ON SYMMETRIC SCHEMES AND DIFFERENTIAL-ALGEBRAIC EQUATIONS

Uri Ascher

Technical Report 88-12

June 1988

Abstract

An example is given which demonstrates a potential risk in using symmetric difference schemes for initial value differential-algebraic equations (DAEs) or for very stiff ODEs. The basic difficulty is that the stability of the scheme is controlled by the stability of an auxiliary (ghost) ODE problem which is not necessarily stable even when the given problem is.

The stability of symmetric schemes is better understood in the context of boundary value problems. In this context, such schemes are more naturally applied as well. For initial value problems, better alternatives may exist. A computational algorithm is proposed for boundary value index-1 DAEs.

Subject classification: AMS(MOS): 65L10.

Keywords: Differential algebraic equations, symmetric schemes, initial value problems, boundary value problems, index.

1. Introduction

The possibility of using a symmetric difference scheme, like the midpoint scheme (collocation at one Gauss point), for solving differential-algebraic equations (DAEs), has been recently considered in the literature. This appears to be particularly attractive for fully implicit index 1 boundary value problems (BVPs). However, as we demonstrate in §2, a careless use of such schemes, even for initial value problems (IVPs), can be dangerous.

We consider the linear DAE

$$E(t)\mathbf{x}' = A(t)\mathbf{x} + \mathbf{q}(t), \quad 0 < t < 1, \tag{1}$$

where $E(t)$ is a singular matrix with constant rank. For a nonlinear DAE

$$\boldsymbol{\phi}(t,\mathbf{x},\mathbf{x}') = 0, \quad 0 < t < 1,$$

we have in mind a quasilinearization method (see, e.g., [AMR, §2.3.4]), which yields at each iteration a linear DAE like (1) with

$$E(t) = \frac{\partial \boldsymbol{\phi}(t,\hat{\mathbf{x}}(t),\hat{\mathbf{x}}'(t))}{\partial \mathbf{x}'},$$

$\hat{\mathbf{x}}(t)$ being the current iterate.

Following the example, we will analyze the situation in §§ 3, 4. Conclusions are offered in §5. These suggest that for IVPs, better alternatives than using symmetric schemes may exist. But for BVPs, pursuing symmetric schemes is more worthwhile, at least for DAEs with index 1. We propose an algorithm for this class of problems.

2. Example

Consider the IVP (1) with

$$E = \begin{pmatrix} 1 & -t \\ 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} -1 & 1+t \\ \beta & -1-\beta t \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} 0 \\ \sin t \end{pmatrix}, \tag{2}$$

β a parameter, and $x_1(0) = 1$.

This problem is an extension of the example of Petzold [Pe] who considered it for the parameter value β=0. The transformation of variables

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 - t x_2 \end{pmatrix} \equiv T^{-1}\mathbf{x}, \quad T^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -t \end{pmatrix} \tag{3}$$

gives the equations

$$0 = -y + \beta z + \sin t \tag{4a}$$

$$z' = -z \tag{4b}$$

with $z(0)$=1. The unique solution is therefore

$$z = e^{-t}, \quad y = \beta e^{-t} + \sin t, \quad x_2 = y, \quad x_1 = t \sin t + (1+\beta t) e^{-t}. \tag{5}$$

To apply the midpoint scheme

$$E(t_{i+1/2}) \frac{\mathbf{x}_{i+1} - \mathbf{x}_i}{h_i} = A(t_{i+1/2}) \frac{\mathbf{x}_i + \mathbf{x}_{i+1}}{2} + \mathbf{q}(t_{i+1/2}), \quad 1 \le i \le N, \tag{6}$$

on a mesh

$$\pi : 0 = t_1 < t_2 < \cdots < t_N < t_{N+1} = 1 \tag{7}$$

$$h_i := t_{i+1} - t_i, \quad h := \max_{1 \le i \le N} h_i \quad t_{i+1/2} := t_i + \frac{1}{2} h_i,$$

we need another side value (in addition to the given one on x_1), which we take as the exact ini-

tial value

$$x_2(0) = \beta.$$

In Tables 1 and 2 we list the errors and computed rates of convergence at t=1 when uniform step sizes $h = 1/N$ are used for various values of β. The error magnitudes in $x_j(1)$ are listed under err$j$, the convergence rates under rate$j$, j=1,2. For each value of β for which solutions are computed we list in a separate table results where the coarsest mesh is with $h = 0.2/|\beta|$ (h=0.2 for β=0 as well), and then refining by halving the step size a number of times. While all meshes used are uniform, there appears to be no reason to take nonuniform meshes here. The computations were performed on a SUN 3 running a UNIX f77 compiler in double precision.

For Tables 1(a-f) we use the midpoint scheme (6).

Table 1(a) β=0, $x_1(1)$=0.121+01, $x_2(1)$=0.841+00

h	err1	err2	rate1	rate2
0.200+00	0.212-02	0.422-02		
0.100+00	0.524-03	0.105-02	0.202+01	0.200+01
0.500-01	0.131-03	0.263-03	0.200+01	0.200+01
0.250-01	0.326-04	0.657-04	0.200+01	0.200+01
0.125-01	0.816-05	0.164-04	0.200+01	0.200+01

Table 1(b) β=1, $x_1(1)$=0.158+01, $x_2(1)$=0.121+01

h	err1	err2	rate1	rate2
0.200+00	0.115-02	0.380-02		
0.100+00	0.143-03	0.798-03	0.302+01	0.225+01
0.500-01	0.416-04	0.205-03	0.178+01	0.196+01
0.250-01	0.108-04	0.517-04	0.195+01	0.199+01
0.125-01	0.272-05	0.129-04	0.199+01	0.200+01

Table 1(c) β=10, $x_1(1)$=0.489+01, $x_2(1)$=0.452+01

h	err1	err2	rate1	rate2
0.200-01	0.202+03	0.202+03		
0.100-01	0.498+02	0.498+02	0.202+01	0.202+01
0.500-02	0.124+02	0.124+02	0.201+01	0.201+01
0.250-02	0.310+01	0.310+01	0.200+01	0.200+01
0.125-02	0.774+00	0.774+00	0.200+01	0.200+01

Table 1(d) β=50, $x_1(1)$=0.196+02, $x_2(1)$=0.192+02

h	err1	err2	rate1	rate2
0.400-02	0.594+20	0.594+20		
0.200-02	0.132+20	0.132+20	0.217+01	0.217+01
0.100-02	0.321+19	0.321+19	0.204+01	0.204+01
0.500-03	0.796+18	0.796+18	0.201+01	0.201+01
0.250-03	0.199+18	0.199+18	0.200+01	0.200+01

Table 1(e) β=100, $x_1(1)$=0.380+02, $x_2(1)$=0.376+02

h	err1	err2	rate1	rate2
0.200-02	0.368+42	0.368+42		
0.100-02	0.721+41	0.721+41	0.235+01	0.235+01
0.500-03	0.170+41	0.170+41	0.209+01	0.209+01
0.250-03	0.418+40	0.418+40	0.202+01	0.202+01
0.125-03	0.104+40	0.104+40	0.201+01	0.201+01
0.625-04	0.260+39	0.260+39	0.200+01	0.200+01
0.313-04	0.650+38	0.650+38	0.200+01	0.200+01
0.156-04	0.162+38	0.162+38	0.200+01	0.200+01
0.781-05	0.406+37	0.406+37	0.200+01	0.200+01
0.391-05	0.102+37	0.102+37	0.200+01	0.200+01

Table 1(f) β=–100, $x_1(1)$=–0.356+02, $x_2(1)$=–0.359+02

h	err1	err2	rate1	rate2
0.200-02	0.283-04	0.916-05		
0.100-02	0.703-05	0.233-05	0.201+01	0.198+01
0.500-03	0.175-05	0.584-06	0.200+01	0.199+01
0.250-03	0.438-06	0.146-06	0.200+01	0.200+01
0.125-03	0.110-06	0.358-07	0.199+01	0.203+01

Tables 1(a,b) indicate good results for small values of β. But Tables 1(c-e) show an exponential increase in the error size as β is increased. The computed rates of convergence, especially in Table 1(e), clearly show that the difficulty is *not* in just roundoff error accumulation. Table 1(f) shows that the difficulty is not an approximation question either, as it does not occur for $\beta<0$.

To emphasize the point and clarify the difficulty further, we list in Tables 2 results of comparable runs, made with the backward Euler scheme

$$E(t_{i+1})\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{h_i} = A(t_{i+1})\mathbf{x}_{i+1} + \mathbf{q}(t_{i+1}), \tag{8}$$

instead of the midpoint scheme.

Table 2(a) β=0, $x_1(1)$=0.121+01, $x_2(1)$=0.841+00

h	err1	err2	rate1	rate2
0.200+00	0.131+00	0.111-15		
0.100+00	0.671-01	0.111-15	0.961+00	
0.500-01	0.340-01	0.111-15	0.979+00	
0.250-01	0.171-01	0.222-15	0.989+00	
0.125-01	0.860-02	0.888-15	0.995+00	

Table 2(b) β=1, $x_1(1)$=0.158+01, $x_2(1)$=0.121+01

h	err1	err2	rate1	rate2
0.200+00	0.135+00	0.673-01		
0.100+00	0.659-01	0.329-01	0.103+01	0.103+01
0.500-01	0.326-01	0.163-01	0.101+01	0.101+01
0.250-01	0.162-01	0.811-02	0.101+01	0.101+01
0.125-01	0.809-02	0.405-02	0.100+01	0.100+01

Table 2(c) β=10, $x_1(1)$=0.489+01, $x_2(1)$=0.452+01

h	err1	err2	rate1	rate2
0.200-01	0.723+00	0.657+00		
0.100-01	0.345+00	0.313+00	0.107+01	0.107+01
0.500-02	0.168+00	0.153+00	0.103+01	0.103+01
0.250-02	0.831-01	0.756-01	0.102+01	0.102+01
0.125-02	0.413-01	0.376-01	0.101+01	0.101+01

Table 2(d) β=50, $x_1(1)$=0.196+02, $x_2(1)$=0.192+02

h	err1	err2	rate1	rate2
0.400-02	0.399+01	0.391+01		
0.200-02	0.190+01	0.186+01	0.107+01	0.107+01
0.100-02	0.926+00	0.907+00	0.104+01	0.104+01
0.500-03	0.457+00	0.448+00	0.102+01	0.102+01
0.250-03	0.227+00	0.223+00	0.101+01	0.101+01

Table 2(e) β=100, $x_1(1)$=0.380+02, $x_2(1)$=0.376+02

h	err1	err2	rate1	rate2
0.200-02	0.806+01	0.798+01		
0.100-02	0.383+01	0.379+01	0.107+01	0.107+01
0.500-03	0.187+01	0.185+01	0.104+01	0.104+01
0.250-03	0.923+00	0.914+00	0.102+01	0.102+01
0.125-03	0.459+00	0.454+00	0.101+01	0.101+01
0.625-04	0.229+00	0.226+00	0.100+01	0.100+01
0.313-04	0.114+00	0.113+00	0.100+01	0.100+01
0.156-04	0.570-01	0.564-01	0.100+01	0.100+01
0.781-05	0.285-01	0.282-01	0.100+01	0.100+01
0.391-05	0.142-01	0.141-01	0.100+01	0.100+01

Table 2(f) β=–100, $x_1(1)$=–0.356+02, $x_2(1)$=–0.359+02

h	err1	err2	rate1	rate2
0.200-02	0.673+01	0.679+01		
0.100-02	0.353+01	0.357+01	0.930+00	0.930+00
0.500-03	0.181+01	0.183+01	0.964+00	0.964+00
0.250-03	0.916+00	0.925+00	0.982+00	0.982+00
0.125-03	0.461+00	0.466+00	0.991+00	0.991+00

The results in Tables 2(a-f) demonstrate the first order convergence rate of the backward Euler scheme. (An exception is in Table 2(a), where $x_2(1)$ is reproduced because the backward Euler scheme reproduces the algebraic equations (4a) at mesh points, and when β=0 this determines y=x_2 exactly.) Thus, it is not surprising to see that the second order midpoint scheme performs much more accurately for β = 0,1,–100. However, for larger and positive values of β the backward Euler scheme obviously does not share whatever it is that is bothering the midpoint scheme, and so produces better results for the range of h listed.

The demonstrated difficulty is not restricted to DAEs: If we replace (4a) by the ODE

$$\epsilon y' = -y + \beta z + \sin t \tag{9a}$$

$0 < \epsilon \ll 1$, and then transform to $\mathbf{x}$, yielding a replacement of $E(t)$ in (2) by

$$E(t;\epsilon) = \begin{pmatrix} 1 & -t \\ 0 & \epsilon \end{pmatrix}, \tag{9b}$$

then a very stiff ODE is obtained. The IVP for this ODE exhibits a similar phenomenon to that depicted in Tables 1(a-f), so long as $\epsilon \ll h$. This, even though the scheme used is not only A-stable, but also D-stable [Ve] and algebraically stable [BuBu].

3. Analysis

It is important to identify the source of the difficulty demonstrated above. In recent discussions emphasis has been placed on roundoff error accumulation: März [Ma] has observed the merely marginal stability of symmetric schemes which allows for a linear roundoff error accumulation; Ascher & Weiss [AsWe1], [AsWe2], [AsWe3], [We], [As1] analyzed related stiff problems and computed many solutions which did not display any difficulty with roundoff error; and Burrage & Petzold [BuPe] recently also observed the same lack of pronounced roundoff error effect. Indeed, asymptotically the roundoff error accumulation here is similar in order to that obtained when discretizing directly with a uniform step size a 2nd order ODE, and the latter rarely (though not never) causes difficulties in practice. Roundoff error accumulation is *not* a cause for concern in the above example either, as the computed rates of convergence clearly indicate (i.e., it is dominated by the discretization error).

The computational difficulty in this example arises because the *stability constant* of the discretization method becomes exponentially large in β. This has been analyzed by Ascher [As2, §3.1], where it is shown that for a general DAE (1) of index 1 the discretization tends as $h \to 0$ to approximate an auxiliary or "ghost" differential problem. The stability constant of the discretization method therefore depends on the stability (or conditioning - see [AMR, §3.2]) constant of this ghost problem, which is not necessarily of moderate size.

In detail, given the DAE (1) and assuming that $E(t)$ can be written as

$$E = S\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} T^{-1}, \tag{10}$$

where for each t, $S(t)$ and $T(t)$ are smooth and nonsingular, consider the *semi-explicit form*

$$0 = U^{11}(t)\mathbf{y} + U^{12}(t)\mathbf{z} + \mathbf{g}^1(t) \tag{11a}$$

$$\mathbf{z}' = U^{21}(t)\mathbf{y} + U^{22}(t)\mathbf{z} + \mathbf{g}^2(t) \tag{11b}$$

obtained from (1) through the transformation of variables

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \hat{\mathbf{x}} = T^{-1}\mathbf{x}, \tag{12}$$

with

$$\begin{pmatrix} U^{11} & U^{12} \\ U^{21} & U^{22} \end{pmatrix} = U = S^{-1}AT - \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} T^{-1}T', \qquad 0 \le t \le 1. \tag{13}$$

In (11a) there are n_y equations and in (11b) there are n_z equations, $n_y+n_z=n$. For the example in §2, a semi-explicit form is (4), with $n=2$, $n_y=n_z=1$,

$$S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad U = \begin{pmatrix} -1 & \beta \\ 0 & -1 \end{pmatrix} \qquad T = \begin{pmatrix} t & 1 \\ 1 & 0 \end{pmatrix}. \tag{14}$$

The general linear DAE (1) has (global) index 1 if $U^{11}(t)$ is nonsingular for all t, as we shall assume. Then

$$\mathbf{y}(t) = -[U^{11}(t)]^{-1}[U^{12}(t)\mathbf{z}(t) + \mathbf{g}^1(t)], \quad 0 \le t \le 1. \tag{15}$$

An IVP for (11), or (1), is therefore completely specified by specifying $\mathbf{z}(0)$.

First, consider a midpoint scheme for (11), with $\mathbf{y}_\pi$, $\mathbf{z}_\pi$ denoting the approximation for $\mathbf{y}$, $\mathbf{z}$, respectively. The correct scheme for this simple case is obtained by recognizing that $\mathbf{y}$ is less smooth than $\mathbf{z}$, being defined similarly to $\mathbf{z}'$. Thus we let $\mathbf{y}_\pi$ be a piecewise constant function

$$\mathbf{y}_\pi(t) = \mathbf{y}_{i+1/2}, \quad t_i \le t < t_{i+1}$$

and $\mathbf{z}_\pi$ be a continuous piecewise linear function

$$\mathbf{z}_\pi(t) = \mathbf{z}_i + \frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i}(t-t_i), \quad t_i \le t \le t_{i+1}$$

(so $\mathbf{z}_\pi{}'$ is in the same space as $\mathbf{y}_\pi$), obtaining

$$0 = U^{11}(t_{i+1/2})\mathbf{y}_{i+1/2} + U^{12}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^1(t_{i+1/2}) \tag{16a}$$

$$\frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i} = U^{21}(t_{i+1/2})\mathbf{y}_{i+1/2} + U^{22}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^2(t_{i+1/2}). \tag{16b}$$

The usual collocation theory then applies with a slight twist, see [As2, §2].

Unfortunately, in (1) we cannot distinguish between $\mathbf{y}$ and $\mathbf{z}$ without an explicit transformation, so we next consider the midpoint scheme for (11a,b) in case that $\mathbf{y}_\pi$ is from the same approximation space as $\mathbf{z}_\pi$,

$$0 = U^{11}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{12}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^1(t_{i+1/2}) \tag{17a}$$

$$\frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i} = U^{21}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{22}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^2(t_{i+1/2}). \tag{17b}$$

In contrast to (16), we now must specify side conditions on $\mathbf{y}_\pi$, in addition to those which are imposed on $\mathbf{z}_\pi$. We assume for now that $\mathbf{y}_1$ is specified such that (15) holds at $t=0$.

An analysis for (17) was carried out in [We], [AsWe2], and we only mention essential points here. Thus, we can eliminate $\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2}$ from the first equation and substitute into the second, obtaining an ordinary midpoint scheme for $\mathbf{z}$. Stability and second order convergence for $\mathbf{z}_i$ then follow, as usual for an ODE. To obtain results for $\mathbf{y}_i$ as well, discussion reduces to the

initial value problem

$$\mathbf{y}_{i+1} = -\mathbf{y}_i + 2\mathbf{f}(t_{i+1/2}), \quad \mathbf{y}_1 \text{ given}, \tag{18a}$$

as an approximation to the problem

$$\mathbf{y} = \mathbf{f}(t). \tag{18b}$$

The solution of the recursion is

$$\mathbf{y}_{i+1} = (-1)^i \mathbf{y}_1 + 2\sum_{j=1}^{i} (-1)^{i-j} \mathbf{f}(t_{j+1/2}). \tag{18c}$$

In this we see the unfortunate properties of the scheme, namely, that no error is damped. Hence the error is not localized. There is also a linear growth of roundoff error which usually is only of theoretical concern. Still, if $\mathbf{y}_1 = \mathbf{f}(0)$ (corresponding to setting $\mathbf{y}(0)$ explicitly via (15) in terms of $\mathbf{z}(0)$) and $\mathbf{f}$ is smooth then the error is $O(h)$, and it is $O(h^2)$ if the mesh satisfies

$$h_{i+1} = h_i(1+O(h_i)) \quad \text{for all i odd} \quad \textit{or} \quad \text{for all i even}. \tag{19}$$

Now we may consider the general case, by observing how the discretization approximates the decoupling transformation. With the midpoint scheme (6), using (10) at $t_{i+1/2}$ and multiplying through by $S^{-1}(t_{i+1/2})$, we obtain for

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{z}_i \end{pmatrix} \equiv \hat{\mathbf{x}}_i := T^{-1}(t_i)\mathbf{x}_i$$

the system

$$0 = U^{11}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{12}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^1(t_{i+1/2}) + \tag{20a}$$

$$+ \frac{h_i}{4} P_y (UT^{-1}T')(t_{i+1/2})(\hat{\mathbf{x}}_{i+1}-\hat{\mathbf{x}}_i) + O(h_i^2)\hat{\mathbf{x}}_i + O(h_i^2)\hat{\mathbf{x}}_{i+1},$$

$$\frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i} = U^{21}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{22}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^2(t_{i+1/2}) +$$

$$+ O(h_i)\hat{\mathbf{x}}_i + O(h_i)\hat{\mathbf{x}}_{i+1} ,$$

where

$$P_y := \begin{pmatrix} I_{n_y} & 0 \end{pmatrix} \in \mathbb{R}^{n_y \times n}. \tag{20b}$$

Comparing this to (17) and considering the marginal stability implied from (18c), attention reduces to the IVP for

$$0 = \frac{\mathbf{y}_i + \mathbf{y}_{i+1}}{2} + \frac{h_i}{4}M(t_{i+1/2})(\mathbf{y}_{i+1} - \mathbf{y}_i) + \mathbf{f}_{i+1/2}, \quad 1 \le i \le N \tag{21}$$

where

$$M := [U^{11}]^{-1}P_y U T^{-1} T' P_y^T = [U^{11}]^{-1}(S^{-1}AT')^{11} \in \mathbb{R}^{n_y \times n_y} \tag{22}$$

(if $U^{12} = 0$ then simply $M = (T^{-1}T')^{11}$), and $\mathbf{f}_{i+1/2}$ is an inhomogeneity. For

$$\hat{\mathbf{y}}_i := (-1)^i \mathbf{y}_i \tag{23a}$$

(a trick due to Kreiss: see [As1] and references therein), (21) gives

$$0 = \frac{\hat{\mathbf{y}}_{i+1} - \hat{\mathbf{y}}_i}{h_i} + M(t_{i+1/2})\frac{\hat{\mathbf{y}}_{i+1} + \hat{\mathbf{y}}_i}{2} + (-1)^{i+1}2h_i^{-1}\mathbf{f}_{i+1/2}, \quad 1 \le i \le N. \tag{23b}$$

The homogeneous part of (23b) is just a midpoint discretization for the ODE

$$\hat{\mathbf{y}}' = -M(t)\hat{\mathbf{y}}. \tag{23c}$$

If the IVP for (23c) has a stability constant K then there is a constant $\tilde{K}$ of size comparable to K such that for h small enough,

$$\max_i |\mathbf{y}_i| \le \tilde{K}\,(|\mathbf{y}_1| + 2|\sum_{j=1}^{N}(-1)^j \mathbf{f}_{j+1/2}|\,). \tag{24}$$

Subsequently, the results for (17) may be retrieved here too: In $\mathbf{f}_{j+1/2}$ we have $O(h_j)$ terms in $\mathbf{z}$ which sum up to a bounded quantity, terms like $F(t_{j+1/2})\frac{\mathbf{z}_j+\mathbf{z}_{j+1}}{2}$ which also sum up to a bounded quantity because of the sign alternation and the smoothness of $F(t)$, and $O(h_j^2)$ terms in $\mathbf{y}$ which are handled by a contraction argument. For a k-stage Gauss collocation (the 1-stage scheme is just the midpoint one (6)), we obtain convergence with possible order reduction,

$$\max_i|\mathbf{y}_i - \mathbf{y}(t_i)| \leq \textit{const}\ h^{k+q} \tag{25}$$

where q=1 if k is odd and the mesh satisfies (19), q=0 otherwise. ***But the constant const in (25) depends on K, and not just on the stability constant of the original IVP for (1) (!)***

4. Example explained and other symmetric schemes.

For the Example in §2 we now consider the ghost IVP (23). A short calculation gives

$$M(t) \equiv -\beta.$$

Equation (23c) here is therefore the scalar, constant coefficient ODE

$$\hat{y}' = \beta\hat{y} \qquad 0<t<1$$

and the stability constant of the IVP for (23) is

$$K = max(1,e^{\beta}).$$

The ***discretization*** error with a uniform mesh is $\sim Kh^2$. If $\beta \leq 0$ then $K = 1$ and good results are obtained. But if e.g. $\beta = 100$ then we have to reckon with a stability constant of $\approx e^{100}$ for the numerical method. Note that the problem itself is well-conditioned (its stability constant grows linearly in $|\beta|$, and the underlying ODE is not even stiff), as can be seen from (4). The poor approximation effect (exponential in β) is caused by the symmetric discretization scheme.

As noted before, a similar phenomenon occurs for very stiff ODEs. Curiously, related effects have been considered more often in the BVP literature (see references in [As1]). In our DAE context, if (23c) is subject to boundary conditions then in general it may not have a solution at all, in which case (25) does not hold for any constant *const.* No such danger arises for the linear IVP, and this perhaps has caused unawareness to the stability question hitherto.

As mentioned above, collocation schemes at Gaussian points all have similar stability properties. Other symmetric schemes (or any other Runge-Kutta scheme with a damping factor which is not strictly less than 1) cannot be expected to do better in general. In fact, collocation at Gaussian points is in some sense the most stable among symmetric schemes for very stiff ODEs (see [AsBa]).

In particular, suppose we wish to extend the usual trapezoidal scheme for (1). If we consider the scheme

$$\frac{E(t_{i+1})\mathbf{x}_{i+1}-E(t_i)\mathbf{x}_i}{h_i} = \frac{A(t_i)\mathbf{x}_i+A(t_{i+1})\mathbf{x}_{i+1}}{2} + \frac{\mathbf{q}(t_{i+1})+\mathbf{q}(t_i)}{2}, \quad 1 \le i \le N, \qquad (26)$$

then it is clear that the expression on the left hand side approximates $(E(t)\mathbf{x}(t))'$ instead of $E(t)\mathbf{x}'(t)$, so the approximation is meaningless if $E'(t)\mathbf{x}(t)$ is not very small in magnitude. Writing (1) as

$$(E(t)\mathbf{x})' = (A(t)+E'(t))\mathbf{x} + \mathbf{q}(t), \quad 0 < t < 1,$$

and applying a discretization like (26) to this form, yields a correct trapezoidal scheme, and this can be generalized to higher order Lobatto schemes. However, terms like $E'(t_i)\mathbf{x}_i$ need to be further approximated if practical use is contemplated.

One (second order) possibility gives

$$\frac{1}{2}(E(t_{i+1})+E(t_i))\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{h_i} = \frac{A(t_i)\mathbf{x}_i+A(t_{i+1})\mathbf{x}_{i+1}}{2} + \frac{\mathbf{q}(t_{i+1})+\mathbf{q}(t_i)}{2}, \quad 1 \le i \le N. \qquad \textbf{(27a)}$$

Another idea could be a hybrid scheme between the trapezoidal and the midpoint schemes,

$$E(t_{i+1/2})\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{h_i} = \frac{A(t_i)\mathbf{x}_i+A(t_{i+1})\mathbf{x}_{i+1}}{2} + \frac{\mathbf{q}(t_{i+1})+\mathbf{q}(t_i)}{2}, \quad 1 \le i \le N. \qquad \textbf{(27b)}$$

These are symmetric schemes whose generalization to higher order (like the Gaussian collocation schemes for midpoint, or the Lobatto collocation schemes for trapezoidal) is less obvious. Yet, they may look attractive at a first glance, because as it turns out they work very well for the Example of §2. For instance, with β=100 and h=0.001, the errors are err 1=.310–05, err 2=.307–5 with a 2nd order convergence rate (cf. second rows of Tables 1(e) and 2(e)).

However, this improvement in the computed results is not general: For these schemes, too, the stability constant is controlled by a ghost IVP, which may or may not be stable even when the original IVP is. The ghost ODE is simply a different one than for the midpoint scheme. Instead of (22), (23c), it can be shown that we now get

$$\hat{\mathbf{y}}' = -\tilde{M}(t)\hat{\mathbf{y}}, \qquad \textbf{(28a)}$$

with

$$\tilde{M} := -(S\Gamma^{-1}AT)^{11}[U^{11}]^{-1}. \qquad \textbf{(28b)}$$

(Note that if $(S^{-1}AT)^{21} = 0$ then $\tilde{M} = (S^{-1}S\Gamma)^{11}$.) For the previous example we get $\tilde{M} = 0$ because $S\Gamma$=0, so no stability problems arise. But other examples can be constructed.

Another example

Consider (1) with

$$E(t) = \begin{pmatrix} 0 & \alpha t \\ 0 & 1 \end{pmatrix}, \quad A(t) = \begin{pmatrix} \alpha t+1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{29a}$$

We can choose

$$T = I, \ S = \begin{pmatrix} 1 & \alpha t \\ 0 & 1 \end{pmatrix}, \tag{29b}$$

(it is easy to show that the considerations given here are independent of the particular choice of S and T satisfying (10)) and this time the midpoint scheme performs well because $T'=0$, so $M=0$ regardless of α. On the other hand, a short calculation gives

$$\tilde{M} = \alpha. \tag{29c}$$

The underlying ODE is

$$x_2' = (\alpha t+1)x_2 + q_2, \tag{29d}$$

so for $\alpha \ll -1$ the given IVP is stable, but the ghost IVP for (28a), (29c) is not. Qualitatively similar results to those displayed in Table 1 are now obtained for any of the schemes (27).

□

5. Using symmetric schemes

In the context of IVPs, not only are symmetric schemes dangerous, but there are also good alternatives like backward differentiation formulae (BDF). Thus, there appears to be little incentive to use them. Still, if such use is contemplated then the previous discussion suggests

how to go about it: Finding $T(0)$ (e.g. using an RQ-factorization), we identify $\mathbf{z}(0)$ which should be specified as part of the given DAE problem, and using (15) we may then specify $\mathbf{y}(0)$ as well. (This is generally necessary for other one-step schemes, too.) Finding also $T'(0)$ (using numerical differentiation) and the eigenvalues of $M(0)$, we may hope for accurate approximations to be obtained by collocation at Gaussian points using reasonable step sizes *if* none of these eigenvalues has a large negative real part.

Symmetric schemes become more attractive in the context of BVPs, where a well-conditioned problem may have both fast decreasing and fast increasing modes, so BDF schemes become dangerous to use because they do not preserve the dichotomy - see, e.g. [AMR, Chs. 3,10].

The BVP context appears to be more suitable for symmetric schemes also for a different reason. Given that the stability of the scheme is controlled by a ghost problem we can *choose* boundary conditions on $\mathbf{y}$ such that the ghost problem becomes a well-conditioned BVP, provided that the ghost ODE has a dichotomy. For IVPs this means possibly solving a BVP instead, and this again is not competitive. But if the original problem is a BVP then we can provide an efficient, usually stable algorithm using a symmetric scheme for a fully implicit index 1 DAE. We concentrate on the midpoint scheme as an instance.

Thus, consider the DAE (1) subject to consistent BC

$$B_0\mathbf{x}(0) + B_1\mathbf{x}(1) = \beta. \tag{30}$$

Algorithm (linear BVPs for index 1 DAEs)

Step 1: Find T and (an approximation for) T' at the two interval ends.

Using T we may now isolate the solution components $\mathbf{y}$ and $\mathbf{z}$, and form U^{11}, U^{12}, $\mathbf{g}^1$ at each interval end.

Step 2: Using the relation (15), obtain from (30) a set of n_z BC on $\mathbf{z}$ alone.

If there are more than n_z BC in (30) then they may be projected appropriately, as described in [As2].

Step 3: Form the matrices $M(0)$ and $M(1)$, and analyze their eigenvalues using the QR algorithm. Let

$$V(t) = Q^T(t)M(t)Q(t) \qquad \text{at } t=0,1$$

where Q is orthogonal and V is upper triangular with the eigenvalues arranged in increasing order of real parts, from large negative to large positive. Let n_0^- (n_0^+) be the number of eigenvalues of $M(0)$ which have a large negative (positive) real part, and similarly define n_1^-, n_1^+ for $M(1)$.

If $n_0^- + n_1^+ > n_y$ or $n_0^+ + n_1^- > n_y$ *then exit:* another method (e.g. transforming explicitly to (11) everywhere first) should be used.

Otherwise set $k := \max(n_0^-, n_1^-)$; set the last n_y–k components of $Q^T\mathbf{y}$ at t=0 according to (15), and set the first k components of $Q^T\mathbf{y}$ at t=1 according to (15). Transform back to obtain BC on $\mathbf{x}(0)$ and $\mathbf{x}(1)$.

This completes specification of the BC for the application of the midpoint scheme.

Step 4: Solve the discretized equations (6) with the obtained BC.

□

This algorithm gives a general solution method which often works very well. The utility of Step 3 depends on an assumption that, in case that $M(t)$ has eigenvalues with large real part (the only case where this matters), its variation in t is slow compared to the size of such eigenvalues. Then the eigenvalue analysis approximates Lyapunov's equation well. (The large size eigenvalues approximate the large size kinematic eigenvalues well, and $Q^T\mathbf{y}$ are decoupled

variables.) This step includes a check for a dichotomy of the ghost ODE, with k being the number of increasing modes (which render an IVP for the ghost ODE unstable). Note that except at the end points we never find $T(t)$ or form $U(t)$. For the example of §2, if $\beta \leq 1$ then $k=0$ and we have an IVP as before. But with $\beta \gg 1$ we get $k=1$, so the side condition that Step 3 above dictates is at $t=1$, and it reads

$$-\beta x_1(1) + (1+\beta)x_2(1) = \sin 1.$$

In addition, the original condition $x_1(0)=1$ remains unchanged through Step 2. Solving the discretization equations (6) under these boundary conditions is a stable process, and the obtained errors are much smaller than those listed in Tables 2(a-e) for (8).

It should be realized that this algorithm does not always produce moderate stability constants. In fact, a priori there is no guarantee that the ghost ODE has a dichotomy. If it does not then playing with boundary conditions will not be of much help.

A third example

Consider the IVP (1) with

$$E = \begin{pmatrix} 0 & 0 \\ -1 & \beta t \end{pmatrix}, \quad A = \begin{pmatrix} 1/2-t & \beta t(t-1/2)-1 \\ 1 & -\beta(1+t) \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} \sin t \\ 0 \end{pmatrix},$$

β a parameter, and $x_1(0) = -1$.

Here we can let $S = I$, $T^{-1} = \begin{pmatrix} t-1/2 & 1-\beta t(t-1/2) \\ -1 & \beta t \end{pmatrix}$. Then for $\begin{pmatrix} y \\ z \end{pmatrix} = T^{-1}\mathbf{x}$ we have

$$y = \sin t$$

$$z' = -z,$$

i.e. $U = -I$. The initial condition gives $z(0)=1$, yielding $z = e^{-t}$. This problem has index 1 and

solution

$$x_1 = (\beta t(t-1/2)-1)e^{-t} + \beta t \sin t,$$
$$x_2 = \sin t + (t-1/2)e^{-t}.$$

(Note that, while the particular semi-explicit form above is very simple, the conditioning of the problem does grow linearly with $|\beta|$, and also $||T^{-1}T'|| \sim |\beta|$.)

For the ghost ODE (23c) we obtain

$$M = (T^{-1}T')^{11} = \beta(t-1/2)$$

so it has no dichotomy. Computing solutions with the midpoint scheme we get for $|\beta|$ large ($\approx$1000) poor results as in Table 1(e), regardless of whether we plant the side condition on y at 0 or at 1.

Note that for $\beta \gg 1$ we have $n_0^- = n_1^+ = 1$, and for $\beta \ll -1$ we have $n_0^+ = n_1^- = 1$. In both cases the check in Step 3 of the algorithm discovers the potential trouble.

□

While we expect our algorithm to handle most problems well, it is certainly possible that its Step 3 would not discover a lack of dichotomy in a given problem. In such a case, the algorithm may lead to a computed solution with large errors. Therefore, one should compute an error estimate along with the solution by one of the usual techniques. If the error is estimated to be too large, and the step size needed to meet a given tolerance is deemed too small, then a different solution method may be switched to.

References

[As1] U. Ascher, "On some difference schemes for singular singularly perturbed boundary value problems", Numer. Math. 46 (1986), 1-30.

[As2] U. Ascher, "On numerical differential algebraic problems with application to semiconductor device simulation", SIAM J. Numer. Anal. to appear.

[AsBa] U. Ascher and G. Bader, "Stability of collocation at Gaussian points", SIAM J. Numer. Anal. 23 (1986), 412-422.

[AMR] U. Ascher, R. Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, 1988.

[AsWe1] U. Ascher and R. Weiss, "Collocation for singular perturbation problem I: First order systems with constant coefficients", SIAM J. Num. Anal. 20 (1983), 537-557.

[AsWe2] U. Ascher and R. Weiss, "Collocation for singular perturbation problems II: Linear first order systems without turning points", Math. Comp. 43 (1984), 157-187.

[AsWe3] U. Ascher and R. Weiss, "Collocation for singular perturbation problems III: Nonlinear problems without turning points", SIAM J. Scient. Stat. Comp. 5 (1984), 811-829.

[BuBu] K. Burrage and J.C. Butcher, "Stability criteria for implicit Runge-Kutta methods", SIAM J. Numer. Anal. 16 (1979), 46-57.

[BuPe] K. Burrage and L. Petzold, "On order reduction for Runge-Kutta methods applied to differential/algebraic systems and to stiff systems of ODEs", Lawrence Livermore UCRL-98046 preprint.

[Ma] R. März, "On difference and shooting methods for boundary value problems in differential-algebraic equations", ZAMM 64 (1984), 463-473.

[Pe] L.R. Petzold, "Order results for implicit Runge-Kutta methods applied to differential/algebraic systems", SIAM J. Numer. Anal. 23 (1986), 837-852.

[Ve] M. van Veldhuisen, "D-stability", SIAM J. Numer. Anal. 18 (1981), 45-64.

[We] R. Weiss, "An analysis of the box and trapezoidal schemes for linear singularly perturbed boundary value problems", Math. Comput. 42 (1984), 41-68.