Electronic Transactions on Numerical Analysis.
Volume 26, pp. 385-420, 2007.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

# STABILITY PROPERTIES OF DIFFERENTIAL-ALGEBRAIC EQUATIONS AND SPIN-STABILIZED DISCRETIZATIONS*

PETER KUNKEL† AND VOLKER MEHRMANN‡

**Abstract.** Classical stability properties of solutions that are well-known for ordinary differential equations (ODEs) are generalized to differential-algebraic equations (DAEs). A new test equation is derived for the analysis of numerical methods applied to DAEs with respect to the stability of the numerical approximations. Moreover, a stabilization technique is developed to improve the stability of classical DAE integration methods. The stability regions for these stabilized discretization methods are determined and it is shown that they much better reproduce the stability properties known for the ODE case than in the unstabilized form. Movies that depict the stability regions for several methods are included for interactive use.

**Key words.** nonlinear differential-algebraic equations, stability, asymptotic stability, Lyapunov stability, spin-stabilized discretization, test equation, strangeness index

**AMS subject classifications.** 65L80, 65L20, 34D20, 34D23

## 1. Introduction and survey of previous results.

In this paper we study different stability concepts for differential-algebraic equations (DAEs) as well as stabilization techniques for numerical methods. In particular, we consider initial value problems for general implicit systems of DAEs

$$F(t, x, \dot{x}) = 0, \tag{1.1}$$

with an initial condition

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

on the unbounded interval $\mathbb{I} = [t_0, \infty)$, with $F \in C^0(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^n)$ sufficiently smooth and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open sets.

DAEs like (1.1) arise in constrained multibody dynamics [9], electrical circuit simulation [11, 12], chemical engineering [7, 8], and many other applications, in particular when the dynamics of a system is constrained to a manifold or when different physical models are coupled together [28].

While DAEs provide a very convenient modeling concept, many numerical difficulties arise due to the fact that the dynamics is constrained to a manifold, which often is only given implicitly; see [31] or the recent textbook [21]. These difficulties are typically characterized by one of many index concepts that exist for DAEs; see [2, 10, 13, 21]. The fact that the dynamics of DAEs is constrained also requires a modification of the classical stability concepts that were developed for ODEs. Appropriate stability concepts for DAEs have been discussed already in several publications. The extension of the classical Lyapunov stability theory for linear DAEs with constant coefficients has been studied in [36, 37, 38]. For particular classes of DAEs, the classical stability concepts known for ODEs and for the corresponding integration methods have been analyzed in [1, 15, 16, 25, 27, 33, 34, 39]. Often

*Received September 12, 2006. Accepted for publication June 27, 2007. Recommended by S. Vandewalle. The research was supported by the *Mathematisches Forschungsinstitut Oberwolfach* Research-in-Pairs Program.

†Mathematisches Institut, Universität Leipzig, Augustusplatz 10–11, D-04109 Leipzig, Fed. Rep. Germany (kunkel@mathematik.uni-leipzig.de).

‡Institut für Mathematik, MA 4-5, Technische Universität Berlin, D-10623 Berlin, Fed. Rep. Germany (mehrmann@math.tu-berlin.de). Supported by *Deutsche Forschungsgemeinschaft*, through MATHEON, the DFG Research Center "Mathematics for Key Technologies" in Berlin.

ETNA
Kent State University
etna@mcs.kent.edu



this leads to modifications of the DAEs by global transformations to some canonical form to avoid instabilities in the numerical methods.

All these papers deal with special classes or special formulations of DAEs and usually some restrictions on the size of the index of the DAE. In this paper we extend the classical stability concepts for ordinary differential equations to general DAEs of the form (1.1) of arbitrary index; see Section 3.

The second topic of this paper is the development of stable integration methods for DAEs, where stability problems arise that cannot be observed for ODEs, as, e.g., the following example taken from [27] demonstrates.

EXAMPLE 1.1. Consider the linear DAE

$$\begin{bmatrix} \delta-1 & \delta t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\eta(\delta-1) & -\eta\delta t \\ \delta-1 & \delta t-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

with real parameters $\eta$ and $\delta \neq 1$. This system of differentiation index 1 has the solution

$$x_1(t) = (\delta-1)^{-1}(1-\delta t)x_2(t), \quad x_2(t) = e^{(\delta-\eta)t}x_2(0).$$

Obviously, $x(t) \to 0$ as $t \to \infty$ independently of $x_2(0)$ for $\delta < \eta$. On the other hand, using a constant stepsize $h$, the implicit Euler method yields numerical approximations

$$x_{i,1} = (\delta-1)^{-1}(1-\delta t_i)x_{i,2}, \quad x_{i,2} = \frac{1+h\delta}{1+h\eta}x_{i-1,2},$$

which satisfy $x_i \to 0$ as $i \to \infty$ independently of $x_{0,2}$ if and only if $|1+h\delta| < |1+h\eta|$. Hence, there exist parameter values $(\delta, \eta)$ for which the exact solution asymptotically goes to zero while the numerical solution grows unboundedly.

Example 1.1 demonstrates that for DAEs instabilities may arise that cannot be observed for ODEs. One can show that this instabilitity is caused by the time-dependence of the kernel of the leading coefficient matrix. Such a time-dependence easily arises when a higher index problem (even in semi-explicit form) is turned into a problem with reduced index by differentiation. Since these effects do not occur in ordinary differential equations, the classical test equation

$$\dot{x} = \lambda x, \quad \lambda \in \mathbb{C}, \tag{1.3}$$

is not sufficient to analyze this instability.

For this reason and in order to allow a better comparison of different integration methods for DAEs, in Section 4 we will take up on Example 1.1 and suggest a new linear test equation for DAEs which generalizes (1.3). This new test equation is

$$\begin{bmatrix} 1 & -\omega t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \lambda & \omega(1-\lambda t) \\ -1 & 1+\omega t \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

and combines the classical test equation with an algebraic equation in such a way that the kernel of the corresponding matrix function $E$ spins and $\omega$ is a measure for the size of the time derivative of a kernel function.

We will show that with the variation of these two parameters many stability properties of classical DAE integration methods can be tested and compared. A comparison of well-known DAE integration methods for this test equation is presented in Section 5, where also DAE stability functions for these methods are derived.

Finally, in Section 6 we derive a new stabilization technique for general DAE integration methods (which we call *spin-stabilization*). We analyze the stability behavior of several classical DAE integrators and show that with this technique more appropriate stability regions can be achieved.

## 2. Preliminaries, notation, and definitions.

**2.1. Notation.** For $x_0$ in some vector space $\mathbb{X}$ and $\varrho > 0$, we denote the open ball with radius $\varrho$ around $x_0$ in $\mathbb{X}$ by $\mathcal{B}(x_0, \varrho)$, i.e.

$$\mathcal{B}(x_0, \varrho) = \{x \in \mathbb{X} \mid \|x - x_0\| < \varrho\},$$

and the corresponding closed ball by $\overline{\mathcal{B}}(x_0, \varrho)$, i.e.,

$$\overline{\mathcal{B}}(x_0, \varrho) = \overline{\mathcal{B}(x_0, \varrho)} = \{x \in \mathbb{X} \mid \|x - x_0\| \leq \varrho\}.$$

By $\langle \cdot, \cdot \rangle$, we denote the *Euclidian scalar product* and by $\|x\|_2$ the associated *Euclidian norm* in $\mathbb{R}^n$ as well as the associated *spectral norm* for matrices in $\mathbb{R}^{n,n}$.

If (1.1) together with (1.2) possesses a unique solution on $\mathbb{I}$, then we denote it by $x(t; t_0, x_0)$ when we want to stress its dependence on the initial condition.

**2.2. DAE theory.** In this section we briefly recall some concepts from the theory of differential-algebraic equations, see [2, 10, 21, 30]. We follow [21] in notation and style of presentation.

DEFINITION 2.1. *Consider system (1.1) with sufficiently smooth $F$. A function $x : \mathbb{I} \to \mathbb{R}^n$ is called a* solution *of (1.1) if $x \in C^1(\mathbb{I}, \mathbb{R}^n)$ and $x$ satisfies (1.1) pointwise. It is called a* solution of the initial value problem *(1.1)–(1.2) if $x$ is a solution of (1.1) and satisfies (1.2). An initial condition (1.2) is called* consistent *if the corresponding initial value problem has at least one solution.*

It is possible to weaken this solution concept [22, 26, 29], but we will not consider such weaker solution concepts in this paper.

For the DAE system (1.1), as in [4, 5, 19], we introduce a nonlinear derivative array of the form

$$F_\ell(t, x, \dot{x}, \ldots, x^{(\ell+1)}) = 0,$$

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

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

Partial derivatives of $F_\ell$ with respect to selected variables $p$ from the vector $z_\ell = (t, x, \dot{x}, \ldots, x^{(\ell+1)})$ are denoted by $F_{\ell;p}$, e. g.,

$$F_{\ell;x} = \tfrac{\partial}{\partial x} F_\ell, \quad F_{\ell;\dot{x},\ldots,x^{(k+1)}} = [\, \tfrac{\partial}{\partial \dot{x}} F_\ell \ \cdots \ \tfrac{\partial}{\partial x^{(\ell+1)}} F_\ell \,].$$

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

In order to analyze existence and uniqueness of solutions, we introduce the *solution set* of the nonlinear algebraic equation associated with the derivative array $F_\mu$ for some integer $\mu$, given by

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

We make the following hypothesis; see [21].

HYPOTHESIS 2.2. *Consider the general system of nonlinear differential-algebraic equations (1.1). There exist integers $\mu$, $r$, $a$, $d$, and $v$ such that $\mathbb{L}_\mu$ is not empty and for every point $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ there exists a (sufficiently small) neighborhood in which the following properties hold:*

1. *We have* $\operatorname{rank} F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} = (\mu+1)n - a$ *on* $\mathbb{L}_\mu$ *such that there exists a smooth full rank matrix function* $Z_2$ *of size* $(\mu+1)m \times a$ *satisfying*

$$Z_2^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} = 0$$

*on* $\mathbb{L}_\mu$.

2. *We have* $\operatorname{rank} Z_2^T F_{\mu;x} = a$ *on* $\mathbb{L}_\mu$ *such that there exists a smooth full rank matrix function* $T_2$ *of size* $n \times (n-a)$ *satisfying*

$$Z_2^T F_{\mu;x} T_2 = 0.$$

3. *We have* $\operatorname{rank} F_{\dot{x}} T_2 = d = n - a$ *such that there exists a smooth full rank matrix function* $Z_1$ *of size* $n \times d$ *satisfying*

$$\operatorname{rank} Z_1^T F_{\dot{x}} T_2 = d.$$

As in [19, 21], we call the smallest possible $\mu$ for which Hypothesis 2.2 is valid the *strangeness index* of (1.1). Systems with vanishing strangeness index are called *strangeness-free*.

It has been shown in [20] that Hypothesis 2.2 implies locally (via the implicit function theorem) the existence of a *reduced system* such that the solutions are in one-to-one correspondence and the differential and algebraic part contained in the given DAE are separated. This result can be globalized when we start with a solution $x$ in the sense that we have path

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

with some $\mathcal{P} \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+2)n})$. In the present context, where stability questions are concerned, we must take care that the involved transformations do not alter the behavior of the solution as $t \to \infty$. We therefore sketch the construction of the reduced system along the lines of [21] and pay special attention to the conservation of the stability properties of the given DAE.

Due to Hypothesis 2.2 there exist

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

with the described properties. Since Gram-Schmidt orthonormalization is a smooth process, we may assume without loss of generality that the columns of these matrix functions are pointwise orthonormalized. Let then

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

be such that $[\, Z_2' \;\; Z_2 \,]$, $[\, T_2' \;\; T_2 \,]$, $[\, Z_1' \;\; Z_1 \,]$ are pointwise orthogonal. Furthermore, there exist

$$T_1 \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)n,a}), \quad T_1' \in C^0(\mathbb{I}, \mathbb{R}^{(\mu+1)n,(\mu+1)n-a})$$

such that $[\, T_1' \;\; T_1 \,]$ is pointwise orthogonal and

$$Z_2'(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x(t), \mathcal{P}(t)) T_1(t) = 0 \text{ for all } t \in \mathbb{I}.$$

If we define a function $\mathcal{H}$ via

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

ETNA
Kent State University
etna@mcs.kent.edu



then

$$\begin{array}{ll}\text{(a)} & \mathcal{H}(t,x(t),\mathcal{P}(t),0)=0,\\ \text{(b)} & \mathcal{H}_{p,\phi}(t,x(t),\mathcal{P}(t),0)=\left[\begin{array}{cc} F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t,x(t),\mathcal{P}(t)) & Z_2(t)\\ T_1(t)^T & 0\end{array}\right].\end{array}$$

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

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

It can then be shown that the equation

$$\hat{F}_2(t,x)=0 \tag{2.1}$$

is just the requirement that $x$ satisfies all constraints that are contained in (1.1) for time $t$.

With the change of variables

$$x=T_2x_1+T_2'x_2,\quad x_1=T_2^Tx,\quad x_2={T_2'}^Tx,$$

the equation (2.1) turns into

$$\hat{F}_2(t,T_2(t)x_1+T_2'(t)x_2)=0. \tag{2.2}$$

Note that this transformation and the corresponding back-transformation preserve the Euclidian norm of the unknown functions at every point $t\in\mathbb{I}$. If we set $x_1(t)=T_2^T(t)x(t)$, $x_2(t)=T_2'(t)^Tx(t)$ then it follows that for all $t\in\mathbb{I}$

$$\begin{array}{ll}\text{(a)} & \hat{F}_2(t,T_2(t)x_1(t)+T_2'(t)x_2(t))=0,\\ \text{(b)} & \hat{F}_{2;x}(t,x(t))T_2'(t)\text{ is nonsingular.}\end{array}$$

Thus, we can solve (2.2) for $x_2$ as $x_2=\mathcal{R}(t,x_1)$ and we have

$$x_2(t)=\mathcal{R}(t,x_1(t))\text{ for all }t\in\mathbb{I}. \tag{2.3}$$

Besides (2.3) we have

$$p_2(t)=\mathcal{R}_t(t,x_1(t))+\mathcal{R}_{x_1}(t,x_1(t))p_1(t), \tag{2.4}$$

where we use the partition

$$[\,I_n\;\;0\;\;\cdots\;\;0\,]\hat{\mathcal{P}}(t,x(t))=\left[\begin{array}{c}p_1(t)\\ p_2(t)\end{array}\right];$$

compare the proof of Theorem 4.13 in [21]. We then obtain

$$\begin{array}{l}Z_1(t)^TF(t,T_2(t)x_1(t)+T_2'(t)x_2(t),\\ \qquad\dot{T}_2(t)x_1(t)+T_2(t)p_1(t)+\dot{T}_2'(t)x_2(t)+T_2'(t)x_2(t))=0\text{ for all }t\in\mathbb{I},\end{array} \tag{2.5}$$

in which we can eliminate $x_2,p_2$ via (2.3) and (2.4), respectively. If we define

$$\begin{array}{l}\hat{F}_1(t,x_1,p_1)=Z_1(t)^TF(t,T_2(t)x_1+T_2'(t)\mathcal{R}(t,x_1),\\ \qquad\dot{T}_2(t)x_1+T_2(t)p_1(t)+\dot{T}_2'(t)\mathcal{R}(t,x_1)+T_2'(t)(\mathcal{R}_t(t,x_1)+\mathcal{R}_{x_1}(t,x_1)p_1)),\end{array}$$

ETNA
Kent State University
etna@mcs.kent.edu



then $(t, x_1(t), \dot{x}_1(t))$ solves $\hat{F}_1(t, z_1, p_1) = 0$. Furthermore,

$$\hat{F}_{1;p_1}(t, x_1(t), p_1(t)) = Z_1(t)^T F_{\dot{x}}(t, x(t), p(t))(T_2(t) + T_2'(t)\mathcal{R}_{x_1}(t, x_1(t))),$$

where $[\, I_n \;\; 0 \;\; \cdots \;\; 0\,]\mathcal{P} = p$. To determine $\mathcal{R}_{x_1}(t, x_1(t))$ one observes that from

$$\hat{F}_2(t, T_2(t)x_1(t) + T_2'(t)\mathcal{R}_{x_1}(t, x_1(t))) = 0 \text{ for all } t \in \mathbb{I},$$

it follows that

$$\hat{F}_{2;x}(t, x(t))(T_2(t) + T_2'(t)\mathcal{R}_{x_1}(t, x_1(t))) = 0 \text{ for all } t \in \mathbb{I}$$

and hence, using (2.1) we obtain

$$Z_2(t)^T F_{\mu;x}(t, x(t), \mathcal{P}(t))(T_2(t) + T_2'(t)\mathcal{R}_{x_1}(t, x_1(t))) = 0 \text{ for all } t \in \mathbb{I}.$$

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

$$\mathcal{R}_{x_1}(t, x_1(t)) = 0 \text{ for all } t \in \mathbb{I}$$

and that $\hat{F}_{1;p_1}(t, x_1(t), p_1(t))$ is nonsingular for all $t \in \mathbb{I}$. Thus, we can solve $\hat{F}_1(t, x_1, p_1) = 0$ for $p_1$ according to

$$p_1 = \mathcal{L}(t, x_1).$$

If we require that $x_1$ is continuously differentiable and that that the part $p_1$ of $\mathcal{P}$ satisfies $p_1(t) = \dot{x}_1(t)$ for all $t \in \mathbb{I}$, then we see that the given $x$ solves the DAE,

$$\text{(2.6)} \qquad \begin{array}{ll} \text{(a)} & \dot{x}_1 = \mathcal{L}(t, x_1), \\ \text{(b)} & x_2 = \mathcal{R}(t, x_1). \end{array}$$

Summarizing the above construction, we observe that we only have applied one transformation of the variable $x$. This transformation together with its inverse are pointwise orthogonal such that it preserves the behavior of the solution as $t \to \infty$. For the applications of the implicit function theorem, however, we must require that the corresponding neighborhoods do not shrink to a point as $t \to \infty$. Sufficient for this is the additional assumption that there exists a set $\mathbb{V} \subseteq \mathbb{I} \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ such that $(t, x(t), \mathcal{P}(t)) \in \mathbb{V}$ for sufficiently large $t$ and that the implicit function theorem can always applied in the whole set $\mathbb{V}$. Note that this condition is trivially satisfied when we study an equilibrium solution $x$ of (1.1) given by the property that $x(t) = x^* \in \mathbb{R}^n$ and $(t, x^*, 0) \in \mathbb{L}_{\mu+1}$ for all $t \in \mathbb{I}$. Instead of (1.1) we can then concentrate on the investigation of (2.6) due to the fact that under mild assumptions the solutions of (1.1) and (2.6) are locally in one-to-one correspondence; see [21].

In the special case of a linear DAE

$$\text{(2.7)} \qquad E(t)\dot{x} = A(t)x + f(t),$$

where $E, A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and $f \in C^0(\mathbb{I}, \mathbb{R}^n)$ are sufficiently smooth, the corresponding reduced DAE (2.6) is linear as well and of the form

$$\text{(2.8)} \qquad \begin{array}{ll} \text{(a)} & \dot{x}_1 = A_{11}(t)x_1 + f_1(t), \\ \text{(b)} & x_2 = A_{21}(t)x_1 + f_2(t). \end{array}$$

This also shows that if the DAE belonging to a pair $(\hat{E}, \hat{A})$ of matrix functions is strangeness-free then there is a pointwise nonsingular matrix function $P \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and a pointwise orthogonal matrix function $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, such that

$$\text{(2.9)} \qquad P\hat{E}Q = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \quad P\hat{A}Q - P\hat{E}\dot{Q} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & -I_a \end{bmatrix}.$$

**ETNA**
Kent State University
etna@mcs.kent.edu



Summarizing the above construction, every DAE satisfying Hypothesis 2.2 can be transformed to a strangeness-free DAE,

$$Z_1(t)^T F(t,x,\dot{x}) = 0, \quad \hat{F}_2(t,x) = 0, \tag{2.10}$$

which possesses the same solutions as the original DAE. Even more, we can transform to a specially structured DAE (2.6) such that stability properties of the solutions are preserved. Thus, w.l.o.g. we can study stability questions for (2.6) instead of the original DAE (1.1). Moreover, also in the numerical treatment we may assume w.l.o.g. that the DAE is given in the form (2.10), since this form can locally be determined numerically. Hence, dealing with DAEs both theoretically concerning stability questions and numerically, we may assume that the given DAE is strangeness-free.

**2.3. Stability concepts for ODEs.** In this section, we briefly recall classical stability concepts for ordinary differential equations

$$\dot{x} = f(t,x), \quad t \in \mathbb{I}. \tag{2.11}$$

See, e.g., [17, 35] for more details on this topic. We include proofs when we need the notation and parts of them when we discuss similar results for DAEs.

DEFINITION 2.3. *A solution $x : t \mapsto x(t;t_0,x_0)$ of (2.11) is called*

1. stable *if for every $\varepsilon > 0$ there exists $\delta > 0$ such that*
   (a) *the initial value problem (2.11) with initial condition $x(t_0) = \hat{x}_0$ is solvable on $\mathbb{I}$ for all $\hat{x}_0 \in \mathbb{R}^n$ with $\|\hat{x}_0 - x_0\| < \delta$;*
   (b) *the solution $x(t;t_0,\hat{x}_0)$ satisfies $\|x(t;t_0,\hat{x}_0) - x(t;t_0,x_0)\| < \varepsilon$ on $\mathbb{I}$.*
2. asymptotically stable *if it is stable and there exists $\varrho > 0$ such that*
   (a) *the initial value problem (2.11) with initial condition $x(t_0) = \hat{x}_0$ is solvable on $\mathbb{I}$ for all $\hat{x}_0 \in \mathbb{R}^n$ with $\|\hat{x}_0 - x_0\| < \varrho$;*
   (b) *the solution $x(t;t_0,\hat{x}_0)$ satisfies $\lim_{t\to\infty} \|x(t;t_0,\hat{x}_0) - x(t;t_0,x_0)\| = 0$.*
3. exponentially stable *if it is stable and* exponentially attractive*, i.e., if there exist $\delta > 0$, $L > 0$, and $\gamma > 0$ such that*
   (a) *the initial value problem (2.11) with initial condition $x(t_0) = \hat{x}_0$ is solvable on $\mathbb{I}$ for all $\hat{x}_0 \in \mathbb{R}^n$ with $\|\hat{x}_0 - x_0\| < \delta$;*
   (b) *the solution satisfies the estimate $\|x(t;t_0,\hat{x}_0) - x(t;t_0,x_0)\| < Le^{-\gamma(t-t_0)}$ on $\mathbb{I}$.*

Note that we can transform the ODE (2.11) in such a way that a given solution $x(t;t_0,x_0)$ is mapped to the trivial solution by simply shifting the arguments according to

$$\dot{\tilde{x}} = \tilde{f}(t,\tilde{x}) = f(t,\tilde{x} + x(t;t_0,x_0)) - \tfrac{\partial}{\partial t} x(t;t_0,x_0). \tag{2.12}$$

When studying the stability of a selected solution, we may therefore assume without loss of generality that the selected solution is the trivial solution. This also applies to DAEs. In the following, we will concentrate on *equilibrium solutions* $x^*$, i.e., solutions with $x(t;t_0,x_0) = x^*$ independent of $t$, although we may simply set $x^* = 0$.

We will also study further concepts which are not related to a selected solution such as contractivity and dissipativity.

DEFINITION 2.4. *The ODE (2.11) is called* contractive *if for any two solutions $x, y$ the scalar function $d : \mathbb{I} \to \mathbb{R}_0^+$ defined by $d(t) = \|x(t) - y(t)\|_2^2$ is monotonically non-increasing. It is called* exponentially contractive *if $d$ decays exponentially.*

DEFINITION 2.5. *The ODE (2.11) is called* dissipative *if there exists a bounded set $\mathbb{B} \subseteq \mathbb{R}^n$ with the property that for any bounded set $\mathbb{E} \subseteq \mathbb{R}^n$ there exists $\hat{t} \geq t_0$ with $x(t;t_0,\hat{x}_0) \in \mathbb{B}$ for all $\hat{x}_0 \in \mathbb{E}$ and $t > \hat{t}$. In this case the set $\mathbb{B}$ is called* absorbing.

We start our survey of stability results with the special case of linear ODEs. In view of (2.12) it is sufficient to study homogeneous equations

$$\dot{x} = A(t)x. \tag{2.13}$$

Since we obtain (2.13) no matter which solution we want to look at, the stability properties of Definition 2.3 are merely properties of the given linear ODE. In particular, the initial value problem

$$\tfrac{\partial}{\partial t}\Phi(t,t_0) = A(t)\Phi(t,t_0), \quad \Phi(t_0,t_0) = I_n$$

possesses a solution $t \mapsto \Phi(t,t_0)$ on $\mathbb{I}$, so-called *fundamental solution*, and the solution $x$ of (2.13) with $x(t_0) = x_0$ can be written as $x(t) = \Phi(t,t_0)x_0$. The following characterizations are then straightforward.

THEOREM 2.6. *The trivial solution of the linear homogeneous ODE (2.13)*

1. *is stable if and only if there exists a constant $L > 0$ with $\|\Phi(t,t_0)\| \leq L$ on $\mathbb{I}$;*
2. *is asymptotically stable if and only if $\|\Phi(t,t_0)\| \to 0$ for $t \to \infty$;*
3. *is exponentially stable if there exists $L > 0$ and $\gamma > 0$ such that $\|\Phi(t,t_0)\| \leq Le^{-\gamma(t-t_0)}$ on $\mathbb{I}$.*

In the general nonlinear case, we can only expect sufficient conditions that guarantee the specific stability properties. The classical result is given in the so-called Lyapunov stability theorems; see, e.g., [17]. In this context, we use the notation $X > Y$ ($X \geq Y$) for symmetric (Hermitian) matrices $X, Y$ to denote that $X - Y$ is positive (semi-)definite.

DEFINITION 2.7. *Let $\mathbb{U}$ be an (open) neighborhood of an equilibrium solution $x^*$ of the ODE (2.11). A function $V \in C^1(\mathbb{I} \times \mathbb{U}, \mathbb{R}_0^+)$ is called* Lyapunov function *associated with $x^*$ if*

1. *$V(t,x^*) = 0$ for all $t \in \mathbb{I}$,*
2. *$\dot{V}(t,x) \leq 0$ for all $(t,x) \in \mathbb{I} \times \mathbb{U}$, where $\dot{V}(t,x) = V_x(t,x)f(t,x) + V_t(t,x)$,*
3. *there exists a continuous function $W : \mathbb{U} \to \mathbb{R}_0^+$ with $W(x) > 0$ for all $x \in \mathbb{U}\backslash\{x^*\}$ and $V(t,x) \geq W(x)$ for all $(t,x) \in \mathbb{I} \times \mathbb{D}$.*

THEOREM 2.8. *Let $V$ be a Lyapunov function associated with an equilibrium solution $x^*$ of (2.11). Then $x^*$ is stable.*

THEOREM 2.9. *Let $V$ be a Lyapunov function associated with an equilibrium solution $x^*$ of (2.11) satisfying*

1. *for all $\varepsilon < 0$ there exists $\delta > 0$ such that $V(t,x) < \varepsilon$ for all $t \in \mathbb{I}$ and all $x \in \mathbb{U}$ with $\|x - x^*\| < \delta$;*
2. *there exists a continuous function $\tilde{W} : \mathbb{U} \to \mathbb{R}_0^+$ with $\tilde{W}(x) > 0$ for all $x \in \mathbb{U}\backslash\{x^*\}$, $\tilde{W}(x^*) = 0$, and $\dot{V}(t,x) \geq -\tilde{W}(x)$ for all $(t,x) \in [t_0,\infty) \times \mathbb{D}$.*

*Then $x^*$ is asymptotically stable.*

We turn now to stability properties which are not associated with a particular solution. Recall for the following that we assume that $f \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^n)$ is sufficiently smooth. Moreover, we suppose that the interesting domain $\mathbb{U}$ is sufficiently large.

THEOREM 2.10. *Let $f \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^n)$ satisfy a one-sided Lipschitz condition with constant $c \in \mathbb{R}$, i.e. let*

$$\langle f(t,x) - f(t,y), x - y\rangle \leq c\|x - y\|_2^2 \text{ for all } t \in \mathbb{I} \text{ and } x, y \in \mathbb{U}.$$

*If $c = 0$, then (2.11) is contractive. If $c < 0$, then (2.11) is exponentially contractive.*

*Proof*. For two solutions $x, y$ of (2.11), we have

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}\|x(t) - y(t)\|_2^2 &= \langle \dot{x}(t) - \dot{y}(t), x(t) - y(t)\rangle \\ &= \langle f(t,x(t)) - f(t,y(t)), x(t) - y(t)\rangle \leq c\|x(t) - y(t)\|_2^2.\end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu



Setting $d(t) = \|x(t) - y(t)\|_2^2$, this relation reads $\dot{d}(t) \leq 2cd(t)$ and Gronwall's lemma (see, e.g., [17]) yields

$$d(t) \leq e^{2c(t-t_0)} d(t_0)$$

in both cases. □

THEOREM 2.11. *Let $f \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^n)$ satisfy*

$$\langle f(t,x), x\rangle \leq \alpha - \beta\|x\|_2^2 \text{ for all } t \in \mathbb{I} \text{ and } x \in \mathbb{U},$$

*with constants $\alpha \geq 0$ and $\beta > 0$. Then the ODE (2.11) is dissipative with absorbing set $\mathbb{B} = \mathcal{B}(0, \sqrt{\alpha/\beta + \varepsilon})$ for arbitrary $\varepsilon > 0$.*

*Proof*. Let $x$ be a solution of (2.11). Since

$$\tfrac{1}{2}\tfrac{d}{dt}\|x(t)\|_2^2 = \langle f(t, x(x)), x(t)\rangle \leq \alpha - \beta\|x(t)\|_2^2,$$

Gronwall's lemma yields

$$\|x(t)\|_2^2 \leq \alpha/\beta + e^{-2\beta t}(\|x(t_0)\|_2^2 - \alpha/\beta) \leq \max\{\|x(t_0)\|_2^2, \alpha/\beta\},$$

such that

$$\|x(t)\|_2 \leq \max\{\|x(t_0)\|_2, \sqrt{\alpha/\beta}\}.$$

Hence, $\mathbb{B}$ is positive invariant, i.e.,

$$x(t; t_0, \hat{x}_0) \in \mathbb{B} \text{ for all } t \geq t_0,\ \hat{x}_0 \in \mathbb{B}.$$

Let

$$\varrho = \sup_{\hat{x}_0 \in \mathbb{E}} \|\hat{x}_0\|_2.$$

The estimate

$$\|x(t)\|_2 \leq \alpha/\beta + e^{-2\beta t}(\varrho^2 - \alpha/\beta) \leq \alpha/\beta + \varepsilon$$

finally gives

$$e^{-2\beta\hat{t}}(\varrho^2 - \alpha/\beta) \leq \varepsilon$$

as condition on $\hat{t}$ in Definition 2.5. □

THEOREM 2.12. *Let $f \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^n)$ satisfy*

$$\langle f(t,x), x\rangle < 0 \text{ for all } t \in \mathbb{I} \text{ and } x \in \mathbb{U} \text{ with } \|x\|_2 > \varrho.$$

*Then the ODE (2.11) is dissipative with absorbing set $\mathbb{B} = \mathcal{B}(0, \varrho + \varepsilon)$ for arbitrary $\varepsilon > 0$.*

*Proof*. A solution $x$ of (2.11) satisfies

$$\tfrac{1}{2}\tfrac{d}{dt}\|x(t)\|_2^2 = \langle f(t, x(x)), x(t)\rangle.$$

If $x(t) \in \mathbb{R}^n \setminus \mathbb{B}$, then $\frac{d}{dt}\|x(t)\|_2 < 0$ and therefore

$$\|x(t)\|_2 < \max\{\|x(t_0)\|_2, \varrho + \varepsilon\} \text{ for all } t > t_0.$$

ETNA
Kent State University
etna@mcs.kent.edu



Hence, $\mathbb{B}$ is positive invariant. Let

$$r > \max\{\sup_{\hat{x}_0\in\mathbb{E}} \|\hat{x}_0\|_2, \varrho+\varepsilon\}$$

and let $\hat{\mathbb{B}} = \overline{\mathcal{B}}(0,r)$. Because of $\hat{\mathbb{B}} \supseteq \mathbb{B}$, we have $\mathbb{R}^n\setminus\hat{\mathbb{B}} \subseteq \mathbb{R}^n\setminus\mathbb{B}$ and therefore $\frac{d}{dt}\|x(t)\|_2^2 < 0$, as long as $x(t)\in\hat{\mathbb{B}}\setminus\mathbb{B}$. Hence, $\hat{\mathbb{B}}$ is positive invariant as well. Moreover, $\hat{\mathbb{B}}\setminus\mathbb{B}$ is compact and $\langle f(t,x(x)),x(t)\rangle < 0$ on $\hat{\mathbb{B}}\setminus\mathbb{B}$. Due to the continuity of $f$, there exists $\delta>0$ with

$$\tfrac{d}{dt}\|x(t)\|_2^2 < -\delta \text{ on } \hat{\mathbb{B}}\setminus\mathbb{B},$$

as long as $x(t)\in\hat{\mathbb{B}}\setminus\mathbb{B}$. For $\hat{x}_0\in\mathbb{E}$ it then follows that

$$x(t;t_0,\hat{x}_0)\in\mathbb{B} \text{ for all } t>\hat{t}=(r^2-(\varrho+\varepsilon)^2)/\delta,$$

with $\hat{t}$ as required in Definition 2.5. □

## 3. Stability results for DAEs.

In this section we generalize the classical ODE stability results that we have reviewed in Section 2.3 to DAEs. The key idea to obtain these analytical results is to consider first the transformation to the reduced system (2.6) which has the same solution set and consider the stability results in this framework. After this has been done we then transform back to the original system.

### 3.1. Linear DAEs.

We begin our analysis with linear DAEs (2.7) with variable coefficients. The stability analysis for such equations has been studied for systems of tractability index up to 2 in [14, 15, 16, 27, 39], we study here the general case.

In the case of linear DAEs, the reduced system has the form (2.8) with

$$x = Q\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\quad Q=[\,T_2'\ T_2\,], \tag{3.1}$$

according to the notation of Section 2.2. For the homogeneous system

$$E(t)\dot{x} = A(t)x,\quad x(t_0)=x_0, \tag{3.2}$$

with consistent $x_0$, we then have an explicit representation of the solution $x$ as

$$x(t) = Q(t)\begin{bmatrix} I_d \\ A_{2,1}(t)\end{bmatrix}\hat{\Phi}(t,t_0)\begin{bmatrix} I_d & 0\end{bmatrix}Q(t_0)^Tx_0,$$

where $\hat{\Phi}(t,t_0)$ is a fundamental solution of the so-called *inherent ODE* associated with (2.7) given by

$$\dot{x}_1 = A_{1,1}(t)x_1. \tag{3.3}$$

In particular, $\hat{\Phi}(t,t_0)$ solves the linear matrix differential equation

$$\tfrac{\partial}{\partial t}\hat{\Phi}(t,t_0) = A_{1,1}(t)\hat{\Phi}(t,t_0),\quad \hat{\Phi}(t_0,t_0)=I_d.$$

It follows that the fundamental solution $\Phi(t,t_0)$ of the homogeneous case (3.2), in the sense that the solution $x$ can be written as $x(t)=\Phi(t,t_0)x_0$, is given by

$$\Phi(t,t_0) = Q(t)\begin{bmatrix} I_d \\ A_{2,1}(t)\end{bmatrix}\hat{\Phi}(t,t_0)\begin{bmatrix} I_d & 0\end{bmatrix}Q(t_0)^T,$$

ETNA
Kent State University
etna@mcs.kent.edu



with

$$\|\Phi(t,t_0)\|_2 = \left\| \begin{bmatrix} I_d \\ A_{2,1}(t) \end{bmatrix} \hat{\Phi}(t,t_0) \begin{bmatrix} I_d & 0 \end{bmatrix} \right\|_2,$$

since $Q$ is pointwise orthogonal. Thus, we have

$$\|\Phi(t,t_0)\|_2 \geq \|\hat{\Phi}(t,t_0)\|_2,$$

and the implications

$$\begin{array}{rcl} \|\Phi(t,t_0)\|_2 \leq L & \Longrightarrow & \|\hat{\Phi}(t,t_0)\|_2 \leq L, \\ \|\Phi(t,t_0)\|_2 \to 0 & \Longrightarrow & \|\hat{\Phi}(t,t_0)\|_2 \to 0, \\ \|\Phi(t,t_0)\|_2 \leq Le^{-\gamma(t-t_0)} & \Longrightarrow & \|\hat{\Phi}(t,t_0)\|_2 \leq Le^{-\gamma(t-t_0)} \end{array}$$

hold. From this, it is clear that for the different stability concepts to extend to DAEs it is necessary that the inherent ODE (3.3) satisfies the corresponding stability concepts in the classical sense.

On the other hand, since

$$\|\Phi(t,t_0)\|_2^2 \leq \left\| \begin{bmatrix} I_d \\ A_{2,1}(t) \end{bmatrix} \right\|_2^2 \|\hat{\Phi}(t,t_0)\|_2^2 \leq (1 + \|A_{2,1}(t)\|_2^2)\, \|\hat{\Phi}(t,t_0)\|_2^2,$$

we have the implications

$$\begin{array}{rcl} \|\hat{\Phi}(t,t_0)\|_2 \leq L,\ \|A_{2,1}(t)\|_2 \leq c & \Longrightarrow & \|\Phi(t,t_0)\|_2 \leq \sqrt{1+c^2}L, \\ \|\hat{\Phi}(t,t_0)\|_2 \to 0,\ \|A_{2,1}(t)\|_2 \leq c & \Longrightarrow & \|\Phi(t,t_0)\|_2 \to 0, \\ \|\hat{\Phi}(t,t_0)\|_2 \leq Le^{-\gamma(t-t_0)},\ \|A_{2,1}(t)\|_2^2 \leq c(t-t_0)^k & \Longrightarrow & \|\Phi(t,t_0)\|_2 \leq \tilde{L}e^{-\tilde{\gamma}(t-t_0)}, \end{array}$$

where $k \geq 0$ is an arbitrary integer and $\tilde{L}, \tilde{\gamma} > 0$ are appropriate constants. We thus have obtained the following sufficient conditions.

THEOREM 3.1. *Consider system (2.7) and its reduced form (2.8) with inherent ODE (3.3).*

1. *If the inherent ODE is stable and $\|A_{2,1}(t)\|_2 \leq c$ holds with some constant $c > 0$ for all $t \in \mathbb{I}$, then (2.7) is stable in the sense that $\|\Phi(t,t_0)\| < \tilde{L}$ on $\mathbb{I}$ for some positive constant $\tilde{L}$.*
2. *If the inherent ODE is asymptotically stable and $\|A_{2,1}(t)\|_2 \leq c$ holds for some constant $c > 0$ for all $t \in \mathbb{I}$, then (2.7) is asymptotically stable in the sense that $\Phi(t,t_0) \to 0$ as $t \to \infty$.*
3. *If the inherent ODE is exponentially stable and $\|A_{2,1}(t)\|_2 \leq c(t-t_0)^k$ holds for some constant $c > 0$ and integer $k \geq 0$ for all $t \in \mathbb{I}$, then (2.7) is exponentially stable in the sense that $\|\Phi(t,t_0)\| < \tilde{L}e^{-\tilde{\gamma}(t-t_0)}$ on $\mathbb{I}$ for some constants $\tilde{L}, \tilde{\gamma} > 0$.*

**3.2. Nonlinear DAEs.** We turn now to the general case of a nonlinear DAE (1.1) with corresponding reduced problem (2.6). As in the linear case, the unknowns are connected by the transformation (3.1) such that it is again sufficient to study the reduced problem. Corresponding to the condition $\|A_{21}(t)\|_2 \leq c$ for all $t \in \mathbb{I}$ we require here that the function $\mathcal{R}$ is globally Lipschitz continuous on a sufficiently large domain $\mathbb{U}$ for $x_1$, i.e.,

$$\|\mathcal{R}(t,x_1) - \mathcal{R}(t,y_1)\|_2 \leq L\|x_1 - y_1\|_2 \text{ for all } t \in \mathbb{I} \text{ and all } x_1, y_1 \in \mathbb{U}, \tag{3.4}$$

with some constant $L > 0$. It is then clear that stability and asymptotic stability of the inherent ODE $\dot{x}_1 = \mathcal{L}(t,x_1)$ carry over to the whole reduced DAE (2.6). In particular, we have obtained the following result for an equilibrium solution $(x_1^*, x_2^*)$ of (2.6).

THEOREM 3.2. *Consider the nonlinear DAE (1.1) and its associated reduced system (2.6) and assume that (3.4) holds.*

**ETNA**
Kent State University
etna@mcs.kent.edu



1. *If $V$ satisfies the assumptions of Theorem 2.8 for the inherent ODE $\dot{x}_1 = \mathcal{L}(t, x_1)$, then $(x_1^*, x_2^*)$ is stable in the sense of Definition 2.3 with $\hat{x}_0$ restricted to be consistent.*
2. *If $V$ satisfy the assumptions of Theorem 2.9 for the inherent ODE $\dot{x}_1 = \mathcal{L}(t, x_1)$, then $(x_1^*, x_2^*)$ is asymptotically stable in the sense of Definition 2.3 with $\hat{x}_0$ restricted to be consistent.*

Contractivity and dissipativity for nonlinear DAEs have been studied for special cases in [15, 16]. We now discuss the general case.

In view of Theorem 2.10 we first require that $\mathcal{L}$ of the inherent ODE (2.6a) satisfies a one-sided Lipschitz condition, i.e.,

$$\langle \mathcal{L}(t, x_1) - \mathcal{L}(t, y_1), x_1 - y_1 \rangle_2 \le c\|x_1 - y_1\|_2^2 \text{ for all } t \in \mathbb{I} \text{ and } x_1, y_1 \in \mathbb{U}. \tag{3.5}$$

Then for two solutions $x_1, y_1$ of (3.3) and $d_1(t) = \|x_1(t) - y_1(t)\|_2^2$, we obtain

$$\begin{aligned} \tfrac{1}{2}\tfrac{d}{dt} d_1(t) &= \tfrac{1}{2}\tfrac{d}{dt}\|x_1(t) - y_1(t)\|_2^2 = \langle \dot{x}_1(t) - \dot{y}_1(t), x_1(t) - y_1(t) \rangle_2 \\ &= \langle \mathcal{L}(t, x_1(t)) - \mathcal{L}(t, y_1(t)), x_1(t) - y_1(t) \rangle_2 \le c\|x_1(t) - y_1(t)\|_2^2. \end{aligned}$$

As in Section 2.3, by Gronwall's lemma, the relation

$$\dot{d}_1(t) \le 2c d_1(t)$$

yields

$$\dot{d}_1(t) \le e^{2c(t-t_0)} d_1(t_0).$$

Introducing $d(t) = \|x(t) - y(t)\|_2^2 = \|x_1(t) - y_1(t)\|_2^2 + \|x_2(t) - y_2(t)\|_2^2$ and using (3.4), we obtain

$$\begin{aligned} d(t) &\le \|x_1(t) - y_1(t)\|_2^2 + \|\mathcal{R}(t, x_1(t)) - \mathcal{R}(t, y_1(t)))\|_2^2 \\ &\le \|x_1(t) - y_1(t)\|_2^2 + L^2\|x_1(t) - y_1(t)\|_2^2 \\ &\le (1 + L^2) e^{2c(t-t_0)} d_1(t_0). \end{aligned}$$

Thus, we have shown the following result.

THEOREM 3.3. *Consider the nonlinear DAE (1.1) and its associated reduced system (2.6). Let $\mathcal{L} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^d)$ satisfy a one-sided Lipschitz condition with constant $c \in \mathbb{R}$ according to (3.5) and let $\mathcal{R} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^a)$ be Lipschitz continuous according to (3.4). If $c = 0$, then (2.6) is contractive in the sense that $\|x(t) - y(t)\|_2$ is monotonically nonincreasing for two solutions $x, y$ of (2.6). If $c < 0$, then (2.6) is exponentially contractive in the sense that $\|x(t) - y(t)\|_2$ decays exponentially for two solutions $x, y$ of (2.6).*

To study dissipativity, we first require that

$$\langle \mathcal{L}(t, x_1), x_1 \rangle \le \alpha - \beta\|x_1\|_2^2 \text{ for all } t \in \mathbb{I} \text{ and } x \in \mathbb{U}, \tag{3.6}$$

with $\alpha \ge 0$ and $\beta > 0$. Then

$$\tfrac{1}{2}\tfrac{d}{dt}\|x_1(t)\|_2^2 = \langle \dot{x}_1(t), x_1(t) \rangle = \langle \mathcal{L}(t, x_1), x_1 \rangle \le \alpha - \beta\|x_1\|_2^2,$$

and as in Theorem 2.11 we obtain

$$x_1(t) \in \mathcal{B}(0, \sqrt{\alpha/\beta + \varepsilon}) \text{ for } t > \hat{t}.$$

With the natural requirement that $\mathcal{R}$ is bounded according to

$$\|x_1\|_2 < \alpha/\beta + \varepsilon \quad \Longrightarrow \quad \|x_2\|_2 < M \text{ for } t > \hat{t}, \tag{3.7}$$

ETNA
Kent State University
etna@mcs.kent.edu



where $M > 0$ is a suitable constant depending on $\varepsilon$, we obtain

$$\|x(t)\|_2^2 = \|x_1(t)\|_2^2 + \|x_2(t)\|_2^2 < \alpha/\beta + \varepsilon + M^2$$

and thus,

$$\|x(t)\|_2 \in \mathcal{B}(0, \sqrt{\alpha/\beta + \varepsilon + M^2}) \text{ for } t > \hat{t}.$$

Thus, we have proved the following theorem.

THEOREM 3.4. *Consider the nonlinear DAE (1.1) and its associated reduced system (2.6). Let $\mathcal{L} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^d)$ satisfy (3.6) and let $\mathcal{R} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^a)$ satisfy (3.4) and (3.7). Then the DAE (2.6) is dissipative in the sense of Definition 2.5 with $\hat{x}_0$ restricted to be consistent. An absorbing set is given by $\mathbb{B} = \mathcal{B}(0, \sqrt{\alpha/\beta + \varepsilon + M^2})$ for arbitrary $\varepsilon > 0$.*

Finally, we assume that

$$\langle \mathcal{L}(t, x_1), x_1 \rangle < 0 \text{ for all } t \in \mathbb{I} \text{ and } x_1 \in \mathbb{U} \text{ with } \|x_1\|_2 > \varrho. \tag{3.8}$$

As in Theorem 2.12 we obtain that

$$x_1(t) \in \mathcal{B}(0, \varrho + \epsilon) \text{ for } t > \hat{t},$$

and we can proceed as for (3.6), completing the proof of the following theorem.

THEOREM 3.5. *Consider the nonlinear DAE (1.1) and its associated reduced system (2.6). Let $\mathcal{L} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^d)$ satisfy (3.8) and let $\mathcal{R} \in C^0(\mathbb{I} \times \mathbb{U}, \mathbb{R}^a)$ satisfy (3.4) and (3.7). Then the DAE (2.6) is dissipative in the sense of Definition 2.5 with $\hat{x}_0$ restricted to be consistent. An absorbing set is given by $\mathbb{B} = \mathcal{B}(0, \sqrt{(\varrho + \varepsilon)^2 + M^2})$ for arbitrary $\varepsilon > 0$.*

Recall that the domain $\mathbb{U}$ must be sufficiently large to ensure that $x(t)$ does not leave the domain of definition in finite time, i.e., one has to assume that the solution exists at least until $\hat{t}$ and that the desired absorbing set is contained in $\mathbb{U}$. Besides these technical assumptions we can observe that also in the nonlinear case the various stability concepts for DAEs require the corresponding properties to hold for the inherent ODE and sufficient conditions are obtained under natural assumptions on the algebraic constraints.

**4. A test equation for DAEs.** In this section we propose and investigate a new test equation for differential-algebraic equations. As we have mentioned at the end of Section 2.2, it is sufficient to consider strangeness-free DAEs. To get an idea how a suitable test equation should look like, we must understand the reasons for the instabilities in Example 1.1.

Suppose that we discretize the linear homogeneous problem (3.2) with the implicit Euler method, i.e.,

$$(E_i - hA_i)x_i = E_i x_{i-1},$$

where $E_i = E(t_i)$, $A_i = A(t_i)$, and $x_i$ is an approximation to $x(t_i)$. If we scale the equation by a pointwise nonsingular matrix function $P \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and the solution by a pointwise nonsingular matrix function $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, then the transformed equation reads

$$\tilde{E}(t)\dot{\tilde{x}} = \tilde{A}(t)\tilde{x}, \tag{4.1}$$

where

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

Setting $P_i = P(t_i)$, $Q_i = Q(t_i)$, $\dot{Q}_i = \dot{Q}(t_i)$, and defining $\tilde{x}_i$ by $x_i = Q_i\tilde{x}_i$, we obtain

$$(\tilde{E}_i - h\tilde{A}_i - h\tilde{E}_i Q_i^{-1}\dot{Q}_i)\tilde{x}_i = \tilde{E}_i Q_i^{-1} Q_{i-1}\tilde{x}_{i-1}.$$

Since $Q_{i-1} = Q_i - h\dot{Q}_i + \mathcal{O}(h^2)$, we can rewrite this as

$$(\tilde{E}_i(I - hQ_i^{-1}\dot{Q}_i) - h\tilde{A}_i)\tilde{x}_i = \tilde{E}_i(I - hQ_i^{-1}\dot{Q}_i + \mathcal{O}(h^2))\tilde{x}_{i-1}.$$

If we would directly discretize the equation (4.1), then we would instead obtain

$$(\tilde{E}_i - h\tilde{A}_i)\tilde{x}_i = \tilde{E}_i\tilde{x}_{i-1}.$$

Example 1.1 shows that these perturbations to $\tilde{E}_i$ may have the effect that the numerical method is unstable even though the DAE itself is asymptotically stable. Obviously, to have an effect on the solution behavior, the perturbation $h\tilde{E}_iQ_i^{-1}\dot{Q}_i$ must be reasonably large. In order to simulate this behavior in a test equation, we consider for $x_1$ the classical test equation (1.3), which is (allowing here as usual for complex solutions) asymptotically stable if $Re(\lambda) < 0$. Moreover, since we solve DAEs by discretizing a numerically available strangeness-free formulation, it is sufficient that the test equation is also strangeness-free.

As we have seen in Section 3, we still obtain asymptotic stability if in (2.6) the entry $A_{2,1}$ is bounded, e.g., $A_{2,1}(t) = -1$. This corresponds to the pair of matrix functions

$$(\hat{E}(t), \hat{A}(t)) = \left( \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \lambda & 0 \\ -1 & 1 \end{bmatrix} \right).$$

In order to simulate the effect that the kernel of $E(t)$ is changing and, therefore, to have a nontrivial transformation with a derivative that depends on a parameter that can be used to control the rate of change, we will choose

$$R(t) = \begin{bmatrix} 1 & \omega t \\ 0 & 1 \end{bmatrix}, \tag{4.2}$$

with a real parameter $\omega$, to transform $(\hat{E}, \hat{A})$ to

$$(E, A) = (\hat{E}R^{-1}, \hat{A}R^{-1} - \hat{E}\frac{d}{dt}(R^{-1})).$$

A simple calcultation yields

$$E(t) = \begin{bmatrix} 1 & -\omega t \\ 0 & 0 \end{bmatrix}, \quad A(t) = \begin{bmatrix} \lambda & \omega(1 - \lambda t) \\ -1 & 1 + \omega t \end{bmatrix}.$$

In the following we, therefore, consider the test equation

$$\begin{bmatrix} 1 & -\omega t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \lambda & \omega(1 - \lambda t) \\ -1 & 1 + \omega t \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{4.3}$$

With initial data $x_1(0) = 1$, $x_2(0) = 1$, equation (4.3) has the solution

$$x(t) = \begin{bmatrix} 1 & \omega t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e^{\lambda t} \\ e^{\lambda t} \end{bmatrix} = \begin{bmatrix} (1 + \omega t)e^{\lambda t} \\ e^{\lambda t} \end{bmatrix},$$

which is asymptotically stable for $\mathrm{Re}(\lambda) < 0$ and $\omega$ arbitrary. In particular, asymptotic stability of (4.3) does not depend on $\omega$. Note that all transformations of $x$ such that the transforming matrix function and its pointwise inverse are polynomially bounded for $t \to \infty$ preserve the asymptotic stability of the solution.

ETNA
Kent State University
etna@mcs.kent.edu



Since we will need it later in the course of this paper, we describe the transformation of (4.3) to the reduced form corresponding to (2.8). With

$$Q(t) = \frac{1}{\sqrt{1+\omega^2t^2}}\begin{bmatrix} 1 & \omega t \\ -\omega t & 1 \end{bmatrix}, \quad \dot{Q}(t) = \frac{\omega}{(1+\omega^2t^2)^{3/2}}\begin{bmatrix} -\omega t & 1 \\ -1 & -\omega t \end{bmatrix}, \tag{4.4}$$

according to (3.1) we obtain that

$$\begin{aligned} EQ &= \frac{1}{\sqrt{1+\omega^2t^2}}\begin{bmatrix} 1+\omega^2t^2 & 0 \\ 0 & 0 \end{bmatrix}, \\ AQ - E\dot{Q} &= \frac{1}{\sqrt{1+\omega^2t^2}}\begin{bmatrix} \lambda - \omega^2 t + \lambda\omega^2t^2 & 0 \\ -1-\omega t - \omega^2t^2 & 1 \end{bmatrix}, \end{aligned}$$

and thus, by scaling with a diagonal matrix from the left, the pair $(E, A)$ is equivalent to the pair

$$(\tilde{E}, \tilde{A}) = \left( \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \lambda - \frac{\omega^2 t}{1+\omega^2t^2} & 0 \\ 1+\omega t+\omega^2t^2 & -1 \end{bmatrix} \right), \tag{4.5}$$

which is the required reduced form (2.8) of the test equation (4.3). Since the pointwise orthogonal transformation does not alter the asymptotic stability of the solution, the inherent ODE of (4.5) is asymptotically stable and part 3 of Theorem 3.1 applies.

Note that there is an important difference between this new test equation and the standard test equation (1.3) for ODEs. Due to the requirement that the new test equation must involve a changing kernel of $E$, it cannot be autonomous. As a consequence, the difference equation for the numerical solution which is typically obtained by discretization will explicitly include time positions.

REMARK 4.1. It should be noted that $R(t)$ in (4.2) is not pointwise orthogonal. An orthogonal variation of this transformation would be to choose

$$\hat{R}(t) = \frac{1}{\sqrt{1+\omega^2t^2}}\begin{bmatrix} 1 & \omega t \\ -\omega t & 1 \end{bmatrix}$$

or the case of a rotation with frequency $\omega$

$$\hat{R}(t) = \begin{bmatrix} \sin(\omega t) & \cos(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{bmatrix}.$$

The problem with these two orthogonal transformations is that the analysis of the stability regions of different numerical methods becomes very technical analytically. Numerical tests, however, show that there is no essential difference in the corresponding stability regions.

**5. DAE integration methods and DAE stability functions.** To demonstrate the properties of the test equation (4.3) let us apply some of the well-known DAE integration methods to this equation. In analogy to the classical stability functions $R(h\lambda) = R(z)$ for ODEs (see [13]), we will introduce *DAE stability functions* of the form $R(h\lambda, h\omega) = R(z, w)$, using the abbreviations $z = h\lambda$, $w = h\omega$. We will present several plots of stability functions. In all cases the plots depict the region given by $(z, w) \in [-9, 9]^2$. The color coding is chosen so that the dark regions are those with $|R(z, w)| \leq 1$ and the shading is according to the modulus of $R(z, w)$.

**5.1. Implicit Euler method.** Applying the implicit Euler method to the test equation (4.3), we obtain the iteration

$$\begin{bmatrix} 1-h\lambda & -\omega t_i-\omega h+\omega h\lambda t_i \\ -1 & 1+\omega t_i \end{bmatrix}\begin{bmatrix} x_{1,i} \\ x_{2,i} \end{bmatrix}=\begin{bmatrix} 1 & -\omega t_i \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_{1,i-1} \\ x_{2,i-1} \end{bmatrix}.$$

The coefficient matrix on the left side has determinant

$$D=1-h(\lambda+\omega)$$

and, thus, for $D\neq 0$ the linear system has a unique solution given by

$$\begin{aligned}\begin{bmatrix} x_{1,i} \\ x_{2,i} \end{bmatrix}&=\frac{1}{D}\begin{bmatrix} 1+\omega t_i & \omega t_i+\omega h-\omega h\lambda t_i \\ 1 & 1-h\lambda \end{bmatrix}\begin{bmatrix} 1 & -\omega t_i \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_{1,i-1} \\ x_{2,i-1} \end{bmatrix}\\ &=\frac{1}{D}\begin{bmatrix} 1+\omega t_i & -\omega t_i(1+\omega t_i) \\ 1 & -\omega t_i \end{bmatrix}\begin{bmatrix} x_{1,i-1} \\ x_{2,i-1} \end{bmatrix}.\end{aligned}$$

Since $x_{1,i-1}=(1+\omega t_{i-1})x_{2,i-1}$, we obtain

$$\begin{aligned}\begin{bmatrix} x_{1,i} \\ x_{2,i} \end{bmatrix}&=\frac{1}{D}\begin{bmatrix} 1+\omega t_i & -\omega t_i(1+\omega t_i) \\ 1 & -\omega t_i \end{bmatrix}\begin{bmatrix} (1+\omega t_{i-1})x_{2,i-1} \\ x_{2,i-1} \end{bmatrix}\\ &=\frac{1}{D}\begin{bmatrix} (1-\omega h)(1+\omega t_i) \\ 1-\omega h \end{bmatrix}x_{2,i-1}\\ &=\frac{1-\omega h}{1-(\lambda+\omega)h}\begin{bmatrix} 0 & 1+\omega t_i \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_{1,i-1} \\ x_{2,i-1} \end{bmatrix}\\ &=\left(\frac{1-\omega h}{1-(\lambda+\omega)h}\right)^i\begin{bmatrix} 0 & 1+\omega t_i \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}.\end{aligned}$$

We see that the stability behavior of the equation depends on the *DAE stability function*

$$R(z,w)=\frac{1-w}{1-z-w}.$$

Note that for $w=0$ the DAE stability function $R(z,w)$ reduces to the stability function $R(z)=(1-z)^{-1}$ of the ODE case. A plot of this function is given in Figure 5.1.

**5.2. Radau IIa method with two stages.** Applying the 2-stage Radau IIa method (see, e.g., [13]) given by the Butcher tableau

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

to (4.3), we obtain the iteration

$$\begin{aligned} x_{1,i}&=x_{1,i-1}+\tfrac{3}{4}h\dot{X}_{1,1}+\tfrac{1}{4}h\dot{X}_{2,1},\\ x_{2,i}&=x_{2,i-1}+\tfrac{3}{4}h\dot{X}_{1,2}+\tfrac{1}{4}h\dot{X}_{2,2},\end{aligned}$$

where the stage values and derivatives satisfy

$$\dot{X}_{1,1}-\omega(t_{i-1}+\tfrac{h}{3})\dot{X}_{1,2}=\lambda X_{1,1}+\omega(1-\lambda(t_{i-1}+\tfrac{h}{3}))X_{1,2},$$

**ETNA**
Kent State University
etna@mcs.kent.edu



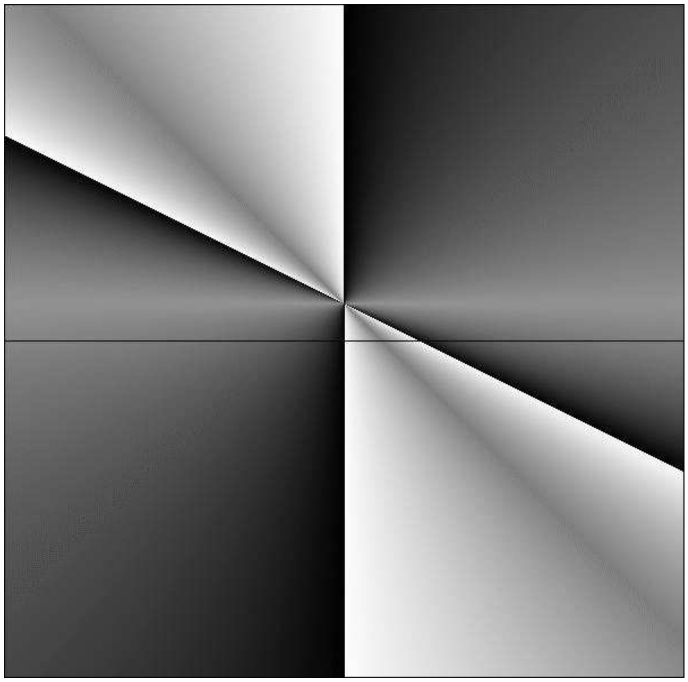

FIG. 5.1. *DAE stability function for the implicit Euler method for* $(z, w) \in [-9, 9]^2$

$$
\begin{aligned}
0 &= -X_{1,1} + (1 + \omega(t_{i-1} + \tfrac{h}{3}))X_{1,2}, \\
\dot{X}_{2,1} - \omega t_i \dot{X}_{2,2} &= \lambda X_{2,1} + \omega(1 - \lambda t_i)X_{2,2}, \\
0 &= -X_{2,1} + (1 + \omega t_i)X_{2,2}, \\
X_{1,1} &= x_{1,i-1} + \tfrac{5}{12}h\dot{X}_{1,1} - \tfrac{1}{12}h\dot{X}_{2,1}, \\
X_{1,2} &= x_{2,i-1} + \tfrac{5}{12}h\dot{X}_{1,2} - \tfrac{1}{12}h\dot{X}_{2,2}, \\
X_{2,1} &= x_{1,i-1} + \tfrac{3}{4}h\dot{X}_{1,1} + \tfrac{1}{4}h\dot{X}_{2,1}, \\
X_{2,2} &= x_{2,i-1} + \tfrac{3}{4}h\dot{X}_{1,2} + \tfrac{1}{4}h\dot{X}_{2,2}.
\end{aligned}
$$

Since the Radau IIa methods are stiffly accurate, they yield consistent approximations. Using therefore $x_{1,i-1} = (1 + \omega t_{i-1})x_{2,i-1}$ shows that all quantities are multiples of $x_{2,i-1}$. Gaussian elimination and simplification finally leads to

$$x_{2,i} = X_{2,2} = -\frac{2(2h\omega h\lambda + 2h\omega - h\lambda - 3)}{2h\lambda h\omega + (h\lambda)^2 - 4h\omega - 4h\lambda + 6}x_{2,i-1}.$$

Thus, the DAE stability function for the 2-stage Radau IIa method reads

$$R(z, w) = \frac{6 - 4w + 2z - 2zw}{6 - 4z - 4w + z^2 + 2zw}.$$

A plot of this function is given in Figure 5.2.

ETNA
Kent State University
etna@mcs.kent.edu



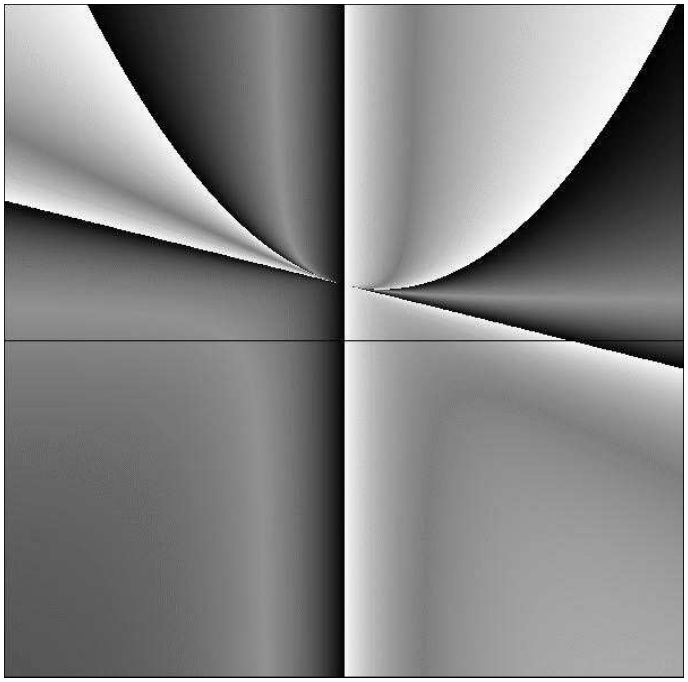

FIG. 5.2. *DAE stability function for the Radau IIa method with two stages for* $(z, w) \in [-9, 9]^2$

**5.3. Projected implicit midpoint rule.** Applying the implicit midpoint rule, i.e., the Gauß method with $s = 1$, given by the Butcher tableau

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

(see [13]) to (4.3), we obtain the following iteration for the stage values and stage derivatives

$$\begin{aligned} \dot{X}_1 - \omega(t_{i-1} + \tfrac{1}{2}h)\dot{X}_2 &= \lambda X_1 + \omega(1 - \lambda(t_{i-1} + \tfrac{1}{2}h))X_2, \\ 0 &= -X_1 + (1 + \omega(t_{i-1} + \tfrac{1}{2}h))X_2, \\ X_1 &= x_{i-1,1} + \tfrac{1}{2}h\dot{X}_1, \\ X_2 &= x_{i-1,2} + \tfrac{1}{2}h\dot{X}_2, \\ x_{i,1} &= x_{i-1,1} + h\dot{X}_1, \\ x_{i,2} &= x_{i-1,2} + h\dot{X}_2. \end{aligned}$$

Elimination of the stage values gives the linear system

$$\begin{bmatrix} 1 - \frac{1}{2}h\lambda & -\omega(t_{i-1} + \frac{1}{2}h) - \frac{1}{2}h\omega(1 - \lambda(t_{i-1} + \frac{1}{2}h)) \\ \frac{1}{2}h & -\frac{1}{2}h(1 + \omega(t_{i-1} + \frac{1}{2}h)) \end{bmatrix} \begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \end{bmatrix} = \begin{bmatrix} \lambda x_{i-1,1} + \omega(1 - \lambda(t_{i-1} + \frac{1}{2}h))x_{i-1,2} \\ -1 + (1 + \omega(t_{i-1} + \frac{1}{2}h))x_{i-1,2} \end{bmatrix}.$$

ETNA
Kent State University
etna@mcs.kent.edu



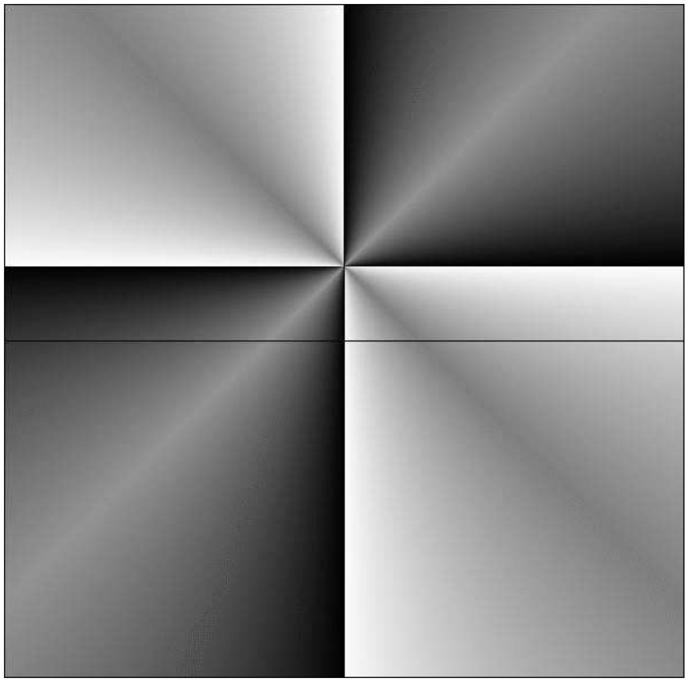

FIG. 5.3. *DAE stability function for the projected implicit midpoint rule for* $(z, w) \in [-9, 9]^2$

Using $x_{1,i-1} = (1 + \omega t_{i-1})x_{2,i-1}$ and, hence, assuming that we work with consistent approximations (e.g., by projecting in every step), one derives that $x_{2,i} = R(z, w)x_{2,i-1}$ with

$$R(z, w) = \frac{2 + z - w}{2 - z - w}. \tag{5.1}$$

A plot of this function is given in Figure 5.3.

**5.4. Projected implicit trapezoidal rule.** Applying the implicit trapezoidal rule, i.e., the 2-stage Lobatto method (see [13]) given by the Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

to (4.3), we obtain the relations

$$\begin{aligned} \dot{X}_{1,1} - \omega t_{i-1}\dot{X}_{1,2} &= \lambda X_{1,1} + \omega(1 - \lambda t_{i-1})X_{1,2}, \\ 0 &= -X_{1,1} + (1 + \omega t_{i-1})X_{1,2}, \\ \dot{X}_{2,1} - \omega t_i \dot{X}_{2,2} &= \lambda X_{2,1} + \omega(1 - \lambda t_i)X_{2,2}, \\ 0 &= -X_{2,1} + (1 + \omega t_i)X_{2,2}, \\ X_{1,1} &= x_{1,i-1}, \\ X_{1,2} &= x_{2,i-1}, \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu



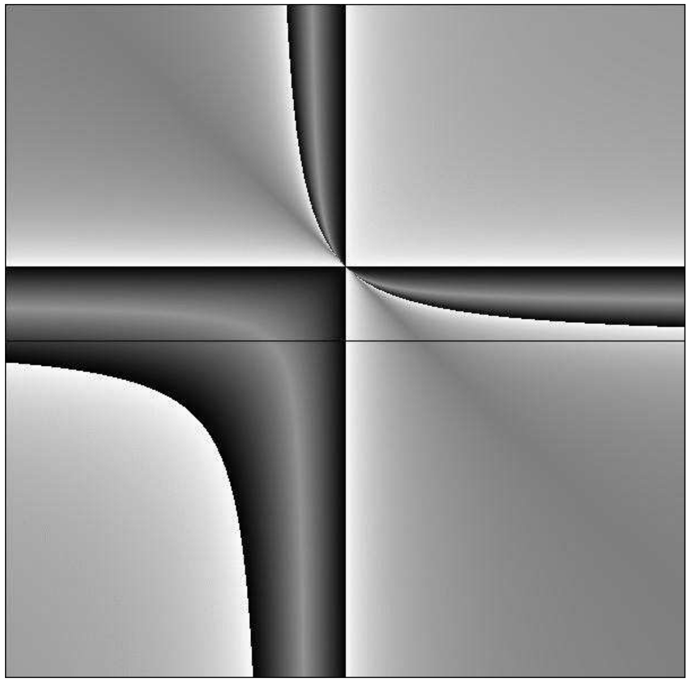

FIG. 5.4. *DAE stability function for the projected implicit trapezoidal rule for* $(z, w) \in [-9, 9]^2$

$$X_{2,1} = x_{1,i-1} + \tfrac{1}{2}h\dot{X}_{1,1} + \tfrac{1}{2}h\dot{X}_{2,1},$$
$$X_{2,2} = x_{2,i-1} + \tfrac{1}{2}h\dot{X}_{1,2} + \tfrac{1}{2}h\dot{X}_{2,2},$$

for the stage values and derivatives. Eliminating the stage values and using as before $x_{1,i-1} = (1 + \omega t_{i-1})x_{2,i-1}$ under the assumption that we work with consistent approximations, we obtain that $x_{2,i} = X_{2,2} = R(z, w)x_{2,i-1}$ with

$$R(z, w) = \frac{2 + z - w - zw}{2 - z - w}.$$

A plot of this function is given in Figure 5.4.

In the following sections we consider classes of numerical methods which are common in the treatment of DAEs in order to show the typical structure of the corresponding DAE stability functions.

**5.5. Stiffly accurate Runge-Kutta methods.** Applying a general stiffly accurate Runge-Kutta method (see [13]) given by the Butcher tableau

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

with $\mathcal{A}$ invertible and $b^T\mathcal{A}^{-1}e = 1$, $e = [\,1 \ \cdots \ 1\,]^T$ to (4.3), we obtain the relations

$$\text{(5.2)} \quad \begin{array}{ll} \text{(a)} & \dot{X}_{j,1} - \omega(t_{i-1} + c_j h)\dot{X}_{j,2} = \lambda X_{j,1} + \omega(1 - \lambda(t_{i-1} + c_j h))X_{j,2}, \\ \text{(b)} & 0 = -X_{j,1} + (1 + \omega(t_{i-1} + c_j h))X_{j,2}, \\ \text{(c)} & X_{j,1} = x_{1,i-1} + h\sum_{l=1}^{s} a_{j,l}\dot{X}_{l,1}, \\ \text{(d)} & X_{j,2} = x_{2,i-1} + h\sum_{l=1}^{s} a_{j,l}\dot{X}_{l,2}, \end{array}$$

ETNA
Kent State University
etna@mcs.kent.edu



for $j = 1, \ldots, s$. Obviously, all stage values are consistent due to

$$X_{j,1} = (1 + \omega(t_{i-1} + c_j h))X_{j,2},$$

and so all numerical approximations due to $x_{1,i-1} = (1 + \omega t_{i-1})x_{2,i-1}$. Using the vectors of stage values and derivatives defined by

$$X_1 = \begin{bmatrix} X_{1,1} \\ \vdots \\ X_{s,1} \end{bmatrix}, \quad X_2 = \begin{bmatrix} X_{1,2} \\ \vdots \\ X_{s,2} \end{bmatrix}, \quad \dot{X}_1 = \begin{bmatrix} \dot{X}_{1,1} \\ \vdots \\ \dot{X}_{s,1} \end{bmatrix}, \quad \dot{X}_2 = \begin{bmatrix} \dot{X}_{1,2} \\ \vdots \\ \dot{X}_{s,2} \end{bmatrix},$$

the relations (5.2c,d) yield

$$\dot{X}_1 = \tfrac{1}{h}\mathcal{A}^{-1}(X_1 - e\,x_{1,i-1}), \quad \dot{X}_2 = \tfrac{1}{h}\mathcal{A}^{-1}(X_2 - e\,x_{2,i-1}).$$

Eliminating then $\dot{X}_1$, $\dot{X}_2$ in (5.2a) and multiplying by $h$ gives

$$\begin{aligned} \mathcal{A}^{-1}(X_1 - e\,x_{1,i-1}) - \begin{bmatrix} \omega\hat{t}_1 & & \\ & \ddots & \\ & & \omega\hat{t}_s \end{bmatrix} \mathcal{A}^{-1}(X_2 - e\,x_{2,i-1}) \\ = \lambda X_1 + \begin{bmatrix} \omega(1-\lambda\hat{t}_1) & & \\ & \ddots & \\ & & \omega(1-\lambda\hat{t}_s) \end{bmatrix} X_2, \end{aligned}$$

with $\hat{t}_j = t_{i-1} + c_j h$, $j = 1, \ldots, s$. Utilizing finally the consistency relations, we obtain the linear equation

$$\begin{aligned} \begin{bmatrix} v_{1,1} - z - w & v_{1,2}(1 + w(c_2 - c_1)) & \ldots & v_{1,s}(1 + w(c_s - c_1)) \\ v_{2,1}(1 + w(c_1 - c_2)) & v_{2,2} - z - w & \ldots & v_{2,s}(1 + w(c_s - c_2)) \\ \vdots & & \ddots & \vdots \\ v_{s,1}(1 + w(c_1 - c_s)) & v_{s,2}(1 + w(c_2 - c_s)) & \cdots & v_{s,s} - z - w \end{bmatrix} \begin{bmatrix} X_{1,2} \\ X_{2,2} \\ \vdots \\ X_{s,2} \end{bmatrix} \\ = \begin{bmatrix} d_1(1 - c_1 w)x_{2,i-1} \\ d_2(1 - c_2 w)x_{2,i-1} \\ \vdots \\ d_s(1 - c_s w)x_{2,i-1} \end{bmatrix}, \end{aligned}$$

with $\mathcal{A}^{-1} = (v_{j,l})$ and $d_j = v_{j,1} + \cdots + v_{j,s}$. Since $x_{2,i} = X_{2,s}$, this in particular shows that $x_{2,i} = R(z,w)x_{2,i-1}$ with a rational stability function $R(z,w)$ only depending on the parameters defining the Runge-Kutta method.

**5.6. Gauß-Lobatto methods.** Applying the Gauß-Lobatto method collocation method (see [23, 24]) with $k = 1$ to (4.3), we obtain the iteration

$$\begin{aligned} \frac{x_{1,i} - x_{1,i-1}}{h} - \omega(t_i - \tfrac{1}{2}h)\frac{x_{2,i} - x_{2,i-1}}{h} \\ = \lambda\frac{x_{1,i} + x_{1,i-1}}{2} + \omega(1 - \lambda(t_i - \tfrac{1}{2}h))\frac{x_{2,i} + x_{2,i-1}}{2}, \\ 0 = -x_{1,i} + (1 + \omega t_i)x_{2,i}, \end{aligned}$$

which yields

$$\begin{aligned} &\left[(1 + \omega t_i) - \omega(t_i - \tfrac{1}{2}h) - \tfrac{1}{2}h\lambda(1 + \omega t_i) - \tfrac{1}{2}h\omega(1 - \lambda(t_i - \tfrac{1}{2}h))\right] x_{2,i} \\ &= \left[(1 + \omega t_{i-1}) - \omega(t_{i-1} + \tfrac{1}{2}h) + \tfrac{1}{2}h\lambda(1 + \omega t_{i-1}) + \tfrac{1}{2}h\omega(1 - \lambda(t_{i-1} - \tfrac{1}{2}h))\right] x_{2,i-1}. \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu



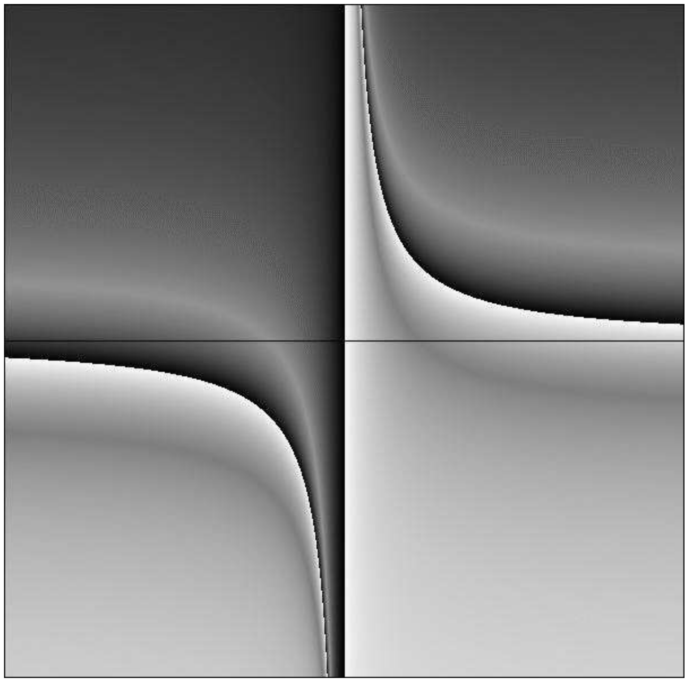

FIG. 5.5. *DAE stability function for the Gauß-Lobatto method with* $k = 1$ *for* $(z, w) \in [-9, 9]^2$

Simplifying the bracketed expressions, we obtain $x_{2,i} = R(z,w)x_{2,i-1}$ with the DAE stability function

$$R(z,w) = \frac{4 + 2z - zw}{4 - 2z - zw}.$$

A plot of this function is given in Figure 5.5.

For the general Gauß-Lobatto collocation method applied to (4.3), we obtain the relations

$$\frac{1}{h}\sum_{l=0}^{k} v_{j,l}(X_{l,1} - \omega(t_{i-1} + \varrho_j h)X_{l,2} - \sum_{l=0}^{k} u_{j,l}(\lambda X_{l,1} + \omega(1 - \lambda(t_{i-1} + \varrho_j h))X_{l,2}) = 0,$$
$$-X_{j,1} + (1 + \omega(t_{j-1} + \sigma_j h))X_{j,2} = 0,$$

for $j = 1, \ldots, k$. Here, $\varrho_1, \ldots, \varrho_k$ denote the Gauß nodes and $\sigma_0, \ldots, \sigma_k$ the Lobatto nodes with the corresponding number of stages. If $L_l$ denote the Lagrange polynomials in the Lobatto nodes, then $v_{j,l} = \dot{L}_l(\varrho_j)$ and $u_{j,l} = L_l(\varrho_j)$ for $l = 0, \ldots, k$. Furthermore, $x_{2,i} = X_{k,2}$, $X_{0,1} = x_{1,i-1}$, and $X_{0,2} = x_{2,i-1}$. Since these methods yield consistent approximations, we have that $x_{1,i-1} = (1 + \omega t_{i-1})x_{2,i-1}$. Combining all these, we obtain the formulation

$$\sum_{l=0}^{k}(v_{j,l} - u_{j,l}h\lambda)X_{l,1} - \sum_{l=0}^{k}\Big[v_{j,l}\omega(t_{i-1} + \varrho_j h) + u_{j,l}h\omega(1 - \lambda(t_{i-1} + \varrho_j h))\Big]X_{l,2} = 0,$$

**ETNA**
Kent State University
etna@mcs.kent.edu



which is equivalent to

$$\sum_{l=0}^{k}\Big[(v_{j,l}-u_{j,l}h(\lambda+\omega)+v_{j,l}h\omega(\sigma_l-\varrho_j)-u_{j,l}h\omega h\lambda(\sigma_l-\varrho_j)\Big]X_{l,2}=0.$$

The latter relation shows that the values $X_{l,2}$, $l=1,\dots,k$, satisfy a linear system of equations with a right hand side containing the factor $X_{0,2}=x_{2,i-1}$. Moreover, besides the quantities $(z,w)$ the relation only contains coefficients describing the specific method. Hence, $x_{2,i}=R(z,w)x_{2,i-1}$ with a rational stability function $R(z,w)$.

**5.7. BDF methods.** Applying a BDF method (see, e.g., [2, 13]) to (4.3), we obtain the iteration

$$\frac{1}{h}\sum_{l=0}^{k}\alpha_{k-l}x_{1,i-l}-\omega t_i\frac{1}{h}\sum_{l=0}^{k}\alpha_{k-l}x_{2,i-l}=\lambda x_{1,i}+\omega(1-\lambda t_i)x_{2,i},$$
$$0=-x_{1,i}+(1+\omega t_i)x_{2,i}.$$

Due to the latter relation, the BDF method yields consistent approximations. Utilizing, therefore, that all past approximations are consistent, we obtain

$$\sum_{l=0}^{k}\alpha_{k-l}[(1+\omega t_{i-l})-\omega t_i]x_{2,i-l}=[h\lambda(1+\omega t_i)+h\omega(1-\lambda t_i)]x_{2,i}.$$

On an equidistant grid, this yields the homogeneous difference equation

$$(\alpha_k-h\lambda)x_{2,i}+\sum_{l=1}^{k}\alpha_{k-l}(1-lh\omega)x_{2,i-l}=0.$$

Requiring that all solutions of the difference equations are bounded is equivalent to requiring that the associated polynomial

$$(\alpha_k-h\lambda)\varrho^i+\sum_{l=1}^{k}\alpha_{k-l}(1-lh\omega)\varrho^{i-l}=0.$$

satisfies the so-called root condition, namely that all roots are bounded by one in modulus and those of modulus one are simple; see again [2, 13]. Note that this property only depends on $(z,w)$. The dark regions in Figure 5.6 for $k=2$ are those points $(z,w)$ where the root condition holds. The shading is related to the largest modulus of the roots.

**5.8. Summary of DAE stability functions.** Table 5.1 summarizes all DAE stability functions that we have obtained by applying classical DAE one-step methods to the test equation (4.3). Moreover, we have included some DAE stability functions for higher order methods which were obtained with the help of the formula manipulation package `MAPLE`. Note also that methods which do not yield consistent numerical approximations are considered together with projection.

Obviously, for $w=0$ the obtained DAE stability function reduces to the classical stability function for this method applied to the standard test function (1.3) for ODEs. As $\lambda$ describes eigenvalues in the system, one is interested in complex values of $z=h\lambda$. Of course, the above results are still valid for a parameter $z\in\mathbb{C}$. Instead of the plots given in the previous sections, we can think of stability regions in the complex $z$-plane parameterized

**ETNA**
Kent State University
etna@mcs.kent.edu



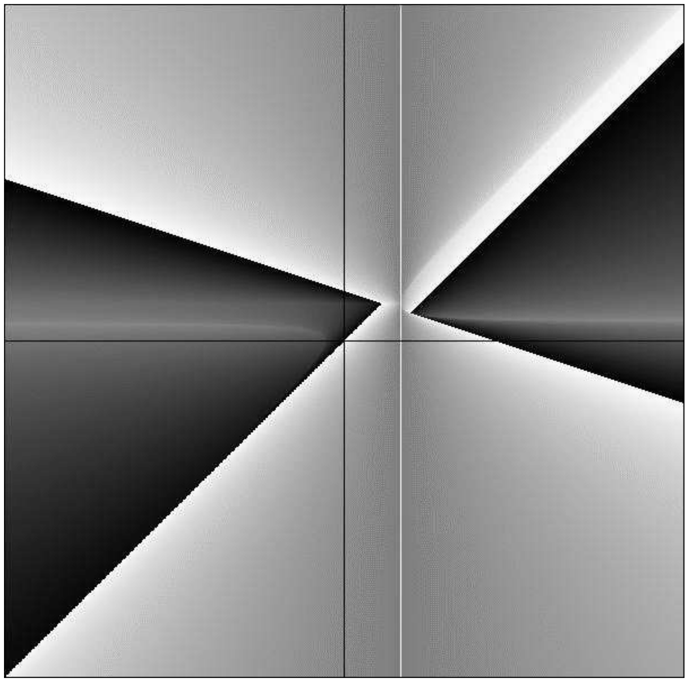

FIG. 5.6. *DAE stability function for the BDF method with* $k = 2$ *for* $(z, w) \in [-9, 9]^2$

by a real parameter $w$. Such objects can be visualized by movies. For the methods discussed here such movies can be found in the supplement of this paper,

http://etna.math.kent.edu/vol.26.2007/pp385-420.dir/stab.html.

They show the $z$-plane in the range $\mathrm{Re}\, z, \mathrm{Im}\, z \in [-9, 9]$ with the time running over $w \in [-5, 5]$.

Comparing the stability domains of the various methods, one recognizes that they behave differently with the sign of $w$. In the case of a negative eigenvalue $\lambda$ in (4.3), the Radau IIa method with $s = 2$ for example stays stable for arbitrary negative $w$ but may exhibit difficulties for a certain positive $w$, whereas for the Gauß-Lobatto method with $k = 1$ it is just the other way around. It remains, however, unclear how this behavior can be exploited in applications.

**6. Spin-stabilized discretizations.** As we have seen in Section 4, numerical schemes may become unstable when they are applied to DAEs with a spinning kernel of $\hat{E}$, where $\hat{E}$ in general is the linearization of a reduced formulation (2.10) of the given DAE (1.1) with respect to $\dot{x}$, i.e.,

$$\hat{E}(t) = \begin{bmatrix} Z_1(t)^T F_{\dot{x}}(t, x(t), \dot{x}(t)) \\ 0 \end{bmatrix}.$$

In particular, we expect such effects, when the transformation $Q$ involved in (2.9) yields a large term $[\, I_d \;\; 0\,]Q^T\dot{Q}$. Example 1.1 for $\eta = 0$ shows that discretizing a given DAE with the implicit Euler method actually results in discretizing the inherent ODE with the explicit

ETNA
Kent State University
etna@mcs.kent.edu



TABLE 5.1
*DAE stability functions*

| Method | DAE-stability function $R(z,w)$ |
|---|---|
| Implicit Euler | $R(z,w) = \dfrac{1-w}{1-z-w}$ |
| Radau IIa $s=2$ | $R(z,w) = \dfrac{6-4w+2z-2zw}{6-4z-4w+z^2+2zw}$ |
| Radau IIa $s=3$ | $R(z,w) = \dfrac{60-36w+24z-18zw+3z^2-3z^2w}{60-36w-36z+18zw+9z^2-z^3-3z^2w}$ |
| Implicit midpoint rule | $R(z,w) = \dfrac{2+z-w}{2-z-w}$ |
| Gauß $s=2$ | $R(z,w) = \dfrac{12-6w+6z-4zw+z^2}{12-6w-6z+2zw+z^2}$ |
| Gauß-Lobatto $k=1$ | $R(z,w) = \dfrac{4+2z-zw}{4-2z-zw}$ |
| Gauß-Lobatto $k=2$ | $R(z,w) = \dfrac{24+12z-2zw+2z^2-z^2w}{24-12z-2zw+2z^2+z^2w}$ |
| Implicit trapezoidal rule | $R(z,w) = \dfrac{2+z-w-zw}{2-z-w}$ |

Euler method. If in such a case the inherent ODE is stiff, then it is necessary to apply stable discretization methods. A possibility to overcome these difficulties would be to determine a smooth transformation $Q$ to get rid of the spinning kernel. Although this could be performed numerically (see, e.g., [3, 6, 18, 32, 40] or [21, Cor. 3.10]), such a procedure in general would be too costly. In the following, we therefore present an alternative approach.

As in the treatment of stiff ODEs, where it is assumed that the stiffness is contained in the linearized equation, we assume that the spin-effect is covered by the linearization of $Q$. The idea then is to use a linear approximation

$$\tilde{Q}(t) = \tilde{Q}(\hat{t}) + (t-\hat{t})\dot{\tilde{Q}}, \quad \dot{\tilde{Q}} \in \mathbb{R}^{n,n},\ \hat{t} \in \mathbb{I} \text{ fixed}, \tag{6.1}$$

such that in the $i$-th step of a $k$-step method with stepsize $h$

$$\tilde{Q}(t) - Q(t) = \mathcal{O}(h^2), \quad \dot{\tilde{Q}} - \dot{Q}(t) = \mathcal{O}(h) \tag{6.2}$$

holds for all $t \in [t_i, t_{i+k}]$ with small constants in the remainder terms. We then use this linear approximation to transform the given DAE before we discretize it. A possible choice is given by

$$\tilde{Q}(t) = Q(t_{i+k}) + (t-t_{i+k})\dot{\tilde{Q}}, \quad \dot{\tilde{Q}} = \frac{1}{h}(Q(t_{i+k}) - Q(t_{i+k-1})). \tag{6.3}$$

In the numerical computations, one must be aware that $Q$ is not unique and that we therefore do not get a smooth representation of $Q$. The selection can be made unique by freezing the pivoting and all other decisions performed during the computation of $Q(t_{i+k})$ say by QR-decomposition, when we determine $Q(t_{i+k-1})$.

ETNA
Kent State University
etna@mcs.kent.edu



In the stability analysis of the resulting numerical methods we will make use of the following properties of the linear transformation described by $\tilde{Q}$.

LEMMA 6.1. *Let $\tilde{Q}$ satisfy (6.1) with (6.2) and let $(E, A)$ be the matrix functions of the test equation (4.3). For $\tau_1, \tau_2 \to \infty$ with $\tau_1, \tau_2 \in [t_i, t_{i+k}]$ and $\tau_2 - \tau_1 = \mathcal{O}(h)$, the constants here and in (6.2) being independent of $\tau_1, \tau_2$, the limits*

$$\lim_{\tau_1,\tau_2\to\infty} E(\tau_2)\tilde{Q}(\tau_2)\tilde{Q}(\tau_1)^{-1}\begin{bmatrix} 1+\omega\tau_1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \mathcal{O}(h^2)$$

*and*

$$\lim_{\tau_1\to\infty} E(\tau_1)\dot{\tilde{Q}}\tilde{Q}(\tau_1)^{-1}\begin{bmatrix} 1+\omega\tau_1 \\ 1 \end{bmatrix} = \begin{bmatrix} \omega \\ 0 \end{bmatrix} + \mathcal{O}(h)$$

*hold.*

*Proof*. The properties (6.2) imply that

$$\begin{aligned}\tilde{Q}(\tau_2)\tilde{Q}(\tau_1)^{-1} &= (Q(\tau_2) + \mathcal{O}(h^2))(Q(\tau_1) + \mathcal{O}(h^2))^{-1} \\ &= (Q(\tau_1) + (\tau_2-\tau_1)\dot{Q}(\tau_1) + \mathcal{O}(h^2))(Q(\tau_1) + \mathcal{O}(h^2))^{-1} \\ &= I + (\tau_2-\tau_1)\dot{Q}(\tau_1)Q(\tau_1)^{-1} + \mathcal{O}(h^2)\end{aligned}$$

and

$$\dot{\tilde{Q}}\tilde{Q}(\tau_1)^{-1} = (\dot{Q}(\tau_1) + \mathcal{O}(h))(Q(\tau_1) + \mathcal{O}(h^2))^{-1} = \dot{Q}(\tau_1)Q(\tau_1)^{-1} + \mathcal{O}(h),$$

where

$$\dot{Q}(\tau_1)Q(\tau_1)^{-1} = \frac{\omega}{1+\omega^2\tau_1^2}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

by a simple calculation. Multiplying with $E(\tau_2)$ and $E(\tau_1)$, respectively, from the left gives

$$E(\tau_2)\tilde{Q}(\tau_2)\tilde{Q}(\tau_1)^{-1} = \begin{bmatrix} 1 & -\omega\tau_2 \\ 0 & 0 \end{bmatrix} + (\tau_2-\tau_1)\frac{\omega}{1+\omega^2\tau_1^2}\begin{bmatrix} \omega\tau_2 & 1 \\ 0 & 0 \end{bmatrix} + \mathcal{O}(h^2)$$

and

$$E(\tau_1)\dot{\tilde{Q}}\tilde{Q}(\tau_1)^{-1} = \frac{\omega}{1+\omega^2\tau_1^2}\begin{bmatrix} \omega\tau_1 & 1 \\ 0 & 0 \end{bmatrix} + \mathcal{O}(h)$$

such that

$$\begin{aligned}&E(\tau_2)\tilde{Q}(\tau_2)\tilde{Q}(\tau_1)^{-1}\begin{bmatrix} 1+\omega\tau_1 \\ 1 \end{bmatrix} \\ &\quad = \begin{bmatrix} 1-\omega(\tau_2-\tau_1) + (\tau_2-\tau_1)\frac{\omega}{1+\omega^2\tau_1^2}(1+\omega\tau_2+\omega^2\tau_1\tau_2) \\ 0 \end{bmatrix} + \mathcal{O}(h^2)\end{aligned}$$

and

$$E(\tau_1)\dot{\tilde{Q}}\tilde{Q}(\tau_1)^{-1}\begin{bmatrix} 1+\omega\tau_1 \\ 1 \end{bmatrix} = \frac{\omega}{1+\omega^2\tau_1^2}\begin{bmatrix} 1+\omega\tau_1+\omega^2\tau_1^2 \\ 0 \end{bmatrix} + \mathcal{O}(h).$$

The limits are then obvious. □

According to [21], we are allowed to restrict ourselves to the case of strangeness-free DAEs. In the following we also concentrate mainly on linear problems.

**ETNA**
Kent State University
etna@mcs.kent.edu



**6.1. A general convergence result.** In the following, we study the convergence properties of methods that are obtained by including a transformation before a given convergent method is applied. We use the notation of [21, Ch. 5] but have to slightly modify the general approach given there. As usual we restrict ourselves to equidistant grids.

Let $\tilde{\mathfrak{X}}_i$ represent the numerical approximation and let $\tilde{\mathfrak{X}}(t_i)$ represent the corresponding true solution at time $t_i = t_0 + ih$. We start with a basic numerical method given by

$$\tilde{\mathfrak{X}}_{i+1} = \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}_i; h) \tag{6.4}$$

representing any classical integration method for DAEs. We assume that (6.4) is *consistent of order* $p$ according to

$$\|\tilde{\mathfrak{X}}(t_{i+1}) - \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h)\| \le Ch^{p+1} \tag{6.5}$$

and *stable* according to

$$\|\mathfrak{R}_{i+1}(\tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h) - \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}_i; h))\| \le (1 + hK)\|\mathfrak{R}_i(\tilde{\mathfrak{X}}(t_i) - \tilde{\mathfrak{X}}_i)\|. \tag{6.6}$$

In the latter estimate, the quantities $\mathfrak{R}_i$ are matrices which are required to satisfy

$$\begin{array}{ll} \text{(a)} & \|\mathfrak{R}_i\|,\ \|\mathfrak{R}_i^{-1}\| \le M, \\ \text{(b)} & \mathfrak{R}_{i+1}\mathfrak{R}_i^{-1} = I + \mathcal{O}(h). \end{array} \tag{6.7}$$

Moreover, all involved constants are assumed to be independent of $i$ and $h$. Then, the estimate

$$\begin{aligned} &\|\mathfrak{R}_{i+1}(\tilde{\mathfrak{X}}(t_{i+1}) - \tilde{\mathfrak{X}}_{i+1})\| \\ &\qquad = \|\mathfrak{R}_{i+1}(\tilde{\mathfrak{X}}(t_{i+1}) - \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h) + \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h) - \tilde{\mathfrak{X}}_{i+1})\| \\ &\qquad \le MCh^{p+1} + (1 + hK)\|\mathfrak{R}_i(\tilde{\mathfrak{X}}(t_i) - \tilde{\mathfrak{X}}_i)\| \end{aligned}$$

holds and, hence, the method is convergent.

With the help of this basic method, we define a new method by applying in each step first a transformation, then the integration step by the basic method in the transformed system, and finally a back-transformation. Thus, the so obtained new method has the form

$$\mathfrak{X}_{i+1} = \mathfrak{F}(t_i, \mathfrak{X}_i; h),$$

with

$$\mathfrak{F}(t_i, \mathfrak{X}_i; h) = \mathfrak{Q}_{i+1}\tilde{\mathfrak{F}}(t_i, \mathfrak{Q}_i^{-1}\mathfrak{X}_i; h).$$

The quantities $\mathfrak{Q}_i$ will describe the mentioned spin-stabilization but at the moment they may represent any suitable transformations. Note that we omit a subscript $i$ although $\mathfrak{F}$ is defined differently in each integration step. According to (6.7) we require that

$$\begin{array}{ll} \text{(a)} & \|\mathfrak{Q}_i\|,\ \|\mathfrak{Q}_i^{-1}\| \le M, \\ \text{(b)} & \mathfrak{Q}_{i+1}\mathfrak{Q}_i^{-1} = I + \mathcal{O}(h). \end{array} \tag{6.8}$$

With the relations $\tilde{\mathfrak{X}}_i = \mathfrak{Q}_i^{-1}\mathfrak{X}_i$ and $\tilde{\mathfrak{X}}(t_i) = \mathfrak{Q}_i^{-1}\mathfrak{X}(t_i)$, we then have that

$$\begin{aligned} &\|\mathfrak{X}(t_{i+1}) - \mathfrak{F}(t_i, \mathfrak{X}(t_i); h)\| \\ &\qquad = \|\mathfrak{X}(t_{i+1}) - \mathfrak{Q}_{i+1}\tilde{\mathfrak{F}}(t_i, \mathfrak{Q}_i^{-1}\mathfrak{X}(t_i); h)\| = \|\mathfrak{Q}_{i+1}\tilde{\mathfrak{X}}(t_{i+1}) - \mathfrak{Q}_{i+1}\tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h)\| \\ &\qquad \le \|\mathfrak{Q}_{i+1}\|\, \|\tilde{\mathfrak{X}}(t_{i+1}) - \tilde{\mathfrak{F}}(t_i, \tilde{\mathfrak{X}}(t_i); h)\| \le MCh^{p+1} \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu



and that

$$\begin{aligned}&\|\mathfrak{R}_{i+1}\mathfrak{Q}_{i+1}^{-1}(\mathfrak{F}(t_i,\mathfrak{X}(t_i);h)-\mathfrak{F}(t_i,\mathfrak{X}_i;h))\|\\&\quad=\|\mathfrak{R}_{i+1}(\tilde{\mathfrak{F}}(t_i,\tilde{\mathfrak{X}}(t_i);h)-\tilde{\mathfrak{F}}(t_i,\tilde{\mathfrak{X}}_i;h))\|\\&\quad\le(1+hK)\|\mathfrak{R}_i(\tilde{\mathfrak{X}}(t_i)-\tilde{\mathfrak{X}}_i;h)\|=(1+hK)\|\mathfrak{R}_i\mathfrak{Q}_i^{-1}(\mathfrak{X}(t_i)-\mathfrak{X}_i)\|.\end{aligned}$$

Hence, if the basic method is convergent, then the new method that first transforms, then applies the basic method, and finally transforms back is convergent as well.

In the special case of the DAE integration methods that we will consider together with the spin-stabilization according to (6.1) for the transformations, we will be in the situation that

$$\mathfrak{X}_i=\begin{bmatrix}x_{i+k-1}\\x_{i+k-2}\\\vdots\\x_i\end{bmatrix},\quad\mathfrak{X}(t_i)=\begin{bmatrix}x(t_{i+k-1})\\x(t_{i+k-2})\\\vdots\\x(t_i)\end{bmatrix}\tag{6.9}$$

and

$$\mathfrak{Q}_i=\begin{bmatrix}\tilde{Q}(t_{i+k-1})&&&\\&\tilde{Q}(t_{i+k-2})&&\\&&\ddots&\\&&&\tilde{Q}(t_i)\end{bmatrix},$$

where we again omit a subscript $i$ at $\tilde{Q}$, which also differs from step to step. Since we stay close to a (continuous) path $Q(t)$ of orthogonal matrices on a compact interval when we deal with convergence, it is clear that the properties (6.8) hold.

The numerical method given by $\tilde{\mathfrak{F}}$ in (6.4) is then applied to integrate the transformed DAE with coefficient functions

$$\tilde{E}=E\tilde{Q},\quad\tilde{A}=A\tilde{Q}-E\dot{\tilde{Q}}.$$

In the following section we discuss the spin-stabilization approach for two classes of standard DAE integrators.

**6.2. Spin-stabilized stiffly accurate Runge-Kutta methods.** In this section we discuss the use of spin-stabilization within stiffly accurate Runge-Kutta methods possessing an invertible coefficient matrix $\mathcal{A}$. For this, let a linear DAE (2.7) be given which is already strangeness-free such that we do not need to perform an index reduction.

A Runge-Kutta method for the integration of (2.7) has the form

$$\begin{array}{lrll}\text{(a)}&x_{i+1}=x_i+h\sum_{j=1}^s\beta_j\dot{X}_j,&\\\text{(b)}&X_j=x_i+h\sum_{l=1}^s\alpha_{j,l}\dot{X}_l,&j=1,\ldots,s,\\\text{(c)}&E_j\dot{X}_j=A_jX_j+f_j,&j=1,\ldots,s,\end{array}\tag{6.10}$$

with

$$E_j=E(t_i+\gamma_jh),\quad A_j=A(t_i+\gamma_jh),\quad f_j=f(t_i+\gamma_jh).$$

For convenience, we use the short hand notation

$$\operatorname{diag}(E_j)=\begin{bmatrix}E_1&&\\&\ddots&\\&&E_s\end{bmatrix},\quad\operatorname{col}(f_j)=\begin{bmatrix}f_1\\\vdots\\f_s\end{bmatrix}$$

ETNA
Kent State University
etna@mcs.kent.edu



which also applies to other arguments. Using the Kronecker product, as it is common in the treatment of Runge-Kutta methods, we can solve (6.10b) according to

$$\dot{X} = \frac{1}{h}(\mathcal{A}^{-1} \otimes I_n)(X - (e \otimes x_i)),$$

where $X = \operatorname{col}(X_j)$ and $\dot{X} = \operatorname{col}(\dot{X}_j)$. Writing (6.10c) as

$$\operatorname{diag}(E_j)\dot{X} = \operatorname{diag}(A_j)X + \operatorname{col}(f_j),$$

we can eliminate $\dot{X}$ to obtain

$$\operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n)(X - (e \otimes x_i)) = h\operatorname{diag}(A_j)X + h\operatorname{col}(f_j)$$

and thus

$$\begin{aligned}[\operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n) - h\operatorname{diag}(A_j)]X \\ = \operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n)(e \otimes x_i) + h\operatorname{col}(f_j).\end{aligned} \tag{6.11}$$

Observing that the leading matrix is invertible for sufficiently small $h$ and that the numerical solution $x_{i+1}$ is given by the last block entry of $X$ in the case of stiffly accurate Runge-Kutta schemes, we obtain

$$\begin{aligned}x_{i+1} = (e_s^T \otimes I_n)[\operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n) - h\operatorname{diag}(A_j)]^{-1} \\ \cdot [\operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n)(e \otimes x_i) + h\operatorname{col}(f_j)],\end{aligned}$$

where $e_s = [\,0 \ \cdots \ 0 \ 1\,]^T \in \mathbb{R}^s$. In view of (6.6), we must consider the matrix

$$W = (e_s^T \otimes I_n)[\operatorname{diag}(E_j)(\mathcal{A}^{-1} \otimes I_n) - h\operatorname{diag}(A_j)]^{-1}\operatorname{diag}(E_j)(d \otimes I_n),$$

where $d = \mathcal{A}^{-1}e$ as in Section 5.5. Let $P_j, Q_j$ denote matrices that transform $(E_j, A_j)$ to Weierstraß canonical form (see [2, 21]) according to

$$P_jE_jQ_j = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \quad P_jA_jQ_j = \begin{bmatrix} C_j & 0 \\ 0 & I_a \end{bmatrix}.$$

Then $W$ can be represented as

$$\begin{aligned}W = (e_s^T \otimes I_n)\operatorname{diag}(Q_j) \\ \cdot [\operatorname{diag}(P_jE_j)(\mathcal{A}^{-1} \otimes I_n)\operatorname{diag}(Q_j) - h\operatorname{diag}(P_jA_jQ_j)]^{-1} \\ \cdot \operatorname{diag}(P_jE_jQ_j)\operatorname{diag}(Q_j^{-1})(d \otimes I_n).\end{aligned}$$

Utilizing that $P_jE_j$ has already a vanishing second block row, we see that

$$\begin{aligned}&\operatorname{diag}(P_jE_j)(\mathcal{A}^{-1} \otimes I_n)\operatorname{diag}(Q_j) - h\operatorname{diag}(P_jA_jQ_j) \\ &= \left[\begin{array}{c|c|c} v_{1,1}P_1E_1Q_1 - hP_1A_1Q_1 & \cdots & v_{1,s}P_1E_1Q_s \\ \hline \vdots & \ddots & \vdots \\ \hline v_{s,1}P_sE_sQ_1 & \cdots & v_{s,s}P_sE_sQ_s - hP_sA_sQ_s \end{array}\right] \\ &= \left[\begin{array}{cc|c|cc} v_{1,1}I_d - hC_1 & 0 & \cdots & v_{1,1}I_d + \mathcal{O}(h) & \mathcal{O}(h) \\ 0 & -hI_a & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline v_{s,1}I_d + \mathcal{O}(h) & \mathcal{O}(h) & \cdots & v_{s,s}I_d - hC_s & 0 \\ 0 & 0 & \cdots & 0 & -hI_a \end{array}\right].\end{aligned} \tag{6.12}$$

ETNA
Kent State University
etna@mcs.kent.edu



The inverse of this matrix must be applied to

$$\operatorname{diag}(P_j E_j Q_j) = \operatorname{diag}(J), \quad J = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}.$$

Because of its zero block rows, in (6.12) we can replace the entries consisting only of $\mathcal{O}(h)$ by zero and the entries $-hI_a$ by $v_{j,l}I_a$ without altering the resulting $W$. Hence,

$$\begin{aligned} W &= (e_s^T \otimes I_n) \operatorname{diag}(Q_j)((\mathcal{A}^{-1} \otimes I_n) + \mathcal{O}(h))^{-1} \operatorname{diag}(J) \operatorname{diag}(Q_j^{-1})(d \otimes I_n) \\ &= Q_s(e_s^T \otimes I_n)((\mathcal{A} \otimes I_n) + \mathcal{O}(h)) \operatorname{diag}(J) \operatorname{diag}(Q_j^{-1})(d \otimes I_n). \end{aligned}$$

Observing, furthermore, that

$$\begin{aligned} &\operatorname{diag}(J) \operatorname{diag}(Q_j^{-1})(d \otimes I_n) = \operatorname{diag}(J) \operatorname{diag}(Q_j^{-1}) \operatorname{col}(d_j Q_j Q_0^{-1} + \mathcal{O}(h)) \\ &\quad = \operatorname{diag}(J) \operatorname{diag}(d_j I_n) \operatorname{col}(Q_0^{-1} + \mathcal{O}(h)) = (d \otimes I_n) \operatorname{diag}(J) \operatorname{col}(I_n + \mathcal{O}(h)) Q_0^{-1}, \end{aligned}$$

with $Q_0$ belonging to the transformation of $(E(t_i), A(t_i))$ to Weierstraß canonical form and using that $e_s^T \mathcal{A} d = e_s^T \mathcal{A}\mathcal{A}^{-1} e = 1$, we finally arrive at

$$\begin{aligned} W &= Q_s(e_s^T \otimes I_n)((\mathcal{A} \otimes I_n) + \mathcal{O}(h))(d \otimes I_n) \operatorname{diag}(J) \operatorname{col}(I_n + \mathcal{O}(h)) Q_0^{-1} \\ &= Q_s((e_s^T \mathcal{A} d \otimes I_n) + \mathcal{O}(h)) \operatorname{diag}(J) \operatorname{col}(I_n + \mathcal{O}(h)) Q_0^{-1} \\ &= Q_s(I_n + \mathcal{O}(h)) \operatorname{diag}(J) \operatorname{col}(I_n + \mathcal{O}(h)) Q_0^{-1}. \end{aligned}$$

Comparing with (6.6) we have stability with

$$\mathfrak{R}_i = Q_0^{-1}, \quad \mathfrak{R}_{i+1} = Q_s^{-1}.$$

Together with the known consistency, we get convergence of any transformation method that is based on stiffly accurate Runge-Kutta methods, in particular of the spin-stabilized stiffly accurate Runge-Kutta methods, i.e., we have proved the following convergence result.

THEOREM 6.2. *A spin-stabilized stiffly accurate Runge-Kutta method based on a stiffly accurate Runge-Kutta method of order p with invertible $\mathcal{A}$ as $\tilde{\mathfrak{F}}$ together with the transformation (6.1) is convergent of order p.*

In order to study the stability properties of a spin-stabilized stiffly accurate Runge-Kutta method concerning its long-time behavior, we apply it to the test equation (4.3). Let $(E, A)$ denote the coefficients of the test equation, let $P, Q$ denote matrix functions that transform $(E, A)$ to the canonical form of (4.5), and let $\tilde{Q}$ be the stabilizing transformation according to (6.1).

Setting $x = \tilde{Q}\tilde{x}$, we have to integrate the DAE

$$E(t)\tilde{Q}(t)\dot{\tilde{x}} = (A(t)\tilde{Q}(t) - E(t)\dot{\tilde{Q}})\tilde{x}.$$

Using, furthermore, the quantities $\hat{t}_j = t_i + \gamma_j h$ and $\tilde{Q}_j = \tilde{Q}(\hat{t}_j)$, the spin-stabilized Runge-Kutta method has the form

$$\text{(6.13)} \quad \begin{aligned} &\text{(a)} && \tilde{Q}(t_{i+1})^{-1} x_{i+1} = \tilde{Q}(t_i)^{-1} x_i + h \textstyle\sum_{j=1}^s \beta_j \dot{\tilde{X}}_j, \\ &\text{(b)} && \tilde{X}_j = \tilde{Q}(t_i)^{-1} x_i + h \textstyle\sum_{l=1}^s \alpha_{j,l} \dot{\tilde{X}}_l, \quad j = 1, \ldots, s, \\ &\text{(c)} && E_j \tilde{Q}_j \dot{\tilde{X}}_j = (A_j \tilde{Q}_j - E_j \dot{\tilde{Q}}) \tilde{X}_j, \quad j = 1, \ldots, s. \end{aligned}$$

Due to (6.13c) and the special form of the test equation, the scaled stage values $\tilde{Q}_j \tilde{X}_j$ are consistent at time $\hat{t}_j$. Writing down (6.11) for the present situation, we obtain that

$$\begin{aligned} &\Big[\operatorname{diag}(E_j \tilde{Q}_j)(\mathcal{A}^{-1} \otimes I_n) - h \operatorname{diag}(A_j \tilde{Q}_j - E_j \dot{\tilde{Q}})\Big] \operatorname{diag}(\tilde{Q}_j^{-1}) \operatorname{col}(\tilde{Q}_j \tilde{X}_j) \\ &\qquad = \Big[\operatorname{diag}(E_j \tilde{Q}_j)(\mathcal{A}^{-1} \otimes I_n)(e \otimes I_n)\Big] \tilde{Q}(t_i)^{-1} x_i \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu



or

$$\Big[\mathrm{diag}(E_j\tilde{Q}_j)(\mathcal{A}^{-1}\otimes I_n)\,\mathrm{diag}(\tilde{Q}_j^{-1}) - h\,\mathrm{diag}(A_j - E_j\dot{\tilde{Q}}\tilde{Q}_j^{-1})\Big]\,\mathrm{col}(\tilde{Q}_j\tilde{X}_j)$$
$$= \mathrm{col}(d_jE_j\tilde{Q}_j\tilde{Q}(t_i)^{-1}x_i).$$

The diagonal entries of the leading block matrix are given by $v_{j,j}E_j - h(A_j - E_j\dot{\tilde{Q}}\tilde{Q}_j^{-1})$, whereas the off-diagonal entries have the form $v_{j,l}E_j\tilde{Q}_j\tilde{Q}_l^{-1}$. The third term, which has to be considered is $E_j\tilde{Q}_j\tilde{Q}(t_i)^{-1}$ in the right hand side. Using the consistency of the numerical approximations $x_i$ according to

$$x_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \end{bmatrix} = \begin{bmatrix} 1+\omega t_i \\ 1 \end{bmatrix} x_{2,i}, \tag{6.14}$$

and similarly that of the stage values $\tilde{Q}_j\tilde{X}_j$, Lemma 6.1 yields that

$$\lim_{t_i\to\infty} E_j\tilde{Q}_j\tilde{Q}(t_i)^{-1}\begin{bmatrix} 1+\omega t_i \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \mathcal{O}(h^2),$$
$$\lim_{t_i\to\infty} E_j\tilde{Q}_j\tilde{Q}_l^{-1}\begin{bmatrix} 1+\omega \hat{t}_l \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \mathcal{O}(h^2),$$
$$\lim_{t_i\to\infty}\Big(v_{j,j}E_j - h(A_j - E_j\dot{\tilde{Q}}\tilde{Q}_j^{-1})\Big)\begin{bmatrix} 1+\omega \hat{t}_j \\ 1 \end{bmatrix} = \begin{bmatrix} v_{j,j}-h\lambda \\ 0 \end{bmatrix} + \mathcal{O}(h^2).$$

Since $x_{i+1}$ coincides with $\tilde{Q}_s\tilde{X}_s$, we altogether have derived the representation

$$x_{i+1,2} = (e_s^T(\mathcal{A}^{-1} - h\lambda I)^{-1}d + \mathcal{O}(h^2))x_{i,2}$$

in the limit $t_i \to \infty$. Comparing with Section 5.5, we immediately see that

$$e_s^T(\mathcal{A}^{-1} - h\lambda I)^{-1}d = R(z,0),$$

where $R(z,w)$ is the stability function derived there. Moreover, $R(z,0)$ is nothing else than the classical stability function for ODEs. Hence, under the assumption that the constant in the remainder term (which depends on the choice of $\tilde{Q}$) is small, we see that in the limit $t_i \to \infty$ the influence of the parameter $\omega$ on the stability of the discretization has been removed.

**6.3. Spin-stabilized BDF methods.** In this section we discuss the use of spin-stabilization within BDF methods. As in the previous section, we consider a strangeness-free DAE (2.7).

A BDF method for the integration of (2.7) has the form

$$\frac{1}{h}E_i\sum_{l=0}^{k}\alpha_{k-l}x_{i-l} = A_ix_i + f_i, \tag{6.15}$$

with

$$E_i = E(t_i), \quad A_i = A(t_i), \quad f_i = f(t_i).$$

We assume that the method is normalized to have the leading coefficient $\alpha_k = 1$. The relation (6.15) then yields

$$x_i = (E_i - h\beta_kA_i)^{-1}\Big[h\beta_kf_i - E_i\sum_{l=1}^{k}\alpha_{k-l}x_{i-l}\Big].$$

In view of (6.6), we must consider the matrix

$$W = \begin{bmatrix} -\alpha_{k-1}D_i & \cdots & -\alpha_1 D_i & -\alpha_0 D_i \\ I_n & & & \\ & \ddots & & \\ & & I_n & \end{bmatrix}$$

with

$$D_i = (E_i - h\beta_k A_i)^{-1} E_i.$$

Let $P_i, Q_i$ transform $(E_i, A_i)$ to Weierstraß canonical form. Then $D_i$ has the form

$$\begin{aligned} D_i &= Q_i(P_iE_iQ_i - k\beta_k P_iA_iQ_i)^{-1} P_iE_iQ_iQ_i^{-1} \\ &= Q_i \begin{bmatrix} I_d - h\beta_k C_i & 0 \\ 0 & -hI_a \end{bmatrix} \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} Q_i^{-1} \\ &= Q_i \begin{bmatrix} I_d + \mathcal{O}(h) & 0 \\ 0 & 0 \end{bmatrix} Q_i^{-1}, \end{aligned}$$

implying that

$$\begin{aligned} &\operatorname{diag}(Q_i^{-1}, \ldots, Q_i^{-1}) W \operatorname{diag}(Q_i, \ldots, Q_i) \\ &= \left[\begin{array}{cc|cc|cc|cc} -\alpha_{k-1}I_d & 0 & \cdots & \cdots & -\alpha_1 I_d & 0 & -\alpha_0 I_d & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & 0 \\ \hline I_d & 0 & & & & & & \\ 0 & I_a & & & & & & \\ \hline & & \ddots & & & & & \\ & & & \ddots & & & & \\ \hline & & & & I_d & & & \\ & & & & & I_a & & \end{array}\right] + \mathcal{O}(h). \end{aligned}$$

Hence, if the BDF method is D-stable, see [13], then there is a vector norm such that the latter matrix is bounded by $1 + hK$ in the corresponding matrix norm with a suitable constant $K$. Comparing with (6.6) and observing (6.9) we have stability with

$$\mathfrak{R}_i = \operatorname{diag}(Q_{i+k}, \ldots, Q_{i+k}).$$

Thus, we have the following theorem.

THEOREM 6.3. *A spin-stabilized BDF method based on a BDF method of order $k$, $1 \leq k \leq 6$ together with the transformation (6.1) is convergent of order $k$.*

In order to study the stability properties of a spin-stabilized BDF method concerning its long-time behavior, we apply it to the test equation (4.3). Let $(E, A)$ denote the coefficients of the test equation, let $P, Q$ denote matrix functions that transform $(E, A)$ to the canonical form of (4.5), and let $\tilde{Q}$ be the stabilizing transformation according to (6.1).

Setting $x = \tilde{Q}\tilde{x}$ and $x_{i-l} = \tilde{Q}_{i-l}\tilde{x}_{i-l}$ with $\tilde{Q}_{i-l} = \tilde{Q}(t_{i-l})$, we have to integrate the DAE

$$E(t)\tilde{Q}(t)\dot{\tilde{x}} = (A(t)\tilde{Q}(t) - E(t)\dot{\tilde{Q}})\tilde{x}.$$

Hence, the spin-stabilized BDF method has the form

$$\frac{1}{h} E_i \tilde{Q}_i \sum_{l=0}^{k} \alpha_{k-l} \tilde{Q}_{i-l}^{-1} x_{i-l} = (A_i\tilde{Q}_i - E_i\dot{\tilde{Q}})\tilde{Q}_i^{-1} x_i,$$

ETNA
Kent State University
etna@mcs.kent.edu



leading to the difference equation

$$\left[E_i - hA_i + hE_i\dot{\tilde{Q}}\tilde{Q}_i^{-1}\right]x_i + E_i\tilde{Q}_i\sum_{l=1}^{k}\alpha_{k-l}\tilde{Q}_{i-l}^{-1}x_{i-l} = 0. \tag{6.16}$$

Since the BDF methods yield consistent numerical approximations, we can again utilize (6.14). Lemma 6.1 yields that

$$\lim_{t_i\to\infty} E_i\tilde{Q}_i\tilde{Q}_{i-l}^{-1}\begin{bmatrix}1+\omega t_{i-l}\\ 1\end{bmatrix} = \begin{bmatrix}1\\ 0\end{bmatrix} + \mathcal{O}(h^2),$$
$$\lim_{t_i\to\infty} E_i\dot{\tilde{Q}}_j\tilde{Q}_i^{-1}\begin{bmatrix}1+\omega t_i\\ 1\end{bmatrix} = \begin{bmatrix}\omega\\ 0\end{bmatrix} + \mathcal{O}(h).$$

Hence, in the limit $t_i \to \infty$, this difference equation reads

$$\left[(1+\omega t_i) - \omega t_i - h\lambda(1+\omega t_i) - h\omega(1-\lambda t_i) + h\omega + \mathcal{O}(h^2)\right]x_{2,i} + \sum_{l=1}^{k}\alpha_{k-l}(1+\mathcal{O}(h^2))x_{2,k-l} = 0$$

which reduces to

$$(1 - z + \mathcal{O}(h^2))x_{2,i} + \sum_{l=1}^{k}\alpha_{k-l}(1+\mathcal{O}(h^2))x_{2,k-l} = 0.$$

But this is nothing else than a perturbation of the difference equation which we obtain when we apply the BDF method to the standard ODE test equation. Thus, provided the constants involved in the remainder terms (which depend on the choice of $\tilde{Q}$) are small, we can expect the same stability properties of the spin-stabilized BDF methods as in the ODE case.

**6.4. A numerical experiment.** We have implemented the standard implicit Euler method and its spin-stabilized version choosing $\tilde{Q}$ as in (6.3). Using a constant stepsize, we applied both methods to the problem of Example 1.1 for a range of parameter values $(\delta, \eta)$ and checked numerically the stability of the numerical solutions. The results can be seen in Figure 6.1 for the standard implicit Euler method and in Figure 6.2 for the spin-stabilized implicit Euler method. Both figures were obtained with a stepsize of $h = 0.1$ and cover the range $(\delta, \eta) \in [-3, 3]^2$. The shading is based on a numerical estimate of the limit factor between the norms of $x_i$ and $x_{i+1}$.

In Figure 6.1, one can recognize the stability restriction $|1 + h\delta| < |1 + h\eta|$, whereas Figure 6.2 shows that the spin-stabilized implicit Euler method is stable in the region $\delta < \eta$, where the actual solution is stable. We also see the superstability of the implicit Euler method, i. e., the stability of the numerical solution of the implicit Euler method in regions where the actual solution is not stable.

Applying one of the other methods gives similar results. In particular, as we have shown in Section 6, the stability region of the spin-stabilized version is approximately given by the condition

$$|R(\delta - \eta, 0)| \le 1,$$

where $R$ is the corresponding DAE stability function, and can be seen for the implicit Euler method in Figure 6.2.

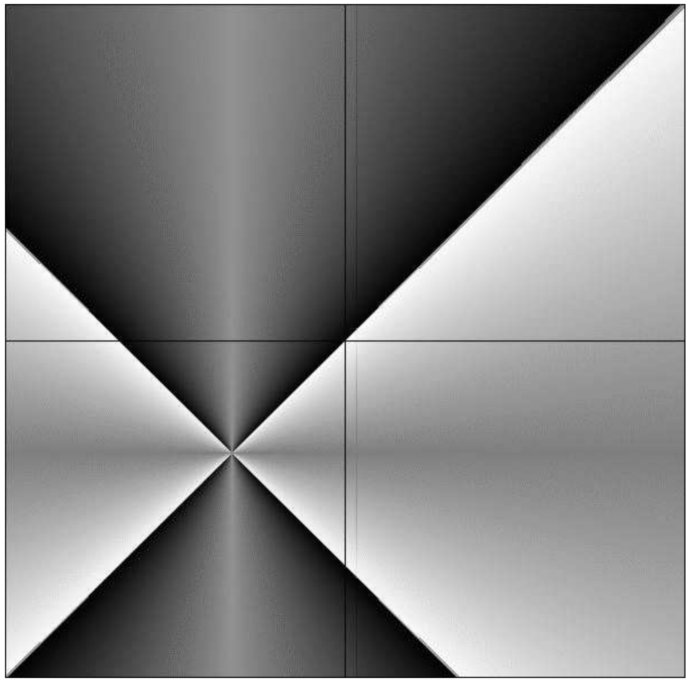

FIG. 6.1. *Numerical stability region for the standard implicit Euler method for* $(\delta, \eta) \in [-3, 3]^2$

**7. Conclusion.** We have analyzed the stability properties of general differential-algebraic equations of arbitrary index and related them to those of the corresponding inherent ordinary differential equation.

We have presented a new test equation for differential-algebraic equations that takes into account that the kernel of $F_{\dot{x}}$ in a strangeness-free formulation of a given DAE may spin along the solution. We have analyzed the stability of classical numerical integration methods for differential-algebraic equations on the basis of this new test equation and introduced the concept of DAE stability functions.

In order to deal with rapidly spinning kernels we have derived a new stabilization method that can be used together with all classical integrators. We have shown that this approach which in every integration step first transforms the equation, then carries out the integration step by the given method, and finally transforms back, leads to the same convergence results for stiffly accurate Runge-Kutta and BDF methods as for the unstabilized methods, while getting more appropriate regions of numerical stability. Moreover, we have demonstrated our new approach with a numerical example.

**Acknowledgment.** We thank two anonymous referees for several comments and suggestions that improved the readability of the paper.

REFERENCES

[1] U. M. ASCHER AND L. R. PETZOLD, *Stability of computation for constrained dynamical systems*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 95–120.

**ETNA**
**Kent State University**
**etna@mcs.kent.edu**



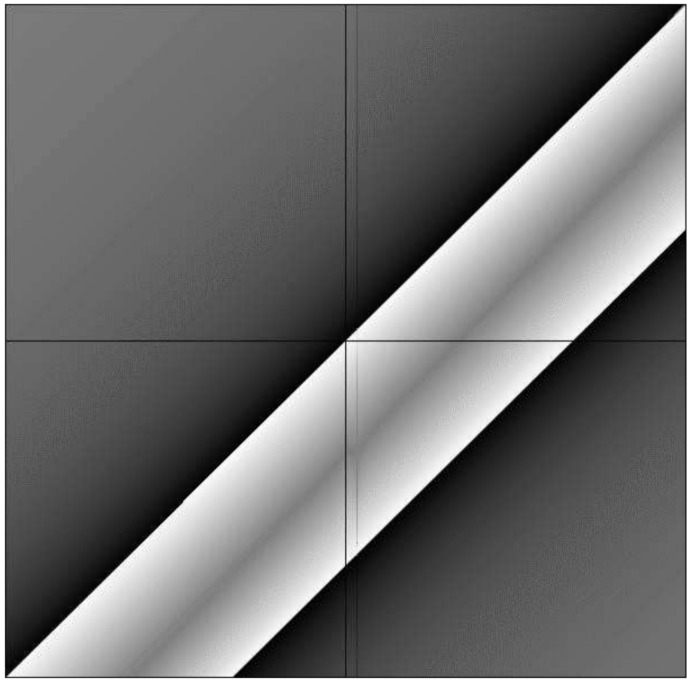

FIG. 6.2. *Numerical stability region for the spin-stabilized implicit Euler method for* $(\delta, \eta) \in [-3, 3]^2$

[2] K. E. BRENAN, S. L. CAMPBELL, AND L. R. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, Second ed., SIAM Publications, Philadelphia, PA, 1996.
[3] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. K. NICHOLS, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.
[4] S. L. CAMPBELL, *Comment on controlling generalized state-space (descriptor) systems*, Internat. J. Control, 46 (1987), pp. 2229–2230.
[5] S. L. CAMPBELL AND C. W. GEAR, *The index of general nonlinear DAEs*, Numer. Math., 72 (1995), pp. 173–196.
[6] L. DIECI, R. D. RUSSELL, AND E. S. VAN VLECK, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal., 31 (1994), pp. 261–281.
[7] M. DIEHL, D. B. LEINEWEBER, A. SCHÄFER, H. G. BOCK, AND J. P. SCHLÖDER, *Optimization of multiple-fraction batch distillation with recycled waste cuts*, AIChE Journal, 48 (2002), pp. 2869–2874.
[8] M. DIEHL, I. USLU, R. FINDEISEN, S. SCHWARZKOPF, F. ALLGÖWER, H. G. BOCK, T. BÜRNER, E. D. GILLES, A. KIENLE, J. P. SCHLÖDER, AND E. STEIN, *Real-time optimization for large scale processes: Nonlinear model predictive control of a high purity distillation column*, in Online Optimization of Large Scale Systems: State of the Art, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer, 2001, pp. 363–384.
[9] E. EICH-SOELLNER AND C. FÜHRER, *Numerical Methods in Multibody Systems*, Teubner Verlag, Stuttgart, Germany, 1998.
[10] E. GRIEPENTROG AND R. MÄRZ, *Differential-Algebraic Equations and their Numerical Treatment*, Teubner Verlag, Leipzig, Germany, 1986.
[11] M. GÜNTHER AND U. FELDMANN, *CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations*, Surv. Math. Ind., 8 (1999), pp. 97–129.
[12] ———, *CAD-based electric-circuit modeling in industry II. Impact of circuit configurations and parameters*, Surv. Math. Ind., 8 (1999), pp. 131–157.
[13] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Second ed., Springer, Berlin, Germany,1996.
[14] M. HANKE, E. I. MACANA, AND R. MÄRZ, *On asymptotics in case of linear index-2 differential-algebraic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1326–1346.

[15] I. HIGUERAS, R. MÄRZ, AND C. TISCHENDORF, *Stability preserving integration of index-1 DAEs*, Appl. Numer. Math., 45 (2003), pp. 175–200.

[16] ———, *Stability preserving integration of index-2 DAEs*, Appl. Numer. Math., 45 (2003), pp. 201–229.

[17] D. HINRICHSEN AND A. J. PRITCHARD, *Mathematical Systems Theory I. Modelling, State Space Analysis, Stability and Robustness*, Springer, New York, NY, 2005.

[18] P. KUNKEL AND V. MEHRMANN, *Smooth factorizations of matrix valued functions and their derivatives*, Numer. Math., 60 (1991), pp. 115–132.

[19] ———, *Regular solutions of nonlinear differential-algebraic equations and their numerical determination*, Numer. Math., 79 (1998), pp. 581–600.

[20] ———, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.

[21] ———, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006.

[22] P. KUNKEL, V. MEHRMANN, M. SCHMIDT, I. SEUFER, AND A. STEINBRECHER, *Weak formulations of linear differential-algebraic systems*, Technical Report 2006-16, Institut für Mathematik, TU Berlin, Berlin, Germany, 2006. url: http://www.math.tu-berlin.de/preprints/.

[23] P. KUNKEL, V. MEHRMANN, AND R. STÖVER, *Symmetric collocation for unstructured nonlinear differential-algebraic equations of arbitrary index*, Numer. Math., 98 (2004), pp. 277–304.

[24] P. KUNKEL AND R. STÖVER, *Symmetric collocation methods for linear differential-algebraic boundary value problems*, Numer. Math., 91 (2002), pp. 475–501.

[25] R. MÄRZ, *Criteria for the trivial solution of differential algebraic equations with small nonlinearities to be asymptotically stable*, J. Math. Anal. Appl., 225 (1998), pp. 587–607.

[26] R. MÄRZ, *Solvability of linear differential algebraic equations with properly stated leading terms*, Results Math., 45 (2004), pp. 88–105.

[27] R. MÄRZ AND A. R. RODRIGUEZ-SANTIESTEBAN, *Analyzing the stability behaviour of solutions and their approximations in case of index-2 differential-algebraic systems*, Math. Comp., 71 (2001), pp. 605–632.

[28] M. OTTER, H. ELMQVIST, AND S. E. MATTSON, *Multi-domain modeling with modelica*, Chap. 36, CRC Handbook of Dynamic System Modeling, P. Fishwick, ed., Chapman & Hall/CRC Press, Boca Raton, Fl, 2007, pp. 36.1–36.27.

[29] P. J. RABIER AND W. C. RHEINBOLDT, *Classical and generalized solutions of time-dependent linear differential-algebraic equations*, Linear Algebra Appl., 245 (1996), pp. 259–293.

[30] ———, *Theoretical and Numerical Analysis of Differential-Algebraic Equations*, vol. VIII of Handbook of Numerical Analysis, P. Ciarlet and J. L. Lions, eds., Elsevier Publications, Amsterdam, The Netherlands, 2002.

[31] W. C. RHEINBOLDT, *Differential-algebraic systems as differential equations on manifolds*, Math. Comp., 43 (1984), pp. 473–482.

[32] ———, *On the computation of multi-dimensional solution manifolds of parameterized equations*, Numer. Math., 53 (1988), pp. 165–181.

[33] R. RIAZA, *Stability issues in regular and non-critical singular DAEs*, Acta Appl. Math., 73 (2002), pp. 243–261.

[34] R. RIAZA AND C. TISCHENDORF, *Topological analysis of qualitative features in electrical circuit theory*, Tech. Report 04-18, Institut für Mathematik, Humboldt Universität zu Berlin, Berlin, Germany, 2004.

[35] A. M. STUART AND A. R. HUMPHRIES, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

[36] T. STYKEL, *Analysis and Numerical Solution of Generalized Lyapunov Equations*, Dissertation, Institut für Mathematik, TU Berlin, Berlin, Germany, 2002.

[37] ———, *On criteria for asymptotic stability of differential-algebraic equations*, ZAMM Z. Angew. Math. Mech., 92 (2002), pp. 147–158.

[38] ———, *Stability and inertia theorems for generalized lyapunov equations*, Linear Algebra Appl., 355 (2002), pp. 297–314.

[39] C. TISCHENDORF, *On stability of solutions of autonomous index-1 tractable and quasilinear index-2 tractable DAE's*, Circuits Systems Signal Process., 13 (1994), pp. 139–154.

[40] K. WRIGHT, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math., 3 (1992), pp. 283–295.