

# BLOCK IMPLICIT METHODS FOR ODES.\*

LUIGI BRUGNANO<sup>†</sup> AND DONATO TRIGIANTE<sup>‡</sup>

*Dedicated to Professor Ilio Galligani on the occasion of his 65th birthday.*

**Abstract.** In this paper we study some important features of block implicit methods for ODEs. By using a unified framework, properties such as stability, blending of methods,  $A$ -convergence and parallelism are characterized. Major attention has been devoted to the implementation properties in view of their use in solving large-size problems.

**Key words.** Numerical Methods for ODEs, Stiff Problems, Implicit Methods, Iterative Solution of Linear Systems, Parallel Computing.

**AMS subject classifications.** 65L06, 65L05, 65L20, 65H10, 65F10.

**1. Introduction.** The numerical approximation of ODEs is still a very active field of research, as shown by the very rich amount of significant contributions in the last forty years. The demand for new methods well-suited for particular classes of applicative problems is relevant, as well as the need for the efficient implementation of the methods on modern computers, including parallel computers.

Many of the obtained results have been collected in several books, among which we quote [2, 7, 10, 12, 15, 17, 23]. Across the years, the required properties for the numerical methods have had an interesting evolution. Until the fifties, major attention was devoted to the order of accuracy. Later on, stability and convergence became more important. In the last years, properties of the methods such as the definition of efficient splittings, degree of parallelism, etc., more relevant for the implementation purposes, have become focal. This because such features are essential when solving large-size problems. In this paper we shall deal with such topics, namely the definition of methods that, in addition to classical requirements, do have favorable properties from the point of view of the implementation. In particular, we focus our attention on *r-block implicit methods* for approximating the solution of the problem

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

over the discrete set of points  $t_i = t_0 + ih$ ,  $i = 1, 2, \dots, N$ , where  $h = (T - t_0)/N$  is the stepsize. An  $r$ -block method is a method which provides the approximations at the first  $r$  mesh-points as the solution of the following discrete problem,

$$(2) \quad F(\mathbf{y}) \equiv A \otimes I_m \mathbf{y} - hB \otimes I_m \mathbf{f} + \mathbf{a} \otimes y_0 - h\mathbf{b} \otimes f_0 = \mathbf{0},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_r \end{pmatrix},$$

---

\*Work supported by GNIM-INDAM and MURST.

<sup>†</sup>Dipartimento di Matematica "U. Dini", Viale Morgagni 67/A, 50134 Firenze (Italy),  
E-mail: na.brugnano@na-net.ornl.gov

<sup>‡</sup>Dipartimento di Energetica "S. Stecco", Via C. Lombroso 6/17, 50134 Firenze (Italy),  
E-mail: na.dtrigliante@na-net.ornl.gov

$y_i$  is the approximation to  $y(t_i)$ , and  $f_i = f(t_i, y_i)$ . Finally,  $A$  and  $B$  are  $r \times r$  matrices,  $\mathbf{a}$ ,  $\mathbf{b}$  are vectors in  $\mathbb{R}^r$ , and in general, for any integer  $j$ ,  $I_j$  will denote the identity matrix of size  $j$ . Moreover, in the following we shall also consider the augmented matrices

$$(3) \quad \hat{A} = [\mathbf{a} \ A], \quad \hat{B} = [\mathbf{b} \ B].$$

Methods in the form (2) are a generalization of those proposed originally in [3, 4, 27], where the normalization  $A = I_r$  is considered. With such a normalization, they are also related to Runge-Kutta methods (see, for example, [5, 8, 9, 10, 16, 17]). We prefer, however, to consider the more general formulation (2) for two main reasons:

- in such formulation they are also related to block Boundary Value Methods (see [7]),
- it presents some advantages in discussing the implementation issues.

The latter point will be discussed in Sections 5 and 6. For such purpose, it will be hereafter assumed that both the two matrices  $A$  and  $B$  are nonsingular. Before that, in Sections 2 and 3 we collect some results, some of them known, concerning the construction of the methods. Additional considerations are reported in Section 4. Most results in the two sections dedicated to the implementation of the methods are new. Finally, in Section 7 we summarize the main facts and provide some final remark.

**2. Construction of block implicit methods.** An implicit  $r$ -block method is characterized by the fact that the rows of the augmented matrices (3) are defined through the characteristic polynomials of a suitable set of (distinct)  $r$ -step LMFs:

$$(4) \quad \hat{A} = \begin{pmatrix} \alpha_0^{(1)} & \dots & \alpha_r^{(1)} \\ \vdots & & \vdots \\ \alpha_0^{(r)} & \dots & \alpha_r^{(r)} \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} \beta_0^{(1)} & \dots & \beta_r^{(1)} \\ \vdots & & \vdots \\ \beta_0^{(r)} & \dots & \beta_r^{(r)} \end{pmatrix}.$$

On the coefficients of such matrices we shall impose order conditions. Let us consider at first the case where all such methods have (at least) order  $p \geq 1$ . The following result then holds true.

**THEOREM 1.** *Let the matrices (3) satisfy the following set of equations,*

$$(5) \quad \hat{A} \hat{\mathbf{q}}_i = i \hat{B} \hat{\mathbf{q}}_{i-1}, \quad i = 0, \dots, p,$$

where

$$(6) \quad \hat{\mathbf{q}}_{-1} = \mathbf{0}, \quad \hat{\mathbf{q}}_i = \begin{pmatrix} 0^i \\ 1^i \\ \vdots \\ r^i \end{pmatrix} \equiv \begin{pmatrix} 0^i \\ \mathbf{q}_i \end{pmatrix}, \quad i = 0, 1, \dots$$

Then the LMFs defining the block method have a truncation error which is at least  $O(h^{p+1})$ .

*Proof.* The equations (5) are nothing but the usual order  $p$  conditions for LMFs, simultaneously imposed for all the  $r$  LMFs corresponding to (4).  $\square$

From the above result, when all methods in (4) are consistent ( $p \geq 1$ ), one obtains relations between the first columns of the augmented matrices  $\hat{A}$  and  $\hat{B}$ , and the square matrices  $A$  and  $B$ , respectively. In fact, from (5) we obtain, for  $i = 0, 1$ , the following conditions:

$$(7) \quad \mathbf{a} = -A\mathbf{e}, \quad \mathbf{b} = A\mathbf{q}_1 - B\mathbf{e},$$

where  $\mathbf{e} \equiv \mathbf{q}_0$  denotes the vector with all unit entries (see (6)). Having spent the consistency conditions to relate the vectors  $\mathbf{a}$  and  $\mathbf{b}$  to the matrices  $A$  and  $B$ , we can use the subsequent order conditions to define the entries of such matrices. In particular, it is an easy matter to verify the following result.

**COROLLARY 1.** *Let the matrices defined in (3) satisfy (7) and the following set of equations,*

$$(8) \quad A\mathbf{q}_i = iB\mathbf{q}_{i-1}, \quad i = 2, \dots, p.$$

*Then the LMFs defining the block method have a truncation error which is at least  $O(h^{p+1})$ .*

Note that if (8) holds also true for  $i = 1$ , from (7) it follows that  $\mathbf{b} = \mathbf{0}$ . Such condition would simplify the derivation of the methods, moreover it implies the  $L$ -stability of  $A$ -stable methods. Nevertheless, we prefer not to impose it directly, in order to obtain a more general framework.

Let now define the following matrices:

$$D_j = \text{diag} ( 1 \quad 2 \quad \dots \quad j ), \quad Q_j = ( \mathbf{q}_1 \quad \dots \quad \mathbf{q}_j ), \quad j = 1, 2, \dots$$

As a consequence, the order conditions (8) can be rewritten as

$$(9) \quad AD_r Q_{p-1} = BQ_{p-1} (I_{p-1} + D_{p-1}).$$

The following result then holds true (see also [6]).

**THEOREM 2.** *If  $p = r + 1$ , the matrix  $A^{-1}B$  is uniquely determined.*

*Proof.* In fact, when  $p = r + 1$ , the matrix  $Q_r$  in (9) is a nonsingular Vandermonde matrix. Consequently, one obtains that

$$(10) \quad A^{-1}B = D_r Q_r (I_r + D_r)^{-1} Q_r^{-1},$$

whose right-hand side only depends on  $r$ .  $\square$

We observe that the matrix  $A^{-1}B$  in (10) is similar to a Frobenius-type matrix. In fact, we have that

$$(11) \quad A^{-1}B = Q_r (Q_r^{-1} D_r Q_r (I_r + D_r)^{-1}) Q_r^{-1}$$

and

$$(12) \quad Q_r^{-1}D_rQ_r = \begin{pmatrix} & & -\alpha_0 \\ 1 & & -\alpha_1 \\ & \ddots & \vdots \\ & & 1 & -\alpha_{r-1} \end{pmatrix},$$

where

$$\sum_{i=0}^r \alpha_i x^i \equiv (x-1)(x-2)\cdots(x-r)$$

is the  $r$ -th order Wilkinson's polynomial. Consequently,

$$(13) \quad Q_r^{-1}D_rQ_r(I_r + D_r)^{-1} = \begin{pmatrix} & & -\frac{\alpha_0}{r+1} \\ \frac{1}{2} & & -\frac{\alpha_1}{r+1} \\ & \ddots & \vdots \\ & & \frac{1}{r} & -\frac{\alpha_{r-1}}{r+1} \end{pmatrix} = G^{-1} \begin{pmatrix} & & -\frac{1!\alpha_0}{(r+1)!} \\ 1 & & -\frac{2!\alpha_1}{(r+1)!} \\ & \ddots & \vdots \\ & & 1 & -\frac{r!\alpha_{r-1}}{(r+1)!} \end{pmatrix} G,$$

where

$$(14) \quad G = \text{diag} ( 1! \quad 2! \quad \dots \quad r! ),$$

From (2) and (7), it follows that methods having the same matrix  $A^{-1}B$  provide the same discrete solution. For this reason, we shall call as *equivalent* methods sharing the same matrix

$$(15) \quad C = A^{-1}B.$$

In what follows we shall exploit the possibility of writing a specific method by using different equivalent forms. Moreover, it is important to point out that equivalent methods do have the same order and stability properties. For an  $r$ -block method, it is customary to define such properties by taking into account only the last entry of the solution vector  $\mathbf{y}$ , because it represents the starting point for subsequent applications of the same method. In particular, concerning the stability, if we consider the usual test equation

$$(16) \quad y' = \lambda y, \quad y(t_0) = \eta \neq 0,$$

the application of the block method (2) generates the discrete problem

$$(17) \quad (A - qB)\mathbf{y} = (q\mathbf{b} - \mathbf{a})\eta, \quad q = h\lambda.$$

From such equation one obtains that the usual property of  $A$ -stability is equivalent to require, for all  $q \in \mathbf{C}^-$ ,

$$|y_r(q)| \equiv |\mathbf{e}_r^T (A - qB)^{-1} (q\mathbf{b} - \mathbf{a})\eta| < |\eta|,$$

where  $\mathbf{e}_r$  is the last unit vector in  $\mathbb{R}^r$ , i.e.

$$(18) \quad \operatorname{Re}(q) < 0 \quad \Rightarrow \quad g(q) \equiv |\mathbf{e}_r^T (A - qB)^{-1} (q\mathbf{b} - \mathbf{a})| < 1.$$

A necessary requirement for this purpose, is to have problem (17) well-posed for all such  $q$ . For this reason, we give the following definition:

DEFINITION 1. A block method is said to be *pre-stable* if the spectrum of the corresponding matrix pencil  $A - \mu B$  is contained in  $\mathbb{C}^+$ .

This fact implies that the result of Theorem 2 is useful only to define pre-stable methods up to  $r = 8$ ; as matter of fact, by direct inspection one verifies that the matrix on the right-hand side of (10) has eigenvalues with negative real part, when  $r \geq 9$ . Consequently, the corresponding method cannot be pre-stable: in fact, the spectrum of the pencil  $(A - \mu B)$  coincides with that of  $C^{-1}$  (see (15)), since both the two matrices  $A$  and  $B$  are assumed to be nonsingular.

In order to obtain alternative criteria for choosing the matrix  $C$ , we shall relax the order conditions for the LMFs on each row of the block method. In particular, it will be convenient to impose only an order  $r$  condition: i.e. (see (9))

$$(19) \quad AD_r Q_{r-1} = BQ_{r-1} (I_{r-1} + D_{r-1}).$$

It remains one more condition to be imposed. It will be used to fix the spectrum of the matrix  $C$ . Such a condition will be written as

$$(20) \quad AQ_r \hat{\mathbf{d}} = BQ_r \mathbf{e}_r,$$

where  $\hat{\mathbf{d}}$  is a suitably chosen vector. The two conditions (19)-(20) can be also written as (see (15))

$$Q_r \left( \begin{array}{c} \mathbf{0}^T \\ (I_{r-1} + D_{r-1})^{-1} \end{array} \right) = CQ_r \left( \begin{array}{c} I_{r-1} \\ \mathbf{0}^T \end{array} \right),$$

and

$$Q_r \hat{\mathbf{d}} = CQ_r \mathbf{e}_r,$$

respectively. Altogether, we then obtain:

$$Q_r \left( \left( \begin{array}{c} \mathbf{0}^T \\ (I_{r-1} + D_{r-1})^{-1} \end{array} \right) \middle| \hat{\mathbf{d}} \right) = CQ_r \left( \left( \begin{array}{c} I_{r-1} \\ \mathbf{0}^T \end{array} \right) \middle| \mathbf{e}_r \right) = CQ_r.$$

Consequently, considering that  $Q_r$  is a nonsingular Vandermonde matrix, we have that (see (14))

$$(21) \quad C = Q_r \left( \left( \begin{array}{c} \mathbf{0}^T \\ (I_{r-1} + D_{r-1})^{-1} \end{array} \right) \middle| \hat{\mathbf{d}} \right) Q_r^{-1} = Q_r G^{-1} F G Q_r^{-1}$$

where

$$(22) \quad F = \left( \left( \begin{array}{c} \mathbf{0}^T \\ I_{r-1} \end{array} \right) \middle| -\mathbf{d} \right), \quad \mathbf{d} \equiv \begin{pmatrix} d_0 \\ \vdots \\ d_{r-1} \end{pmatrix} = -\frac{1}{r!} G \hat{\mathbf{d}}.$$

One then concludes that the characteristic polynomial of  $C$  is given by

$$(23) \quad d(z) = \sum_{i=0}^r d_i z^i, \quad d_r = 1.$$

Moreover, from (22) it follows that the vector  $\hat{\mathbf{d}}$  in (20) and (21) is given by

$$\hat{\mathbf{d}} = -r! G^{-1} \mathbf{d} \equiv -r! \begin{pmatrix} \frac{d_0}{1!} \\ \vdots \\ \frac{d_{r-1}}{r!} \end{pmatrix}.$$

By the way, we observe that the choice (see (10)–(13))

$$\hat{\mathbf{d}} = \frac{-1}{r+1} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{r-1} \end{pmatrix}$$

corresponds to the method obtained by imposing the order  $r+1$  conditions on each row.

Let us now study the problem of appropriately choosing the characteristic polynomial  $d(z)$  in (23). We surely will choose it in order to have all the roots contained in  $\mathbb{C}^+$ , so that the method is pre-stable. This is not enough, however, to define a “good” method. An additional requirement may be obtained by considering that, in order the Absolute stability region of the method (see (18)),

$$\mathcal{D} = \{q \in \mathbb{C} : g(q) < 1\},$$

to be unbounded, it would be preferable to have the point at  $\infty$  inside  $\mathcal{D}$ . From (18), such a requirement is easily proved to be equivalent to

$$(24) \quad g(\infty) \equiv |\mathbf{e}_r^T B^{-1} \mathbf{b}| < 1.$$

The next result relates  $g(\infty)$  to the characteristic polynomial (23).

THEOREM 3.

$$(25) \quad g(\infty) = \left| \frac{1}{d_0} \sum_{i=0}^r \frac{r^i}{i!} d_i \right|.$$

*Proof.* From (7), (15), (14), (21) and (22), we obtain that

$$\begin{aligned} B^{-1}\mathbf{b} &= B^{-1}A A^{-1}\mathbf{b} = C^{-1}(\mathbf{q}_1 - C\mathbf{e}) = C^{-1}D_r\mathbf{e} - \mathbf{e} = Q_r G^{-1} F^{-1} G Q_r^{-1} D_r \mathbf{e} - \mathbf{e} \\ &= Q_r G^{-1} F^{-1} G \begin{pmatrix} \mathbf{q}_0 & \dots & \mathbf{q}_{r-1} \end{pmatrix}^{-1} \mathbf{e} - \mathbf{e} = Q_r G^{-1} F^{-1} G \mathbf{e}_1 - \mathbf{e} \\ &= Q_r G^{-1} F^{-1} \mathbf{e}_1 - \mathbf{e} = -Q_r G^{-1} \begin{pmatrix} \frac{d_1}{d_0} \\ \vdots \\ \frac{d_r}{d_0} \end{pmatrix} - \mathbf{e} = -Q_r \begin{pmatrix} \frac{d_1}{1!d_0} \\ \vdots \\ \frac{d_r}{r!d_0} \end{pmatrix} - \mathbf{e}. \end{aligned}$$

Consequently, one obtains

$$\begin{aligned} g(\infty) &= |\mathbf{e}_r^T B^{-1}\mathbf{b}| = \left| \begin{pmatrix} r^1 & \dots & r^r \end{pmatrix} \begin{pmatrix} \frac{d_1}{1!d_0} \\ \vdots \\ \frac{d_r}{r!d_0} \end{pmatrix} + 1 \right| = \left| 1 + \frac{1}{d_0} \sum_{i=1}^r \frac{r^i}{i!} d_i \right| \\ &= \left| \frac{1}{d_0} \sum_{i=0}^r \frac{r^i}{i!} d_i \right|. \end{aligned}$$

□

The above result then provides an additional condition for the characteristic polynomial (23) of the matrix  $C$ , namely (see (24)-(25))

$$(26) \quad \left| \frac{1}{d_0} \sum_{i=0}^r \frac{r^i}{i!} d_i \right| < 1.$$

A further requirement for choosing the polynomial  $d(z)$  follows by observing that, from (7) and (15)–(17), one obtains

$$(27) \quad y_r(q) = \frac{\det(M(q))}{\det(I_r - qC)} \eta \approx e^{rq} \eta,$$

where  $M(q)$  is the matrix obtained from the matrix  $I_r - qC$ , whose last column has been substituted by  $\mathbf{e} + q(\mathbf{q}_1 - C\mathbf{e})$ . From the above equation, one obtains

$$(28) \quad e^{rz} \approx \frac{\det(M(z))}{\det(I_r - zC)} = \frac{\varphi(z)}{z^r d(z^{-1})} \equiv \frac{\varphi(z)}{\mu(z)},$$

where  $\varphi(z) = \det(-M(z))$  is a polynomial of maximum degree  $r$ , and

$$(29) \quad \mu(z) = \sum_{i=0}^r d_i z^{r-i}, \quad d_r = 1,$$

is a polynomial of exact degree  $r$ , since we assume  $(I_r - qC)$  to be nonsingular. The above arguments can then be summarized by saying that the characteristic polynomial of the matrix  $C$  (see (23)) coincides with the inverse of the polynomial at the denominator of a rational approximation to the exponential. By considering that, from (19), the above approximation to  $e^{rz}$  is at least  $O(z^{r+1})$ , the following result easily follows.

**THEOREM 4.** *Let the denominator  $\mu(z)$  of the rational approximation (28) be given (i.e., the characteristic polynomial (23) is given). Then, the polynomial  $\varphi(z)$  is uniquely determined.*

We observe that, from (27)-(29), one obtains that (24) is always satisfied, when  $\deg(\varphi) < \deg(\mu) \equiv r$ : in such a case, in fact, one obtains  $g(\infty) = 0$  and  $A$ -stability implies  $L$ -stability. A complementary result is the following (see also [27]).

**THEOREM 5.** *If  $\mu(z) \equiv \varphi(-z)$  and all the roots of  $\mu(z)$  have positive real part, then the corresponding block method is perfectly  $A$ -stable and  $g(\infty) = 1$ .*

In other words, the above result states that, under the above condition of symmetry for  $\mu(z)$  and  $\varphi(z)$ , pre-stability and  $A$ -stability are equivalent. Moreover, this result is complementary to the condition (26).

All the above facts then imply that a partial list of good criteria for choosing the characteristic polynomial  $d(z)$  in (23) are the following:

- it must have all the roots contained in  $\mathbb{C}^+$  (pre-stability);
- $z^r d(z^{-1})$  must be the denominator of a (at least)  $O(z^{r+1})$  rational approximation to  $e^{rz}$ ;
- the polynomial should satisfy (26) (point at  $\infty$  inside the absolute stability region  $\mathcal{D}$ ). Alternatively, the conditions of Theorem 5 should be satisfied.

In view of the above criteria, and in particular of the second one, in the next section we shall review some rational approximations to the exponential. Before that, we conclude this section by considering the problem of determining the order of a block method. For this purpose, let us denote by

$$\hat{\mathbf{y}} = \begin{pmatrix} y(t_1) \\ \vdots \\ y(t_r) \end{pmatrix}, \quad \hat{\mathbf{f}} = \begin{pmatrix} f(t_1, y(t_1)) \\ \vdots \\ f(t_r, y(t_r)) \end{pmatrix},$$

where  $y(t)$  is the solution of problem (1). From (2) one then obtains that

$$A \otimes I_m \hat{\mathbf{y}} - hB \otimes I_m \hat{\mathbf{f}} + \mathbf{a} \otimes y_0 - h\mathbf{b} \otimes f_0 = \boldsymbol{\tau},$$

where  $\boldsymbol{\tau}$  is the vector with the truncation errors of the method. By assuming that  $y(t)$  is suitably smooth, the entries of the latter vector are given by

$$\begin{aligned} \tau_i &= \sum_{j>r} \frac{y^{(j)}(t_n)}{j!} h^j \left( \sum_{k=0}^r k^{j-1} (k\alpha_k^{(i)} - j\beta_k^{(i)}) \right) \\ (30) \quad &\equiv \sum_{j>r} y^{(j)}(t_n) h^j v_{ji}, \quad i = 1, \dots, r, \end{aligned}$$

because of the conditions (19). Consequently, we obtain that

$$A \otimes I_m(\hat{\mathbf{y}} - \mathbf{y}) - hB \otimes I_m(\hat{\mathbf{f}} - \mathbf{f}) = \boldsymbol{\tau}.$$

By introducing the vector  $\hat{\mathbf{e}} = \hat{\mathbf{y}} - \mathbf{y}$ , one then concludes that it satisfies the equation

$$(31) \quad (A \otimes I_m - hB \otimes I_m \hat{\mathbf{J}}) \hat{\mathbf{e}} = \boldsymbol{\tau},$$

where

$$(32) \quad \hat{\mathbf{J}} = \begin{pmatrix} \hat{J}_1 & & \\ & \ddots & \\ & & \hat{J}_r \end{pmatrix},$$

$$\hat{J}_i = \int_0^1 J(t_i, sy(t_i) + (1-s)y_i) ds \equiv J_0 + O(h),$$

$J(t, y) = \frac{\partial}{\partial y} f(t, y)$  and  $J_0 = J(t_0, y_0)$ . The order of the block method is then defined according to the next definition.

**DEFINITION 2.** The block method corresponding to (31) has order  $p$  provided that  $\hat{e}_r = O(h^{p+1})$ , where  $\hat{e}_r$  is the last entry of  $\hat{\mathbf{e}}$ .

It is obvious that, from (30) and (31), we have that the order of the method is  $p \geq r$ . In general, the relations between the order conditions (30) and the global order of the method may be very entangled, as the Butcher theory for Runge-Kutta methods (see, for example, [12]) shows. Nevertheless, in case we look for values of  $p$  only slightly greater than  $r$ , the following result may be useful.

**THEOREM 6.** Consider the following possible cases for the method corresponding to (30)-(31)

- 0)  $\mathbf{e}_r^T A^{-1} \mathbf{v}_{r+1} \neq 0$ ;
- 1)  $\mathbf{e}_r^T A^{-1} \mathbf{v}_{r+1} = 0$ , and  $\mathbf{e}_r^T A^{-1} \mathbf{v}_{r+2} \neq 0$  or  $\mathbf{e}_r^T C A^{-1} \mathbf{v}_{r+1} \neq 0$ ;
- 2)  $\mathbf{e}_r^T A^{-1} \mathbf{v}_{r+1} = 0$ ,  $\mathbf{e}_r^T A^{-1} \mathbf{v}_{r+2} = 0$ , and  $\mathbf{e}_r^T C A^{-1} \mathbf{v}_{r+1} = 0$ ,

where  $\mathbf{v}_j = (v_{j1} \ \dots \ v_{jr})^T$ . Then the global order of the method is exactly  $p = r+i$  in cases  $i = 0, 1$ , and  $p \geq r+2$  in case 2.

*Proof.* From (15), (30) and (31)-(32), by posing

$$\hat{\mathbf{y}}^{(j)} = (y^{(j)}(t_1) \ \dots \ y^{(j)}(t_r))^T,$$

one obtains

$$\begin{aligned} \hat{\mathbf{e}} &= (I_r \otimes I_m - hC \otimes I_m \hat{\mathbf{J}})^{-1} A^{-1} \otimes I_m \boldsymbol{\tau} \\ &= h^{r+1} (A^{-1} \mathbf{v}_{r+1}) \otimes I_m \hat{\mathbf{y}}^{(r+1)} + \\ &\quad h^{r+2} \left( (A^{-1} \mathbf{v}_{r+2}) \otimes I_m \hat{\mathbf{y}}^{(r+2)} + C \otimes I_m \hat{\mathbf{J}} (A^{-1} \mathbf{v}_{r+1}) \otimes I_m \hat{\mathbf{y}}^{(r+1)} \right) + O(h^{r+3}) \\ &= h^{r+1} (A^{-1} \mathbf{v}_{r+1}) \otimes I_m \hat{\mathbf{y}}^{(r+1)} + \\ &\quad h^{r+2} \left( (A^{-1} \mathbf{v}_{r+2}) \otimes I_m \hat{\mathbf{y}}^{(r+2)} + (C A^{-1} \mathbf{v}_{r+1}) \otimes J_0 \hat{\mathbf{y}}^{(r+1)} \right) + O(h^{r+3}), \end{aligned}$$

from which, in view of Definition 2, the thesis easily follows.  $\square$

**3. Approximating the exponential.** In this section we review some classical rational approximation to the exponential. In particular, we look for polynomials described by (29), i.e. for characteristic polynomials (23), suitable to define “good” block methods, according to the criteria previously introduced.

**3.1. Padé approximation.** One of the most classical rational approximation to the exponential is the Padé  $(\nu, r)$ , where (see (28))

$$(33) \quad \varphi(z) \equiv \varphi_{\nu,r}(rz), \quad \mu(z) \equiv \mu_{\nu,r}(rz)$$

are polynomials of degree  $\nu$  and  $r$ , respectively, such that

$$\varphi_{\nu,r}(z) = \mu_{\nu,r}(z)e^z + O(z^{\nu+r+1}).$$

The expression of the two polynomial is well-known, and is given by

$$(34) \quad \begin{aligned} \varphi_{\nu,r}(z) &= \sum_{i=0}^{\nu} \frac{(\nu+r-i)! \nu!}{(\nu+r)! i! (\nu-i)!} z^i, \\ \mu_{\nu,r}(z) &= \sum_{i=0}^r (-1)^i \frac{(\nu+r-i)! r!}{(\nu+r)! i! (r-i)!} z^i. \end{aligned}$$

Such choice has been already considered in [27]. The following properties hold true for the polynomials  $\varphi_{\nu,r}$  and  $\mu_{\nu,r}$  (see [24] and the references therein).

**THEOREM 7.** *For all  $\nu, r \geq 0$ :*

1.  $\mu_{\nu,r}(z) \equiv \varphi_{r,\nu}(-z)$ ;
2. if  $r \geq 1$ , all the zeros of the polynomial  $\mu_{\nu,r}(z)$  lie in the annulus

$$(r+\nu)\xi < |z| < r+\nu+4/3,$$

where  $\xi \approx 0.278465$  is the unique positive root of  $x e^{x+1} = 1$ .

Moreover, it is also known (*Ehle conjecture*, see also [26]) that, for  $\nu \in \{r-2, r-1, r\}$ , one obtains  $A$ -stable methods. Such methods are also  $L$ -stable when  $\nu < r$ , conversely, when  $\nu = r$  one obtains  $g(\infty) = 1$ , since the result of Theorem 5 applies.

Concerning the order of convergence of the corresponding block methods, by applying the criteria in Theorem 6, it can be verified that it is given by

$$(35) \quad \begin{aligned} p &= r+1, & \text{for } r \text{ odd,} \\ p &\geq r+2, & \text{for } r \text{ even,} \end{aligned} \quad \begin{aligned} r &\geq 3, & \nu = r-2, r-1, r. \end{aligned}$$

We shall consider again such a choice in the next sections.

**3.2. Bernoulli numbers.** One of the way of defining Bernoulli numbers is through the following formal power series:

$$P(x) = \sum_{i=0}^{\infty} \frac{x^i}{(i+1)!},$$

whose sum is  $x^{-1}(e^x - 1)$ . Its reciprocal, in fact, is given by

$$P^{-1}(x) = \sum_{i=0}^{\infty} \frac{B_i}{i!} x^i,$$

where  $\{B_i\}$  is the sequence of Bernoulli numbers. Consequently, by means of simple calculations, one obtains that

$$(36) \quad e^{rq} = \frac{P^{-1}(-rq)}{P^{-1}(rq)} = \frac{\sum_{i=0}^{\infty} \frac{B_i}{i!} (-rq)^i}{\sum_{i=0}^{\infty} \frac{B_i}{i!} (rq)^i}.$$

Consequently (see (28)-(29)), we choose the polynomial (23) as

$$d(z) = \sum_{i=0}^r \frac{B_i}{i!} (rz)^{r-i},$$

obtaining corresponding block methods. By considering that  $B_{2i+1} = 0$ , for  $i \geq 1$ , we can take into account only methods with blocks of even size. In such a case, by using the result of Theorem 6, it can be verified that the order of the corresponding methods is  $p \geq r + 2$ , for all even values of  $r$ .

However, there is a severe drawback concerning such methods. In fact, for  $r \geq 4$ , the characteristic polynomials have roots with negative real parts, so that the corresponding block methods are not pre-stable. For this reason, it is not appropriate to use such an approximation if high order methods are needed, and we shall not consider it further.

**3.3. RITZ fractions.** In this case, the rational approximation to the exponential is obtained by using a truncated continuous fraction. To begin with, let us look for an approximation defined through a continuous fraction [18]

$$(37) \quad e^z = \frac{a_0}{b_0 + \frac{za_1}{b_1 + \frac{za_2}{b_2 + \dots}}},$$

where the sequences  $\{a_i\}$ , and  $\{b_i\}$  are suitably defined. In more detail, by defining the Moebius transformation

$$(38) \quad \Phi_i : u \rightarrow \frac{a_i}{b_i + zu}, \quad i \geq 0,$$

we obtain that the right-hand side in (37) can be written as

$$\Phi_0 \circ \Phi_1 \circ \Phi_2 \circ \dots (0).$$

Since the transformation (38) involves a rotation (R), an inversion (I), a translation (T) and, moreover, it depends on a complex parameter  $z$ , the corresponding fraction (37) is called a *RITZ fraction*. The entries of the two sequences  $\{a_i\}$  and  $\{b_i\}$  are determined in order the  $i$ th truncated fraction,

$$(39) \quad \frac{p_i(z)}{q_i(z)} \equiv \Phi_0 \circ \Phi_1 \circ \dots \circ \Phi_{i-1}(0) = \frac{a_0}{b_0 + \frac{za_1}{b_1 + \frac{\ddots}{b_{i-2} + \frac{za_{i-1}}{b_{i-1}}}}},$$

to be the best possible approximation to  $e^z$ . It can be shown that this lead to this (not unique) choice for the two sequences in (37):

$$a_i = (-1)^i, \quad b_0 = 1, \quad b_{2i+1} = 2i + 1, \quad b_{2i+2} = 2, \quad i = 0, 1, \dots$$

Since (38) are Moebius transformations, the corresponding matrix of the transformation,

$$M_i = \begin{pmatrix} 0 & a_i \\ 1 & b_i \end{pmatrix}, \quad i \geq 0,$$

provides the following recursion for the polynomials (39):

$$\begin{aligned} p_{i+1}(z) &= za_i p_{i-1}(z) + b_i p_i(z), \\ q_{i+1}(z) &= za_i q_{i-1}(z) + b_i q_i(z), \quad i \geq 1, \end{aligned}$$

with the initial conditions  $p_0 = 0, q_0 = 1, p_1 = a_1, q_1 = b_1$ . An easy induction argument shows that

$$(40) \quad \nu \equiv \deg(p_i) = \lfloor \frac{i-1}{2} \rfloor, \quad r \equiv \deg(q_i) = \lfloor \frac{i}{2} \rfloor, \quad q_i(0) = 1, \quad i \geq 1.$$

However, in such a way we obtain again Padé approximations: in fact, one verifies that

$$\frac{p_i(z)}{q_i(z)} \equiv \frac{\varphi_{\nu,r}(z)}{\mu_{\nu,r}(z)}, \quad i \geq 2,$$

where  $\nu$  and  $r$  are defined according to (40) and the polynomials  $\varphi_{\nu,r}(z)$  and  $\mu_{\nu,r}(z)$  are respectively the numerator and denominator of the Padé approximation (see (34)). It is an easy matter to verify that, for  $i \geq 2$ , the couples  $(\nu, r)$  assume the values  $(r-1, r), (r, r), r \geq 1$ .

**4. Conditioning and nonlinear iteration.** From the results in the previous section, we can conclude that the choice of the Padé approximation for constructing  $A$ -stable (or  $L$ -stable) block methods of arbitrary high order is appropriate. We now shall study aspects related to the conditioning of the discrete problem, which make such a choice appealing from this point of view as well. Such aspects are related to the solution of the discrete problem (2), which is usually carried out by means of the Newton method (or its variants), typically through the iteration

$$(41) \quad \begin{aligned} (A \otimes I_m - hB \otimes J_0) \boldsymbol{\delta}^{(i)} &= F(\mathbf{y}^{(i)}), \\ \mathbf{y}^{(i+1)} &= \mathbf{y}^{(i)} - \boldsymbol{\delta}^{(i)}, \quad i = 0, 1, \dots \end{aligned}$$

We have, therefore, the problem of having the iteration (41) converging as fast as possible, as well as to accurately solve the linear system at each iteration. By means of a suitable modification of the Newton-Kantorovich theorem (see, for example, [22]) and of standard arguments of Numerical Linear Algebra, one obtains that both the two facts can be accomplished by requiring the condition number of the matrix  $(A \otimes I_m - hB \otimes J_0)$ , which we rewrite in the equivalent form (see (15))

$$(I_r \otimes I_m - hC \otimes J_0),$$

to be as small as possible. Obviously, when  $h$  is small there are no problems, since such matrix reduces to the identity matrix. Conversely, when  $h$  is large, the main contribute is due to  $hC \otimes J_0$ . By considering that

$$\kappa(hC \otimes J_0) = \kappa(C)\kappa(J_0),$$

(where obviously  $\kappa$  denotes the condition number of the specified matrix) we then conclude that it would be very appreciable to have the matrix  $C$  with a small condition number. We observe that the case where the stepsize  $h$  is large is a significant one, since it corresponds to the use of large stepsizes when, for example, in a stiff problem an asymptotically stable stationary solution has been approached. In such a case, in fact, the solution is approximately constant and, consequently, any (appropriate) nonlinear iteration should converge rapidly.

Moreover, a necessary condition for having the matrix  $C$  well-conditioned is to have its eigenvalues bounded away from zero and infinity, as the size  $r$  increases. In this respect, the block methods derived by the Padé  $(\nu, r)$  have this property. In fact, from (33)-(34) and the result of point 2 in Theorem 7, one obtains that the eigenvalues of  $C$  (which have positive real part) belong to the annulus

$$(42) \quad \frac{r + \nu}{r} \xi < |z| < \frac{r + \nu + 4/3}{r}, \quad \xi \approx 0.278465.$$

If, in addition, we consider that the best values for  $\nu$  are  $r - 1$  and  $r$ , we conclude that all the eigenvalues of  $C$  have positive real part and are approximately contained in the annulus

$$(43) \quad 2\xi < |z| < 2, \quad r \gg 0.$$

In Table 1 we list the condition numbers of the matrix  $C$  corresponding to the block methods obtained from the Padé  $(r, r)$  and  $(r - 1, r)$ ,  $r = 1, \dots, 12$ .

TABLE 1

Condition numbers of the matrix  $C$  for the block methods derived from the Padé  $(r, r)$  and  $(r - 1, r)$ .

$r$	1	2	3	4	5	6
$(r, r)$	1.00E0	6.88E0	1.49E1	3.10E1	5.62E1	1.15E2
$(r - 1, r)$	1.00E0	4.19E0	8.15E0	1.67E1	3.03E1	6.08E1
$r$	7	8	9	10	11	12
$(r, r)$	2.19E2	5.61E2	1.28E3	4.23E3	1.14E4	4.08E4
$(r - 1, r)$	1.17E2	2.95E2	6.86E2	2.26E3	6.16E3	2.23E4

**5. Implementation of block implicit methods: blended schemes.** We now consider the problem of the efficient implementation of the block methods (2), namely we shall devise procedures for the solution of the corresponding discrete problem. This topic is focal for a numerical method to be competitive and has been, therefore, extensively studied for various classes of numerical methods (see, e.g., [1, 6, 11, 19, 20, 25]), even when implemented on different computer platforms. Concerning the the latter field, we mention Galligani's contributions [13, 14].

From the point of view of the implementation, the main problem is the efficient solution of the linear systems required by the iterative solution of the discrete problem. According to the general ways of solving linear systems, i.e. by using direct or iterative procedures, it is possible to classify the currently used approaches in two main categories:

- diagonalization (or block diagonalization) of the matrix  $C$  [11, 17];
- definition of a suitable splitting [1, 6, 19, 20, 21].

Concerning the first category, it would be preferable to have the matrix  $C$  with all real eigenvalues. This because in such a way it is possible to solve real linear systems of the same size of the continuous problem. Nevertheless, there is some drawback on this point, since the matrix which diagonalize the matrix  $C$  may be ill-conditioned. As matter of fact, from (21), by considering that (assuming for simplicity all distinct eigenvalues)

$$F = V^{-1}\Lambda V,$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}, \quad V = \begin{pmatrix} \lambda_1^0 & \dots & \lambda_1^{r-1} \\ \vdots & & \vdots \\ \lambda_r^0 & \dots & \lambda_r^{r-1} \end{pmatrix},$$

one obtains that the matrix

$$(44) \quad W = VG^{-1}Q_r^{-1}$$

diagonalizes the matrix  $C$ . If we assume, for example, that the eigenvalues are equally spaced in the interval  $[2\xi, 2]$  (see (42)-(43)), i.e.

$$(45) \quad \lambda_i = 2\xi + (i - 1)\frac{2(1 - \xi)}{r - 1}, \quad i = 1, \dots, r,$$

TABLE 2

Condition numbers for the matrix (44) in the case of the real eigenvalues (45) (first column) and for the block methods derived from the Padé  $(r-1, r)$  (second column).

$r$	$\kappa(W)$	
2	4.66E0	3.03E0
3	6.65E1	1.01E1
4	3.71E2	3.47E1
5	1.34E4	1.37E2
6	1.24E5	6.16E2
7	3.49E6	3.18E3
8	5.29E7	1.82E4
9	9.58E8	1.08E5
10	2.26E10	6.51E5

we obtain the condition numbers listed in the first column of Table 2. One may observe that the matrix  $W$  becomes rapidly ill-conditioned, as  $r$  grows. This feature is less evident when the eigenvalues are chosen more appropriately: for example, in the second column we list the condition numbers of the matrices  $W$  corresponding to the block methods derived from the Padé  $(r-1, r)$ ,  $r = 2, \dots, 10$ .

Concerning the second approach, in general there is the problem of finding a splitting having satisfactory convergence properties.

The first strategy is used, for example, in the code RADAU5 [17], whereas the second strategy is used in the code GAM [21]. In this paper, after the introduction of the notion of “blended block method”, we shall obtain a natural way to define appropriate splittings for the methods. Our approach, which is derived from the results in [6], is based on the definition of methods obtained as the combination of a couple of equivalent block methods. For this reason, such methods will be called *blended block methods*.

In order not to unnecessarily complicate the notation, we shall derive the new methods when the continuous problem to be solved is the usual test equation (16). Then, we look for methods generating a discrete problem in the following form,

$$(46) \quad M(q)\mathbf{y} \equiv (A(q) - qB(q))\mathbf{y} = (q\mathbf{b}(q) - \mathbf{a}(q))\eta \equiv \mathbf{g},$$

where, being  $\theta = \theta(q)$  a suitable “weight” matrix,  $M(q)$  is the  $r \times r$  matrix defined by

$$(47) \quad A(q) = \theta A_1 + (I_r - \theta)A_2, \quad B(q) = \theta B_1 + (I_r - \theta)B_2,$$

and

$$(48) \quad \mathbf{a}(q) = \theta \mathbf{a}_1 + (I_r - \theta)\mathbf{a}_2, \quad \mathbf{b}(q) = \theta \mathbf{b}_1 + (I_r - \theta)\mathbf{b}_2.$$

The couples of augmented matrices

$$(49) \quad \hat{A}_1 = [\mathbf{a}_1 | A_1], \quad \hat{B}_1 = [\mathbf{b}_1 | B_1], \quad \hat{A}_2 = [\mathbf{a}_2 | A_2], \quad \hat{B}_2 = [\mathbf{b}_2 | B_2],$$

define two suitable equivalent block methods. In particular, we start considering the following choices:

$$(50) \quad A_1 = I_r, \quad B_1 = C, \quad A_2 = C^{-1}, \quad B_2 = I_r, \quad \theta(q) = (1 - q)^{-1}I_r,$$

with the remaining vectors obtained through the consistency conditions (7). We shall choose the block methods in (49) so that the discrete problem (46) can be solved by using an iterative procedure,

$$(51) \quad N(q)\mathbf{y}^{(i+1)} = (N(q) - M(q))\mathbf{y}^{(i)} + \mathbf{g}, \quad N(q) = A_1 - qB_2.$$

In view of the choice (50), we obtain

$$(52) \quad N(q)^{-1} = (1 - q)^{-1}I_r \equiv \theta(q),$$

i.e. (51) defines a diagonal splitting. Moreover, the diagonal entries are equal, so that the resulting computational cost per iteration is relatively low.

The iteration (51) converges to the solution of (46) if and only if the spectral radius, say  $\rho(q)$ , of the iteration matrix

$$(53) \quad I_r - N(q)^{-1}M(q)$$

is smaller than 1. According to [19, 20], the *region of convergence* of the iteration (51) is given by

$$\Gamma = \{q \in \mathbb{C} : \rho(q) < 1\}.$$

Moreover, the iteration is said to be *A-convergent* if  $\mathbb{C}^- \subseteq \Gamma$  (*A*( $\alpha$ )-convergence is similarly defined). From the definition of  $N(q)$ , *A*-convergence is equivalent to require that

$$\rho^* \equiv \sup_{x>0} \rho(ix) < 1,$$

where, as usual,  $i$  is the imaginary unit. Another interesting property of the iteration (51) is that from the definition of  $N(q)$  one obtains

$$\rho(q) \rightarrow 0, \quad \text{as } q \rightarrow \infty.$$

Such property is very welcome, in order to have the iteration rapidly converging when the method is applied to stiff problems [19, 20].

A simple expression for the parameter  $\rho^*$  may be derived from (51)-(53). In fact, if  $\lambda$  is an eigenvalue of  $C$ , then

$$(54) \quad \zeta(\lambda) = \frac{q(\lambda - 1)^2}{\lambda(q - 1)^2}$$

is an eigenvalue of the iteration matrix (53). Consequently, by assuming  $q = ix$ , we obtain that

TABLE 3  
Value of  $\rho^*$  for the blended block methods corresponding to the Padé  $(r, r)$  and  $(r - 1, r)$ .

$r$	1	2	3	4	5	6	7	8	9	10	11	12
$(r, r)$	.250	.289	.419	.522	.600	.661	.710	.749	.783	.811	.835	.856
$(r - 1, r)$	.0	.204	.386	.508	.595	.661	.712	.753	.787	.815	.840	.861

$$(55) \quad \rho^* = \max_{x>0} \max_{\lambda \in \sigma(C)} g(x; \lambda),$$

where, by posing  $\lambda = a + ib$ ,

$$(56) \quad g(x; \lambda) = \frac{x((1-a)^2 + b^2)}{\sqrt{a^2 + b^2}(x^2 + 1)}$$

is the modulus of  $\zeta(\lambda)$ . By considering that  $g(0; \lambda) = g(\infty; \lambda) = 0$ , one concludes that if there is one stationary point, this must be a maximum. By differentiating  $g(x; \lambda)$ , we obtain

$$g'(x; \lambda) = \frac{(1-x^2)((1-a)^2 + b^2)}{\sqrt{a^2 + b^2}(x^2 + 1)^2},$$

from which we conclude that  $x = 1$  is the point of maximum for  $g(x; \lambda)$ . As a consequence, we get

$$(57) \quad \rho^* = \max_{\lambda \in \sigma(C)} \frac{(1-a)^2 + b^2}{2\sqrt{a^2 + b^2}},$$

which is easy to compute, since we have the possibility of choosing the spectrum of  $C$ . In Table 3 we list the value of  $\rho^*$  for the block methods obtained from the Padé approximations  $(r, r)$  and  $(r - 1, r)$ ,  $r = 1, \dots, 12$ . One then concludes that the iterations corresponding to all such methods are  $A$ -convergent. Moreover, it can be verified that this property holds true at least up to  $r = 24$ , for both type of methods.

An improvement can be obtained by considering an alternative choice to (50), still providing us with a diagonal splitting, namely

$$(58) \quad A_1 = I_r, \quad B_1 = C, \quad A_2 = DC^{-1}, \quad B_2 = D, \quad \theta(q) = (I_r - qD)^{-1},$$

where  $D$  is a diagonal matrix with positive real entries. The first possibility that we consider is the following:

$$(59) \quad D = \alpha I_r, \quad \alpha > 0,$$

which reduces to the previous choice (50) in the case  $\alpha = 1$ . In such a case, (54), (55), (56) and (57) become, respectively,

$$\zeta(\lambda; \alpha) = \frac{q(\lambda - \alpha)^2}{\lambda(\alpha q - 1)^2}, \quad \rho^* = \max_{x>0} \max_{\lambda \in \sigma(C)} g(x; \lambda, \alpha), \quad g(x; \lambda, \alpha) = \frac{x((\alpha - a)^2 + b^2)}{\sqrt{a^2 + b^2}(\alpha x^2 + 1)},$$

TABLE 4

Values of  $\rho^*$  for the modified blended block methods corresponding to the Padé  $(r, r)$ .

$r$	1	2	3	4	5	6	7	8	9	10	11	12
$\alpha_r$	.5	.5774	.5902	.5901	.5867	.5826	.5786	.5748	.5713	.5682	.5653	.5628
$\rho_{\alpha_r}^*$	.0	.134	.277	.379	.454	.511	.556	.592	.622	.647	.668	.687

TABLE 5

Values of  $\rho^*$  for the modified blended block methods corresponding to the Padé  $(r-1, r)$ .

$r$	1	2	3	4	5	6	7	8	9	10	11	12
$\alpha_r$	1.	.8165	.7387	.6952	.6671	.6471	.6322	.6205	.6111	.6033	.5968	.5911
$\rho_{\alpha_r}^*$	.0	.184	.340	.442	.512	.564	.605	.637	.663	.685	.703	.720

and, considering that  $x = \alpha^{-1}$  is the point of maximum for  $g(x; \lambda, \alpha)$ ,

$$(60) \quad \rho_\alpha^* = \max_{\lambda \in \sigma(C)} \frac{(\alpha - a)^2 + b^2}{2\sqrt{a^2 + b^2}}.$$

Consequently, the convergence properties of the iteration (51) now depend on the parameter  $\alpha > 0$ , which can be chosen in order to minimize the corresponding value of  $\rho_\alpha^*$ . In Tables 4 and 5 we list, respectively, the optimum parameters  $\alpha_r$ , along with the corresponding parameter  $\rho_{\alpha_r}^*$ , for the block methods obtained from the Padé  $(r, r)$  and  $(r-1, r)$ ,  $r = 1, \dots, 12$ . One verifies that there has been an improvement over the previous values in Table 3. A different choice for the matrix  $D$  will be considered in the next section.

**6. Implementation of block implicit methods: further analysis.** We now consider a particular case of the above approach, namely

$$(61) \quad \lambda_i(C) = \alpha, \quad i = 1, \dots, r,$$

where  $\alpha$  is the same parameter in (59). From equation (60), we then obtain  $\rho_\alpha^* = 0$ . This conclusion may appear astonishing, since it implies that the iteration (51) converges in one iteration, when  $q \in \mathbb{C}^-$  (actually, for all  $q \neq \alpha^{-1}$ ). Nevertheless, the choice of a  $r$ -fold eigenvalue for the matrix  $C$  has a severe drawback, since the matrix  $C$  becomes very ill-conditioned, for increasing values of  $r$ . In fact, let us first consider the problem of choosing appropriately the parameter  $\alpha$ . From (23) and (61), one obtains that

$$d(z) = \sum_{i=0}^r \binom{r}{i} (-\alpha)^{r-i} z^i,$$

i.e.

$$d_i = \binom{r}{i} (-\alpha)^{r-i}, \quad i = 0, \dots, r.$$

In order to obtain a method with  $\infty \in \mathcal{D}$ , we impose (see (24)-(25))

$$(62) \quad 0 = \frac{(-\alpha)^{-r}}{d_0} \sum_{i=0}^r \frac{r^i}{i!} d_i = \sum_{i=0}^r \frac{(-1)^i}{i!} \binom{r}{i} \left(\frac{r}{\alpha}\right)^i \equiv w(r\alpha^{-1}).$$

TABLE 6  
Condition numbers for the matrix  $C$  obtained from (61)–(63).

$r$	$\bar{\alpha}$	$\kappa(C)$
2	.58578643762690497	3.08E0
3	1.3075995645253780	4.19E1
4	2.2912642499285432	2.15E3
5	1.3902692056822543	2.16E3
6	2.0048542024083011	6.92E4
7	2.7259875357196162	3.98E6
8	1.8749852768446345	3.86E6

In such a case, in fact,  $A$ -stability implies  $L$ -stability. The polynomial  $w(z)$  turns out to have all positive and real zeros. Among them, we obviously shall choose the one which generates the method with the best stability properties. Let it be  $\bar{z}$ . In correspondence of such root, we get the parameter

$$(63) \quad \bar{\alpha} = r \bar{z}^{-1}.$$

When more than one zero fulfills the above requirement, in view of the arguments in Section 4, we shall choose the one which minimizes the condition number of the matrix  $C$ . As an example, for  $r = 2$  from (62)–(63) one obtains, after a few calculations, that the two values of  $\alpha$  are

$$\alpha_1 = \frac{2}{2 + \sqrt{2}}, \quad \alpha_2 = 2 + \sqrt{2},$$

both generating  $A$ -stable methods. The parameter which minimizes the value of  $\kappa(C)$  is  $\alpha_1$ . From such value, one obtains an  $L$ -stable block method of order 2 having the following matrix

$$A^{-1} \hat{B} = \begin{pmatrix} 4.9264068711928521\text{E} - 1 & 5.1471862576142957\text{E} - 1 & -7.3593128807148411\text{E} - 3 \\ 6.5685424949238014\text{E} - 1 & 6.8629150101523972\text{E} - 1 & 6.5685424949238014\text{E} - 1 \end{pmatrix},$$

for which (see (3) and (15))  $\kappa(C) \approx 3$ . Nonetheless, for increasing values of  $r$  one realizes that the condition number of the matrix  $C$  grows very rapidly with  $r$ , as it is shown in Table 6. For each listed value of  $r$ , the corresponding block method has order  $r$  and is  $A$ -stable (or  $A(\alpha)$ -stable). Nevertheless, due to the growth of  $\kappa(C)$ , it is advisable to use such methods only for  $r$  very small.

REMARK 1. We observe that the method corresponding, for example, to  $r = 8$  in Table 6 is  $A$ -stable (see Figure 1). Nevertheless, the corresponding matrix  $C$  is ill-conditioned (see the third column in Table 6). The reason for this is the large value of departure from normality of the matrix, which can be measured by

$$\Delta(C) = \|C C^T - C^T C\|_2.$$

In fact, for this method one obtains  $\Delta(C) \approx 3.0\text{E}9$ , whereas the corresponding value for the method derived from the Padé (7,8) (see Table 1,  $r = 8$ ) is  $\Delta(C) \approx 66$ .

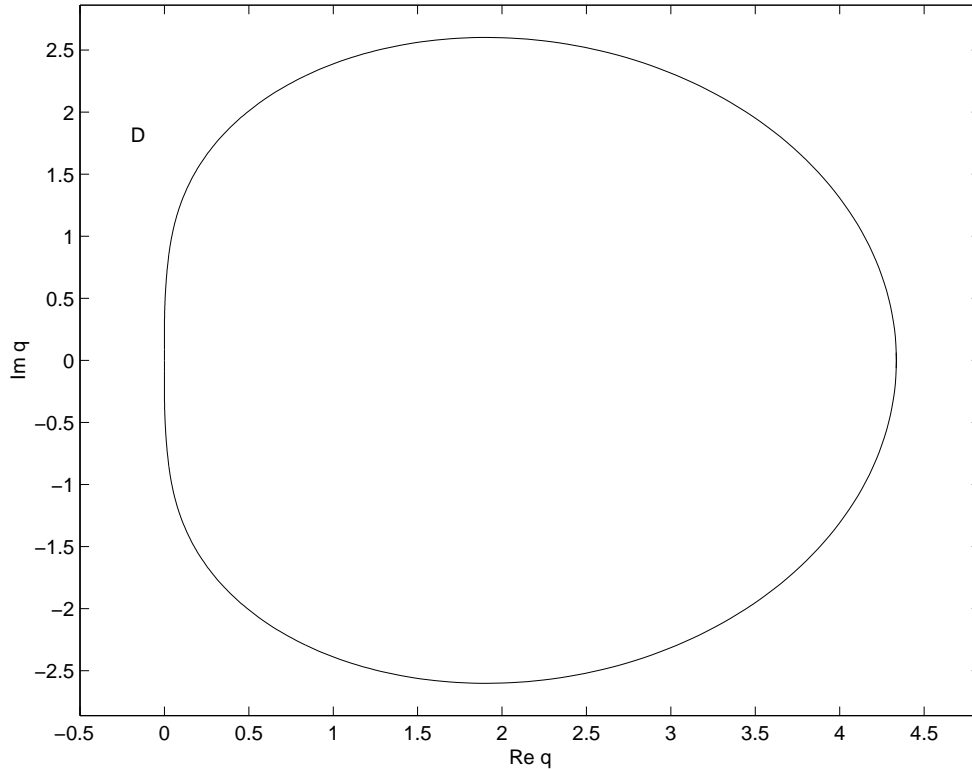


FIG. 1. Stability region  $\mathcal{D}$  for the block method corresponding to  $r = 8$  in Table 6.

Let us now consider, instead of (59), the more general choice

$$(64) \quad D = \text{diag} ( \gamma_1 \quad \dots \quad \gamma_r ), \quad \gamma_i > 0, \quad i = 1, \dots, r,$$

for the iteration (51)-(58). The diagonal entries of  $D$  will be chosen in order to reduce the value of the parameter  $\rho^*$ . Unfortunately, the spectral analysis is not as simple as in the case of the choice (59) and we must resort to a computational approach to obtain such entries. In particular, we have used the Matlab function `fmins` to minimize  $\rho^* \equiv \rho^*(\gamma_1, \dots, \gamma_r)$ , for the block methods obtained from the Padé  $(r, r)$  and  $(r-1, r)$ ,  $r = 1, \dots, 10$ . The obtained diagonal entries are listed in Tables 7 and 8. The corresponding values of the parameter  $\rho^*$  are listed in Table 9: in all cases we obtain an improvement over the corresponding values listed in Tables 3 and 4-5. Moreover, for comparison we also plot in Figure 2 the function  $\rho(ix)$  for  $x > 0$ , for the block methods obtained from the Padé  $(7, 8)$  with the choices (50) and (58)-(59), corresponding respectively to  $\alpha = 1$  and  $\alpha = 0.6205$  (see Table 5), and to the choice (58)-(64), where the diagonal entries of  $D$  are those listed in Table 8,  $r = 8$ . From the plots, one may infer that the choice (58)-(64) is definitely an improvement over (50), since the curve in solid is always below the dashed one. Conversely, the dotted curve, corresponding to the choice (58)-(59) has a maximum greater than that of the curve in solid; nevertheless, for  $x < 0.77$  it is under such curve. This means that, for the corresponding values of  $q$ , the iteration (58)-(59) is faster than (58)-(64), even though for larger values of  $x$  the roles are reversed.

Let us now end the section by studying the application of a blended block method to the more general problem (1). In such a case, the discrete problem corresponding

TABLE 7

Diagonal entries of the matrix  $D$  for the blended block methods derived from the Padé  $(r, r)$ .

$r$	$\gamma_i, \quad i = 1, \dots, r$				
2	0.66275204106065	0.99880381565137			
3	0.79177839434600	0.66142441654254	1.09728375380713		
4	0.71248323670683	0.68572979240044	1.00835280450325	0.94336189309863	
5	0.73794709503502	0.69667939234304	0.94678950632936	0.96925988167893	1.08556894280691
6	0.78960666760857	0.77764317616699	1.10379439358622	1.12287112979631	1.04425410677717
	0.97747272393049				
7	0.90328518460926	0.93588189061015	1.27378358368408	1.06482337773942	0.95699568328484
	0.87962577636270	0.88647177187620			
8	0.90189079913764	0.96029017717008	1.20044261367058	1.06182151859105	0.98279767187713
	0.95130581186793	0.95477396330476	0.97709719034535		
9	0.78890956014085	0.98692274344950	1.16031736214545	1.06429394889840	0.97558449511996
	0.97973125397795	0.97616792047847	1.01092465767460	1.02624249573684	
10	0.98194772536412	0.91245934163159	1.04481326127609	1.00603860548456	1.00089107693710
	0.99217500074166	1.00032852544433	1.00860473013924	1.02306653120170	1.05905836178089

TABLE 8

Diagonal entries of the matrix  $D$  for the blended block methods derived from the Padé  $(r - 1, r)$ .

$r$	$\gamma_i, \quad i = 1, \dots, r$				
2	0.96936165526039	1.50848352835328			
3	0.81868602814689	0.87495096575515	1.44998332454753		
4	0.35337640949037	0.80121932984835	1.27762338976231	1.66446327607499	
5	0.63946725983491	0.82968839663997	0.91909524862064	1.42126433313185	1.52443079963375
6	0.79474199557416	0.69885470794190	0.96328076538840	1.17906063317010	0.96006333245803
	1.34062294955658				
7	0.76154788274216	0.77052101658868	1.03651805522176	1.06062392757843	0.98001202471669
	1.07157541314092	1.33204542634207			
8	0.90280588616216	0.94667898522012	1.16837318551763	0.97151531383934	0.94628652026249
	0.92599286317791	1.01107530401679	1.15188588598478		
9	0.95021396531815	0.96824923001551	1.16031142832145	1.00905109392683	0.98542777497508
	0.95758015675129	0.97574516430175	0.97853576014775	0.97686463476444	
10	0.89639578999006	0.95322083486291	1.09782034111785	1.01308990374127	0.97688329481101
	0.97154048045308	0.96942938230334	0.99731505715738	0.97199798518272	1.08135584615257

TABLE 9

Values of  $\rho^*$  for the blended block methods corresponding to the Padé  $(r, r)$  and  $(r - 1, r)$  with the diagonal scaling  $D$ .

$r$	2	3	4	5	6	7	8	9	10
$(r, r)$	.055	.230	.273	.359	.414	.444	.479	.527	.622
$(r - 1, r)$	.086	.280	.317	.450	.501	.536	.515	.541	.593

to the application of the method (46), (47)-(48), (58) over the first  $r$  mesh points is given by

$$(65) \quad \tilde{A} \mathbf{y} - h \tilde{B} \mathbf{f} = h \tilde{\mathbf{b}} f(t_0, y_0) - \tilde{\mathbf{a}} y_0,$$

where, by setting  $I = I_r \otimes I_m$ ,

$$\begin{aligned} N &= I - hD \otimes J_0 \\ \tilde{A} &= N^{-1}I + (I - N^{-1}) ((DC^{-1}) \otimes I_m), \\ \tilde{B} &= N^{-1} (C \otimes I_m) + (I - N^{-1}) (D \otimes I_m), \\ \tilde{\mathbf{a}} &= N^{-1} (\mathbf{e} \otimes I_m) + (I - N^{-1}) ((DC^{-1}\mathbf{e}) \otimes I_m), \\ \tilde{\mathbf{b}} &= N^{-1} ((\mathbf{q}_1 - C\mathbf{e}) \otimes I_m) + (I - N^{-1}) ((D(C^{-1}\mathbf{q}_1 - \mathbf{e})) \otimes I_m). \end{aligned}$$

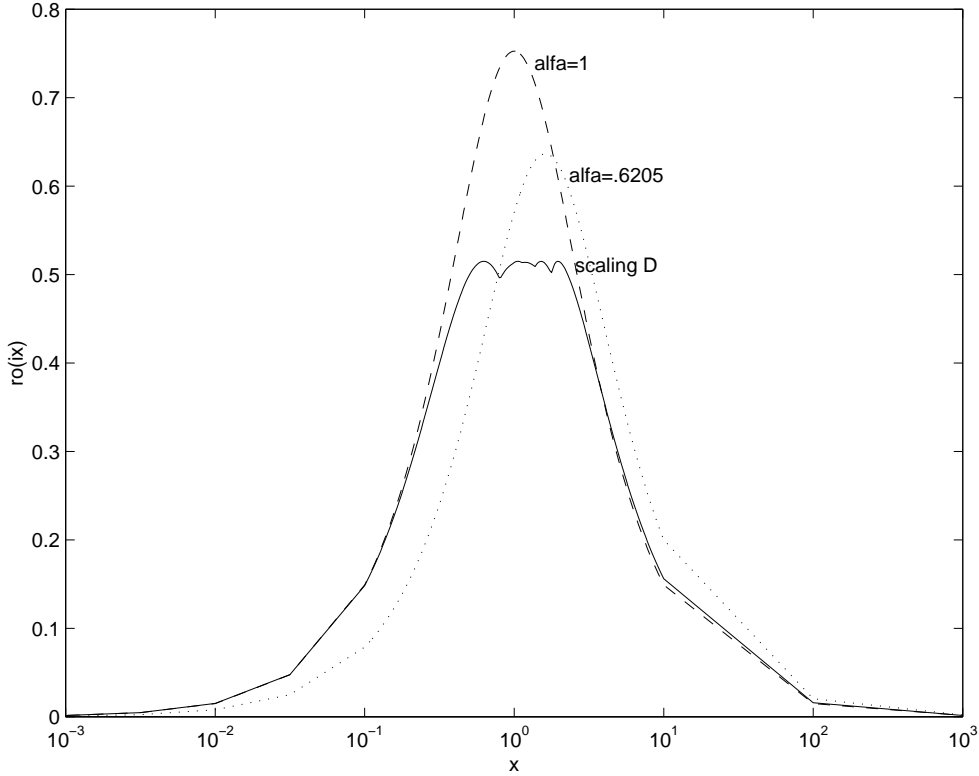


FIG. 2. Spectral radius of the iteration matrix for the blended block method derived from the Padé (7, 8), in its original formulation (dashed line), with the scaling  $\alpha_s I_s$  (dotted line) and with the scaling  $D$  (solid line).

From the above expressions, it is evident that the application of the method requires the factorization of the matrix  $N$ , i.e. of the block

$$(66) \quad I_m - h\alpha J_0$$

in case we use the method corresponding to (59); alternatively, when (64) is considered, we need to factorize the blocks

$$(67) \quad I_m - h\gamma_i J_0, \quad i = 1, \dots, r.$$

It is customary to solve the equation (65) by using the modified Newton method, then solving linear systems with the matrix

$$M = \tilde{A} - h\tilde{B}(I_s \otimes J_0).$$

In place of such linear systems, we solve an *inner iteration* similar to (51), thus involving only linear systems with the matrix  $N$ , which has already been factored.

We then conclude that, leaving aside for simplicity function and Jacobian evaluations, the arithmetic complexity for solving (65) when  $\ell$  Newton iterations are performed, each requiring  $\mu$  inner iterations, amounts to

$$\frac{2}{3}m^3 + O(\ell\mu r m^2) \quad \text{floating operations,}$$

in the case of (59), or

$$\frac{2}{3}rm^3 + O(\ell\mu rm^2) \quad \text{floating operations,}$$

in the case of (64). The leading term is obviously due to the factorizations of the matrices (66) and (67), respectively. In the latter case, however, by observing that the  $r$  factorizations are independent each other, it is possible to use  $r$  parallel processors to execute them concurrently. Consequently, the choice (64) will result in a method having a natural *parallelism across the method*.

**7. Conclusions.** In this paper we have reviewed the derivation of  $r$ -block methods for ODEs. The derivation of such methods has been done in a unified framework, which allows to discuss many important theoretical properties of the methods.

Among the possible choices, it has been confirmed that the methods derived from the Padé approximation to the exponential have good stability properties. In more detail, the methods derived from the Padé  $(r, r)$  are perfectly  $A$ -stable and those derived from the Padé  $(r - 1, r)$  are  $L$ -stable, for all  $r \geq 1$ . Moreover, from (35) one obtains that even values of  $r$  are preferable, because they provide higher order methods than the subsequent odd values.

The implementation of the block methods as blended block methods has also been introduced. Such an implementation exploits the possibility of writing a given block method in different equivalent forms. The blended block methods here studied are characterized by a diagonal splitting, which is  $A$ -convergent for all values of  $r$  of practical interest. Improvements of the basic splitting have been also considered, which can be tailored for an efficient implementation either on sequential or parallel computers. In both cases, we obtain methods with a (possibly parallel) complexity whose leading term is  $2m^3/3$  flops per integration step (apart function and Jacobian evaluations), when the size  $m$  of the continuous problem is large.

#### REFERENCES

- [1] P. Amodio, L. Brugnano. A Note on the Efficient Implementation of Implicit Methods for ODEs, *Jour. Comput. Appl. Math.* **87** (1997) 1–9.
- [2] U. M. Ascher, L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.
- [3] O. Axelsson. Global Integration of Differential Equations Through Lobatto Quadrature, *BIT* **4** (1964) 69–86.
- [4] O. Axelsson. A Class of  $A$ -stable Methods, *BIT* **9** (1969) 185–199.
- [5] C. Bendsten. On Implicit Runge-Kutta Methods with High Stage Order, *BIT* **37** (1997) 212–217.
- [6] L. Brugnano. Blended Block BVMs ( $B_3$ VMS): a Family of Economical Implicit Methods for ODEs, *Jour. Comput. Appl. Math.* (1999) in press.
- [7] L. Brugnano, D. Trigiantè. *Solving Differential Problems by Multistep Initial and Boundary Value Methods*, Gordon and Breach, Amsterdam, 1998.
- [8] K. Burrage. A Special Family of Runge-Kutta Methods for Solving Stiff Differential Equations, *BIT* **18** (1978) 22–41.
- [9] K. Burrage. High Order Algebraically Stable Runge-Kutta Methods, *BIT* **18** (1978) 373–383.
- [10] K. Burrage. *Parallel and Sequential Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, 1995.
- [11] J. C. Butcher. On the Implementation of Implicit Runge-Kutta Methods, *BIT* **6** (1976) 237–240.
- [12] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta Methods and General Linear Methods*, John Wiley, Chichester, 1987.
- [13] I. Galligani. Splitting Methods for Solving Large Systems of Linear Ordinary Differential Equations on a Vector Computer, *WSSIAA* **2** (1993) 165–176.

- [14] I. Galligani, V. Ruggiero. Solving Large Systems of Linear Ordinary Differential Equations on a Vector Computer. *Parallel Comput.* **9** (1989) 359–365.
- [15] E. Hairer, S. P. Norsett, G. Wanner. *Solving Ordinary Differential Equations I*, 2<sup>nd</sup> ed., Springer Series in Computational Mathematics, vol. 8, Springer-Verlag, Berlin, 1993.
- [16] E. Hairer, G. Wanner. Algebraically Stable and Implementable Runge-Kutta Methods of High Order, *SIAM J. Numer. Anal.* **18** (1981) 1098–1108.
- [17] E. Hairer, G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.
- [18] P. Henrici. *Applied and Computational Complex Analysis, Vol. 2*, John Wiley & Sons, New York, 1977.
- [19] P. J. van der Houwen, J. J. B. de Swart. Triangularly Implicit Iteration Methods for ODE-IVP Solvers, *SIAM J. Sci. Comput.* **18** (1997) 41–55.
- [20] P. J. van der Houwen, J. J. B. de Swart. Parallel Linear System Solvers for Runge-Kutta Methods, *Adv. Comput. Math.* **7**,1-2 (1997) 157–181.
- [21] F. Iavernaro, F. Mazzia. Solving Ordinary Differential Equations by Generalized Adams Methods: Properties and Implementation Techniques. *Appl. Numer. Math.* **28** (1998) 107–126.
- [22] V. Lakshmikantham, D. Trigiante. *Theory of Difference Equations: Numerical Methods and Applications*, Series “Mathematics is Science and Engineering”, vol. 181, Academic Press, San Diego, 1988.
- [23] J. D. Lambert. *Numerical methods for Ordinary Differential Systems*, John Wiley & Sons, New York, 1991.
- [24] E. B. Saff, R. S. Varga. On Zeros and Poles of Padé Approximants to  $e^z$ . III, *Numer. Math.* **30** (1978) 241–266.
- [25] R. D. Skeel, A. K. Kong. Blended Linear Multistep Methods, *ACM TOMS* **3** (1977) 326–345.
- [26] G. Wanner, E. Hairer, S. P. Nørsett. Order Stars and Stability Theorems, *BIT* **18** (1978) 475–489.
- [27] H. A. Watts, L. F. Shampine. A-stable Block One-step Methods, *BIT* **12** (1972) 252–266.