NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2005; **12**:305–314
Published online 4 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/nla.417

Some linear algebra issues concerning the implementation of blended implicit methods

Luigi Brugnano*,† and Cecilia Magherini‡

Dipartimento di Matematica "U. Dini", Viale Morgagni 67/A, 50134 Firenze, Italy

SUMMARY

In this paper we discuss some linear algebra issues concerning the implementation of *blended implicit methods* (*J. Comput. Appl. Math.* 2000; **116**:41–62, *Appl. Numer. Math.* 2002; **42**:29–45, *J. Comput. Appl. Math.* 2004; **164–165**:145–158, In Recent Trends in Numerical Analysis, Trigiante D (ed.), Nova Science Publication Inc.: New York, 2001; 81–105) for the numerical solution of ODEs. In particular, we describe the strategies, used in the numerical code BiM (*J. Comput. Appl. Math.* 2004; **164–165**:145–158), for deciding whether re-evaluating the Jacobian and/or the factorization involved in the non-linear splitting for solving the discrete problem. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: numerical methods for ODEs; stiff problems; iterative solution of linear systems; nonlinear splittings; computational codes

1. INTRODUCTION

Computational codes represent an outstanding technological aspect of the Mathematical Sciences. Moreover, these codes constitute basic tools for *problem solving* in applied fields. The construction of such codes requires, in turn, the systematic solution of a number of related sub-problems, which constitute the intermediate steps to reach the desired goal. This aspect of Numerical Mathematics is often underestimated and considered to be only of secondary importance. On the contrary, it is a source of new trends of investigation, and a necessary *building block* to make Mathematics usable from people involved in solving real-life problems.

With this premise, our attention will be devoted to the solution of several specific sub-problems which are met when constructing a numerical code for solving stiff IVPs for ODEs.

*Correspondence to: Luigi Brugnano, Dipartimento di Matematica "U. Dini", Viale Morgagni 67/A, 50134 Firenze, Italy.
†E-mail: brugnano@math.unifi.it
‡E-mail: magherini@math.unifi.it

Contract/grant sponsor: CNR
Contract/grant sponsor: GNCS-INdAM

Received 7 October 2002
Revised 25 April 2003

Copyright © 2004 John Wiley & Sons, Ltd.

Indeed the efficient solution of such problems requires to properly address (at least) the following points:

- the choice of appropriate methods,
- the selection of an appropriate mesh, in order to meet a prescribed accuracy requirement,
- the construction of a suitable discrete problem,
- the solution of the discrete problem itself.

Our attention will be mainly devoted to further sub-problems related to the last point. In more detail, when solving the problem

$$y' = f(t, y), \quad t \in [t_0, T], \quad y(t_0) = y_0 \in \mathbb{R}^m \tag{1}$$

by means of an implicit difference scheme, the discrete solution of the local discrete problem, approximating locally the continuous one, usually requires the evaluation of the Jacobian of f at the most recently known point, as well as the factorization of a matrix, involved in the non-linear iteration, which depends on the Jacobian itself and from the current stepsize. Often, such operations are not performed at each integration step, in order to lower the computational cost. Essentially, this is done when the non-linear iteration has performed well in the last step, but the decision whether to re-evaluate the Jacobian and/or to factorize the matrix is usually made according to some heuristics. This is, indeed, the case for the most efficient codes currently available for solving problem (1). For the methods implemented in the code `BiM` [1], namely *blended implicit methods* [1–4], such a decision is supported by a linear analysis of convergence, which can be carried out because of the particular form of the discrete problem. This analysis will be the main concern of the paper. For the remaining computational details of the code `BiM`, we refer to Reference [1].

The organization of the paper is as follows: in Section 2, we recall the basic facts about the blended implicit methods implemented in the code `BiM`; Sections 3 and 4 are devoted to the analysis of the convergence properties of the iterative procedure for solving the corresponding discrete problem when, respectively, the Jacobian and the factorization are not updated, in order to provide a practical criterion.

2. BLENDED IMPLICIT METHODS

Block implicit methods are difference methods for the numerical integration of problem (1) which provide a discrete problem in the form,

$$F(\mathbf{y}_n) \equiv A \otimes I_m \mathbf{y}_n - h_n B \otimes I_m \mathbf{f}_n - \boldsymbol{\eta}_n = \mathbf{0} \tag{2}$$

where A and B are $r \times r$ non-singular matrices defining the method, the block vectors

$$\mathbf{y}_n = (y_{1n}, \ldots, y_{rn})^{\mathrm{T}}, \quad \mathbf{f}_n = (f_{1n}, \ldots, f_{rn})^{\mathrm{T}}, \quad f_{jn} = f(t_{jn}, y_{jn})$$

contain the discrete solution, and the vector $\boldsymbol{\eta}_n$ only depends on values of the solution at previous mesh points. Finally, following a standard notation, y_{in} denotes the approximation to $y(t_{in})$, where $t_{in} = t_{0n} + ih_n$, $i = 1, \ldots, r$, $t_{0n} \equiv t_{r,n-1}$, $n \geqslant 1$, $(t_{01} \equiv t_0)$ and h_n is the stepsize used in the current nth *block*. Instances of methods falling in this class are the majority of implicit

Copyright © 2004 John Wiley & Sons, Ltd.

Runge–Kutta methods, a number of general linear methods and, more recently, block BVMs [5]. In References [1,3] it is possible to find all details concerning the methods implemented in the code BiM. In order to describe their *blended implementation*, let us apply the methods to the standard test equation,

$$y' = \mu y, \quad y(t_0) = y_0 \in \mathbb{R}, \quad \mathrm{Re}(\mu) < 0 \tag{3}$$

for which, by setting as usual $q_n = h_n \mu$, the discrete problem (2) assumes the simpler form:

$$(A - q_n B)\mathbf{y}_n = \boldsymbol{\eta}_n$$

By setting I the identity matrix of size $r \times r$ and $C = A^{-1}B$, the previous equation is equivalent to the following ones:

$$(I - q_n C)\mathbf{y}_n = \boldsymbol{\eta}_n^{(1)} \equiv A^{-1}\boldsymbol{\eta}_n, \quad (C^{-1} - q_n I)\mathbf{y}_n = \boldsymbol{\eta}_n^{(2)} \equiv B^{-1}\boldsymbol{\eta}_n \tag{4}$$

By the way, we observe that in general a loss of sparsity could result when considering the matrices C and C^{-1} in place of A and B. This happens when the latter matrices are both sparse. However, for the methods we are interested in, either A or B are full matrices. By introducing now the function

$$\theta(q_n) = (I - q_n \gamma I)^{-1}, \quad \gamma > 0 \tag{5}$$

and by weighting the equations in (4) with weights $\theta(q_n)$ and $I - \theta(q_n)$, respectively, we then obtain the following equivalent formulation.

$$\begin{aligned} M(q_n)\mathbf{y}_n - \boldsymbol{\eta}(q_n) &\equiv (A(q_n) - q_n B(q_n))\mathbf{y}_n - \boldsymbol{\eta}(q_n) \\ &\equiv ((\theta(q_n)I + (I - \theta(q_n))\gamma C^{-1}) - q_n(\theta(q_n)C + (I - \theta(q_n))\gamma I))\mathbf{y}_n \\ &\quad -(\theta(q_n)\boldsymbol{\eta}_n^{(1)} + (I - \theta(q_n))\gamma\boldsymbol{\eta}_n^{(2)}) = \mathbf{0} \end{aligned} \tag{6}$$

which defines the *blended implicit method* associated with the block method (2). This name is due to the fact that the discrete problem is obtained as the 'blending' of two equivalent forms of the same basic block method. The key point concerning a blended implicit method is that its structure naturally induces the choice of a splitting for iteratively solving (6). In fact, one easily verifies that, for $q_n \approx 0$, $M(q_n) \approx I$, and, for $|q_n| \gg 1$, $M(q_n) \approx -q_n \gamma I$. Consequently, instead of solving (6), one may think to solve iteratively

$$N(q_n)\mathbf{y}_n^{(i+1)} = (N(q_n) - M(q_n))\mathbf{y}_n^{(i)} + \boldsymbol{\eta}(q_n), \quad i = 0, 1, \ldots \tag{7}$$

where

$$N(q_n) = I - q_n \gamma I \equiv \theta(q_n)^{-1} \tag{8}$$

For the methods described in References [1,3], this iteration has been proved to converge for all $\mathrm{Re}(q_n) \leqslant 0$, since, for all such values of q_n the spectral radius of the iteration matrix,

$$I - N(q_n)^{-1}M(q_n) \tag{9}$$

Copyright © 2004 John Wiley & Sons, Ltd.

say $\rho(q_n)$, is smaller than 1. As matter of fact, the *maximum amplification factor*, $\rho^* = \max_{x>0} \rho(\mathrm{i}x)$, with i denoting the imaginary unit, is smaller than 1. We observe that, from (5) to (8), one obtains that $\rho(0) = 0$, and $\rho^{(\infty)} \equiv \lim_{q_n \to \infty} \rho(q_n) = 0$, since in both cases the iteration matrix is the zero matrix. Consequently, one has that, because of the second property, iteration (7) is well-suited for stiff problems, since the *stiff amplification factor* $\rho^{(\infty)}$ is 0. Moreover,

$$\rho(q_n) \approx \tilde{\rho} q_n \quad \text{for } q_n \approx 0 \tag{10}$$

where $\tilde{\rho}$ is the *non-stiff amplification factor*. In Reference [3] the parameter γ has been chosen in order to minimize the maximum amplification factor ρ^*, thus giving $\rho^* < 1$ for all methods implemented in the code [1]. Moreover, it has been proved [3] that the eigenvalues of the iteration matrix (9) are given by

$$\frac{q_n(\lambda - \gamma)^2}{\lambda(1 - q_n\gamma)^2}, \quad \lambda \in \sigma(C) \tag{11}$$

where

$$\gamma = |\lambda_1| \equiv \min_{\lambda \in \sigma(C)} |\lambda|, \quad \rho^* = 1 - \cos\zeta_1, \quad \tilde{\rho} = 2|\lambda_1|\rho^* \tag{12}$$

with $\sigma(C)$ denoting the spectrum of the matrix $C \equiv A^{-1}B$, and ζ_1 the argument of λ_1. Finally, again from the arguments in Reference [3], one obtains that the spectral radius of the iteration matrix (9) is given by (see (12))

$$\rho(q_n) = \left| \frac{q_n(\lambda_1 - \gamma)^2}{\lambda_1(1 - q_n\gamma)^2} \right| \tag{13}$$

Coming back to problem (1), the blended iteration (7) generated by a blended implicit method now becomes:

$$\begin{aligned}\mathbf{y}_n^{(i+1)} = \mathbf{y}_n^{(i)} - \theta_n[\theta_n \left((I - \gamma C^{-1}) \otimes I_m \mathbf{y}_n^{(i)} - h_n(C - \gamma I) \otimes I_m \mathbf{f}_n^{(i)}\right) \\ + \gamma(C^{-1} \otimes I_m \mathbf{y}_n^{(i)} - h_n I \otimes I_m \mathbf{f}_n^{(i)}) + \boldsymbol{\eta}_n], \quad i = 0, 1, \ldots\end{aligned} \tag{14}$$

where $\mathbf{y}_n^{(i)} = (y_{1n}^{(i)}, \ldots, y_{rn}^{(i)})^{\mathrm{T}}$, $\mathbf{f}_n^{(i)} = (f_{1n}^{(i)}, \ldots, f_{rn}^{(i)})^{\mathrm{T}}$, $f_{jn}^{(i)} = f(t_{jn}, y_{jn}^{(i)})$, and the vector $\boldsymbol{\eta}_n$ only depends on (t_{0n}, y_{0n}). Finally,

$$\theta_n = I \otimes \Omega_n^{-1}, \quad \Omega_n = (I_m - h_n\gamma J_n) \tag{15}$$

where J_n is the Jacobian of f at (t_{0n}, y_{0n}). Consequently, if v iterations are performed to obtain convergence, the overall computational cost is approximately made up from four components:

- the evaluation of the Jacobian matrix J_n,
- the factorization of the $m \times m$ matrix Ω_n (see (15)),
- rv function evaluations, and
- $2rv$ system solvings with the factors of the matrix Ω_n.

Obviously, the relative computational cost of the first two entries with respect to the overall computational cost depends on the continuous problem and on v. In particular, their relative cost increases when v decreases. Therefore, when the blended iteration (14) converges rapidly, the overall computational cost of the iteration can be reduced significantly by means of one of the following approximations: $J_n \approx J_{n-1}$, and/or $\Omega_n \approx \Omega_{n-1}$. It is clear that (see (15)) in both cases a perturbation is introduced in the matrix θ_n and, therefore, the spectral radius of the corresponding iteration matrix turns out to be affected. In the following two sections, we shall study this aspect by means of a linear analysis, which relies on the particular structure of the discrete problem.

3. LINEAR ANALYSIS FOR THE BLENDED ITERATION WITH APPROXIMATE JACOBIAN

Let us consider the application of the method defined by (14) to the test problem:

$$y' = \mu(t)y, \quad y(t_0) = y_0 \in \mathbb{R}, \quad \mathrm{Re}(\mu(t)) < 0$$

If we set

$$\mu_n \equiv \mu(t_{0n}) = \mu_{n-1}(1 + \delta_n), \quad \delta_n \in \mathbb{C} \tag{16}$$

the approximate blended iteration, corresponding to the use of the previous Jacobian, is

$$\begin{aligned}\mathbf{y}_n^{(i+1)} = \mathbf{y}_n^{(i)} - \hat{\theta}_n[&\hat{\theta}_n((I - \gamma C^{-1})\mathbf{y}_n^{(i)} - h_n(C - \gamma I)\mathbf{f}_n^{(i)}) \\ &+ \gamma(C^{-1}\mathbf{y}_n^{(i)} - h_n\mathbf{f}_n^{(i)}) + \boldsymbol{\eta}_n], \quad i = 0, 1, \ldots\end{aligned} \tag{17}$$

where

$$\hat{\theta}_n = (1 - \gamma\hat{q}_n)^{-1}I, \quad \hat{q}_n \equiv h_n\mu_{n-1} \tag{18}$$

We shall consider the additional first order approximation $\mathbf{f}_n^{(i)} \approx \mu_n\mathbf{y}_n^{(i)}$ so that the iteration (17) can be rewritten as

$$\begin{aligned}\mathbf{y}_n^{(i+1)} &= \mathbf{y}_n^{(i)} - \hat{\theta}_n[(\hat{\theta}_n(I - \gamma C^{-1} - q_n(C - \gamma I)) + \gamma(C^{-1} - q_nI))\mathbf{y}_n^{(i)} + \boldsymbol{\eta}_n] \\ &= \mathbf{y}_n^{(i)} - \hat{\theta}_n[(\hat{\theta}_n(I - \gamma C^{-1} - \hat{q}_n(1 + \delta_n)(C - \gamma I)) \\ &\quad + \gamma(C^{-1} - \hat{q}_n(1 + \delta_n)I))\mathbf{y}_n^{(i)} + \boldsymbol{\eta}_n], \quad i = 0, 1, \ldots\end{aligned} \tag{19}$$

where $q_n \equiv h_n\mu_n = \hat{q}_n(1+\delta_n)$. The spectral radius of the corresponding iteration matrix depends, therefore, on both $\hat{q}_n$ and δ_n: let it be $\hat{\rho}(\hat{q}_n, \delta_n)$. The following result holds.

Copyright © 2004 John Wiley & Sons, Ltd.

Theorem 1
If (11)–(13) holds true, and $|\delta_n|$ is sufficiently small, then the spectral radius $\hat{\rho}(\hat{q}_n, \delta_n)$ of the iteration matrix corresponding to (19) is given by

$$\hat{\rho}(\hat{q}_n, \delta_n) = \left| \frac{\hat{q}_n}{\lambda_1(1-\gamma\hat{q}_n)^2} \left((\lambda_1-\gamma)^2 + \delta_n\lambda_1(\lambda_1 - \gamma^2\hat{q}_n)\right) \right|$$

Proof
The iteration matrix corresponding to (19) is given by (see (18))

$$\begin{aligned} I - \hat{\theta}_n^2(I - \gamma C^{-1} - \hat{q}_n(1+\delta_n)(C-\gamma I) + \gamma\hat{\theta}_n^{-1}(C^{-1} - \hat{q}_n(1+\delta_n)I)) \\ = \frac{\hat{q}_n}{(1-\gamma\hat{q}_n)^2} C^{-1}((C-\gamma I)^2 + \delta_n C(C - \gamma^2\hat{q}_n I)) \end{aligned}$$

Therefore, the corresponding spectral radius is given by

$$\hat{\rho}(\hat{q}_n, \delta_n) = \max_{\lambda \in \sigma(C)} \left| \frac{\hat{q}_n}{\lambda(1-\gamma\hat{q}_n)^2}((\lambda-\gamma)^2 + \delta_n\lambda(\lambda - \gamma^2\hat{q}_n)) \right|$$

We observe that, for any fixed $\lambda \in \sigma(C)$ (which is contained in $\mathbb{C}^-$) and for any fixed $\hat{q}_n \in \mathbb{C}^-$, the function

$$\left| \frac{\hat{q}_n}{\lambda(1-\gamma\hat{q}_n)^2}((\lambda-\gamma)^2 + \delta_n\lambda(\lambda - \gamma^2\hat{q}_n)) \right|$$

is analytical at $\delta_n = 0$ so that, for δ_n sufficiently small, the result follows from (11) and (13) since $\hat{\rho}(\hat{q}_n, 0) = \rho(\hat{q}_n)$. □

The previous theorem immediately implies that

$$\hat{\rho}(0, \delta_n) = 0, \quad \hat{\rho}^\infty(\delta_n) \equiv \lim_{\hat{q}_n \to \infty} \hat{\rho}(\hat{q}_n, \delta_n) = |\delta_n| \tag{20}$$

Consequently, even though in general $\hat{\rho}^\infty(\delta_n) > 0$, one is still able, by estimating $|\delta_n|$, to control the convergence properties of such iteration when $|\hat{q}_n| \gg 1$. On the other hand, when $\hat{q}_n \approx 0$ the following result holds true.

Theorem 2
If $\hat{q}_n \approx 0$, $\alpha > 0$ is a fixed parameter and (see (12)–(13))

$$|\delta_n| \leqslant \frac{\tilde{\rho}\alpha}{(1+\alpha)\tilde{\rho} + \gamma} \tag{21}$$

then $\hat{\rho}(\hat{q}_n, \delta_n)$ is approximately bounded by $\rho(q_n)(1+\alpha)$.

Copyright © 2004 John Wiley & Sons, Ltd.

Table I. Parameters of the methods used in the code BiM.

r	p	γ	ρ^*	$\tilde{\rho}$	x_1	x_2	$d_n^{\min}$	$d_n^{\max}$
3	4	0.7387	0.3398	0.5021	−1.4487	2.3593	0.90	1.10
4	6	0.8482	0.5291	0.8975	−1.4983	3.1163	0.91	1.09
6	8	0.7285	0.6299	0.9177	−1.4662	3.5197	0.92	1.08
8	10	0.6745	0.6885	0.9288	−1.4290	3.7538	0.93	1.07
10	12	0.6433	0.7276	0.9361	−1.3964	3.9104	0.94	1.06
12	14	0.6227	0.7560	0.9415	−1.3689	4.0240	0.95	1.05

Proof

From Theorem 1 it follows that, for $\hat{q}_n \approx 0$,

$$\hat{\rho}(\hat{q}_n, \delta_n) \approx |\hat{q}_n| \left| \frac{(\lambda_1 - \gamma)^2}{\lambda_1} + \delta_n \lambda_1 \right| \tag{22}$$

Moreover, since $|\delta_n|$ is bounded, then $q_n = \hat{q}_n(1 + \delta_n) \approx 0$ as well and, therefore, see (10), $\rho(q_n) \approx \tilde{\rho}|q_n| = \tilde{\rho}|\hat{q}_n||1 + \delta_n|$. From (12) and (21)–(22), it then follows that,

$$\hat{\rho}(\hat{q}_n, \delta_n) \approx |\hat{q}_n|\,|\tilde{\rho} + \delta_n\lambda_1| \leqslant |\hat{q}_n|(\tilde{\rho} + |\delta_n\lambda_1|) \leqslant |\hat{q}_n|\tilde{\rho}(1 - |\delta_n|)(1 + \alpha) \leqslant \rho(q_n)(1 + \alpha) \qquad \square$$

An immediate consequence of the previous theorem is that an estimate of $|\delta_n|$ is needed in order to control the perturbation on the spectral radius of the iteration matrix. From (16) we obtain $\delta_n = (\mu_n - \mu_{n-1})/\mu_{n-1}$. Consequently, an estimate of $|\mu_n - \mu_{n-1}|$ and of $|\mu_{n-1}|$ are needed. In general, when we are solving problem (1), we will need to estimate $\delta_n = \|J_n - J_{n-1}\|/\|J_{n-1}\|$. By considering a suitable vector $\mathbf{u} \in \mathbb{R}^m$, having unit norm, we then evaluate the vector $g_n = f(t_{0n}, y_{0n} + s \cdot \mathbf{u}) - f_{0n}$, with $s > 0$ a suitably small parameter, thus obtaining the following estimates:

$$\|J_n\|_\infty \approx \frac{1}{s}\|g_n\|_\infty, \quad \|J_n - J_{n-1}\|_\infty \approx \frac{1}{s}\|g_n - g_{n-1}\|_\infty$$

We observe that, for the linear autonomous equation $y' = Jy$, one obtains $\|g_n - g_{n-1}\|_\infty = 0$, so that the re-evaluation of the Jacobian is not needed, in such case, as one would expect.

Concerning the choice of the parameter α (see (21)) made in the code BiM, if p is the order of the method with blocksize r_p (see Table I) then, by using arguments similar to those used in Reference [1], the corresponding parameter, say α_p, is chosen as follows:

$$\alpha_4 = 5 \times 10^{-2}, \quad \alpha_p = (\alpha_{p-2})^{r_p/r_{p-2}}, \quad p = 6, 8, 10, 12, 14$$

4. THE BLENDED ITERATION WITH APPROXIMATE FACTORIZATION

We now study the case where, at step n, one considers the approximation,

$$\Omega_n \approx \Omega_{n-1} \tag{23}$$

Copyright © 2004 John Wiley & Sons, Ltd.

(see (14)–(15)), in order to not evaluate the new factorization. We shall again resort to a linear analysis, by applying the method to the test problem (3). In such a case, the blended iteration (14) becomes

$$\begin{aligned}\mathbf{y}_n^{(i+1)} &= \mathbf{y}_n^{(i)} - \theta_{n-1}[(\theta_{n-1}(I - \gamma C^{-1} - q_n(C - \gamma I)) + \gamma(C^{-1} - q_n I))\mathbf{y}_n^{(i)} + \boldsymbol{\eta}_n] \\ &= \mathbf{y}_n^{(i)} - \theta_{n-1}[(\theta_{n-1}(I - \gamma C^{-1} - q_{n-1}d_n(C - \gamma I)) \\ &\quad + \gamma(C^{-1} - q_{n-1}d_n I))\mathbf{y}_n^{(i)} + \boldsymbol{\eta}_n], \quad i = 0, 1, \ldots \end{aligned} \tag{24}$$

where

$$q_n \equiv h_n \mu = \left(\frac{h_n}{h_{n-1}}\right) q_{n-1} \equiv d_n q_{n-1} \tag{25}$$

Therefore, the spectral radius, say $\bar{\rho}$, of the corresponding iteration matrix will now depend on both q_{n-1} and d_n. The following theorem holds true.

Theorem 3
If (11)–(13) holds true and $|d_n - 1|$ is sufficiently small, then the spectral radius $\bar{\rho}$ of the iteration matrix corresponding to (24) is given by

$$\bar{\rho}(q_{n-1}, d_n) = \left| \frac{q_{n-1}}{\lambda_1(1 - \gamma q_{n-1})^2} \left((\lambda_1 - \gamma)^2 + (d_n - 1)\lambda_1(\lambda_1 - \gamma^2 q_{n-1})\right) \right|$$

Proof
We observe that iteration (24) formally coincides with iteration (18)–(19) with the substitutions $\hat{q}_n \leftarrow q_{n-1}$ and $\delta_n \leftarrow d_n - 1$. Consequently, from Theorem 1, one immediately obtains $\bar{\rho}(q_{n-1}, d_n) = \hat{\rho}(q_{n-1}, d_n - 1)$, and hence the result follows. □

From the previous theorem one immediately obtains that (see (20))

$$\bar{\rho}(0, d_n) = 0, \qquad \lim_{q_{n-1} \to \infty} \bar{\rho}(q_{n-1}, d_n) = |d_n - 1| \tag{26}$$

so that $d_n \in (0, 2)$, in order to have a satisfactory behaviour for stiff problems. Moreover (compare with (22)), for $q_{n-1} \approx 0$, which we shall assume hereafter, one obtains

$$\bar{\rho}(q_{n-1}, d_n) \approx |q_{n-1}| \left| \frac{(\lambda_1 - \gamma)^2}{\lambda_1} + (d_n - 1)\lambda_1 \right| \equiv |q_{n-1}| \tilde{\rho}(d_n) \tag{27}$$

Finally, if the factors of the matrix θ_n are computed, then the spectral radius of the corresponding iteration matrix is given by $\bar{\rho}(q_n, 1) \equiv \rho(q_n)$ (see (13)).

The following analysis is devoted to provide an estimate of the number, say $\bar{v}$, of iterations in (24), depending on the number of iterations v that would have been required without the approximation (23). The latter number can be estimated from the iteration parameters, as shown in Reference [1]. In order to derive the criterion used in the code BiM, we shall look for values of d_n (see (25)) such that,

$$\bar{v} \leqslant \beta v, \quad \beta = 1 + m(6rv)^{-1} \tag{28}$$

where r is the blocksize of the blended implicit method and m is the size of the continuous problem. Indeed, for such value of the parameter β, one verifies that the cost of the linear algebra involved in the blended iteration with the approximation (23) is less than or equal to the cost of the exact iteration plus the cost to factor Ω_n (evidently, for sake of simplicity, the cost of function and Jacobian evaluations has been neglected). If the same stopping criterion must be satisfied, then $\bar{\rho}(q_{n-1}, d_n)^{\bar{v}} = \bar{\rho}(q_n, 1)^v$ and, therefore,

$$\bar{v} = v \frac{\log \bar{\rho}(q_n, 1)}{\log \bar{\rho}(q_{n-1}, d_n)}$$

Consequently, the inequality in (28) can be written as

$$\frac{\bar{\rho}(q_{n-1}, d_n)^\beta}{\bar{\rho}(q_n, 1)} \leqslant 1 \tag{29}$$

We observe that (see (25)), since d_n is bounded, then $q_{n-1} \approx 0$ implies $q_n \approx 0$ as well. Therefore (see (27)),

$$\bar{\rho}(q_{n-1}, d_n) \approx |q_{n-1}| \tilde{\rho}(d_n) \approx \left(\frac{\rho_{n-1}}{\tilde{\rho}} \right) \tilde{\rho}(d_n), \quad \bar{\rho}(q_n, 1) \approx |q_n| \tilde{\rho} \approx \rho_{n-1} d_n \tag{30}$$

where ρ_{n-1} is the spectral radius of the iteration matrix at the previous integration step. From (29) and (30), we then obtain that d_n must satisfy

$$\frac{\tilde{\rho}(d_n)^\beta}{d_n} \leqslant \rho_{n-1} \left(\frac{\tilde{\rho}}{\rho_{n-1}} \right)^\beta \tag{31}$$

Moreover, from (27), one easily obtains that (see (12))

$$\tilde{\rho}(d_n) \equiv \left| \frac{(\lambda_1 - \gamma)^2}{\lambda_1} + (d_n - 1)\lambda_1 \right| = \gamma (d_n^2 + 2x_1 d_n + x_2)^{1/2} \tag{32}$$

where, $x_1 = (1 - 2\cos\zeta_1)\cos 2\zeta_1 - 2\sin\zeta_1 \sin 2\zeta_1$, and $x_2 = 5 - 4\cos\zeta_1$. The values of x_1 and x_2 for the methods implemented in the code BiM are listed in Table I. From (31) and (32), we then obtain that the stepsizes ratio d_n must satisfy

$$\frac{(d_n^2 + 2x_1 d_n + x_2)^{\frac{\beta}{2}}}{d_n} \leqslant \rho_{n-1} \left(\frac{\tilde{\rho}}{\gamma \rho_{n-1}} \right)^\beta \tag{33}$$

Copyright © 2004 John Wiley & Sons, Ltd.

Only one of the following two cases may then occur:

$$(1) \quad d_n \geqslant 1; \quad (2) \quad d_n < 1$$

In the first case, i.e. when the stepsize has been increased, from Table I it is easy to verify that inequality (33) is satisfied for $\beta = 1$ and $d_n \in [1, 2)$. Clearly, from (28) one obtains that this will hold true for all $\beta \geqslant 1$. Consequently, (see (25)) in the code `BiM` re-factorization is avoided, when the stepsize has been increased, unless $d_n > d_n^{\max}$ (see Table I), where the last inequality is aimed to guarantee fast convergence for stiff problems (see (26)).

In the second case, i.e. when the stepsize has been decreased, we can assume $1 > d_n \geqslant d_n^{\min}$, for a fixed $d_n^{\min} > 0$ (see Table I, for the values used in the code `BiM`). In such a case, one derives that a sufficient condition for (33) to be satisfied is given by

$$d_n^2 + 2x_1 d_n + x_3 \leqslant 0 \tag{34}$$

where

$$x_3 = x_2 - (\bar{d}_n \rho_{n-1})^{\frac{2}{\beta}} (\tilde{\rho}/(\gamma \rho_{n-1}))^2$$

Consequently, in the design of the code `BiM`, re-factorization is avoided, when the stepsize is reduced, unless (34) turns out to be not satisfied or $d_n < d_n^{\min}$.

ACKNOWLEDGEMENTS

The authors are very indebted to two unknown referees, for their useful suggestions and comments.

REFERENCES

1. Brugnano L, Magherini C. The `BiM` code for the numerical solution of ODEs. *Journal of Computational and Applied Mathematics* 2004; **164–165**:145–158. Code available at http://www.math.unifi.it/~brugnano/BiM/index.html
2. Brugnano L. Blended block BVMs (B_3 VMs): a family of economical implicit methods for ODEs. *Journal of Computational and Applied Mathematics* 2000; **116**:41–62.
3. Brugnano L, Magherini C. Blended implementation of block implicit methods for ODEs. *Applied Numerical Mathematics* 2002; **42**:29–45.
4. Brugnano L, Trigiante D. Block Implicit Methods for ODEs. In *Recent Trends in Numerical Analysis*, Trigiante D (ed.). Nova Science Publ. Inc.: New York, 2001; 81–105.
5. Brugnano L, Trigiante D. *Solving Differential Problems by Multistep Initial and Boundary Value Methods*. Taylor & Francis: London, 1998.

Copyright © 2004 John Wiley & Sons, Ltd.