. Especially, the so-called Hessenberg systems are well investigated ([1], [4], [9]). We will generalize these results to the general linear DAE (1.1 a). In order to overcome the inherent difficulties when the nullspace of A varies we introduce a new approach for IRK methods using the projection Q. Doing so we can show that these methods share all properties which have IRK methods when applied to Hessenberg systems even for general DAEs (1.1 a).

Following the idea of MÄRZ [10] we use the matrix functions

$$\begin{aligned} G_1 &:= A + BQ, \\ A_1 &:= G_1 - AP'Q, \\ G_2 &:= A_1 + BPQ_1, \end{aligned}$$

as well as the subspaces

$$\begin{aligned} N(t) &:= \ker(A(t)), \\ S(t) &:= \{x \in \mathbb{R}^m \ : \ B(t)x \in \operatorname{im}(A(t))\}, \\ N_1(t) &:= \ker(A_1(t)), \\ S_1(t) &:= \{x \in \mathbb{R}^m \ : \ B(t)P(t)x \in \operatorname{im}(A_1(t))\}. \end{aligned}$$

Here, $t \in [t_0, T]$ and $Q_1(t)$ denotes a projection onto $N_1(t)$; further $P_1(t) := I - Q_1(t)$. The subspaces N, S, N_1 and S_1 are called the canonical the subspaces of the DAE (1.1 a) (see Lemma A.1).

Definition 1.1 (MÄRZ [11]) *The DAE (1.1 a) is called transferable if $G_1(t)$ is nonsingular for every $t \in [t_0, T]$. The DAE (1.1 a) is called tractable with index 2 if $G_1(t)$ is singular, but $G_2(t)$ is nonsingular for every $t \in [t_0, T]$.*

Note that $G_1(t)$ is nonsingular iff so is $A_1(t)$. Transferable DAEs are well understood and suitably modified numerical ODE methods work well for them. We refer to [1], [3] for this case. Nontransferable DAEs are essentially more complex. Tractability with index 2 characterizes an important class of nontransferable DAEs. Maybe the most important qualitative difference between

transferable and nontransferable DAEs is that the first class of problems remain well-posed in the HADAMARD sense, whereas initial and boundary value nontransferable problems are ill-posed in the sense of TIKHONOV. This fact has to be taken into account for the numerical treatment.

The formulation of appropriate initial conditions for index 2 tractable DAEs depends on the DAE itself. The index 2 tractable DAE (1.1) subject to the condition

$$P(t_0)P_1(t_0)x(t_0) = b \tag{1.1 b}$$

represent an initial value problem (IVP) with consistent initial values.

According to the originally conceived method for the numerical solution of ordinary differential equations (KUTTA, 1901), an IRK method can be realized for the DAE (1.1 a) in the following way: from an approximation x_{n-1} of the solution of the IVP (1.1) at t_{n-1}, these one-step methods construct an approximation x_n at $t_n = t_{n-1} + h$ via the formulas

$$x_n = x_{n-1} + h\sum_{j=1}^{s} b_j X_j', \tag{1.2 a}$$

where X_j' is defined by

$$A(t_{nj})X_j' + B(t_{nj})X_j = q(t_{nj}) \tag{1.2 b}$$

with $t_{nj} := t_{n-1} + c_j h$ and internal stages X_j given by

$$X_j = x_{n-1} + h\sum_{l=1}^{s} a_{jl}X_l', \quad j = 1, 2, \ldots, s. \tag{1.2 c}$$

Here a_{jl}, b_j, c_j are the coefficients determined by the method, and s is the number of stages. Usually, the matrix $\mathcal{A} = (a_{jl})_{j,l=1}^{s}$ and the vectors $b = (b_1, b_2, \ldots, b_s)^T$, $c = (c_1, c_2, \ldots, c_s)^T$ are combined in the so-called Butcher Diagram [2] or Runge-Kutta scheme

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}.$$

The paper is organized as follows. In Section 2 we give conditions which ensure the feasibility of IRK methods. These conditions depend not only on the method, but essentially on the behaviour of the nullspace $N(t)$. This is similar to what is known for the BDF [13]. In Section 3 we consider the case of a constant nullspace $N(t) \equiv N$. We show the weak instability and give order results in terms of the Butcher identities. In Section 4 we show that similar results are not true if the nullspace is variable. Therefore, for this case, we

introduce a new approximation which makes use of the projection Q. In Section 5 we show that these new approximations have the same properties as the original IRK methods in the constant nullspace case. Section 6 contains some numerical examples.

Assumption: *In the following,* $Q_1(1)$ *denotes the canonical projection onto* $N_1(t)$ *along* $S_1(t)$ *and for any matrix function* $M(t)$ *is*

$$D_M = \operatorname{diag}\left(M(t_{n1}), \ldots, M(t_{ns})\right).$$

Moreover, we assume that all matrix functions and projections are sufficiently smooth.

2. Existence and uniqueness of the Runge-Kutta solution

First we study the existence and uniqueness of a Runge-Kutta solution, *i.e.*, we try to answer the question: when does the system (1.2 b) have a unique solution X_j (or X_j') for $j = 1, 2, \ldots, s$?

Suppose that the Runge-Kutta matrix $\mathcal{A}$ is nonsingular. Then, the IRK method (1.2) is equivalent to

$$x_n = \rho x_{n-1} + \sum_{j=1}^{s} \sum_{l=1}^{s} b_j \widehat{a}_{jl} X_l, \tag{2.1 a}$$

$$A(t_{nj}) \sum_{l=1}^{s} \widehat{a}_{jl}(X_l - x_{n-1}) + hB(t_{nj})X_j = hq(t_{nj}), \quad j = 1, 2, \ldots, s, \tag{2.1 b}$$

where $\mathcal{A}^{-1} = (\widehat{a}_{jl})_{j,l=1}^{s}$ and $\rho := 1 - \sum_{i=1}^{s} \sum_{j=1}^{s} b_j \widehat{a}_{ij}$ ([6], [8]).

We now consider the IRK method defined by (2.1). Let us introduce some notation.

$$\begin{aligned}
&D_A = \operatorname{diag}(A(t_{n1}), \ldots, A(t_{ns})), \quad D_B = \operatorname{diag}(B(t_{n1}), \ldots, A(t_{ns})), \\
&D_q = \operatorname{diag}(q(t_{n1}), \ldots, q(t_{ns})), \quad \mathbb{1}_s = (1, \ldots, 1)^T \in \mathbb{R}^s, \\
&\overline{X} = (X_1^T, X_2^T, \ldots, X_s^T)^T \in \mathbb{R}^{ms}.
\end{aligned}$$

Now, system (2.1 b) can be written in the following form

$$[D_A(\mathcal{A}^{-1} \otimes I_m) + hD_B]\overline{X} = hD_q \mathbb{1}_s + D_A(\mathcal{A}^{-1}\mathbb{1}_s \otimes x_{n-1}). \tag{2.2}$$

The IRK method (2.1) is feasible if the matrix $\mathcal{M} := D_A(\mathcal{A}^{-1} \otimes I_m) + hD_B$ is nonsingular. Unfortunately, simple examples show that $\mathcal{M}$ is not always so.

Example 2.1. Consider

$$A(t) = \begin{pmatrix} 0 & 0 & 0 \\ t & e^{-t} & 0 \\ 0 & 0 & t \end{pmatrix}, \quad B(t) = \begin{pmatrix} t & e^{-t} & 0 \\ 0 & 0 & 3 \\ 0 & 0 & t+1 \end{pmatrix}.$$

The related DAE is transferable with index 2 for $t \in (0, \infty)$ and

$$Q = \begin{pmatrix} 0 & -e^{-t}/t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is a projector onto $N(t) = \ker\big(A(t)\big)$. A generic 1-stage IRK method yields

$$\mathcal{M} = A(t)(a^{-1} \otimes I_3) + hB(t) = \begin{pmatrix} ht & he^{-t} & 0 \\ a^{-1}t & a^{-1}e^{-t} & 3h \\ 0 & 0 & a^{-1}t + ht + h \end{pmatrix}.$$

Obviously, this matrix is singular for all $t, h \in (0, \infty)$ and all $a \in \mathbb{R} \setminus \{0\}$. □

The following theorem provides a necessary and sufficient condition for the existence of $\mathcal{M}^{-1}$. We present here a detailed proof which makes frequent use of the statements of Lemma A.2. For a more detailed proof we refer the reader to [7], [8].

Theorem 2.1 *Suppose that the IVP (1.1) is tractable with index 2 and moreover, that $\mathcal{A}$ and $I_{ms} - \mathcal{E}D_{QQ_1P'}$ are nonsingular for all $t \in [t_0, T]$, where $Q_1(t)$ denotes the canonical projection onto $N_1(t)$ along $S_1(t)$. Then the system (2.2) is uniquely solvable for sufficiently small h. Here*

$\mathcal{E} := I_{ms} + [(\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m]$, $\mathcal{C} = (c_{jl})_{j,l=1}^{s} \in L(\mathbb{R}^s)$, $c_{jl} = c_j - c_l$, and $\mathcal{C} \odot \mathcal{A}^{-1} = (c_{jl}\widehat{a}_{jl})_{j,l=1}^{s} \in L(\mathbb{R}^s)$ is the Hadamard product of matrices.

Proof. First we decouple (2.2) in the PP_1–, PQ_1– and Q-components, using the projection matrices defined in Section 1, and Lemma A.2. Finally, we calculate $\mathcal{M}^{-1}$.

System (2.2) is equivalent to

$$\begin{aligned}\big[D_{AP}(\mathcal{A}^{-1} \otimes I_m) + hD_{AQ} + D_{BQP}(\mathcal{A}^{-1} \otimes I_m) + hD_{BQ} + hD_{BP}\big]\overline{X} \\ = hD_q \mathbb{1}_s + D_A(\mathcal{A}^{-1}\mathbb{1}_s \otimes x_{n-1}).\end{aligned}$$

Using G_1, A_1 and G_2 (see Section 1) we can write this equation as:

$$\begin{aligned}\big[(D_{A_1} + D_{BPQ_1})(D_{P_1P}(\mathcal{A}^{-1} \otimes I_m) + hD_{P_1Q} + hD_{Q_1}) + hD_{AP'Q} + hD_{BPP_1}\big]\overline{X} \\ = hD_q \mathbb{1}_s + D_A(\mathcal{A}^{-1}\mathbb{1}_s \otimes x_{n-1}).\end{aligned}$$

Since (1.1) is tractable with index 2, there exists $G_2^{-1} = (A_1 + BPQ_1)^{-1}$. Hence, after multiplying the latter equation by G_2^{-1} we obtain:

$$\begin{aligned}\big[D_{P_1P}(\mathcal{A}^{-1}\otimes I_m)+hD_{P_1PP'Q}+hD_{P_1Q}+hD_{Q_1}+hD_{G_2^{-1}AP'Q}+hD_{G_2^{-1}BPP_1}\big]\overline{X}\\ = hD_{G_2^{-1}q}\mathbb{1}_s + D_{G_2^{-1}A}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1}).\end{aligned}$$

We multiply the latter equation by D_{PP_1}, $\frac{1}{h}D_{QP_1}$ and $\frac{1}{h}D_{Q_1}$ and use Lemma A.2 together with the following identities:

$$\begin{aligned}&D_{Q_1Q}=0,\quad D_{Q_1}=D_{Q_1G_2^{-1}BP},\quad D_{PP_1P}=D_{PP_1},\quad D_{PP_1Q}=0,\\ &D_{QP_1Q}=D_Q,\quad D_{QP_1P}=-D_{QQ_1},\quad D_{Q_1G_2^{-1}A}=0,\end{aligned}$$

Note that most of these relations are evident. For a detailed proof we refer the reader to [8]. See also Lemma A.1.

Thus, we obtain the system

$$\begin{aligned}&((\mathcal{A}^{-1}\otimes I_m)+hD_{PP_1G_2^{-1}B}+R_{PP_1})D_{PP_1}\overline{X}\\ &\quad+R_{PP_1}D_{PQ_1}\overline{X}+(hD_{PP_1P'}+R_{PP_1})D_Q\overline{X} \qquad (2.3\text{ a})\\ &=hD_{PP_1G_2^{-1}q}\mathbb{1}_s+D_{PP_1}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1}),\\ &\qquad D_{Q_1}\overline{X}=D_{Q_1G_2^{-1}q}\mathbb{1}_s, \qquad (2.3\text{ b})\\ &D_{QP_1G_2^{-1}BPP_1}\overline{X}-\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QQ_1}\overline{X}+(I_{ms}-D_{QQ_1P'})D_Q\overline{X}\\ &=D_{QP_1G_2^{-1}q}\mathbb{1}_s-\frac{1}{h}D_{QQ_1}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1})+\frac{1}{h}R_{QQ_1}\overline{X}, \qquad (2.3\text{ c})\end{aligned}$$

where

$$\begin{aligned}R_{PP_1}&=h\big[((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m)D_{(PP_1)'}+O(h)\big],\\ \frac{1}{h}R_{QQ_1}\overline{X}&=\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'}+O(h))\overline{X}\\ &=\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(Q'Q_1+QQ_1')}\big((D_{PP_1}+D_{PQ_1}+D_Q+O(h))\overline{X}.\end{aligned}$$

Inserting (2.3 a) and (2.3 b) into the last equation and using the relations $Q'Q_1PP_1=0$ and $Q_1Q=0$ we obtain

$$\frac{1}{h}R_{QQ_1}\overline{X}=\mathcal{V}+\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{QQ_1'Q}\overline{X}+O(h\|\overline{X}\|), \qquad (2.4)$$

where

$$\begin{aligned}\mathcal{V}=\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{QQ_1'}\big[&h(\mathcal{A}\otimes I_m)D_{PP_1G_2^{-1}q}\mathbb{1}_s\\ &+(\mathcal{A}\otimes I_m)D_{PP_1}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1})\big]\\ &+\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'PQ_1G_2^{-1}q}\mathbb{1}_s\,.\end{aligned}$$

We insert (2.4) into (2.3 c). The equation obtained together with (2.3 a-b) can be written

$$\mathcal{F}\begin{bmatrix} D_{PP_1}\overline{X} \\ D_{Q_1}\overline{X} \\ D_Q\overline{X} \end{bmatrix} = \mathcal{J} + \begin{bmatrix} O(h) & O(h) & -h\mathcal{S}+O(h^2) \\ 0 & 0 & 0 \\ O(h) & O(h) & O(h) \end{bmatrix}\begin{bmatrix} D_{PP_1}\overline{X} \\ D_{PQ_1}\overline{X} \\ D_Q\overline{X} \end{bmatrix}, \quad (2.5)$$

where

$$\mathcal{F} = \begin{bmatrix} I_{ms} & 0 & 0 \\ 0 & I_{ms} & 0 \\ D_{QP_1G_2^{-1}B} & -\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_Q & \mathcal{D} \end{bmatrix},$$

$$\mathcal{J} = \begin{pmatrix} h(\mathcal{A}\otimes I_m)D_{PP_1G_2^{-1}q}\mathbb{1}_s + (\mathcal{A}\otimes I_m)D_{PP_1}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1}) \\ D_{Q_1G_2^{-1}q}\mathbb{1}_s \\ D_{QP_1G_2^{-1}q}\mathbb{1}_s - \frac{1}{h}D_{QQ_1}(\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes x_{n-1}) + \mathcal{V} \end{pmatrix},$$

$$\mathcal{S} = \left[(\mathcal{A}\otimes I_m) + (\mathcal{A}(\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\right]D_{(PP_1)'Q}.$$

Note that in (2.5) the identity $QQ_1'Q = QQ_1'QQ = QQ_1P'Q$ has been used.

The matrix $\mathcal{F}$ is nonsingular iff $\mathcal{D}$ is nonsingular and in this case we can compute $\mathcal{F}^{-1}$:

$$\mathcal{F}^{-1} = \begin{bmatrix} I_{ms} & 0 & 0 \\ 0 & I_{ms} & 0 \\ -\mathcal{D}^{-1}D_{QP_1G_2^{-1}B} & \frac{1}{h}\mathcal{D}^{-1}(\mathcal{A}^{-1}\otimes I_m)D_Q & \mathcal{D}^{-1} \end{bmatrix}.$$

Hence, from (2.5) we have

$$\begin{bmatrix} D_{PP_1}\overline{X} \\ D_{PQ_1}\overline{X} \\ D_Q\overline{X} \end{bmatrix} = \begin{bmatrix} I_{ms} & 0 & 0 \\ 0 & D_P & 0 \\ 0 & 0 & I_{ms} \end{bmatrix}\mathcal{F}^{-1}\mathcal{J} + \begin{bmatrix} O(h) & O(h) & O(h) \\ 0 & 0 & 0 \\ O(h) & O(h) & O(h) \end{bmatrix}\begin{bmatrix} D_{PP_1}\overline{X} \\ D_{PQ_1}\overline{X} \\ D_Q\overline{X} \end{bmatrix}. \quad (2.6)$$

Using Banach's lemma we can solve system (2.6) for sufficiently small h. Then we use the identity $\overline{X} = D_{PP_1}\overline{X} + D_{PQ_1}\overline{X} + D_Q\overline{X}$ to construct the unique solution of (2.2). □

Remarks

- If the nullspace $N(t)$ of $A(t)$ is constant, the IRK method is always feasible, for h sufficiently small, because in this case $P' = 0$, *i.e.*, Theorem 2.1 ensures the feasibility of IRK methods applied to DAEs with index 2 and constant nullspace. This makes this theorem very valuable because many important classes of DAEs have constant nullspace, *e.g.*,

systems in Hessenberg or semi-explicit form. Note that for all these general DAEs, $A(t)$ has the form

$$A(t) = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

so that $P = A$ and trivially $P' = 0$.

- It is clear that before we try to solve numerically a DAE with index 2 we must check the nullspace $N(t)$. If $N(t)$ is variable, we have little hope that the method remain convergent (see next Section). In this case we advise to approximate the DAE using additional strategies: *e.g.*, regularization methods [5], or the modified method proposed in Section 5.
- For 1-stage IRK methods it holds that $I_m - \mathcal{E}D_{QQ_1P'} = I_m - QQ_1P'$, because in this case $\mathcal{C} = 0$. Hence, to verify the solvability of the system (2.2) it is sufficient to check the nonsingularity of the matrix $I_m - QQ_1P'$. In example 2.1,

$$I_3 - QQ_1P' = \begin{pmatrix} 1 & e^{-t}/t & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is always singular.

- In [11] it has been proved that BDF methods for the numerical solution of (1.1) are feasible if the matrix $I_m - QQ_1P'$ is nonsingular and h is sufficiently small. For IRK methods with s-stages, $s \geq 2$, the feasibility of the method depends on the method and the IVP.
- Summarizing, the statements of Theorem 2.1 present an important tool for practical numerical applications as well as for investigations about stability and convergence.

Example 2.2. Let

$$A(t) = \begin{pmatrix} 0 & 0 \\ 1 & \eta t \end{pmatrix}, \qquad B(t) = \begin{pmatrix} 1 & \eta t \\ 0 & 1+\eta \end{pmatrix}$$

with $\eta \in \mathbb{R}$. The related DAE is transferable with index 2

$$Q(t) = \begin{pmatrix} 0 & -\eta t \\ 0 & 1 \end{pmatrix}, \quad A_1(t) = \begin{pmatrix} 0 & 0 \\ 1 & 1+\eta t \end{pmatrix}$$

and

$$Q_1(t) = \begin{pmatrix} 1+\eta t & \eta t(1+\eta t) \\ -1 & -\eta t \end{pmatrix}.$$

It is known that the backward Euler method is not feasible if $\eta = 1$ (see [11]). If we instead apply the 2-stage Lobatto IIIC-formula [2] to this DAE, we have

$$\mathcal{A}^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \mathcal{C} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

$$\mathcal{E} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

$$I_{ms} - \mathcal{E} D_{QQ_1P'} = \begin{pmatrix} 1 & -\eta^2 t_{n1} & 0 & \eta^2 t_{n2} \\ 0 & 1+\eta & 0 & -\eta \\ 0 & \eta^2 t_{n1} & 1 & -\eta^2 t_{n2} \\ 0 & -\eta & 0 & 1+\eta \end{pmatrix}.$$

This latter matrix is singular iff $\eta = -\frac{1}{2}$. Moreover,

$$\mathcal{M} = \begin{pmatrix} h & h\eta t_{n1} & 0 & 0 \\ 1 & \eta t_{n1} + h(1+\eta) & 1 & \eta t_{n1} \\ 0 & 0 & h & h\eta t_{n2} \\ -1 & -\eta t_{n2} & 1 & \eta t_{n2} + h(1+\eta) \end{pmatrix}.$$

and $\mathcal{M}$ is singular iff $\eta = -\frac{1}{2}$. □

Now we will study the stability and convergence of the IRK method (1.2). To that end we consider two cases: constant nullspace and variable nullspace.

3. Stability and convergence for the constant nullspace case

IRK methods become unstable when they are applied to DAEs with index 2. This fact is well-known in the simple case of linear DAEs with constant coefficients and constant stepsize. We refer to [1], [8] for this case. For linear DAEs with variable coefficients but constant nullspace $N(t)$ some terms of the system (2.3) are lost, because $P' = 0$. Solving this system with respect to $D_{PP_1}\overline{X}$, $D_{PQ_1}\overline{X}$ and $D_Q\overline{X}$, we obtain

$$\begin{aligned} D_{PP_1}\overline{X} &= h\big((\mathcal{A}\otimes I_m) + O(h)\big) D_{PP_1G_2^{-1}q} \mathbb{1}_s - h\Big[\big(\mathcal{A}(\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big) D_{PP_1'} \\ &+ O(h)\Big] D_{PQ_1G_2^{-1}q}\mathbb{1}_s + \big(D_{PP_1} + O(h)\big) D_P(\mathbb{1}_s \otimes x_{n-1}), \qquad (3.1\text{ a}) \\ D_{PQ_1}\overline{X} &= D_{PQ_1G_2^{-1}q}\mathbb{1}_s \qquad (3.1\text{ b}) \\ D_Q\overline{X} &= \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m) D_{QQ_1} + D_{QP_1} + \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big) D_{QQ_1'PQ_1} \\ &+ O(h)\Big] D_{G_2^{-1}q}\mathbb{1}_s + \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m) D_{QP_1P} - D_{QP_1G_2^{-1}BPP_1} \qquad (3.1\text{ c}) \\ &+ \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big) D_{QQ_1'}(D_{PP_1} - I_{ms}) + O(h)\Big] D_P(\mathbb{1}_s \otimes x_{n-1}). \end{aligned}$$

Clearly, the nontrivial term $1/h(\mathcal{A}^{-1}\otimes I_m)D_{QP_1P}$ in (3.1 c) grows unboundedly if h tends to zero, *i.e.*, the method remains unstable. Nevertheless we can prove convergence. To make this fact more transparent, we show that in the constant nullspace case the instability is weak and it refers only to the Q-component. To that purpose we consider, in addition to the vectors x_n obtained from the formulas (2.1), the $\tilde{x}_n$ resulting from the perturbed systems

$$\tilde{x}_n = \rho\tilde{x}_{n-1} + \sum_{j=1}^{s}\sum_{l=1}^{s} b_j\hat{a}_{jl}\tilde{X}_l + h\delta_{n,s+1}, \tag{3.2 a}$$

$$\frac{1}{h}A(t_{nj})\sum_{l=1}^{s}\hat{a}_{jl}(\tilde{X}_l - \tilde{x}_{n-1}) + B(t_{nj})\tilde{X}_j = q(t_{nj}) + \delta_{nj};\ j = 1,2,\dots,s. \tag{3.2 b}$$

Assume $\|\delta_{ij}\| \le \delta$ for all $i = 0,1,\dots,n$, $j = 1,2,\dots,s+1$.

Analogously to (2.2), we can write the system (3.2 b) in the following form

$$\mathcal{M}\overline{\tilde{X}} = hD_{(q+\delta_n)}\mathbb{1}_s + D_A(\mathcal{A}^{-1}\mathbb{1}_s\otimes x_{n-1}),$$

where $\overline{\tilde{X}} = (\tilde{X}_1^T, \tilde{X}_2^T, \dots, \tilde{X}_s^T)^T$.

According to theorem 2.1, $\mathcal{M}$ is nonsingular for sufficiently small h, *i.e.*, the system (3.2 b) is uniquely solvable.

Let us introduce some new notations:

$\triangle x_k := \tilde{x}_k - x_k$, $\triangle Px_k := P(\tilde{x}_k - x_k)$, $\triangle PP_1x_k := PP_1(t_k)(\tilde{x}_k - x_k)$,
$\triangle PQ_1x_k := PQ_1(t_k)(\tilde{x}_k - x_k)$, $\triangle Qx_k := Q(\tilde{x}_k - x_k)$,
$\triangle PP_1X_j := PP_1(t_{nj})(\tilde{X}_j - X_j)$, $\triangle PQ_1X_j := PQ_1(t_{nj})(\tilde{X}_j - X_j)$,
$\triangle QX_j := Q(\tilde{X}_j - X_j) \in \mathbb{R}^m$,
$\triangle\overline{X} := \overline{\tilde{X}} - \overline{X} \in \mathbb{R}^{ms}$ for $k = n-1, n$, $j = 1,2,\dots,s$.

For the difference $\triangle\overline{X}$ we obtain (see 3.1):

$$\begin{aligned} D_{PP_1}\triangle\overline{X} = {} & h\big((\mathcal{A}\otimes I_m) + O(h)\big)D_{PP_1G_2^{-1}\delta_n}\mathbb{1}_s - h\Big[(\mathcal{A}(\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m)D_{PP_1'} \\ & + O(h)\Big]D_{PQ_1G_2^{-1}\delta_n}\mathbb{1}_s + \big(D_{PP_1} + O(h)\big)D_P(\mathbb{1}_s\otimes\triangle x_{n-1}), \end{aligned} \tag{3.3 a}$$

$$D_{PQ_1}\triangle\overline{X} = D_{PQ_1G_2^{-1}\delta_n}\mathbb{1}_s, \tag{3.3 b}$$

$$\begin{aligned} D_Q\triangle\overline{X} = {} & \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QQ_1} + D_{QP_1} + \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{QQ_1'PQ_1} \\ & + O(h)\Big]D_{G_2^{-1}\delta_n}\mathbb{1}_s + \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QP_1P} - D_{QP_1G_2^{-1}BPP_1} \\ & + \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{QQ_1'}(D_{PP_1} - I_{ms}) + O(h)\Big]D_P(\mathbb{1}_s\otimes\triangle x_{n-1}). \end{aligned} \tag{3.3 c}$$

Using Taylor's expansion about t_{n-1}, and equating like powers of h, it is easy to see that the differences ΔX_j, $j = 1, 2, \ldots, s$, satisfy:

$$\|\Delta P P_1 X_j\| \leq \|\Delta P P_1 x_{n-1}\| + hC\big(\|\Delta P x_{n-1}\| + \delta\big), \tag{3.4 a}$$

$$\|\Delta P Q_1 X_j\| \leq C\delta, \tag{3.4 b}$$

$$\|\Delta Q X_j\| \leq \frac{C}{h}\big(\|\Delta P Q_1 x_{n-1}\| + h\|\Delta P x_{n-1}\| + \delta\big), \tag{3.4 c}$$

with C a constant.

These estimates reflect the influence of pertubations on the solutions of (2.1 b). To prove the main results of this seccion we will need the following lemma. For the proof we refer to [4], [8].

Lemma 3.1. *Suppose that the two sequences of positive numbers* (y_n), (z_n) *satisfy*

$$y_n \leq y_{n-1} + hC(y_{n-1} + z_{n-1} + D_1),$$
$$z_n \leq |\rho| z_{n-1} + hC\left(y_{n-1} + z_{n-1} + \frac{D_2}{h}\right).$$

Then we have, for $hn \leq C_0$,

$$y_n \leq C_1(y_0 + hy_0 + hz_0 + D_1 + D_2), \quad \text{if } |\rho| \leq 1,$$
$$z_n \leq C_1\big(hy_0 + (|\rho|^n + h)z_0 + hD_1 + D_2\big), \quad \text{if } |\rho| < 1,$$
$$z_n \leq C_1\left(hy_0 + (1+h)z_0 + hD_1 + \frac{D_2}{h}\right), \quad \text{if } |\rho| = 1.$$

Here C, C_0, C_1, D_1, D_2 *are positive constants.*

Theorem 3.1 *Suppose the assumptions of Theorem 2.1 are satisfied and furthermore,* $P' = 0$. *If the coefficients of the IRK method (2.1) satisfy the relation* $|\rho| = |b^T \mathcal{A}^{-1} \mathbb{1}_s - 1| < 1$, *then the IRK method (2.1) applied to the DAE (1.1) is weakly unstable. This weak instability refers to the Q-component of the numerical solution. The P-component is stable.*

Proof. Subtracting (2.1a) from (3.2 a), we obtain

$$\Delta x_n = \rho \Delta x_{n-1} + (b^T \mathcal{A}^{-1} \otimes I_m)\Delta \overline{X} + h\delta_{n,s+1}. \tag{3.5}$$

Multiplying this equation by $PP_1P(t_n) = PP_1(t_n)$, $PQ_1P(t_n) = PQ_1(t_n)$ and Q and using Taylor's expansion about t_{n-1} on the right-hand side of the so

obtained equations, we have

$$\begin{aligned}\triangle PP_1x_n =&\rho\triangle PP_1x_{n-1} + h\rho\big(PP_1'(t_{n-1})O(h)\big)\triangle Px_{n-1}\\ &+ (b^T\mathcal{A}^{-1}\otimes I_m)D_{PP_1}\triangle\overline{X} + h\big((b^T\mathcal{A}^{-1}\otimes I_m)D_{\gamma PP_1'}\\ &+ O(h)\big)D_P\triangle\overline{X} + hPP_1(t_n)\delta_{n,s+1},\end{aligned} \tag{3.6 a}$$

$$\begin{aligned}\triangle PQ_1x_n =&\rho\triangle PQ_1x_{n-1} + h\rho\big(PQ_1'(t_{n-1}) + O(h)\big)\triangle Px_{n-1}\\ &+ (b^T\mathcal{A}^{-1}\otimes I_m)D_{PQ_1}\triangle\overline{X}\\ &+ h\big((b^T\mathcal{A}^{-1}\otimes I_m)D_{\gamma PQ_1'} + O(h)\big)D_P\triangle\overline{X}\\ &+ hPQ_1(t_n)\delta_{n,s+1},\end{aligned} \tag{3.6 b}$$

$$\triangle Qx_n =\rho\triangle Qx_{n-1} + (b^T\mathcal{A}^{-1}\otimes I_m)D_Q\triangle\overline{X} + hQ\delta_{n,s+1}, \tag{3.6 c}$$

where $\gamma N(t_{nj}) = (1-c_j)N(t_{nj})$ for any matrix function N.

Inserting (3.3 a) and (3.3 b) into (3.6 a) and (3.6 b) and using Taylor's expansion about t_{n-1} on the right-hand side of the so obtained equations, we get

$$\begin{aligned}\|\triangle PP_1x_n\| &\le \big\|\rho\triangle PP_1x_{n-1} + (b^T\mathcal{A}^{-1}\otimes I_m)(\mathbb{1}_s\otimes\triangle PP_1x_{n-1})\big\|\\ &\quad + hC_1^{(n)}(\|\triangle Px_{n-1}\| + \delta)\\ &= \big\|(1 - b^T\mathcal{A}^{-1}\mathbb{1}_s)\triangle PP_1x_{n-1} + b^T\mathcal{A}^{-1}\mathbb{1}_s\triangle PP_1x_{n-1}\big\|\\ &\quad + hC_1^{(n)}(\|\triangle Px_{n-1}\| + \delta)\\ &= \|\triangle PP_1x_{n-1}\| + hC^{(n)}(\|\triangle Px_{n-1}\| + \delta).\end{aligned} \tag{3.7 a}$$

$$\begin{aligned}\|\triangle PQ_1x_n\| &\le |\rho|\,\big\|\triangle PQ_1x_{n-1}\big\| + \big\|(b^T\mathcal{A}^{-1}\otimes I_m)D_{PQ_1G_2^{-1}\delta_n}\mathbb{1}_s\big\|\\ &\quad + hC_1^{(n)}(\|\triangle Px_{n-1}\| + \delta)\\ &\le |\rho|\,\|\triangle PQ_1x_{n-1}\| + hC^{(n)}\Big(\|\triangle Px_{n-1}\| + \frac{\delta}{h}\Big),\end{aligned} \tag{3.7 b}$$

where $C^{(n)}$ and $C_1^{(n)}$ are constants.

Applying lemma 3.1 to the recursions (3.7 a-b) with $\|\triangle PP_1x_k\|$, $\|\triangle PQ_1x_k\|$, δ, $\max_{i=1}^n C^{(i)}$ instead of y_k, z_k, D_j, C, respectively, $k = n-1, n$, $j = 1, 2$, we obtain:

If $|\rho| \le 1$, then

$$\|\triangle PP_1x_n\| \le C_1\big(\|\triangle PP_1x_0\| + h\|\triangle PP_1x_0\| + h\|\triangle PQ_1x_0\| + \delta\big). \tag{3.8 a}$$

If $|\rho| < 1$, then

$$\|\triangle PQ_1x_n\| \le C_1\big(h\|\triangle PP_1x_0\| + (|\rho|^n + h)\|\triangle PQ_1x_0\| + \delta\big). \tag{3.8 b}$$

If $|\rho| = 1$, then

$$\|\triangle PQ_1 x_n\| \leq C_1 \left(h\|\triangle PP_1 x_0\| + (1+h)\|\triangle PQ_1 x_0\| + \frac{\delta}{h} \right). \quad (3.8\text{ c})$$

With $\tilde{x}_0 - x_0 := \delta_{0,s+1}$, we arrive at

$$\|\triangle P x_n\| \leq \|\triangle PP_1 x_n\| + \|\triangle PQ_1 x_n\| \leq C\delta, \quad (3.9)$$

where C is constant and $|\rho| < 1$.

Now we consider the Q-component. Inserting (3.3 c) into (3.6 c) and using Taylor's expansion about t_{n-1} on the right-hand side of the so obtained equation, we have

$$\|\triangle Q x_n\| \leq |\rho|\, \|\triangle Q x_{n-1}\| + \frac{C_1}{h}(\|\triangle PQ_1 x_{n-1}\| + h\|\triangle P x_{n-1}\| + \delta).$$

Inserting (3.9) into this equation, we get

$$\begin{aligned} \|\triangle Q x_n\| &\leq |\rho|\, \|\triangle Q x_{n-1}\| + \frac{C_2}{h}\delta \\ &\leq |\rho|^n\, \|\triangle Q x_0\| + \frac{C_2}{h}\delta \sum_{i=0}^{n-1} |\rho|^i \leq \frac{C}{h}\delta \end{aligned} \quad (3.10)$$

with C constant. *I.e.*, the IRK method is weak unstable on the Q-component. □

Remarks

- If $|\rho| = 1$, then the P-component is also unstable because in this case we can only obtain an estimate of the form

$$\|\triangle P x_n\| \leq \frac{C}{h}\delta.$$

  Furthermore, for the Q-component we obtain the estimate

$$\|\triangle Q x_n\| \leq \|\triangle Q x_0\| + \frac{C}{h^2}\left(h\|\triangle PP_1 x_0\| + \|\triangle PQ_1 x_0\| + \frac{\delta}{h} \right).$$

- A very important class of IRK methods, the superconvergent symmetric methods, may exhibit severe order reduction when applied to DAEs with index 2. For example, the so-called Gauss-Legendre methods, which have order $2s$ for nonstiff ODEs, exhibit very bad results in numerical experiments. These methods yield $\rho = (-1)^s$.

Now we study the convergence properties of the IRK method (2.1). We derive necessary and sufficient conditions for the local error of an IRK method when applied to the DAE (1.1). Then we study error propagation and derive estimates for the global error. Before we can state the main result of this subsection, we need some definitions.

Definition 3.1.

- *The local error d_n of the method (2.1) is given by*

$$d_n = \rho x(t_{n-1}) + \sum_{j=1}^{s} b_j \sum_{l=1}^{s} \widehat{a}_{jl} X_l - x(t_n).$$

Here X_l is obtained from (2.1 b) with $x_{n-1} = x(t_{n-1})$.

- *The j-th internal local truncation error $\delta_j^{(n)}$ at t_n of an s-stage IRK method (2.1) is given by*

$$\delta_j^{(n)} = x(t_{n-1}) + h \sum_{l=1}^{s} a_{jl} x'(t_{nl}) - x(t_{nj}), \qquad (3.11\text{ a})$$

$$\delta_{s+1}^{(n)} = x(t_{n-1}) + h \sum_{l=1}^{s} b_l x'(t_{nl}) - x(t_n). \qquad (3.11\text{ b})$$

- *The algebraic conditions $B(p)$ and $C(q)$ (Butcher identities) of an s-stage IRK method are given by*

$$B(p) \quad : \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \text{ for } k = 1, 2, \ldots, p,$$

$$C(q) \quad : \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \text{ for } k = 1, 2, \ldots, q \text{ and all } i.$$

- *The internal stage order K_I of an s-stage IRK method is given by*

$$K_I = \max\{k : B(k) \text{ and } C(k)\}.$$

The numerical solution by the IRK method (2.1) satisfies (3.2 b) where δ_{nj}, $j = 1, 2, \ldots, s$, represent errors in the solution of the system (2.1 b). The true solution $x(t)$ of (1.1) satisfies

$$x(t_n) = \rho x(t_{n-1}) + \sum_{j=1}^{s} \sum_{l=1}^{s} b_j \widehat{a}_{jl} (x(t_{nl}) + \delta_l^{(n)}) - \delta_{s+1}^{(n)}. \qquad (3.12\text{ a})$$

$$\frac{1}{h} A(t_{nj}) \sum_{l=1}^{s} \widehat{a}_{jl} (x(t_{nl}) - x(t_{n-1}) + \delta_j^{(n)}) + B(t_{nj}) x(t_{nj}) = q(t_{nj}). \qquad (3.12\text{ b})$$

where $\delta_j^{(n)}$, $j = 1, 2, \ldots, s+1$, are the internal local truncation errors at t_n.

Subtracting (3.12) from (3.2), we obtain

$$e_n = \rho e_{n-1} + (b^T \mathcal{A}^{-1} \otimes I_m)(\overline{E} - \bar{\delta}^{(n)}) + \delta_{s+1}^{(n)}, \qquad (3.13\text{ a})$$

$$\frac{1}{h} A(t_{nj}) \sum_{l=1}^{s} \widehat{a}_{jl} E_l + B(t_{nj}) E_j = \frac{1}{h} A(t_{nj}) \sum_{l=1}^{s} \widehat{a}_{jl} (\delta_l^{(n)} + e_{n-1}), \quad j = 1, 2, \ldots, s, \qquad (3.13\text{ b})$$

where
$e_k := x_k - x(t_k)$, $k = n-1, n$, $E_j := X_j - x(t_{nj})$, $j = 1, 2, \ldots, s$, $\overline{E} = (E_1^T, E_2^T, \ldots, E_s^T)^T \in \mathbb{R}^{ms}$, $\bar{\delta}^{(n)} = \big((\delta_1^{(n)})^T, (\delta_2^{(n)})^T, \ldots, (\delta_s^{(n)})^T\big)^T \in \mathbb{R}^{ms}$, $\bar{e}_{n-1} = (e_{n-1}^T, e_{n-1}^T, \cdots, e_{n-1}^T)^T \in \mathbb{R}^{ms}$ and $\bar{\delta}_n = (\delta_{n1}^T, \delta_{n2}^T, \ldots, \delta_{ns}^T)^T \in \mathbb{R}^{ms}$.

According to theorem 2.1, for sufficiently small h, and $P' = 0$ the system (3.13 b) is uniquely solvable and the solution $\overline{E}$ satisfies (see 3.1)

$$D_{PP_1}\overline{E} = \big(D_{PP_1} + O(h)\big) D_P \big(\bar{\delta}^{(n)} + \bar{e}_{n-1}\big) + O(h\|\bar{\delta}_n\|), \tag{3.14 a}$$

$$D_{PQ_1}\overline{E} = D_{PQ_1G_2^{-1}}\bar{\delta}_n, \tag{3.14 b}$$

$$D_Q\overline{E} = \Big[\frac{1}{h}(\mathcal{A}^{-1} \otimes I_m) D_{QP_1P} + \big((\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m\big) D_{QQ_1'}(D_{PP_1} - I_{ms}) - D_{QP_1G_2^{-1}BPP_1} + O(h)\Big] D_P(\bar{\delta}^{(n)} + \bar{e}_{n-1}) + O\Big(\frac{1}{h}\|\bar{\delta}_n\|\Big). \tag{3.14 c}$$

Multiplying (3.13 a) by $PP_1(t_n)$, $PQ_1(t_n)$ and Q and using Taylor's expansion about t_{n-1} on the right-hand side of the equations, we obtain

$$\begin{aligned} PP_1e_n =& \rho PP_1e_{n-1} + h\rho\big(PP_1'(t_{n-1}) + O(h)\big)Pe_{n-1} \\ & (b^T\mathcal{A}^{-1} \otimes I_m) D_{PP_1}(\overline{E} - \bar{\delta}^{(n)}) \\ & + h\Big[\big((b^T\mathcal{A}^{-1} \otimes I_m) D_{\gamma PP_1'} + O(h)\big)\Big] D_P(\overline{E} - \bar{\delta}^{(n)}) \\ & + PP_1(t_n)\delta_{s+1}^{(n)}, \end{aligned} \tag{3.15 a}$$

$$\begin{aligned} PQ_1e_n =& \rho PQ_1e_{n-1} + h\rho\big(PQ_1'(t_{n-1}) + O(h)\big)Pe_{n-1} \\ & + (b^T\mathcal{A}^{-1} \otimes I_m) D_{PQ_1}\overline{E} + h\Big[\big((b^T\mathcal{A}^{-1} \otimes I_m) D_{\gamma PQ_1'} + O(h)\big)\Big] D_P\overline{E} \\ & - PQ_1(t_n)\big[(b^T\mathcal{A}^{-1} \otimes I_m)\delta^{(n)} - \delta_{s+1}^{(n)}\big], \end{aligned} \tag{3.15 b}$$

$$Qe_n = \rho Qe_{n-1} + (b^T\mathcal{A}^{-1} \otimes I_m) D_Q(\overline{E} - \bar{\delta}^{(n)}) + Q\delta_{s+1}^{(n)}, \tag{3.15 c}$$

where $PP_1e_k := PP_1(t_k)e_k$, $PQ_1e_k := PQ_1(t_k)e_k$, $k = n-1, n$. Inserting (3.14) into (3.15) we thus arrive at

$$PP_1e_n = PP_1e_{n-1} + O\big(h(\|Pe_{n-1}\| + \|\bar{\delta}^{(n)}\| + \|\bar{\delta}_n\|)\big) + PP_1(t_n)\delta_{s+1}^{(n)} \tag{3.16 a}$$

$$\begin{aligned} PQ_1e_n =& \rho PQ_1e_{n-1} + O\big(h\|Pe_{n-1}\| + h\|\bar{\delta}^{(n)}\| + \|\bar{\delta}_n\|\big) \\ & - PQ_1(t_n)\Big[(b^T\mathcal{A}^{-1} \otimes I_m)\bar{\delta}^{(n)} - \delta_{s+1}^{(n)}\Big], \end{aligned} \tag{3.16 b}$$

$$Qe_n = \rho Qe_{n-1} + (b^T\mathcal{A}^{-1} \otimes I_m)\Big[\frac{1}{h}(\mathcal{A}^{-1} \otimes I_m)D_{QP_1P} - D_{QP_1G_2^{-1}BPP_1}$$
$$+ \big((\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m\big)D_{QQ_1'}(D_{PP_1} - I_m) + O(h)\Big]D_P(\bar{\delta}^{(n)} + \bar{e}_{n-1})$$
$$- (b^T\mathcal{A}^{-1} \otimes I_m)\bar{\delta}^{(n)} + Q\delta_{s+1}^{(n)} + O\Big(\frac{1}{h}\|\bar{\delta}_n\|\Big). \qquad (3.16\ c)$$

The local error of the IRK method (2.1) is given by (3.16) for $x_{n-1} = x(t_{n-1})$, *i.e.*, $e_{n-1} = 0$.

Theorem 3.2. *Let the assumptions of Theorem 2.1 be satisfied and furthermore assume that, $P' = 0$. Suppose the IRK method satisfies the Butcher identities $B(p)$ and $C(q)$ for $p \geq q \geq 1$ and the numerical errors in the solution of the system (2.1 b) are $\delta_{nj} = O(h^{q+\alpha})$. Then the local error satisfies:*

(*a*) $Pd_n = O(h^{q+1})$, $Qd_n = O(h^q)$, *if* $\alpha = 1$.
(*b*) $PP_1(t_n)d_n = O(h^{q+2})$, *for* $p \geq q+1$ *and* $\alpha = 1$.
(*c*) $PQ_1(t_n)d_n = O(h^{q+2})$, *if* $\alpha = 2$ *and the coefficients of the IRK method satisfy*

$$b^T\mathcal{A}^{-1}c^{q+1} = 1$$

where $c^{q+1} = (c_1^{q+1}, c_2^{q+1}, \ldots, c_s^{q+1})^T$.
(*d*) $Pd_n = O(h^{q+2})$, *if* $\alpha = 2$, $b^T\mathcal{A}^{-1}c^{q+1} = 1$ *and* $p \geq q+1$.
(*e*) $Qd_n = O(h^{q+1})$, *if* $\alpha = 2$ *and the coefficients of the IRK method satisfy*

$$b^T\mathcal{A}^{-2}c^{q+1} = q+1.$$

Proof.

(*a*) It follows directly from (3.16) with $e_{n-1} = 0$, $\bar{\delta}^{(n)} = O(h^{q+1})$, $\delta_{s+1}^{(n)} = O(h^{q+1})$, $\bar{\delta}_n = O(h^{q+1})$ and $Pe_n = PP_1e_n + PQ_1e_n$.

(*b*) It follows from (3.16 a) with $e_{n-1} = 0$, $\bar{\delta}^{(n)} = O(h^{q+1})$, $\bar{\delta}_n = O(h^{q+1})$ and $\delta_{s+1}^{(n)} = O(h^{q+2})$ (because $B(q+1)$).

(*c*) Inserting $e_{n-1} = 0$, $\bar{\delta}^{(n)} = O(h^{q+1})$ and $\bar{\delta}_n = O(h^{q+2})$ into (3.16 b), we obtain

$$PQ_1(t_n)d_n = -PQ_1(t_n)\Big[(b^T\mathcal{A}^{-1} \otimes I_m)\bar{\delta}^{(n)} - \delta_{s+1}^{(n)}\Big] + O(h^{q+2}). \qquad (3.17)$$

For $q \leq p$, the following relations

$$\bar{\delta}^{(n)} = h^{q+1}\Big[(\mathcal{A}c^q \otimes x^{(q+1)}(t_{n-1})) - \frac{1}{q+1}(c^{q+1} \otimes x^{(q+1)}(t_{n-1}))\Big]$$
$$+ O(h^{q+2}) \qquad (3.18\ a)$$

and

$$\delta_{s+1}^n = h^{q+1}\Big(b^Tc^q - \frac{1}{q+1}\Big)x^{(q+1)}(t_{n-1}) + O(h^{q+2}). \qquad (3.18\ b)$$

hold. This yields

$$
\begin{aligned}
&(b^T\mathcal{A}^{-1}\otimes I_m)\bar{\delta}^{(n)} - \delta_{s+1}^{(n)} = \\
&\quad h^{q+1}(b^T\mathcal{A}^{-1}\otimes I_m)\Big[(\mathcal{A}c^q)\otimes x^{(q+1)}(t_{n-1}) - \frac{1}{q+1}\big(c^{q+1}\otimes x^{(q+1)}(t_{n-1})\big)\Big] \\
&\qquad - h^{q+1}\left(b^T c^q - \frac{1}{q+1}\right)x^{(q+1)}(t_{n-1}) + O(h^{q+2}) \\
&\qquad\quad = \frac{h^{q+1}}{q+1}(1 - b^T\mathcal{A}^{-1}c^{q+1})x^{(q+1)}(t_{n-1}) + O(h^{q+2}).
\end{aligned}
$$

Inserting this equation in (3.17) and using $b^T\mathcal{A}^{-1}c^{q+1} = 1$ we obtain

$$PQ_1(t_n)d_n = O(h^{q+2}).$$

(*d*) Follows directly from (*b*) and (*c*).

(*e*) Inserting (3.18 a) into (3.16 c) and using $\bar{\delta}^{(n)} = O(h^{q+1})$, $\delta_{s+1}^{(n)} = O(h^{q+1})$, $\bar{\delta}_n = O(h^{q+2})$ and $e_{n-1} = 0$ we obtain

$$Qd_n = h^q QP_1P(t_n)\big((b^T\mathcal{A}^{-1}c^q - \frac{1}{q+1}b^T\mathcal{A}^{-2}c^{q+1})\otimes x^{(q+1)}(t_{n-1})\big) + O(h^{q+1}).$$

Inserting the identity $b^T\mathcal{A}^{-2}c^{q+1} = q+1$ into this equation and using $B(q)$ and $C(q)$ we arrive at

$$Qd_n = O(h^{q+1})\ .\ \square$$

Theorem 3.2 gives conditions on the coefficients of an IRK method so that it attains a given order for the local error when applied to the DAE (1.1) with constant nullspace. The conditions $b^T\mathcal{A}^{-1}c^{q+1} = 1$ and $b^T\mathcal{A}^{-2}c^{q+1} = q+1$ have been studied in [1] when IRK methods are applied to linear constant coefficient systems of index 1 and 2. Furthermore, in [1] BRENNAN, CAMPBELL and PETZOLD derive conditions for the local error of an IRK method applied to linear constant coefficient systems of arbitrary index k.

Now we use the recursions (3.16) to derive estimates for the global error.

Theorem 3.3. *Let the assumptions of theorem 2.1 be satisfied and furthermore assume that* $P' = 0$. *Suppose the IRK method satisfies* $|\rho| < 1$. *For* $q \geq 1$, *let* $e_0 = O(h^{q+\alpha})$ *be the errors in the initial conditions and* $\delta_{ij} = O(h^{q+\alpha})$ *the numerical errors in the solution of the system (2.1 b) at the i-th integration step. Then the global error* e_n *satisfies:*

- $\|Pe_n\| \leq Ch^q$, *if* $B(q)$, $C(q)$ *and* $\alpha = 0$,
- $\|Pe_n\| \leq Ch^{q+1}$, *if* $B(q+1)$, $C(q)$ *and* $\alpha = 1$,
- $\|Qe_n\| \leq Ch^q$, *if* $B(q)$, $C(q)$ *and* $\alpha = 1$,
- $\|Qe_n\| \leq Ch^{q+1}$, *if* $B(q+1)$, $C(q)$, $b^T\mathcal{A}^{-1}c^{q+1} = 1$, $b^T\mathcal{A}^{-2}c^{q+1} = q+1$ *and* $\alpha = 2$.

Proof. From (3.16 a-b) we immediatly obtain the estimates

$$\|PP_1e_n\| \leq \|PP_1e_{n-1}\| + hC^{(n)}\big(\|PP_1e_{n-1}\| + \|PQ_1e_{n-1}\| + D_1\big), \tag{3.19 a}$$

$$\|PQ_1e_n\| \leq |\rho|\,\|PQ_1e_{n-1}\| + hC^{(n)}\Big(\|PP_1e_{n-1}\| + \|PQ_1e_{n-1}\| + \frac{D_2}{h}\Big) \tag{3.19 b}$$

with $C^{(n)}$ constants and

$$D_1 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q+1) \text{ and } C(q), \end{cases} \tag{3.20 a}$$

$$D_2 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q) \text{ and } C(q), \\ O(h^{q+2}), & \text{if } \alpha = 2,\ b^T\mathcal{A}^{-1}c^{q+1} = 1,\ B(q) \text{ and } C(q). \end{cases} \tag{3.20 b}$$

Applying Lemma 3.1 to the recursions (3.19) we obtain

$$\|PP_1e_n\| \leq C_1\big(\|PP_1e_0\| + h\|PP_1e_0\| + h\|PQ_1e_0\| + D_1 + D_2\big), \tag{3.21 a}$$

$$\|PQ_1e_n\| \leq C_1\big(h\|PP_1e_0\| + (|\rho|^n + h)\|PQ_1e_0\| + hD_1 + D_2\big). \tag{3.21 b}$$

Now, the estimates for Pe_n follow from (3.20), (3.21) and the inequality

$$\|Pe_n\| \leq \|PP_1e_n\| + \|PQ_1e_n\|.$$

In order to prove the first estimate for Qe_n, we consider (3.16 c) and (3.21 b). This last inequality together with (3.20) yield

$$PQ_1e_n = O(h^{q+1}), \text{ if } B(q),\ C(q) \text{ and } \alpha = 1.$$

Inserting this last equation together with $Pe_n = O(h^q)$, $\bar{\delta}_n = O(h^{q+1})$, $\bar{\delta}^{(n)} = O(h^{q+1})$ and $\delta_{s+1}^{(n)} = O(h^{q+1})$ into (3.16 c), we obtain

$$Qe_n = \rho Qe_{n-1} + O(h^q).$$

The first estimate for Qe_n follows directly from this recursion. In order to prove the second estimate for Qe_n, we insert (3.20) with $\alpha = 2$, $B(q+1)$, $C(q)$ and $b^T\mathcal{A}^{-1}c^{q+1} = 1$ into (3.21 b). This implies

$$PQ_1e_n = O(h^{q+2}).$$

Inserting this equation together with $Pe_n = O(h^{q+1})$, $\bar{\delta}_n = O(h^{q+2})$, $\delta_{s+1}^{(n)} = O(h^{q+2})$ and $\bar{\delta}^{(n)} = O(h^{q+1})$ into (3.16 c) we obtain

$$Qe_n = \rho Qe_{n-1} + \frac{1}{h}\big(b^T\mathcal{A}^{-2} \otimes I_m\big)D_{QP_1P}\bar{\delta}^{(n)} + O(h^{q+1}).$$

Using the identity $b^T\mathcal{A}^{-2}c^{q+1} = q+1$ we thus arrive at

$$Qe_n = \rho e_{n-1} + O(h^{q+1}).$$

The second estimate for Qe_n follows directly from this recursion. □

For many important methods such as the Gauss-Legendre method, $|\rho| = 1$. This means that the statements of theorem 3.3 are not valid for this class of IRK. The following theorem shows how strong the order reduction is when IRK methods with $|\rho| = 1$ are applied to linear DAEs with index 2 and constant nullspace.

Theorem 3.4. *Assume that the hypothesis of Theorem 2.1 are satisfied. Suppose that the IRK method satisfies* $|\rho| = 1$, *and further assume that, for* $q \geq 1$, $e_0 = O(h^{q+\beta})$ *is the error in the initial condition and* $\delta_{ij} = O(h^{q+\alpha})$ *is the numerical error in the solution of the system (2.1 b) at the ith integration step. Then the global error* e_n *satisfies:*

- $\|Pe_n\| \leq Ch^q$, $\|Qe_n\| \leq Ch^{q-2}$, *if* $\alpha = 1$, $\beta = 0$, $B(q)$ *and* $C(q)$ *for* $q \geq 2$,
- $\|Pe_n\| \leq Ch^{q+1}$, $\|Qe_n\| \leq Ch^{q-1}$, *if* $\alpha = 2$, $\beta = 1$, $B(q+1)$, $C(q)$ *and* $b^T\mathcal{A}^{-1}c^{q+1} = 1$,
- $\|Pe_n\| \leq Ch^{q+1}$, $\|Qe_n\| \leq Ch^{q-1}$, *if* $\alpha = 2$, $\beta = 1$, $B(q+1)$, $C(q)$ *and* $\rho = -1$.

Proof. The first and the second estimate for Pe_n and Qe_n follow directly from the theorem assumptions and by application of lemma 3.1 to the recursions (3.20) with $|\rho| = 1$. In order to prove the last estimates, we consider equation (3.16b), or equivalently,

$$PQ_1(t_n)e_n = \rho\big(I_m + hPQ_1'(t_{n-1})\big)PQ_1(t_{n-1})e_{n-1} + A_n + B_n \qquad (3.22)$$

with

$$\begin{aligned} A_n :&= h\rho PQ_1'PP_1(t_{n-1})e_{n-1} + O(h^2\|Pe_{n-1}\|) \\ &\quad - h\Big[\big((b^T\mathcal{A}^{-1}\otimes I_m)D_{\gamma PQ_1'} + O(h)\big)\Big]\big(D_{PP_1} + O(h)\big)D_P(\bar{\delta}^{(n)} + \bar{e}_{n-1}) \\ &\quad - h\Big[\big((b^T\mathcal{A}^{-1}\otimes I_m)D_{\gamma PQ_1'} + O(h)\big)\Big]D_P\bar{\delta}^{(n)} + PQ_1(t_n)\delta_{s+1}^{(n)} + O(h\|\bar{\delta}_n\|), \\ B_n :&= -\sum_{j=1}^{s}\sum_{l=1}^{s} b_j\widehat{a}_{jl}PQ_1\delta_l^{(n)}. \end{aligned}$$

From the theorem assumptions we have $A_n = O\big(h^{q+2}\big)$. Furthermore, recursion (3.22) together with $\rho = -1$ yield

$$PQ_1(t_n)e_n = (-1)^n\Pi_0PQ_1(t_0)e_0 + \sum_{i=1}^{n}(-1)^{n-i}\Pi_iA_i + \sum_{i=1}^{n}(-1)^{n-i}\Pi_iB_i.$$

where $\Pi_i := \prod_{j=1}^{n-i}\big(I_m + hPQ_1'(t_{n-j})\big)$ for $i = 0, 1, \ldots, n$, *i.e.*, $\Pi_i = O(1)$.

Expanding the last term on the right-hand side of this equation in a Taylor series, grouping the respective terms in the so obtained sums two at a time, and then bounding the resulting sums, we obtain

$$PQ_1 e_n = O(h^{q+1}).$$

Note that we bonus in this process a power of h, because the alternating signs in the first terms of the Taylor sums allow these to be cancelled. This is possible because $\rho = -1$.

For the PP_1-component (3.21 a), $B(q+1)$, $C(q)$, $\alpha = \beta = 1$ and $|\rho| = 1$ imply $PP_1 e_n = O(h^{q+1})$.

The last estimates for $PQ_1 e_n$ and $PP_1 e_n$ together with $\|Pe_n\| \leq \|PP_1 e_n\| + \|PQ_1 e_n\|$ yield

$$\|Pe_n\| \leq Ch^{q+1}.$$

The estimate for Qe_n then follows directly from (3.16 c). □

Lobatto IIIA methods

For these methods the coefficient matrix $\mathcal{A}$ is singular, because in this case $\mathcal{A}$ is expressible in the form

$$\mathcal{A} = \begin{pmatrix} 0 & 0 \\ a & \underline{A} \end{pmatrix}$$

with $a = (a_{21}, a_{31}, \ldots, a_{s1})^T \in \mathbb{R}^{s-1}$, $\underline{A} \in L(\mathbb{R}^{s-1})$, $\underline{A}$ nonsingular.

The best known of such methods is the implicit trapezoidal method. To implement this method for the DAE (1.1) we assume that initial values for the derivates of all the variables are given. So we set in the first step $X'_{01} := x'_0$ and at the end of each step we set $X'_{n1} = X'_{n-1,s}$. Because $X_1 = x_{n-1}$, we need to solve (2.1 b) only for $\widehat{X} = (X_2^T, X_3^T, \ldots, X_s^T)^T \in \mathbb{R}^{m(s-1)}$ in the following system

$$\begin{aligned}[\widehat{D}_A(\underline{A}^{-1} \otimes I_m) + h\widehat{D}_B]\widehat{X} = {} & h\widehat{D}_q \mathbb{1}_{s-1} + \widehat{D}_A(\underline{A}^{-1} \otimes I_m)(\mathbb{1}_{s-1} \otimes x_{n-1}) \\ & + h\widehat{D}_A(\underline{A}^{-1} \otimes I_m)(a \otimes X'_1).\end{aligned}$$

Here and throughout this subsection we define, for a matrix $N(t) \in L(\mathbb{R}^n)$, $\widehat{D}_N := \operatorname{diag}\big(N(t_{n2}), N(t_{n3}), \ldots, N(t_{ns})\big) \in L(\mathbb{R}^{m(s-1)})$.

According to Theorem 2.1 this system is uniquely solvable. The main difference to IRK methods with nonsingular $\mathcal{A}$ is that now the internal stages X_j, $j = 1, 2, \ldots, s$, also depend on the Q-component of x_{n-1}. This fact makes the study of stability and convergence of this class of IRK methods more complicated. By means of the special DAEs (1.1) with constant matrix A we study

here the stability and convergence properties of the method. For this case instead of (3.7), we have the estimates

$$\begin{aligned} \|\triangle PP_1x_n\| &\leq \|\triangle PP_1x_{n-1}\| \\ &\quad + hC^{(n)}\big(\|\triangle PP_1x_{n-1}\| + h\|\triangle Qx_{n-1}\| + \delta\big), \qquad (3.23\text{ a}) \\ \|\triangle PQ_1x_n\| &\leq \delta, \qquad (3.23\text{ b}) \end{aligned}$$

and for the Q-component,

$$\begin{aligned} h\|\triangle Qx_n\| &\leq h\|\triangle Qx_{n-1}\| \\ &\quad + hC^{(n)}\Big(\|\triangle PP_1x_{n-1}\| + h\|\triangle Qx_{n-1}\| + \frac{\delta}{h}\Big). \qquad (3.23\text{ c}) \end{aligned}$$

Applying lemma 3.1 to the recursions (3.23 a) and (3.23 c) we obtain

$$\|\triangle PP_1x_n\| \leq C\delta, \quad h\|\triangle Qx_n\| \leq C\frac{\delta}{h}.$$

These estimates together with (3.23 b) show that Lobatto IIIA methods, applied to the DAE (1.1) with $P' = 0$, are weak unstable on the Q-component, but stable on the P-component.

For the global error e_n, instead of (3.16), we have

$$\begin{aligned} \|PP_1e_n\| &\leq \|PP_1e_{n-1}\| + hC^{(n)}\big(\|PP_1e_{n-1}\| + h\|Qe_{n-1}\| + D_1\big), \qquad (3.24\text{ a}) \\ \|PQ_1e_n\| &\leq C\delta, \qquad (3.24\text{ b}) \\ h\|Qe_n\| &\leq h\|Qe_{n-1}\| + hC^{(n)}\Big(\|PP_1e_{n-1}\| + h\|Qe_{n-1}\| + \frac{D_2}{h}\Big), \qquad (3.24\text{ c}) \end{aligned}$$

with

$$\begin{aligned} D_1 &= \begin{cases} O(h^2), & \text{if } s = 2 \text{ and } \delta = O(h^2), \\ O(h^{s+1}), & \text{if } s \geq 3 \text{ and } \delta = O(h^{s+1}), \end{cases} \\ D_2 &= \begin{cases} O(h^2), & \text{if } s = 2 \text{ and } \delta = O(h^2), \\ O(h^{s+1}), & \text{if } s \geq 2 \text{ and } \delta = O(h^{s+1}). \end{cases} \end{aligned}$$

Applying Lemma 3.1 to the recursions (3.24 a) and (3.24 c) we obtain

$$\begin{aligned} \|PP_1e_n\| &\leq \begin{cases} Ch^2, & \text{if } s = 2 \text{ and } \delta = O(h^2), \\ Ch^{s+1}, & \text{if } s \geq 3 \text{ and } \delta = O(h^{s+1}), \end{cases} \\ \|Qe_n\| &\leq Ch^{s-1} \quad \text{if } \delta = O(h^{s+1}). \end{aligned}$$

When $\delta = O(h^{s+2})$ and $s > 2$ is a even number, we can write (3.24 c) in the form

$$Qe_n = -\big(I_m + O(h)\big)Qe_n + A_n + B_n$$

with $\|A_n\| = O(h^{s+1})$ and $B_n = -\sum_{l=2}^{s} \widehat{a}_{sl}QQ_1(t_{n-1})\delta_l^{(n)}$.

In this case we obtain (in a manner analogous to the proof of the last statement of theorem 3.4) the following estimate for Qe_n:

$$\|Qe_n\| \leq Ch^s.$$

IRK(DAEs) and superconvergence.

For nonlinear DAEs with index 2 in Hessenberg Form, KVÆRNØ [9] has shown that the order of accuracy for some stiffly accurate formulas in the differential component is the same nonstiff order of accuracy. In particular, the Radau IIA and Lobatto IIIC methods retain their nonstiff order of accuracy in the differential component. The analysis makes use of the theory of Butcher series and rooted trees and exploits the Hessenberg structure of the DAE. The same order results based on a similar analysis are given in [4] for the same class of DAEs. These early results make clear that, for any special IRK methods, Theorem 3.2 does not give an optimal estimate for the local error of the P-component. These specially well suited IRK methods for the numerical solution of DAEs are those IRKs for which $c_s = 1$ and $b_j = a_{sj}$, $j = 1, 2, \ldots, s$. Following GRIEPENTROG and MÄRZ [3], we name this class of IRK methods the IRK(DAEs). Note that $\rho = 0$ for IRK(DAEs).

In this subsection we want to give an idea of how one can investigate this interesting property of IRK(DAEs) when they are applied to the index 2 DAE (1.1). For this purpose we suppose that the Runge-Kutta matrix $\mathcal{A}$ is nonsingular and the matrices A, B, P, Q as well as the exact solution $x(t)$ of (1.1) are sufficiently differentiable. Since the relation $e_n = x_n - x(t_n) = X_s - x(t_n) = E_s$ is valid for IRK(DAEs), we only have to consider the equations (3.13 b) to study the local error. Furthermore, (3.14 b) implies that the PQ_1-component is approximated exactly by IRK(DAEs). Hence, to find the conditions for the superconvergence of the P-component we only need to consider the PP_1-component. Multiplying (3.15 b) by $D_{PP_1G_2^{-1}}$ and using the relation $P' = 0$ and Taylor series, we obtain

$$(I_{ms} + N)D_{PP_1}\overline{E} = (\mathcal{A} \otimes I_m)D_{PP_1}(\mathcal{A}^{-1} \otimes I_m)\bar{\delta}^{(n)}. \tag{3.25}$$

Here

$$\begin{aligned} N &= (\mathcal{A} \otimes I_m)\Big[\sum_{l=1}^{k} \frac{h^l}{l!}((\mathcal{A}^{-1} \odot \mathcal{C}^l) \otimes I_m)D_{PP_1^{(l)}} + O(h^{k+1}) + hD_{PP_1G_2^{-1}B}\Big] \\ &= O(h), \end{aligned} \tag{3.26}$$

$\mathcal{C}^l := \underbrace{\mathcal{C} \odot \mathcal{C} \odot \ldots \odot \mathcal{C}}_{l\text{-times}}$, and $\mathcal{C}$ is as in Theorem 2.1.

The equation

$$(I_{ms} + N)^{-1} = \sum_{l=0}^{k} (-1)^l N^l + (-1)^{k+1} (I_{ms} + N)^{-1} N^{k+1}$$

together with $(I_{ms} + N)^{-1} = I_{ms} + O(h)$ implies

$$(I_{ms} + N)^{-1} = \sum_{l=0}^{k} (-1)^l N^l + O(h^{k+1}).$$

Multiplying (3.25) by $(I_{ms} + N)^{-1}$ and using the last equation we arrive at

$$D_{PP_1}\overline{E} = \Big[\sum_{l=0}^{k} (-1)^l N^l (\mathcal{A} \otimes I_m) D_{PP_1} (\mathcal{A}^{-1} \otimes I_m) + O(h^{k+1})\Big] \bar{\delta}^{(n)}. \tag{3.27}$$

Using Taylor's expansion about t_n on the right-hand side of (3.27) we can obtain conditions on the coefficients of the IRK(DAEs) so that an order of accuracy $O(h^{K_I+r})$ for $r > 2$ is attained in the PP_1-component. For the special case $r = 3$ we set $k = 1$ in (3.26) and (3.27). In this case

$$\begin{aligned} D_{PP_1}\overline{E} = \Big[& D_{PP_1} + h(\mathcal{A} \otimes I_m)\big((\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m\big) D_{PP_1'} (I_{ms} - D_{PP_1}) \\ & - h(\mathcal{A} \otimes I_m) D_{PP_1 G_2^{-1} B} + O(h^2)\Big] \bar{\delta}^{(n)}. \end{aligned}$$

The Taylor expansion about t_n yields

$$\begin{aligned} D_{PP_1}\overline{E} = \Big[& D_{PP_1} + h\big[\operatorname{diag}(S(t_n))\big](\mathcal{A}(\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m) \\ & - h\big[\operatorname{diag}(T(t_n))\big](\mathcal{A} \otimes I_m) + O(h^2)\Big] \bar{\delta}^{(n)}, \end{aligned}$$

where $S = PP_1' - PP_1'PP_1$ and $T = PP_1G_2^{-1}B$.

So, the IRK(DAE) applied to the index 2 DAE (1.1) in the P-component has local order $O(h^{q+3})$ when it satisfies the conditions $B(p)$, $C(q)$ with $p \geq q + 2$, $[b^T(\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m]\bar{\delta}^{(n)} = O(h^{q+2})$ and $(b^T \otimes I_m)\bar{\delta}^{(n)} = O(h^{q+2})$.

These conditions are equivalent to $B(p)$, $C(q)$ with $p \geq q + 2$ and

$$b^T(\mathcal{C} \odot \mathcal{A}^{-1})\left(\mathcal{A}c^q - \frac{1}{q+1} c^{q+1}\right) = 0.$$

These conditions are fulfilled for the 3-stage Lobatto IIIC and Radau IIA methods, *i.e.*, for these methods we have local order $O(h^5)$ and $O(h^6)$ respectively. This implies that these methods retain the nonstiff order of accuracy in the P-component (superconvergence).

In an analogous manner we can use (3.27) to deduce conditions for a higher order of accuracy. However, we can see that the calculations for these conditions become very tedious.

Remarks

- For the differential component of DAEs with index 2, Theorems 3.1, 3.2, 3.3 and 3.4 summarize analogous results on stability and convergence of IRK methods as when these are applied to the classical nonstiff ODEs.
- It is well-known that IRK methods applied to DAEs with index 2 remain unstable. Theorem 3.1 shows that only in the nullspace component the stability property gets lost, and that the instability of the Q-component is only weak.
- The results of Theorems 3.1, 3.2, 3.3 and 3.4 are also achieved in [4], but in [4] only systems in Hessenberg form are considered; *i.e.*, we generalize the results of [4] to the class of linear time-varying DAEs with index 2.

4. Stability and convergence for the variable nullspace case

When the IRK (2.1) is feasible, we can arrive to equation (2.6). Solving it with respect to $D_{PP_1}\overline{X}$, $D_{PQ_1}\overline{X}$ and $D_Q\overline{X}$, we obtain

$$\begin{aligned} D_{PP_1}\overline{X} = {} & (I_{ms} + O(h))\left[h(\mathcal{A}\otimes I_m)D_{PP_1G_2^{-1}q}\mathbb{1}_s + D_{PP_1}(\mathbb{1}_s\otimes x_{n-1})\right] \\ & + O(\|D_q\mathbb{1}_s\|) + O(\|D_{QQ_1}(\mathbb{1}_s\otimes x_{n-1})\|) + O(h\|x_{n-1}\|), \end{aligned} \tag{4.1 a}$$

$$D_{PQ_1}\overline{X} = D_{PQG_2^{-1}q}\mathbb{1}_s, \tag{4.1 b}$$

$$\begin{aligned} D_Q\overline{X} = {} & (\mathcal{D}^{-1} + O(h))\Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QQ_1G_2^{-1}q}\mathbb{1}_s \\ & + D_{QP_1G_2^{-1}q}\mathbb{1}_s + \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'PQ_1G_2^{-1}q}\mathbb{1}_s \\ & + O(h\|D_{PP_1G_2^{-1}q}\mathbb{1}_s\|) - \frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QQ_1}(\mathbb{1}_s\otimes x_{n-1}) \\ & - \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'}(\mathbb{1}_s\otimes x_{n-1}) \\ & - D_{QP_1G_2^{-1}BPP_1}(\mathbb{1}_s\otimes x_{n-1}) \\ & + \big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{QQ_1'PP_1}(\mathbb{1}_s\otimes x_{n-1}) + O(h\|x_{n-1}\|)\Big)\Big]. \end{aligned} \tag{4.1 c}$$

System (4.1) is essentially different from (3.1), because now the internal stages X_j, $j = 1, 2, \ldots, s$, also depend on the Q-component of the approximation x_{n-1}. That means that the instability of the Q-component can work out to instability on the P-component, and that the weak instability can be amplified to exponential instability. For example, applying the implicit Euler method to

the DAE of the example 2.2 we obtain the recursions

$$\begin{aligned} x_n^{(1)} &= q_n^{(1)} - \eta t_n x_n^{(2)}, \\ x_n^{(2)} &= \frac{\eta}{1+\eta} x_{n-1}^{(2)} + \frac{1}{1+\eta}\left(q_n^{(2)} - \frac{1}{h}(q_n^{(1)} - q_{n-1}^{(1)})\right), \qquad \eta \neq 1. \end{aligned}$$

Evidently, the last recursion has an exponentially unstable behaviour when $\eta < -1/2$.

This simple example shows that, in general, IRK methods are not suitable for the numerical solution of DAEs with index 2 and variable nullspace. Nevertheless, when we restri ourselves to linear DAEs with index 2 and variable nullspace, we can show that, for these special cases, the instability is weak and the method remains convergent. To confirm this fact, let us restrict ourselves to IRK(DAEs), so we avoid the recursion (2.1 a) to calculate x_n. Furthermore, we put $\mathcal{S} = 0$ in (2.5), so we obtain for the difference $D_{PP_1}\Delta\overline{X}$, instead of a relation of the form (4.1 a), the equation

$$\begin{aligned} D_{PP_1}\Delta\overline{X} =& (I_m + O(h))[h(\mathcal{A}\otimes I_m) D_{PP_1 G_2^{-1}\delta_n} \mathbb{1}_s + D_{PP_1}(\mathbb{1}_s \otimes \Delta x_{n-1}) + \\ & (\mathcal{A}\otimes I_m) R_{PP_1}(\mathbb{1}_s \otimes \Delta x_{n-1})] + O(h\|D_{\delta_n}\mathbb{1}_s\|) + O(h^2\|\Delta x_{n-1}\|). \end{aligned}$$

Taylor's expansion about t_{n-1} on the right-hand side of this equation yields

$$\begin{aligned} D_{PP_1}\Delta\overline{X} = & \left(I_{ms} + O(h)\right)\left[h(\mathcal{A}\otimes I_m) D_{PP_1 G_2^{-1}\delta_n}\mathbb{1}_s + (\mathbb{1}_s \otimes \Delta PP_1(t_{n-1})x_{n-1})\right. \\ & + h\left((\overline{E} - \bar{\delta}^{(n)})\mathrm{diag}\left((PP_1)'(t_{n-1})\right) + O(h)\right)(\mathbb{1}_s \otimes \Delta x_{n-1}) \\ & + h\left((\mathcal{A}((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m)\mathrm{diag}\left((PP_1)'(t_{n-1})\right) + O(h)\right)(\mathbb{1}_s\otimes\Delta x_{n-1})\Big] \\ & + O(h\|D_{\delta_n}\|). \end{aligned}$$

Here $D_c = \mathrm{diag}\,(c_1, c_2, \ldots, c_s) \in L(\mathbb{R}^s)$.

This equation and the identities

$$\mathcal{A}(\mathcal{C}\odot\mathcal{A}^{-1}) = \mathcal{A}D_c\mathcal{A}^{-1} - D_c,\ c_s = 1,\ b_j = a_{sj},\ X_s = x_n,\ \tilde{X}_s = \tilde{x}_n$$

imply that

$$\begin{aligned} \Delta PP_1 x_n =& \Delta PP_1 x_{n-1} + hb^T D_c\mathcal{A}^{-1}\mathbb{1}_s(PP_1)'(t_{n-1})\Delta x_{n-1} \\ & + O(h\|\Delta PP_1 x_{n-1}\|) + O(h^2\|\Delta x_{n-1}\|) \\ & + h\left(I_m + O(h)\right)\sum_{j=1}^{s} a_{sj} PP_1 G_2^{-1}(t_{nj})\delta_{nj}\,. \end{aligned}$$

We aim at finding recursions for $\Delta PP_1 x_n$ and $h\Delta Qx_n$, so that we can apply lemma 3.1 on these recursions. This means that the Q-component of the second term on the right-hand side of the latter equation must vanish. The condition

for this is $b^T D_c \mathcal{A}^{-1} \mathbb{1}_s (PP_1)'Q(t) = 0$. When this condition is fulfilled, we obtain

$$\|\triangle PP_1 x_n\| \leq \|\triangle PP_1 x_{n-1}\| + hC^{(n)}(\|\triangle PP_1 x_{n-1}\| + h\|\triangle Qx_{n-1}\| + \delta), \tag{4.2 a}$$

$$\|\triangle PQ_1 x_n\| \leq C\delta. \tag{4.2 b}$$

For the difference $D_Q \triangle \overline{X}$ we obtain using Taylor's expansion about t_{n-1},

$$h\|\triangle Qx_n\| \leq r_n h\|\triangle Qx_{n-1}\| + hC^{(n)}\Big(\|\triangle PP_1 x_{n-1}\| + h\|\triangle Qx_{n-1}\| + \frac{\delta}{h}\Big). \tag{4.2 c}$$

Here,

$$r_n = \Big\| \sum_{i=1}^{s} \sum_{j=1}^{s} c_j \widehat{a}_{ji} \widehat{d}_{sj} (QQ_1)'Q(t_{n-1}) \Big\|, \widehat{d}_{sj} \in L(\mathbb{R}^m),$$

and

$$\mathcal{D}^{-1} = (\widehat{d}_{ij})_{i,j=1}^{s} \in L(\mathbb{R}^{ms}).$$

Theorem 4.1. *Let the assumptions of Theorem 2.1 be satisfied and suppose that the IVP (1.1) satisfies the conditions $(PP_1)'Q(t) = 0$ and $r_i \leq 1$, $i = 1, 2, \ldots, n$. Then, the IRK(DAE) applied to this IVP is weakly unstable in the Q-component but stable in the P-component. Furthermore, if $e_0 = O(h^{q+\alpha})$ are the errors in the initial conditions for $q \geq 1$ and $\delta_{ij} = O(h^{q+\alpha})$ are the numerical errors in the solution of the system (2.1 b) at the i-th integration step, then the global error e_n satisfies the following estimates:*

- $\|Pe_n\| \leq Ch^q$, *if* $B(q)$, $C(q)$ *and* $\alpha = 0$.
- $\|Pe_n\| \leq Ch^{q+1}$, *if* $B(q+1)$, $C(q)$ *and* $\alpha = 1$.
- $\|Qe_n\| \leq Ch^q$, *if* $B(q)$, $C(q)$, $\alpha = 1$ *and* $r_i < 1$.
- $\|Qe_n\| \leq Ch^{q-1}$, *if* $B(q)$, $C(q)$, $\alpha = 1$ *and* $r_i = 1$, $i = 1, 2, \ldots, n$,

Proof. Condition $(PP_1)'Q(t) = 0$ implies $\mathcal{S} = 0$. Applying lemma 3.1 to the recursions (4.2 a) and (4.2 c) we obtain the stability statement. In order to prove the convergence statement we consider (3.13 b). Under the assumptions of the theorem and using the techniques of section 3 we arrive at

$$\|PP_1 e_n\| \leq \|PP_1 e_{n-1}\| + hC^{(n)}(\|PP_1 e_{n-1}\| + h\|Qe_{n-1}\| + D_1),$$

$$h\|Qe_n\| \leq r_n h\|Qe_{n-1}\| + hC^{(n)}\Big(\|PP_1 e_{n-1}\| + h\|Qe_{n-1}\| + \frac{D_2}{h}\Big),$$

$$D_1 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q+1) \text{ and } C(q), \end{cases}$$

$$D_2 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q) \text{ and } C(q). \end{cases}$$

Applying once again the lemma 3.1 to these recursions we obtain the convergence statements. □

Example 4.1. We consider once again the DAE of Example 2.2. In this case $PP_1 = 0$, so that the first condition in Theorem 4.1 is fulfilled. Applying the implicit Euler method to this DAE we have

$$\mathcal{D} = I_2 - QQ_1P'(t) = \begin{pmatrix} 1 & -\eta^2 t \\ 0 & 1+\eta \end{pmatrix}.$$

For $\eta \neq -1$

$$\mathcal{D}^{-1} = \frac{1}{1+\eta} \begin{pmatrix} 1+\eta & \eta^2 t \\ 0 & 1 \end{pmatrix}.$$

exists. Hence,

$$c_1 \hat{a}_{11} \mathcal{D}^{-1}(QQ_1)'Q = \mathcal{D}^{-1}(QQ_1)'Q = \frac{\eta}{1+\eta} \begin{pmatrix} 0 & \eta t \\ 0 & -1 \end{pmatrix},$$

i.e.,

$$r_n = \left| \frac{\eta}{1+\eta} \right| \left\| \begin{pmatrix} 0 & \eta t_{n-1} \\ 0 & -1 \end{pmatrix} \right\|. \quad \square$$

Although we prove in Theorem 4.1 that the IRK(DAEs) applied to any linear DAEs with index 2 and variable nullspace provides the same stabilty and convergence results as in the case of constant nullspace, the assumptions of this theorem restricts strongly the general DAE (1.1) with index 2. To overcome these drawbacks we will try to look at the IRK methods for the DAE (1.1) in a new way. The new method appears to be promising for the solution of the DAE (1.1) with index 2 and variable nullspace.

5. A modified IRK method to the numerical solution of DAEs with index 2 and variable nullspace

The greatest difficulty in the numerical integration of the DAE (1.1) with variable nullspace by IRK methods lies in the fact that the P-component, which is stable, also depends on the Q-component of the numerical approximation x_{n-1}. For this reason we now first split the DAE in the two components Px and Qx and in each one we apply the IRK method (1.2) separately. For this aim we consider the DAE (1.1) with nonconstant nullspace $N(t)$. Under the assumption that (1.1) is transferable with index 2 we can decouple it in the following way:

$$\begin{aligned} \big[A(t) + \big(B(t) - A(t)P'(t)\big)Q(t)\big]\big[P(t)\big(P(t)x(t)\big)' + Q(t)x(t)\big] \\ + B(t)P(t)x(t) = q(t). \end{aligned}$$

Obviously, this equation is equivalent to (1.1). Now we set $y(t) := P(t)x(t)$ and $z(t) := Q(t)x(t)$, so the DAE (1.1) can be written as

$$A(t)y'(t) + B(t)y(t) + A_1(t)z(t) = q(t). \tag{5.1 a}$$

The new variables y and z satisfy the identities $P(t)y(t) = y(t)$ and $Q(t)z(t) = z(t)$. We rewrite these equations as $y(t) - P(t)y(t) + z(t) - Q(t)z(t) = 0$ or, equivalently,

$$Q(t)y(t) + P(t)z(t) = 0. \tag{5.1 b}$$

Formulation of the method.

To solve the IVP (1.1) we compute the approximations y_{nj} and z_{nj} of the solution in t_{nj} in the way described below and then we construct the approximation x_{nj} of $x(t_{nj})$ via the formula

$$x_{nj} := y_{nj} + z_{nj}\,, \qquad j = 1, 2, \ldots, s.$$

Note, that the exact solution of (1.1) satisfies $x(t) = y(t) + z(t)$.

Numerical approach to the y-component.

We set in the first step $y_0 := P(t_0)x_0$, and in the n-th integration step we construct the approximation $y_n = y_{n0}$ via the formulas

$$y_n = \rho y_{n-1} + \sum_{j=1}^{s}\sum_{l=1}^{s} b_j \widehat{a}_{jl} Y_l \tag{5.2}$$

where Y_l is defined by

$$A(t_{nj})\sum_{l=1}^{s}\widehat{a}_{jl}Y_l + hB(t_{nj})Y_j + hA_1(t_{nj})Z_j = hq(t_{nj}) + A(t_{nj})\sum_{l=1}^{s}\widehat{a}_{jl}y_{n-1}\,, \tag{5.3 a}$$

$$Q(t_{nj})Y_j + P(t_{nj})Z_j = 0. \tag{5.3 b}$$

In the n-th integration step we set $y_{nj} = Y_j$, for $j = 1, 2, \ldots, s$.

Numerical approach to the z-component.

If the system (5.3) is uniquely solvable with respect to (Y_j, Z_j) (we prove this statement below), we obtain a numerical solution for the z-component at t_{nj} directly from the internal stages Z_j, *i.e.*, we can take Z_j as the approximation of $z(t_{nj})$ in the n-th integration step. Furthermore, it is possible to obtain an approximation for the z-component at t_i, $i = 1, 2, \ldots, n$, via the formula

$$z_n = \rho z_{n-1} + \sum_{j=1}^{s}\sum_{l=1}^{s} b_j \widehat{a}_{jl} Z_l.$$

We renounce here to consider the latter approximation, because the calculation of (Y_j, Z_j) from (5.3) does not depend on the approximation z_{n-1}.

Existence and uniqueness of the Runge-Kutta solution.

Using the notation of Section 2, system (5.3) becomes equivalent to

$$\left[D_A(\mathcal{A}^{-1} \otimes I_m) + hD_B\right]\overline{Y} + hD_{A_1}\overline{Z} = hD_q \mathbf{1}_s + D_A(\mathcal{A}^{-1}\mathbf{1}_s \otimes y_{n-1}), \tag{5.4 a}$$

$$D_Q\overline{Y} + D_P\overline{Z} = 0. \tag{5.4 b}$$

Here, $\overline{Y} = (Y_1^T, Y_2^T, \ldots, Y_s^T)^T$ and $\overline{Z} = (Z_1^T, Z_2^T, \ldots, T_s^T)^T$.

Together with (5.4) we consider the perturbed system

$$\left[D_A(\mathcal{A}^{-1} \otimes I_m) + hD_B\right]\overline{\tilde{Y}} + hD_{A_1}\overline{\tilde{Z}} = hD_{(q+\delta_n)}\mathbf{1}_s + D_A(\mathcal{A}^{-1}\mathbf{1}_s \otimes \tilde{y}_{n-1}), \tag{5.5 a}$$

$$D_Q\overline{\tilde{Y}} + D_P\overline{\tilde{Z}} = \overline{\Theta}, \tag{5.5 b}$$

where

$$\overline{\tilde{Y}} := (\tilde{Y}_1^T, \tilde{Y}_2^T, \ldots, \tilde{Y}_s^T)^T,\ \overline{\tilde{Z}} := (\tilde{Z}_1^T, \tilde{Z}_2^T, \ldots, \tilde{Z}_s^T)^T,\ \overline{\Theta} := (\theta_{n1}^T, \theta_{n2}^T, \ldots, \theta_{ns}^T)^T,$$

and we suppose that the pertubations satisfy $\|\delta_{ij}\| \leq \delta$ and $\|\theta_{ij}\| \leq \delta$ for all $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, s$.

Obviously, (5.4) is equivalent to (5.5) if $\delta = 0$.

Theorem 5.1. *Suppose that the IVP (1.1) is tractable with index 2 and that the Runge-Kutta matrix $\mathcal{A}$ is nonsingular. Then systems (5.4) and (5.5) are uniquely solvable with respect to $(\overline{Y}, \overline{Z})$ and $(\overline{\tilde{Y}}, \overline{\tilde{Z}})$, respectively, for sufficiently small h.*

Proof. Using the techniques of Section 2 and under the consideration of (5.5 b) we decouple the system (5.5 a) into the PP_1-, Q_1- and Q-components. This yields:

$$\left[D_{PP_1}(\mathcal{A}^{-1} \otimes I_m) + hD_{PP_1G_2^{-1}BPP_1}\right]\overline{\tilde{Y}} =$$

$$hD_{PP_1G_2^{-1}(q+\delta_n)}\mathbf{1}_s + D_{PP_1}(\mathcal{A}^{-1}\mathbf{1}_s \otimes \tilde{y}_{n-1}) - hD_{(PP_1+PP_1P')}\overline{\Theta}, \tag{5.6 a}$$

$$D_{Q_1}\overline{\tilde{Y}} = D_{Q_1G_2^{-1}(q+\delta_n)}\mathbf{1}_s, \tag{5.6 b}$$

$$\left[-D_{QQ_1}(\mathcal{A}^{-1} \otimes I_m) + hD_{QP_1G_2^{-1}BPP_1}\right]\overline{\tilde{Y}} + hD_Q\overline{\tilde{Z}} = \tag{5.6 c}$$

$$hD_{QP_1G_2^{-1}(q+\delta_n)}\mathbf{1}_s - D_{QQ_1}(\mathcal{A}^{-1}\mathbf{1}_s \otimes \tilde{y}_{n-1}) - hD_{Q(P_1-Q_1'Q)}\overline{\Theta}. \tag{5.6 d}$$

Inserting (5.6 b) into (5.6 a) and using Taylor's expansion, we obtain

$$\begin{aligned}\left[(\mathcal{A}^{-1}\otimes I_m)+O(h)\right]D_{PP_1}\overline{\dot{Y}} &= hD_{PP_1G_2^{-1}(q+\delta_n)}\mathbb{1}_s\\ &\quad - R_{PP_1}D_{PQ_1G_2^{-1}(q+\delta_n)}\mathbb{1}_s + D_{PP_1}(\mathcal{A}^{-1}\mathbb{1}_s\otimes\tilde{y}_{n-1})+O(h\|\overline{\dot{\Theta}}\|).\end{aligned}$$

Here R_{PP_1} is as in (2.3 c).

The matrix on the left-hand side of this equation is invertible if h is sufficiently small. Thus, we can write

$$\begin{aligned}D_{PP_1}\overline{\dot{Y}} = \left[(\mathcal{A}\otimes I_m)+O(h)\right]\big[&hD_{PP_1G_2^{-1}(q+\delta_n)}\mathbb{1}_s - R_{PP_1}D_{PQ_1G_2^{-1}(q+\delta_n)}\mathbb{1}_s\\ &+ D_{PP_1}(\mathcal{A}^{-1}\mathbb{1}_s\otimes\tilde{y}_{n-1})+O(h\|\overline{\dot{\Theta}}\|)\big].\end{aligned}\tag{5.7}$$

Using the equation $\overline{\dot{Y}} = D_P\overline{\dot{Y}} + D_Q\overline{\dot{\Theta}} = D_{PP_1}\overline{\dot{Y}} + D_{PQ_1}\overline{\dot{Y}} + D_Q\overline{\dot{\Theta}}$, together with (5.6 b) and (5.7), we obtain $\overline{\dot{Y}}$. As for $\overline{\dot{Z}}$, it follows directly from (5.6 c) and the identity $D_P\overline{\dot{Z}} = \overline{\dot{Z}} - D_Q\overline{\dot{Z}} = D_P\overline{\dot{\Theta}}$. □

Stability and convergence.

In addition to the vector y_n obtained from (5.2) we consider

$$\tilde{y}_n = \rho\tilde{y}_{n-1} + \sum_{j=1}^{s}\sum_{l=1}^{s} b_j\hat{a}_{jl}\tilde{Y}_l + h\delta_{n,s+1}\,.$$

Subtracting this equation from (5.2) we obtain

$$\Delta y_n = \rho\Delta y_{n-1} + (b^T\mathcal{A}^{-1}\otimes I_m)\Delta\overline{Y} + h\delta_{n,s+1}\,.\tag{5.8}$$

Here $\Delta y_k = \tilde{y}_k - y_k$, $k = n-1, n$ and $\Delta\overline{Y} := \overline{\tilde{Y}} - \overline{Y}$.

Subtracting (5.4) from (5.5) and using the same techniques of the feasibility proof on the equation obtained in this way, we arrive at

$$\begin{aligned}D_{PP_1}\Delta\overline{\dot{Y}} =& h\big((\mathcal{A}\otimes I_m)+O(h)\big)D_{PP_1G_2^{-1}\delta_n}\mathbb{1}_s\\ &- h\left[(\mathcal{A}(\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m)D_{(PP_1)'}+O(h)\right]D_{PQ_1G_2^{-1}\delta_n}\mathbb{1}_s\\ &+\big(D_{PP_1}+O(h)\big)\big(\mathbb{1}_s\otimes\Delta y_{n-1}\big)+O(h\|\overline{\Theta}\|),\end{aligned}\tag{5.9 a}$$

$$D_{PQ_1}\Delta\overline{\dot{Y}} = D_{PQ_1G_2^{-1}\delta_n}\mathbb{1}_s\,,\tag{5.9 b}$$

$$\begin{aligned}\Delta\overline{Z} =& \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QQ_1}+D_{QP_1}+\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'PQ_1}\\ &+O(h)\Big]D_{G_2^{-1}\delta_n}\mathbb{1}_s+\Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{QP_1P}-D_{QP_1G_2^{-1}BPP_1}\\ &+\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'}(D_{PP_1}-I_{ms})+O(h)\Big](\mathbb{1}_s\otimes\Delta y_{n-1})\\ &+O(\|\overline{\Theta}\|).\end{aligned}\tag{5.9 c}$$

Now we can use (5.8) and (5.9) to prove the stability and convergence statements. We will also need the following lemma. For its proof we refer the reader to [8].

Lemma 5.1. *Suppose that the sequences of positive numbers* (y_n), (z_n) *and* (w_n) *satisfy*

$$\begin{aligned} y_n &\le y_{n-1} + hC(y_{n-1} + z_{n-1} + w_{n-1} + D_1), \\ z_n &\le |\rho| z_{n-1} + hC(y_{n-1} + z_{n-1} + w_{n-1} + \frac{D_2}{h}), \\ w_n &\le |\rho| w_{n-1} + hC\Big(z_{n-1} + y_{n-1} + w_{n-1} + \frac{D_3}{h}\Big). \end{aligned}$$

Then we have, for $hn \le c_0$,

$$\begin{aligned} y_n &\le C_1(y_0 + hz_0 + hw_0 + D_1 + D_2 + D_3), \quad \text{if } |\rho| \le 1, \\ z_n &\le C_1(hy_0 + (|\rho|^n + h)z_0 + hw_0 + hD_1 + D_2 + hD_3), \quad \text{if } |\rho| < 1, \\ w_n &\le C_1(hy_0 + hz_0 + (|\rho|^n + h)w_0 + hD_1 + hD_2 + D_3), \quad \text{if } |\rho| < 1. \end{aligned}$$

Furthermore,

$$\begin{aligned} z_n &\le C_1\Big(hy_0 + z_0 + h^2 w_0 + hD_1 + \frac{D_2}{h} + D_3\Big), \quad \text{if } |\rho| = 1, \\ w_n &\le C_1\Big(hy_0 + hz_0 + hw_0 + hD_1 + D_2 + \frac{D_3}{h}\Big), \quad \text{if } |\rho| = 1. \end{aligned}$$

Theorem 5.2. *Let the assumption of the Theorem 5.1 be satisfied. If the coefficients of the IRK method satisfy* $|\rho| < 1$, *then the method (5.2), (5.3) is weakly unstable. This weak instability refers to the z-component of the numerical solution. The y-component is stable.*

Proof. Multiplying (5.8) by $PP_1(t_n)$, $PQ_1(t_n)$ and $Q(t_n)$, and using Taylor's expansion about t_{n-1} on the right-hand side of the equations so obtained, we conclude that

$$\begin{aligned} \triangle PP_1 y_n =& \rho \triangle PP_1 y_{n-1} + h\rho\big((PP_1)'(t_{n-1}) + O(h)\big)\triangle y_{n-1} \\ &+ (b^T \mathcal{A}^{-1} \otimes I_m) D_{PP_1} \triangle \overline{Y} + h\big((b^T \mathcal{A}^{-1} \otimes I_m) D_{\gamma(PP_1)'} + O(h)\big)\triangle \overline{Y} \\ &+ hPP_1(t_n)\delta_{n,s+1}, \end{aligned} \tag{5.10 a}$$

$$\begin{aligned} \triangle PQ_1 y_n =& \rho \triangle PQ_1 y_{n-1} + h\rho\big((PQ_1)'(t_{n-1}) + O(h)\big)\triangle y_{n-1} \\ &+ (b^T \mathcal{A}^{-1} \otimes I_m) D_{PQ_1} \triangle \overline{Y} + h\big((b^T \mathcal{A}^{-1} \otimes I_m) D_{\gamma(PQ_1)'} + O(h)\big)\triangle \overline{Y} \\ &+ hPP_1(t_n)\delta_{n,s+1}, \end{aligned} \tag{5.10 b}$$

$$\begin{aligned}\triangle Qy_n =&\rho\triangle Qy_{n-1} + h\rho\big(Q'(t_{n-1}) + O(h)\big)\triangle y_{n-1}\\ &+ (b^T\mathcal{A}^{-1}\otimes I_m)D_Q\overline{\Theta} + h\big((b^T\mathcal{A}^{-1}\otimes I_m)D_{\gamma Q'} + O(h)\big)\triangle\overline{Y}\\ &+ hQ(t_n)\delta_{n,s+1}\,. \qquad (5.10\text{ c})\end{aligned}$$

Here $\triangle PP_1y_k = PP_1(t_k)(\tilde{y}_k - y_k)$, $\triangle PQ_1y_k = PQ_1(t_k)(\tilde{y}_k - y_k)$ and $\triangle Qy_k = Q(t_k)(\tilde{y}_k - y_k)$, $k = n-1, n$.

Inserting (5.9 a), (5.9 b) into (5.10) and using the equation

$$\triangle\overline{Y} = D_{PP_1}\triangle\overline{Y} + D_{PQ_1}\triangle\overline{Y} + D_Q\overline{\Theta},$$

we obtain the recursions:

$$\|\triangle PP_1y_n\| \le \|\triangle PP_1y_{n-1}\| + hC^{(n)}\big(\|\triangle y_{n-1}\| + \delta\big), \qquad (5.11\text{ a})$$

$$\|\triangle PQ_1y_n\| \le |\rho|\,\|\triangle PQ_1y_{n-1}\| + hC^{(n)}\left(\|\triangle y_{n-1}\| + \frac{\delta}{h}\right), \qquad (5.11\text{ b})$$

$$\|\triangle Qy_n\| \le |\rho|\,\|\triangle Qy_{n-1}\| + hC^{(n)}\left(\|\triangle y_{n-1}\| + \frac{\delta}{h}\right). \qquad (5.11\text{ c})$$

Applying Lemma 5.1 to the recursions (5.11) and using the relation $\|\triangle y_n\| \le \|\triangle PP_1y_n\| + \|\triangle PQ_1y_n\| + \|\triangle Qy_n\|$, we obtain

$$\|\triangle y_n\| \le C\delta. \qquad (5.12)$$

Now equation (5.9 c) implies

$$\|\triangle Z_j\| \le \frac{C_1}{h}\Big(\|\triangle PQ_1y_{n-1}\| + h\|\triangle y_{n-1}\| + \delta\Big).$$

Inserting (5.12) into this relation, we obtain in the n-th integration step

$$\|\triangle Z_j\| \le \frac{C\delta}{h}, \qquad j = 1, 2, \ldots, s. \qquad \square$$

To study the convergence we set $e_n^{(y)} = y_n - y(t_n)$, $E_j^{(y)} = Y_j - y(t_{nj})$ and $E_j^{(z)} = Z_j - z(t_{nj})$. Then $e_n^{(y)}$ satifies the recursion

$$e_n^{(y)} = \rho e_{n-1}^{(y)} + (b^T\mathcal{A}^{-1}\otimes I_m)(\overline{E}^{(y)} - \bar{\delta}^{(n)}) + \delta_{s+1}^{(n)}. \qquad (5.13)$$

Under the assumptions of Theorem 5.1, the differences $E_j^{(y)}$ and $E_j^{(z)}$ satisfy

$$D_{PP_1}\overline{E}^{(y)} = \big(D_{PP_1} + O(h)\big)(\bar{e}_{n-1}^{(y)} + \bar{\delta}^{(n)}) + O(h\|\bar{\delta}_n\| + h\|\overline{\Theta}_n\|), \qquad (5.14\text{ a})$$

$$D_{PQ_1}\overline{E}^{(y)} = D_{PQ_1G_2^{-1}}\bar{\delta}_n\,, \qquad (5.14\text{ b})$$

$$\begin{aligned}\overline{E}^{(z)} =& \Big[\frac{1}{h}(\mathcal{A}^{-1}\otimes I_m)D_{PQ_1P} - D_{QP_1G_2^{-1}BPP_1}\\ &+\big((\mathcal{C}\odot\mathcal{A}^{-1})\otimes I_m\big)D_{(QQ_1)'}(D_{PP_1} - I_{ms}) + O(h)\Big](\bar{e}_{n-1}^{(y)} + \bar{\delta}^{(n)})\\ &+O\Big(\frac{1}{h}\|\bar{\delta}_n\| + \|\overline{\Theta}_n\|\Big), \qquad (5.14\text{ c})\end{aligned}$$

where

$$
\begin{aligned}
\overline{E}^{(y)} &:= \big((E_1^{(y)})^T, (E_2^{(y)})^T, \dots, (E_s^{(y)})^T\big)^T, \\
\overline{E}^{(z)} &:= \big((E_1^{(z)})^T, (E_2^{(z)})^T, \dots, (E_s^{(z)})^T\big)^T, \\
\bar{\delta}_n &= (\delta_{n1}^T, \delta_{n2}^T, \dots, \delta_{ns}^T), \\
\overline{\Theta}_n &= (\theta_{n1}^T, \theta_{n2}^T, \dots, \theta_{ns}^T)^T, \\
\bar{e}_{n-1}^{(y)} &:= (e_{n-1}^{(y)} \otimes \mathbb{1}_s), \\
e_{n-1}^{(y)} &= y_{n-1} - y(t_{n-1}),
\end{aligned}
$$

$\bar{\delta}^{(n)}$ and $\delta_{s+1}^{(n)}$ are as in (3.13 a), $(\delta_{nj}, \theta_{nj})$ are the numerical errors in the solution of the system (5.4) at the n-th integration step, and $\delta_j^{(n)}$ is the j-th internal truncation error of the IRK method.

Theorem 5.3. *Let the assumptions of Theorem 5.1 be satisfied and suppose that $e_0^{(y)} = O(h^{q+\beta})$ is the error in the internal condition for $q \geq 1$, and $\delta_{ij} = O(h^{q+\alpha})$ is the numerical error in the solution of the system (5.4) at the i-th integration step. Then the global error e_n satisfies:*

- $\|e_n^{(y)}\| \leq Ch^q$, *if* $B(q)$, $C(q)$, $\alpha = \beta = 0$ *and* $|\rho| < 1$.
- $\|e_n^{(y)}\| \leq Ch^{q+1}$, *if* $B(q+1)$, $C(q)$, $\alpha = \beta = 1$ *and* $|\rho| < 1$.
- $\|E_j^{(z)}\| \leq Ch^q$, *if* $B(q)$, $C(q)$, $\alpha = \beta = 1$ *and* $|\rho| < 1$.
- $\|E_j^{(z)}\| \leq Ch^{q+1}$, *if* $B(q+1)$, $C(q)$, $b^T\mathcal{A}^{-1}c^{q+1} = 1$, $b^T\mathcal{A}^{-2}c^{q+1} = q+1$, $\alpha = \beta = 2$ *and* $|\rho| < 1$.
- $\|e_n^{(y)}\| \leq Ch^q$, $\|E_j^{(z)}\| \leq Ch^{q-1}$, *if* $\alpha = 1$, $\beta = 0$, $B(q)$, $C(q)$ *and* $|\rho| = 1$.
- $\|e_n^{(y)}\| \leq Ch^{q+1}$, $\|E_j^{(z)}\| \leq Ch^q$, *if* $\alpha = 2$, $\beta = 1$, $B(q+1)$, $C(q)$, $b^T\mathcal{A}^{-1}c^{q+1} = 1$ *and* $|\rho| = 1$, $j = 1, 2, \dots, s$.

Proof. Multiplying (5.13) by $PP_1(t_n)$, $PQ_1(t_n)$ and $Q(t_n)$, inserting (5.15) into the obtained equations, and using the Taylor expansion, we obtain the recursions

$$
\|PP_1e_n^{(y)}\| \leq \|PP_1e_{n-1}^{(y)}\| + hC^{(n)}(\|e_{n-1}^{(y)} + D_1), \tag{5.15 a}
$$

$$
\|PQ_1e_n^{(y)}\| \leq |\rho|\,\|PQ_1e_{n-1}^{(y)}\| + hC^{(n)}\Big(\|e_{n-1}^{(y)} + \frac{D_2}{h}\Big), \tag{5.15 b}
$$

$$
\|Q_1e_n^{(y)}\| \leq |\rho|\,\|Q_1e_{n-1}^{(y)}\| + hC^{(n)}\Big(\|e_{n-1}^{(y)} + \frac{D_3}{h}\Big), \tag{5.15 c}
$$

where

$$D_1 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q+1) \text{ and } C(q), \end{cases}$$

$$D_2 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q) \text{ and } C(q), \\ O(h^{q+2}), & \text{if } \alpha = 2,\ b^T\mathcal{A}^{-1}c^{q+1} = 1,\ B(q) \text{ and } C(q), \end{cases}$$

$$D_3 = \begin{cases} O(h^q), & \text{if } \alpha = 0,\ B(q) \text{ and } C(q), \\ O(h^{q+1}), & \text{if } \alpha = 1,\ B(q) \text{ and } C(q), \\ O(h^{q+2}), & \text{if } \alpha = 2,\ b^T\mathcal{A}^{-1}c^{q+1} = 1,\ B(q) \text{ and } C(q). \end{cases}$$

Applying Lemma 5.1 to the recursions (5.15), we obtain the assertions of the theorem for the y-component. The assertions for the z-component follow from (5.14 c) and the estimates for $e_n^{(y)}$. □

Remarks

- If $|\rho| < 1$, then Theorem 5.3 shows that the modified IRK method (5.2), (5.3), applied to the DAE (1.1) with variable nullspace, is convergent with the same order of the IRK (1.2) applied to the DAE (1.1) with constant nullspace.
- For the case $|\rho| = 1$ we bonus by the modified IRK method (5.2), (5.3) a power of h in the z-component with respect to the statements of theorem 3.4. This fact is not surprising because we have used for the approximation of the z-component the collocation equation (5.3 a) instead of the recursion

$$z_n = \rho z_{n-1} + \sum_{j=1}^{s}\sum_{l=1}^{s} b_j \widehat{a}_{jl} Z_l.$$

- The conditions $|\rho| = 1$ and $b^T\mathcal{A}^{-1}c^{q+1} = 1$ are not usually fulfilled for important methods (for example the Gauss-Legendre method). Thus for these methods we must count with a loss of order of convergence. However we can prove, analogously to the proof of Theorem 3.4, that $\|e_n^{(y)}\| \leq Ch^{q+1}$ and $\|E_j^{(z)}\| \leq Ch^q$, if $\alpha = 2$, $\beta = 1$, $B(q+1)$, $C(q)$ and $\rho = -1$.
- The main difficulty concerning the practical implementation of the modified method is the calculation of the projection $P(t)$. Nevertheless, for special DAEs, we can calculate this projection matrix by means of the Gauss eliminaton method applied to the matrix $A(t)$.
- Summarizing the modified IRK method appears to be promising for the numerical approximation of DAEs with index 2 and variable nullspace. The calculation of the projection $P(t)$ for the general DAE with index 2 is an open problem that deserves further research.

6. Numerical experiments

Finally, we present the results of some numerical experiments which are in agreement with the convergence statements predicted in this paper. The IRK method (2.1) and the modified IRK method (5.2), (5.3) have been implemented. In the experiments to determine the global error, we solve the problems with a sequence of fixed stepsize over the interval $[0, 1]$. Rates of convergence for each method were estimated by comparing the global errors at $t = 1$.

Example 6.1.

The first test problem has constant nullspace. The problem is given by

$$\begin{aligned}
&e^{-t}\big(2x_2'(t) + x_3'(t)\big) + 2e^{-t}\big(\cos t x_1(t) + x_2(t) + (3\cos t - t - \frac{1}{2})x_3(t)\big)\\
&\quad = 4 + e^{-t}\sin t(2e^{-t}\cos t - 1) + e^{-t}\cos t(6\cos t - 2t - 1),\\
&\beta(t)\big(x_1'(t) + 3x_3'(t)\big) - \beta(t)(2e^{-t}x_2(t) - \big(6t + e^{-t}\big)x_3(t)\big) + (t^2+1)x_3(t)\\
&\quad = \beta(t)\big[(2e^{-t} + 6t)\cos t - (3 + e^{-t})\sin t - 2\big] + \cos t(t^2+1),\\
&\beta(t)\big(x_1(t) + 3x_3(t)\big) = \beta(t)\big(e^{-t}\sin t + 3\cos t\big),
\end{aligned}$$

with $\beta(t) = \sin t + 2$.

For this problem, the matriz Q given by

$$Q = \begin{pmatrix} 0 & 6 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 0 \end{pmatrix}$$

is a projection onto $N(t)$. With initial values given by $x_0 = (0, 1, 1)^T$, this problem has the solution

$$\begin{aligned} x_1 &= e^{-t}\sin t, \\ x_2 &= e^{t}, \\ x_3 &= \cos t. \end{aligned}$$

This problem has been solved using eight different stepsizes, $h \in H$,

$$H := \{h \in \mathbb{R} \ : \ h = 2^{-j} \text{ for } j = 3, 4, 5, \ldots, 10\}.$$

Table 6.1 shows the results of the experiments.

Example 6.2.

We solve the DAE in example 2.2, with the following right-hand side

$$\begin{aligned} q_1(t) &= e^{-t}(\sin t + \eta t\cos t), \\ q_2(t) &= e^{-t}(\cos t - \sin t) - \eta t e^{-t}(\cos t + \sin t) + (1+\eta)e^{-t}\cos t. \end{aligned}$$

This problem has the solution

$$\begin{aligned} x_1(t) &= e^{-t}\sin t, \\ x_2(t) &= e^{-t}\cos t. \end{aligned}$$

The problem has been solved for several values of η between 0 and -100, by the modified IRK method. The method has been implemented for all the stepsizes $h \in H$. Table 6.2 shows the results for $\eta = -1$.

Example 6.3.

For all continuos differentiable functions $\beta(t)$ and for $t \in [0, \infty)$ the following DAE is tractable with index 2.

$$\begin{aligned} e^{-t}x_2'(t) + \beta(t)(x_3'(t) + x_3(t)) + \beta'(t)x_3(t) &= (1 - e^{-t})\sin t + \cos t, \\ 2x_1(t) + (\cos t + 2t + 2)x_3(t) &= 2e^t + e^{-t}(\cos t + 2t + 2), \\ 2x_1'(t) + (\cos t + 2t + 2)x_3'(t) &= 2e^t - e^{-t}(\cos t + 2t + 2). \end{aligned}$$

For this DAE

$$Q = \begin{pmatrix} 0 & 0 & -\dfrac{\cos t}{2} - t - 1 \\ 0 & 0 & -\beta(t)e^t \\ 0 & 0 & 1 \end{pmatrix}$$

is a projection onto $N(t)$, and for $\beta(t) = e^t \sin t$, this problem has the solution

$$\begin{aligned} x_1(t) &= e^t, \\ x_2(t) &= \cos t, \\ x_3(t) &= e^{-t}. \end{aligned}$$

This problem has been solved by the modified IRK method, which has been implemented for all the stepsizes $h \in H$. Table 6.3 shows the results of the experiments.

Methods used in the experiments

1.	(MP)	midpoint rule or 1-stage Gauss-Legendre formula
2.	(BE)	backward Euler method
3.	(2R2)	2-stage Radau IIA
4.	(2L3)	2-stage Lobatto IIIC
5.	(2R1)	2-stage Radau IA
6.	(C)	2-stage A-stable SDIRK (Crouzeix)
7.	(A)	2-stage S-stable SDIRK formula (Alexander) with $\alpha = 1 - \frac{1}{2}\sqrt{2}$
8.	(2GL)	2-stage Gauss-Legendre formula
9.	(3R2)	3-stage Radau IIA
10.	(2L3)	3-stage Lobatto IIIC

List of symbols

$-1 < \vartheta = -\frac{1+\sqrt{3}}{2+\sqrt{3}} < 1$

ρ	stability condition
K_d	nonstiff order
K_I	internal stage order
$K_{a,1}$	order of constant coefficient index 1 systems
EK	order predicted by theorems 3.3, 3.4 and 5.3
NK	order of the observed global error
$\underline{GF}$	global error for the largest stepsize $h_{\max} = 2^{-3}$
$\overline{GF}$	global error for the smallest stepsize $h_{\min} = 2^{-10}$
Px	P-component
Qx	Q-component
Pe	P-component of the global error
Qe	Q-component of the global error

Method	ρ	K_d	K_I	$K_{a,1}$	EK		NK		$\underline{GF}$		$\overline{GF}$	
					P_x	Q_x	P_x	Q_x	Pe	Qe	Pe	Qe
1. MP	-1	2	1	1	2	0	2	0	6.4E-2	1.7E-0	2.5E-07	1.7E-00
2. BE	0	1	1	∞	1	1	1	1	1.1E-1	2.6E-4	2.4E-04	7.3E-04
3. 2R2	0	3	2	∞	3	2	3	2	3.4E-4	3.8E-3	2.6E-12	1.6E-08
4. 2L3	0	2	1	∞	2	1	2	1	3.0E-2	3.0E-1	1.2E-07	7.3E-04
5. 2R1	0	3	1	1	2	1	2	1	4.1E-2	5.3E-1	1.7E-07	1.5E-03
6. C	ϑ	3	1	1	2	1	2	1	3.5E-2	2.4E-1	1.5E-07	7.0E-04
7. A	0	2	1	∞	2	1	2	1	7.8E-4	2.2E-1	3.4E-09	5.2E-04
8. 2GL	1	4	2	2	2	0	2	0	3.4E-3	2.5E-1	1.3E-08	2.5E-01
9. 3R2	0	5	3	∞	5	3	5	3	4.1E-3	2.1E-2	1.7E-16	3.2E-10
10. 2L3	0	4	2	∞	4	2	4	2	5.2E-2	4.3E-1	2.5E-13	3.3E-05

Table 6.1

Method	ρ	K_d	K_I	$K_{a,1}$	EK		NK		$\underline{GF}$		$\overline{GF}$	
					P_x	Q_x	P_x	Q_x	Pe	Qe	Pe	Qe
1. MR	-1	2	1	1	2	1	2	2†	6.5E-04	2.4E-3	2.3E-09	9.9E-9
2. BE	0	1	1	∞	1	1	1	1	2.5E-21	1.9E-2	6.6E-24	3.8E-5
3. 2R2	0	3	2	∞	3	2	0*	2	1.0E-20	2.5E-3	2.0E-20	1.3E-8
4. 2L3	0	2	1	∞	2	1	0*	1	6.8E-21	1.9E-2	6.8E-21	3.8E-5
5. 2R1	0	3	1	1	2	1	2	1	8.1E-04	1.3E-2	3.1E-09	2.6E-5
6. C	ϑ	3	1	1	2	1	2	1	4.2E-04	1.1E-2	1.6E-09	2.2E-5
7. A	0	2	1	∞	2	1	0*	1	6.8E-21	1.4E-2	1.4E-20	2.7E-5
8. 2GL	1	4	2	2	2	1	2	1	1.4E-04	3.7E-3	5.2E-10	7.4E-6

* In this case, the expected order of convergence is not attained. This is easy to explain by observing the global error, which lies outside the computer precision limit.

† The order of convergence is better than expected.

Table 6.2

Method	ρ	K_d	K_I	$K_{a,1}$	EK		NK		$\underline{GF}$		$\overline{GF}$	
					P_x	Q_x	P_x	Q_x	Pe	Qe	Pe	Qe
1. MP	-1	2	1	1	2	1	2	2*	5.9E-3	3.1E-2	2.3E-08	1.6E-7
2. BE	0	1	1	∞	1	1	1	1	1.2E-1	1.9E+0	2.3E-04	4.0E-3
3. 2R2	0	3	2	∞	3	2	3	2	2.6E-5	5.4E-2	1.9E-13	2.1E-7
4. 2L3	0	2	1	∞	2	1	2	1	1.9E-3	1.9E+0	7.1E-09	4.0E-3
5. 2R1	0	3	1	1	2	1	2	1	7.5E-3	1.1E+0	3.0E-08	2.6E-3
6. C	ϑ	3	1	1	2	1	2	1	3.9E-3	7.7E-1	1.5E-08	2.3E-3
7. A	0	2	1	∞	2	1	2	1	2.1E-4	1.3E+0	8.7E-10	2.8E-3
8. 2GL	1	4	2	2	2	1	2	1	1.3E-3	3.5E-1	5.0E-09	7.6E-4

* The midpoint rule gives the same order in both components.

Table 6.3

Appendix. Basic linear algebra lemma and the commutation lemma

A basic connection between the spaces appearing at the tractability index and the choice of the corresponding projectors is given by the following Lemma, which may be directly obtained from Theorem A.13. and Lemma A.14 in [3].

Lemma A.1 *Let $\overline{A}, \overline{B}, \overline{Q} \in L(\mathbb{R}^m)$ be given, $\overline{Q}^2 = \overline{Q}$, $im(\overline{Q}) = \ker(\overline{A})$, i.e., let $\overline{Q}$ be a projector onto $\ker(\overline{A})$. Let $\overline{S} := \{z \in \mathbb{R}^m : \overline{B}z \in im(\overline{A})\}$. Then the following conditions are equivalent:*

(*i*) *The matrix $\overline{G} := \overline{A} + \overline{BQ}$ is nonsingular.*
(*ii*) $\mathbb{R}^m = \overline{S} \oplus \ker(\overline{A})$.
(*iii*) $\overline{S} \cap \ker(\overline{A}) = \{0\}$.

If $\overline{G}$ is nonsingular, then

$$\overline{Q}_s = \overline{Q}\overline{G}^{-1}\overline{B}$$

holds for the canonical projetor $\overline{Q}_s$ (canonical means $\overline{Q}_s$ projects $\mathbb{R}^m$ onto $\ker(\overline{A})$ along $\overline{S}$).

Proof.

(*i*) $\Longrightarrow$ (*ii*) The space $\mathbb{R}^m$ can be described as $\overline{S} + \ker(\overline{A})$, because

$$z = (I - \overline{Q}\overline{G}^{-1}\overline{B})z + \overline{Q}\overline{G}^{-1}\overline{B}z =: z_1 + z_2 \qquad (*)$$

holds for any $z \in \mathbb{R}^m$.

Now, z_2 obviously lies in $\ker(\overline{A})$, because $\overline{Q}_s$ is a projector onto $\ker(\overline{A})$. For z_1 we obtain

$$\overline{B}z_1 = (I - \overline{BQG}^{-1})\overline{B}z = \overline{AG}^{-1}\overline{B}z \in \mathrm{im}\,(\overline{A}),$$

i.e., $z_1 \in \overline{S}$.

It remains to show that $\overline{S} \cap \ker(\overline{A}) = \{0\}$. For that purpose, let $x \in \overline{S} \cap \ker(\overline{A})$. Then $x = \overline{Q}x$ holds and there exists a $z \in \mathbb{R}^m$ such that

$$\overline{A}z = \overline{B}x = \overline{BQ}x \qquad \text{and so} \qquad \overline{G}^{-1}\overline{A}z = \overline{G}^{-1}\overline{BQ}x.$$

I.e., $(I - \overline{Q})z = \overline{Q}x$; so, $0 = \overline{Q}x = x$.

$(ii) \Longrightarrow (iii)$ This follows trivially from the definition.

$(iii) \Longrightarrow (i)$ Let $x \in \mathbb{R}^m$ be chosen such that $\overline{G}x = 0$, *i.e.*, such that $\overline{BQ}x = -\overline{A}x$. Then $\overline{Q}x \in \overline{S}$. On the other hand, $\overline{Q}x$ lies in $\ker(\overline{A})$. Thus, $x \in \ker(\overline{Q})$ holds because of the assumption. That means $\overline{A}x = 0$; hence $x \in \operatorname{im}(\overline{Q})$. Therefore $x = 0$ must hold, and $\overline{G}$ is nonsingular.

Because of the uniqueness of partition $(*)$, the latter assertion follows immediately.

Lemma A.2 (Commutation Lemma) *Let* $N : [t_0, T] \longrightarrow L(\mathbb{R}^m)$ *be a continuously differentiable matrix function. Then*

$$D_N(\mathcal{A}^{-1} \otimes I_m) = (\mathcal{A}^{-1} \otimes I_m)D_N + R_N,$$

where $R_N := h[(\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m]D_{N'} + O(h^2)$, $\mathcal{C} = (c_{jl})_{j,l=1}^s \in L(\mathbb{R}^s)$ *with* $c_{jl} = c_j - c_l$; *the* c_j, $j = 1, 2, ..., s$, *being the coefficients of the IRK method and* $\mathcal{C} \odot \mathcal{A}^{-1} = (c_{jl}\widehat{a}_{jl})_{j,l=1}^s \in L(\mathbb{R}^s)$ *being the Hadamard product of matrices.*

Proof. $D_N(\mathcal{A}^{-1} \otimes I_m) = \big(\widehat{a}_{jl}N(t_{nj})\big)_{j,l=1}^s$. Taylor's expansion leads to

$$\widehat{a}_{jl}N(t_{nj}) = \widehat{a}_{jl}N(t_{nl}) + h\widehat{a}_{jl}(c_j - c_l)N'(t_{nl}) + O(h^2).$$

Thus, we obtain

$$\begin{aligned} D_N(\mathcal{A}^{-1} \otimes I_m) &= \big(\widehat{a}_{jl}N(t_{nj})\big)_{j,l=1}^s \\ &= \big(\widehat{a}_{jl}N(t_{nl})\big)_{j,l=1}^s + h\big(\widehat{a}_{jl}c_{jl}N'(t_{nl})\big)_{j,l=1}^s + O(h^2) \\ &= (\mathcal{A}^{-1} \otimes I_m)D_N + h\big[(\mathcal{C} \odot \mathcal{A}^{-1}) \otimes I_m\big]D_{N'} + O(h^2). \quad \square \end{aligned}$$

If N is a constant matrix, then

$$D_N(\mathcal{A}^{-1} \otimes I_m) = (\mathcal{A}^{-1} \otimes I_m)D_N$$

trivially holds.

References

1. K. E. Brenan, S. L. Campbell & L. R. Petzold, *Numerical solution of initial value problems in differential-algebraic equations*, North-Holland, 1989.
2. J. C. Butcher, *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods*, John Wiley & Sons, 1987.
3. E. Griepentrog & R. März, *Differential-algebraic equations and their numerical treatment*, Teubner-Texte zur Mathematik, 1986.

4. E. HAIRER, CH. LUBICH & M. ROCHE, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, Lecture Notes in Mathematics 1409, Springer-Verlag, 1989.
5. M. HANKE, *On the regularization of index 2 differential-algebraic equations*, J. Math. Anal. Appl. **151** no. 1 (1990), 236-253.
6. M. HANKE & E. IZQUIERDO, *Implicit Runge-Kutta methods for general linear index 2 differential-algebraic equations with variable coefficients*, Preprint # 93–11, Humboldt Univ., Fachbereich Mathematik, Berlin, 1993.
7. M. HANKE, E. IZQUIERDO & R. MÄRZ, *On asymptotics in case of linear index 2 DAEs.*, Preprint # 94–5, Humboldt Univ., Fachbereich Mathematik, Berlin, 1994.
8. E. IZQUIERDO, *Numerische Approximation von Algebro-Differentialgleichungen mit Index 2 mittels impliziter Runge-Kutta-Verfahren*, Doctoral thesis, Humboldt Unv., Berlin, 1992.
9. A. KVÆRNØ, *Runge-Kutta methods applied to fully implicit differential-algebraic equations of index 1*, Mathematics of Computation **54** no. 190 (1990), 583–625.
10. R. MÄRZ, *Index-2 differential-algebraic equations*, Preprint 135, Humboldt Univ., Sektion Mathematik, Berlin, 1987.
11. R. MÄRZ, *Higher-index differential-algebraic equations: Analysis and numerical treatment*, Preprint 159, Humboldt Univ., Sektion Mathematik, Berlin, 1987.
12. R. MÄRZ, *On quasilinear index 2 differential-algebraic equations*, Preprint 269, Humboldt Univ., Berlin, 1991.
13. R. MÄRZ, *Numerical methods for differential-algebraic equations. Part I: Characterizing DAEs.*, Preprint 91-32/I, Humboldt Univ., Berlin, 1991.
14. R. MÄRZ, *Numerical methods for differential-algebraic equations. Part II: numerical integration methods*, Preprint 91-32/II, Humboldt Univ., Berlin, 1991.

(Recibido en noviembre de 1992; primera revisión: mayo de 1994; segunda revisión: enero de 1995)

EBROUL IZQUIERDO
HEINRICH HERTZ INSTITUT
ABTEILUNG BILDSIGNALVERARBEITUNG
EINSTEIN UFER 37
10587 BERLIN, GERMANY
e-mail: ebroul@hhi.de