Applied Numerical Mathematics 13 (1993) 291–304
North-Holland

APNUM 447

Stability properties of some boundary value methods *

L. Brugnano and D. Trigiante
Dipartimento di Energetica, Via Lombroso 6/17, 50134 Firenze, Italy

Abstract

Bruguano, L. and D. Trigiante, Stability properties of some boundary value methods, Applied Numerical Mathematics 13 (1993) 291–304.

The boundary value methods (BVMs) are a class of numerical methods for solving initial value problems for ODEs. So far they did not have a broad diffusion, essentially for the following two reasons: their stability properties were not clearly understood and, moreover, they were considered too expensive.

In this paper we shall analyze the stability properties of three particular BVMs when used for solving linear systems of ODEs.

An efficient implementation of these methods will be described in a companion paper [6].

1. Introduction

Let us consider the problem of solving the initial value problem (IVP):

$$y'(t) = Ly(t) + b(t), \quad y(t_0) = y_0, \qquad t \in [t_0, T], \tag{1}$$

where $y(t)$, $b(t):[t_0, T] \to \mathbb{R}^m$ and $L \in \mathbb{R}^{m\times m}$ is a constant matrix. Moreover, we shall assume that all the eigenvalues of L have negative real part and $b(t)$ is a smooth and uniformly bounded function. If a discrete two-step method is used to discretize (1), then a second-order difference equation is obtained which needs an additional condition. Usually, the additional condition is chosen by looking for an approximation of $y(t_0 + h_1)$. This approach originates a discrete IVP. An alternative approach is to consider an approximation of $y(T)$ as additional condition. In this case the continuous IVP is approximated by means of a discrete boundary value problem (BVP).

A three-point BVM [2–4,8,13,14] is obtained by fixing a partitioning of the interval $[t_0, T]$, $t_0 < t_1 < t_2 < \cdots < t_k = T$, such that $t_i = t_{i-1} + h_i$, $i = 1, \ldots, k$. The problem (1) is then discretized by using a two-step method (*main method*), while in the last step it is discretized by using an implicit one-step method (*last-point method*).

Correspondence to: L. Brugnano, Dipartimento di Energetica, Via Lombroso 6/17, 50134 Firenze, Italy. E-mail: udini@vm.cnuce.cnr.it.

* Work supported by MURST (40% project), and CNR (contract #92.00535.CT01 and P.F. "Calcolo Parallelo", sottoprogetto 1).

The idea of solving a given IVP by means of a suitable BVP is not a recent one, since it goes back to Miller [15] and Olver [16] for discrete problems. For continuous problems it was suggested by Fox and Mitchell [10] and Fisher and Usmani [9]. More recently, the approach was used by Carasso [7] and Greenspan [11] for PDEs, and by Cash [8], Axelsson and Verwer [3] for ODEs.

The application of a three-point BVM originates a linear system such as:

$$Ay = c. \tag{2}$$

The matrix A is block tridiagonal,

$$A = \begin{pmatrix} \alpha(L; h_1, h_2) & \gamma(L; h_1, h_2) & & \\ \beta(L; h_2, h_3) & \alpha(L; h_2, h_3) & \gamma(L; h_2, h_3) & \\ \ddots & \ddots & \ddots & \\ & \beta(L; h_{k-1}, h_k) & \alpha(L; h_{k-1}, h_k) & \gamma(L; h_{k-1}, h_k) \\ & & \hat{\beta}(L; h_k) & \hat{\alpha}(L; h_k) \end{pmatrix}, \tag{3}$$

and the vector y is

$$y = (y_1, \ldots, y_k)^T,$$

where y_i is the approximation to $y(t_i)$. The functions α, β, γ, $\hat{\alpha}$, and $\hat{\beta}$ in (3), which are polynomials in L of degree at most 1, as well as the structure of the vector c, depend on both the main and last-point methods chosen. In the simpler case where b is time-independent and a constant stepsize h is used, one obtains

$$c = (2hb - \beta(L; h)y_0, 2hb, \ldots, 2hb, hb)^T.$$

Examples of polynomials α, β, γ, $\hat{\alpha}$, and $\hat{\beta}$ are given by

- *Main methods:*

Mid-point	Simpson	Adams
$\alpha(L; h) = -2hL,$	$\alpha(L; h) = -\frac{4}{3}hL,$	$\alpha(L; h) = I - \frac{2}{3}hL,$
$\beta(L; h) = -I,$	$\beta(L; h) = -I - \frac{1}{3}hL,$	$\beta(L; h) = -I - \frac{5}{12}hL,$
$\gamma(L; h) = I,$	$\gamma(L; h) = I - \frac{1}{3}hL,$	$\gamma(L; h) = \frac{1}{12}hL;$

- *Last-point methods:*

Implicit Euler	Trapezoidal rule
$\hat{\alpha}(L' h) = I - hL,$	$\hat{\alpha}(L; h) = I - \frac{1}{2}hL,$
$\hat{\beta}(L; h) = -I,$	$\hat{\beta}(L; h) = -I - \frac{1}{2}hL.$

For the variable step case see [1].

The stability properties of three-point BVMs are well known only when a constant stepsize h is used [14], and partially known when h varies [1]. Nevertheless, these methods are more effective when a variable stepsize is used. In fact, a variable stepsize permits both to cover large intervals of integration with a relatively small number of steps, and to avoid the large errors

due to the presence of a layer. We shall investigate the following case of variable stepsize:

$$h_{i+1} = rh_i, \quad i = 1, \ldots, k-1, \tag{4}$$

where $r > 1$ is a fixed parameter and the initial step h_1 is given. Concerning h_1, a value

$$h_1 < \| L \|^{-1}$$

is appropriate. When not specified, $\| \cdot \|$ may be $\| \cdot \|_1$, or $\| \cdot \|_2$, or $\| \cdot \|_\infty$. The results obtained for the case (4) can be easily extended to the more general case

$$h_{i+1} = r_i h_i, \qquad r_i > 1, \quad i = 1, \ldots, k-1,$$

provided that

$$\begin{aligned} r_1 &= r_2 = \cdots = r_{k_1}, \\ r_{k_1+1} &= \cdots = r_{k_2}, \\ &\vdots \\ r_{k_s+1} &= \cdots = r_k, \end{aligned}$$

where $1 \leqslant k_1 < k_2 < \cdots < k_s \leqslant k$, and s doesn't depend on k. The stability results will be stated in Section 3.

In Section 2 we shall recall some results about the conditioning of tridiagonal matrices which will be used to obtain the stability results.

2. Conditioning of tridiagonal matrices

If the integration steps are chosen according to (4), then matrix (3) can be written as

$$A = T_1 \otimes I - T_2 \otimes h_1 L, \tag{5}$$

where $\otimes$ denotes the right Kronecker product (see [12]), and the matrices T_1 and T_2 both depend on the main and last-point methods chosen; moreover the right-hand side in (2) can be written as

$$c = c_1 + (D \otimes I) h_1 c_2, \tag{6}$$

where

$$c_1 = \left(-\beta(L; h_1, rh_1) y_0, 0, \ldots, 0\right)^{\mathrm{T}},$$

and

$$c_2 = \left(D^{-1} T_2 \otimes I\right)\left(b(t_1), \ldots, b(t_k)\right)^{\mathrm{T}}$$

is a block vector whose entries depend on the BVM chosen and is always bounded if $b(t)$ is a bounded function. Finally, D is the following diagonal matrix:

$$D = \operatorname{diag}(1, r, r^2, \ldots, r^{k-1}). \tag{7}$$

As an example,

$$T_1 = \begin{pmatrix} (1-r^{-2}) & r^{-2} & & \\ -1 & \ddots & \ddots & \\ & \ddots & (1-r^{-2}) & r^{-2} \\ & & -1 & 1 \end{pmatrix} \tag{8}$$

and

$$T_2 = D\begin{pmatrix} (1+r^{-1}) & & & \\ & \ddots & & \\ & & (1+r^{-1}) & \\ & & & 1 \end{pmatrix}, \tag{9}$$

define the BVM which utilizes the mid-point method as main method and the implicit Euler method as last-point method.

If the Simpson method is used as main method and the trapezoidal rule as last-point method, then T_1 and D are the matrices defined above and

$$T_2 = D\begin{pmatrix} \frac{2}{3}(1+r^{-1}) & \frac{1}{3}r^{-1} & & & \\ \frac{1}{3} & & & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{3} & \frac{2}{3}(1+r^{-1}) & \frac{1}{3}r^{-1} \\ & & & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{10}$$

Lastly, for the Adams method as main method and the trapezoidal rule as last-point method, one obtains:

$$T_1 = \begin{pmatrix} 1 & & & \\ -1 & & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \tag{11}$$

and

$$T_2 = (6r(r+1))^{-1} D\begin{pmatrix} (1+3r)(r+1) & -1 & & & \\ r(2+3r) & & & & \\ & \ddots & \ddots & \ddots & \\ & & r(2+3r) & (1+3r)(r+1) & -1 \\ & & & 3r(r+1) & 3r(r+1) \end{pmatrix}, \tag{12}$$

where D is the same matrix as defined in (7).

From (5) it is evident that the conditioning of the matrix A is closely related to the conditioning of tridiagonal matrices; moreover, as shown in Section 3, the conditioning of the matrix A is related to the stability properties of the corresponding BVM. For this reason we shall now summarize sufficient conditions stated in [5] for a tridiagonal matrix to be *well-conditioned,* according to the following:

Definition 2.1. A nonsingular matrix is said to be *well-conditioned* if its condition number is bounded by a quantity independent on the size of the matrix. If this quantity depends as a polynomial of low degree (1 or 2) on the size of the matrix, then the matrix is said to be *weakly well-conditioned.*

For sake of brevity, we shall consider the following real tridiagonal matrix of size k:

$$\begin{pmatrix} 1 & \tau_1 & & \\ \sigma_1 & 1 & \ddots & \\ & \ddots & \ddots & \tau_{k-1} \\ & & \sigma_{k-1} & 1 \end{pmatrix}, \tag{13}$$

with the assumption that $\sigma_0 = \tau_0 = \sigma_k = \tau_k = 0$, and $\sigma_i \tau_i \neq 0$, $i = 1, \ldots, k-1$. Concerning the invertibility of T, one has:

Theorem 2.2. *Suppose that for all $i = 1, \ldots, k-1$ one has $\sigma_i \tau_i \leqslant \frac{1}{4}$, then matrix* (13) *is invertible.*

Corollary 2.3. *Suppose that for all $i = 1, \ldots, k-1$ one has $|\sigma_i + \tau_i| \leqslant 1$, then matrix* (13) *is invertible.*

Suppose now that the hypotheses of Theorem 2.2 are satisfied. We shall consider the simpler case in which the products $\sigma_i \tau_i$, $i = 1, \ldots, k-1$, have constant sign. More complicated cases may be treated similarly by using the Shermann–Morrison formula, provided that the number of changes of sign of the products $\sigma_i \tau_i$ is independent of the size k of the matrix T. Therefore, two main cases may be considered:

(1) $0 < \sigma_i \tau_i \leqslant \frac{1}{4}$, $i = 1, \ldots, k-1$;
(2) $\sigma_i \tau_i < 0$, $i = 1, \ldots, k-1$.

In the first case the following result holds:

Theorem 2.4. *Suppose that for all $i = 1, \ldots, k-1$ one has $0 < \sigma_i \tau_i \leqslant \frac{1}{4}$ and, moreover, the following set of conditions is satisfied for $i = 1, \ldots, k$:*

$$|\sigma_i| + |\tau_{i-1}| < 1,$$
$$|\tau_i| + |\sigma_{i-1}| < 1,$$

then matrix (13) *is nonsingular and well-conditioned. If the inequalities are not strict, then matrix* (13) *is at least weakly well-conditioned.*

In the case $\sigma_i \tau_i < 0$, the following result holds:

Theorem 2.5. *Suppose that for all $i = 1, \ldots, k-1$ one has $\sigma_i \tau_i < 0$ and, moreover, one of the following sets of conditions is satisfied for $i = 1, \ldots, k$:*

(1) $$\begin{cases} |\sigma_i| - |\tau_{i-1}| < 1, \\ |\tau_i| - |\sigma_{i-1}| < 1, \end{cases}$$

(2) $$\begin{cases} |\tau_{i-1}| - |\sigma_i| < 1, \\ |\sigma_{i-1}| - |\tau_i| < 1, \end{cases}$$

then matrix (13) *is nonsingular and well-conditioned. If the inequalities are not strict, then matrix* (13) *is at least weakly well-conditioned.*

We observe that when the σ_i and the τ_i have constant sign along each diagonal, the following simpler result holds:

Corollary 2.6. *Suppose that for all* $i = 1, \ldots, k-1$ *one has* $\sigma_i \tau_i \leqslant \frac{1}{4}$, σ_i *and* τ_i *have constant sign along each diagonal, and, moreover, the following set of conditions is satisfied for* $i = 1, \ldots, k$:

$$\begin{cases} |\sigma_i + \tau_{i-1}| < 1, \\ |\tau_i + \sigma_{i-1}| < 1, \end{cases}$$

then matrix (13) *is nonsingular and well-conditioned. If the inequalities are not strict, then matrix* (13) *is at least weakly well-conditioned.*

For proofs see [5].

Remark 2.7. The above results can be summarized as follows: if the conditions of Theorem 2.2 (or Corollary 2.3) are satisfied, then there is a set of row conditions and a set of column conditions to be satisfied for matrix (13) to be well-conditioned.

3. Stability results

Concerning the stability properties of BVMs, it is natural to require that the numerical solution is always bounded when the continuous solution is bounded for $t \geqslant t_0$. Therefore, when the eigenvalues of L have negative real part, we shall require that $\| y \|$ must be bounded by a quantity which is independent of the number of time-steps k. (In the following, we shall say only "bounded" without the repetition of "with respect to k".) Since from (2) and (6) it follows that the numerical solution of problem (1), when the variable stepsizes (4) are used, is given by

$$y = A^{-1} c_1 + A^{-1}(D \otimes I) h_1 c_2,$$

the above requirement is equivalent to having, for fixed values of the parameters r and h_1, both $\| A^{-1}(D \otimes I) \|$ and $\| A^{-1} \|$ bounded.

It turns out that $\| A^{-1}(D \otimes I) \|$ bounded implies that $\| A^{-1} \|$ is bounded as well, since (see (7)) $\| D^{-1} \| = 1$. Moreover, $\| A^{-1}(D \otimes I) \|$ is bounded if the matrix $(D^{-1} \otimes I)A$ is well-conditioned, according to Definition 2.1, since $\|(D^{-1} \otimes I)A\|$ is bounded from below, as it is stated in the following lemma.

Lemma 3.1. *If the eigenvalues of matrix* L *have negative real part, then the BVMs given by* (5)–(12) *are such that* $\|(D^{-1} \otimes I)A\|$ *is bounded from below by a positive quantity independent of* k, *for all* $r > 1$.

Proof. We shall give a formal proof only for the norm $\| \cdot \|_2$; the proof for the norms $\| \cdot \|_1$ and $\| \cdot \|_\infty$ is similar. Let λ be an eigenvalue of the matrix L, and let v be the corresponding

eigenvector, $\|v\|_2 = 1$. If e_1 is the first unit vector of the canonical base of $\mathbb{R}^k$, it follows that, denoting by $q = -h_1\lambda$ ($\mathrm{Re}(q)$ is then positive):

$$\begin{aligned}\|(D^{-1}\otimes I)A\|_2 &\geqslant \|(D^{-1}\otimes I)A(e_1\otimes v)\|_2\\ &= \|(D^{-1}T_1 + qD^{-1}T_2)e_1\otimes v\|_2\\ &= \|(D^{-1}T_1 + qD^{-1}T_2)e_1\|_2\\ &> |a_1 + a_2 q| > 0,\end{aligned}$$

where (see (8)–(12)),

$$a_1 = \begin{cases} 1 - r^{-2}, & \text{for the mid-point method,}\\ 1 - r^{-2}, & \text{for the Simpson method,}\\ 1, & \text{for the Adams method,}\end{cases}$$

and

$$a_2 = \begin{cases} 1 + r^{-1}, & \text{for the mid-point method,}\\ \frac{2}{3}(1 + r^{-1}), & \text{for the Simpson method,}\\ \frac{1}{6}(1 + 3r)/r, & \text{for the Adams method.}\end{cases}$$ □

We shall now examine in more detail the conditioning of the matrix $(D^{-1}\otimes I)A$; this will be done firstly in the simpler case when problem (1) is a scalar equation. The obtained results will be then extended to systems.

3.1. Stability and row-scaling

We now apply the BVMs given by (5) to the scalar problem

$$y' = \lambda y, \quad \lambda < 0.$$

By posing $q = -h_1\lambda > 0$, matrix (5) becomes

$$T = T_1 + qT_2 =: \hat{D}\hat{T}, \tag{14}$$

where $\hat{D}$ is diagonal with positive diagonal entries, and

$$\hat{T} = \begin{pmatrix} 1 & \hat{\tau}_1 & & \\ \hat{\sigma}_1 & 1 & \ddots & \\ & \ddots & \ddots & \hat{\tau}_{k-1}\\ & & \hat{\sigma}_{k-1} & 1\end{pmatrix}.$$

The first $k-1$ rows of the matrix $\hat{T}$ depend on the main method used, while the last row depends on the chosen last-point method. Nevertheless, by considering the entries on the last row as a perturbation of those that would be obtained from the main method, by using the Sherman–Morrison formula one has that the conditioning of $\hat{T}$ depends essentially on the main method (see also [1]). Concerning the matrix $\hat{T}$ the following result holds:

Theorem 3.2. *If T_1 and T_2 are the tridiagonal matrices which define the BVMs given by* (5)–(12), *then $\hat{T}$ is invertible and well-conditioned* $\forall r > 1$.

Proof. If we neglect the perturbations on the entries on the last row of T_1 and $D^{-1}T_2$, then these matrices are Toeplitz matrices. Denoting by

$$T_1 = \alpha \begin{pmatrix} 1 & \tau^{(1)} & & \\ \sigma^{(1)} & \ddots & \ddots & \\ & \ddots & \ddots & \tau^{(1)} \\ & & \sigma^{(1)} & 1 \end{pmatrix}$$

and

$$D^{-1}T_2 = \beta \begin{pmatrix} 1 & \tau^{(2)} & & \\ \sigma^{(2)} & \ddots & \ddots & \\ & \ddots & \ddots & \tau^{(2)} \\ & & \sigma^{(2)} & 1 \end{pmatrix},$$

one easily obtains that the off-diagonal entries of $\hat{T}$ are given by

$$\hat{\sigma}_i = \frac{\alpha\sigma^{(1)} + q\beta r^i \sigma^{(2)}}{\alpha + q\beta r^i}, \qquad \hat{\tau}_i = \frac{\alpha\tau^{(1)} + q\beta r^{i-1}\tau^{(2)}}{\alpha + q\beta r^{i-1}},$$

where the quantities α, β, $\sigma^{(1)}$, $\sigma^{(2)}$, $\tau^{(1)}$, and $\tau^{(2)}$ depend on the main method used, according to Table 1.

One realizes that $\hat{\sigma}_i$ and $\hat{\tau}_i$, and the products $\hat{\sigma}_i\hat{\tau}_i$ change their signs at most once. Moreover, the row conditions

$$|\hat{\sigma}_{i-1} + \hat{\tau}_i| < 1, \quad i = 1, \ldots, k,$$

are always satisfied. From Remark 2.7, in order to have $\hat{T}$ well-conditioned one needs to

Table 1
Parameters for the various methods

	Mid-point	Simpson	Adams
α	$1-r^{-2}$	$1-r^{-2}$	1
β	$1+r^{-1}$	$\frac{2}{3}(1+r^{-1})$	$\frac{1+3r}{6r}$
$\sigma^{(1)}$	$\frac{-1}{1-r^{-2}}$	$\frac{-1}{1-r^{-2}}$	-1
$\sigma^{(2)}$	0	$\frac{1}{2(1+r^{-1})}$	$\frac{r(2+3r)}{(1+3r)(r+1)}$
$\tau^{(1)}$	$\frac{r^{-2}}{1-r^{-2}}$	$\frac{r^{-2}}{1-r^{-2}}$	0
$\tau^{(2)}$	0	$\frac{r^{-1}}{2(1+r^{-1})}$	$\frac{-1}{(1+3r)(r+1)}$

impose an additional set of column conditions (according to the results of Theorem 2.4, Theorem 2.5, or Corollary 2.6) and the conditions

$$\hat{\sigma}_i \hat{\tau}_i \leqslant \tfrac{1}{4}, \quad i = 1, \ldots, k-1.$$

We omit, for brevity, a formal proof for the mid-point method, which is straightforward.

For all methods the off-diagonal elements on the ith column of $\hat{T}$ are the components of the vector:

$$\begin{pmatrix} \hat{\sigma}_i \\ \hat{\tau}_{i-1} \end{pmatrix} = \begin{pmatrix} \dfrac{\alpha}{\alpha + q\beta r^i}\sigma^{(1)} \\ \dfrac{\alpha}{\alpha + q\beta r^{i-2}}\tau^{(1)} \end{pmatrix} + \begin{pmatrix} \dfrac{q\beta r^i}{\alpha + q\beta r^i}\sigma^{(2)} \\ \dfrac{q\beta r^{i-2}}{\alpha + q\beta r^{i-2}}\tau^{(2)} \end{pmatrix}$$

$$= \frac{\alpha}{\alpha + q\beta r^i} \begin{pmatrix} \sigma^{(1)} \\ \dfrac{\alpha + q\beta r^i}{\alpha + q\beta r^{i-2}}\tau^{(1)} \end{pmatrix} + \frac{q\beta r^i}{\alpha + q\beta r^i} \begin{pmatrix} \sigma^{(2)} \\ \dfrac{\alpha + q\beta r^i}{\alpha r^2 + q\beta r^i}\tau^{(2)} \end{pmatrix}$$

$$=: \frac{\alpha}{\alpha + q\beta r^i} v_1 + \frac{q\beta r^i}{\alpha + q\beta r^i} v_2.$$

For the Simpson method we observe that the vector $(\hat{\sigma}_i, \hat{\tau}_{i-1})^{\mathrm{T}}$ is a convex combination of the two vectors v_1 and v_2, which are inside the convex region

$$\Sigma = \left\{ (\sigma, \tau)^{\mathrm{T}} \in \mathbb{R}^2 : |\sigma + \tau| \leqslant 1 \right\}$$

$\forall r > 1$, whence the second condition in Corollary 2.6 is satisfied.

Concerning the invertibility and the first condition in Corollary 2.6, we have:

$$\begin{pmatrix} \hat{\sigma}_i \\ \hat{\tau}_i \end{pmatrix} = \begin{pmatrix} \dfrac{\alpha}{\alpha + q\beta r^i}\sigma^{(1)} \\ \dfrac{\alpha}{\alpha + q\beta r^{i-1}}\tau^{(1)} \end{pmatrix} + \begin{pmatrix} \dfrac{q\beta r^i}{\alpha + q\beta r^i}\sigma^{(2)} \\ \dfrac{q\beta r^{i-1}}{\alpha + q\beta r^{i-1}}\tau^{(2)} \end{pmatrix}$$

$$= \frac{\alpha}{\alpha + q\beta r^i} \begin{pmatrix} \sigma^{(1)} \\ \dfrac{\alpha + q\beta r^i}{\alpha + q\beta r^{i-1}}\tau^{(1)} \end{pmatrix} + \frac{q\beta r^i}{\alpha + q\beta r^i} \begin{pmatrix} \sigma^{(2)} \\ \dfrac{\alpha + q\beta r^i}{\alpha r + q\beta r^i}\tau^{(2)} \end{pmatrix}$$

$$=: \frac{\alpha}{\alpha + q\beta r^i} w_1 + \frac{q\beta r^i}{\alpha + q\beta r^i} w_2.$$

The hypothesis $\hat{\sigma}_i \hat{\tau}_i \leqslant \frac{1}{4}$ will be obviously satisfied, from Corollary 2.3, if both w_1 and w_2 are

inside the "strip" Σ defined above. The point w_2 is inside this region; the same will hold for w_1 provided that

$$\sigma^{(1)} + \frac{\alpha + q\beta r^i}{\alpha + q\beta r^{i-1}}\tau^{(1)} < 1.$$

By means of simple calculations, one shows that this last condition is satisfied $\forall r > 1$.

The proof for the Adams method is obtainable by similar arguments. □

Remark 3.3. From Theorem 3.2 it follows that also the matrix $D^{-1}T$ is well-conditioned, since (see (7), (14), and Table 1) $\kappa(D^{-1}\hat{D}) < 1 + \alpha(q\beta)^{-1}$.

3.1.1. Complex eigenvalues

In the case of λ complex, the previous result cannot be extended straightforwardly. In order to analyze the stability properties for complex eigenvalues, we apply the BVM to the problem

$$y' = \begin{pmatrix} \rho & -\theta \\ \theta & \rho \end{pmatrix} y = Cy, \quad \rho,\theta \in \mathbb{R}, \quad \rho < 0.$$

Then, matrix (5) will be given by

$$T_C = T_1 \otimes I - T_2 \otimes h_1 C,$$

which is a block tridiagonal matrix. For such matrix the following result holds:

Theorem 3.4. *If T_1 and T_2 are the tridiagonal matrices which define the BVMs given by* (5)–(12), *then the block tridiagonal matrix T_C is invertible and, if scaled on the rows by the inverse of the main diagonal, well-conditioned $\forall r > 1$, provided that $|\theta|$ is "sufficiently" small.*

Proof. By posing $q = -h_1\rho$, we have:

$$\begin{aligned} T_C &= (T_1 + qT_2) \otimes I - h_1\theta T_2 \otimes \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ &= ((T_1 + qT_2) \otimes I)F, \end{aligned}$$

where

$$F = I \otimes I - (T_1 + qT_2)^{-1} h_1\theta T_2 \otimes \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Considering that:

(1) $(T_1 + qT_2) = \hat{D}\hat{T}$ satisfies the conditions of Theorem 3.2 and therefore $\hat{T}$ is well-conditioned ($\hat{D} \otimes I$ is the main diagonal of T_C),

(2) $(T_1 + qT_2)^{-1} h_1\theta T_2 = h_1\theta \hat{T}^{-1} D(D^{-1}T_2)$ and

- $\hat{T}^{-1}D$ is bounded, as seen in Remark 3.3,
- $(D^{-1}T_2)$ is bounded (see (9), (10) and (12)),

then, if

$$h_1|\theta|\eta < 1, \tag{15}$$

where $\eta = \|\hat{T}^{-1}D\| \, \|D^{-1}T_2\|$, by Banach's lemma it follows that

$$\|F^{-1}\| < \frac{1}{1 - h_1|\theta|\eta}. \qquad \square$$

Let us analyze inequality (15) in more detail, in order to obtain an effective bound on $|\theta|$. By using the same notation as in Theorem 3.4, we define the function

$$f(q, r, k) = \left\| \left(D^{-1}T_1 + qD^{-1}T_2 \right)^{-1} \right\|.$$

We know that

- for $q > 0$, $\kappa(D^{-1}T_1 + qD^{-1}T_2) < g(q, r)$, since the matrix is well-conditioned (see Remark 3.3); then, from Lemma 3.1, since $q > 0$, it follows that

$$f(q, r, k) = \frac{\kappa\left(D^{-1}T_1 + qD^{-1}T_2\right)}{\left\| D^{-1}T_1 + qD^{-1}T_2 \right\|} \leqslant \frac{g(q, r)}{a_1 + a_2 q},$$

  for some positive constants a_1 and a_2, which depend only on r;
- for $q = 0$, one has

$$\begin{aligned} f(0, r, k) &= \left\| T_1^{-1} D \right\| = \left\| DD^{-1}T_1^{-1}D \right\| \\ &=: \left\| D\hat{T}_1^{-1} \right\| \leqslant \left\| \hat{T}_1^{-1} \right\| \| D \|, \end{aligned}$$

  and $\| D \| = r^{k-1}$ (see (7)); moreover, one verifies that $\hat{T}_1 = D^{-1}T_1 D$ either satisfies the hypotheses of Theorem 3.4, or is diagonally dominant both by rows and by columns, and therefore is well-conditioned; it follows that $f(0, r, k) \leqslant h(r) r^{k-1}$;
- for $q \gg 0$, $f(q, r, k) \approx q^{-1} \| T_2^{-1} D \|$, and $\| T_2^{-1} D \| \leqslant l(r)$, since $D^{-1}T_2$ is diagonally dominant both by rows and by columns; therefore it is well-conditioned.

The above considerations allow to state that

$$f(q, r, k) \leqslant \frac{\left\| \left(D^{-1}T_2 \right)^{-1} \right\|}{q + \dfrac{\left\| \left(D^{-1}T_2 \right)^{-1} \right\|}{\left\| \hat{T}_1^{-1} \right\| r^{k-1}}} =: b(q, r, k).$$

This bound has been found to be very sharp numerically. Condition (15) is then satisfied if one requires:

$$h_1 |\theta| b(q, r) \left\| D^{-1}T_2 \right\| < 1.$$

This condition is implied by

$$h_1 |\theta| \frac{\left\| \left(D^{-1}T_2 \right)^{-1} \right\|}{q} \left\| D^{-1}T_2 \right\| \leqslant 1,$$

that is, considering that $q = h_1 |\rho|$,

$$|\theta| \leqslant \frac{|\rho|}{\kappa\left(D^{-1}T_2\right)}.$$

Moreover, if we neglect the perturbations on the last row (due to the last-point method) then, as seen in Theorem 3.2, matrix $D^{-1}T_2$ (see (9), (10), and (12)) is a Toeplitz matrix and, after some calculations, one derives:

$$\kappa\left(D^{-1}T_2\right) \begin{cases} = 1, & \text{if the mid-point method is used,} \\ \leqslant 3, & \text{if the Simpson method is used,} \\ \leqslant 3r + 4, & \text{if the Adams method is used.} \end{cases}$$

3.2. Generalization to systems

Now we shall generalize the previous results to systems of equations. The following result holds:

Lemma 3.5. *If matrix L is normal and all its eigenvalues have negative real part, then, for the BVMs defined by* (5)–(12), *the matrix $(D^{-1}\otimes I)A$ is well-conditioned $\forall r>1$, provided that the imaginary part of each complex eigenvalue is "sufficiently" small.*

Proof. Let us consider the real Schur decomposition of L, $L=Q\Lambda Q^{\mathrm{T}}$, where Λ is a block diagonal matrix, with diagonal blocks of size 1 in correspondence to the real eigenvalues of L, and of size 2 in correspondence to each complex-conjugate pair of eigenvalues of L. Then, instead of the matrix A (see (5)), we can consider the matrix

$$(I\otimes Q^{\mathrm{T}})A(I\otimes Q)=T_1\otimes I-T_2\otimes h_1\Lambda =: A_N. \tag{16}$$

The thesis follows by considering that a permutation matrix $H_{k,m}$ of order km exists such that the permuted matrix $H_{k,m}A_NH_{k,m}^{\mathrm{T}}$ is block diagonal, with diagonal blocks which satisfy either the hypotheses of Theorem 3.2 or of Theorem 3.4. □

The previous result can be extended to a more general matrix L:

Theorem 3.6. *If all the eigenvalues of matrix L have negative real part, then, for the BVMs defined by* (5)–(12), *the matrix $(D^{-1}\otimes I)A$ is well-conditioned $\forall r>1$, provided that the imaginary part of each complex eigenvalue is "sufficiently" small.*

Proof. Let $L=Q(\Lambda+N)Q^{\mathrm{T}}$ be as before the real Schur decomposition of L, where N is strictly upper triangular and nilpotent of order m (the size of the matrix L). Then, instead of the matrix A we can consider the matrix (see (16))

$$\begin{aligned}(I\otimes Q^{\mathrm{T}})A(I\otimes Q)&=T_1\otimes I-T_2\otimes h_1\Lambda-T_2\otimes h_1N\\&=A_N\bigl(I\otimes I-A_N^{-1}(T_2\otimes h_1N)\bigr).\end{aligned}$$

The matrix A_N satisfies the hypotheses of Lemma 3.5, and then $(D^{-1}\otimes I)A_N$ is well-conditioned. Moreover, the matrix $(I\otimes I-A_N^{-1}(T_2\otimes h_1N))$ is invertible and well-conditioned since $A_N^{-1}(T_2\otimes h_1N)$is nilpotent of order m. In fact,

$$\begin{aligned}A_N^{-1}&=(T_1\otimes I-T_2\otimes h_1\Lambda)^{-1}=\bigl(I\otimes I-T_1^{-1}T_2\otimes h_1\Lambda\bigr)^{-1}\bigl(T_1^{-1}\otimes I\bigr)\\&=\left(\sum_{i=0}^{km-1}a_i\bigl(T_1^{-1}T_2\bigr)^i\otimes\Lambda^i\right)\bigl(T_1^{-1}\otimes I\bigr),\end{aligned}$$

for some scalars $a_1,\dots,a_{km-1}$ (see [12, Chapter 9]). It follows that:

$$\begin{aligned}A_N^{-1}(T_2\otimes h_1N)&=h_1\sum_{i=0}^{km-1}a_i\bigl(T_1^{-1}T_2\bigr)^{i+1}\otimes\Lambda^iN\\&=:h_1\sum_{i=0}^{km-1}a_i\bigl(T_1^{-1}T_2\bigr)^{i+1}\otimes\tilde N_i.\end{aligned}$$

The matrix results to be nilpotent of order m, since the matrices $\tilde{N}_i$ are strictly upper triangular. □

Remark 3.7. The previous result states that the condition number of the matrix (5) is proportional to that of the matrix D defined in (7), that is to r^{k-1}. One easily obtains that this quantity is related to the width of the interval of integration divided by the initial step:

$$\kappa(A) \approx \mathrm{O}\big((T - t_0)h_1^{-1}\big).$$

Remark 3.8. It can be shown that the results of Theorem 3.6 still hold for $r = 1$, that is when a constant integration step is used.

4. Conclusions

In this paper the stability properties of three BVMs applied to problem (1), with the variable step sizes defined in (4), have been studied when the eigenvalues of the matrix L have negative real part.

It has been shown that the examined methods are stable for every choice of the parameters r and h_1, provided that every eigenvalue λ of L verifies $|\mathrm{Im}(\lambda)| \leqslant \delta\,|\mathrm{Re}(\lambda)|$, where δ is a constant which depends on the BVM considered.

The related problem of solving numerically the discrete problem (2), usually by means of an iterative solver, is examined in a companion paper [6].

Acknowledgements

We thank the referees for their valuable remarks and suggestions.

References

[1] P. Amodio, F. Mazzia and D. Trigiante, Stability of some boundary value methods for the solution of initial value problems, *BIT* (to appear).

[2] P. Amodio and D. Trigiante, A parallel direct method for solving initial value problems for ordinary differential equations, *Appl. Numer. Math.* 11 (1993) 85–93.

[3] A.O.H. Axelsson and J.G. Verwer, Boundary value techniques for initial value problems in ordinary differential equations, Preprint, Mathemathisch Centrum, Amsterdam (1983).

[4] L. Brugnano, F. Mazzia and D. Trigiante, Parallel implementation of BVM methods, *Appl. Numer. Math.* 11 (1993) 115–124.

[5] L. Brugnano and D. Trigiante, Tridiagonal matrices: invertibility and conditioning, *Linear Algebra Appl.* 166 (1992) 131–150.

[6] L. Brugnano and D. Trigiante, A parallel preconditioning technique for boundary value methods, *Appl. Numer. Math.* 13 (1993) 277–290 (this issue).

[7] A. Carasso, Long-range numerical solution of mildly non-linear parabolic equations, *Numer. Math.* 16 (1971) 304–321.

[8] J.R. Cash, *Stable Recursion* (Academic Press, New York, 1976).

[9] C.F. Fisher and R.A. Usmani, Properties of some tridiagonal matrices and their application to boundary value problems, *SIAM J. Numer. Anal.* 6 (1969) 127–142.
[10] L. Fox and A.R. Mitchell, Boundary value techniques for the numerical solution of initial value problems, *Quart. J. Mech. Appl. Math.* 10 (1957) 232–243.
[11] D. Greenspan, *Discrete Numerical Methods for Physics and Engineering* (Academic Press, New York, 1974).
[12] P. Lancaster and M. Tismenetsky, *The Theory of the Matrices, with applications* (Academic Press, New York, 2nd ed., 1985).
[13] L. Lopez, Two-step boundary value methods in the solution of ODEs, *Comput. Math. Appl.* (to appear).
[14] L. Lopez and D. Trigiante, Boundary methods and BV-stability in the solution of initial value problems, *Appl. Numer. Math.* 11 (1993) 225–239.
[15] J.P.C. Miller, *Bessel Functions, Part II* (Cambridge University Press, Cambridge, England, 1952).
[16] F.W. Olver, Numerical solution of second order linear difference equations, *J. Res. Nat. Bur. Standards Math. Phys.* 71B (1967) 111–129.