Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

Approximating Runge-Kutta matrices by triangular matrices

W. Hoffmann and J.J.B. de Swart

Department of Numerical Mathematics

NM-R9517 1995

Report NM-R9517
ISSN 0169-0388

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Approximating Runge-Kutta Matrices by Triangular Matrices

W. Hoffmann [1] and J.J.B. de Swart [2]

[1] *Department of Mathematics and Computer Science,*
University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam,

[2] *CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands*

Abstract

The implementation of implicit Runge-Kutta methods requires the solution of large systems of non-linear equations. Normally these equations are solved by a modified Newton process, which can be very expensive for problems of high dimension. The recently proposed triangularly implicit iteration methods for ODE-IVP solvers [HSw95] substitute the Runge-Kutta matrix A in the Newton process for a triangular matrix T that approximates A, hereby making the method suitable for parallel implementation. The matrix T is constructed according to a simple procedure, such that the stiff error components in the numerical solution are strongly damped. In this paper we prove for a large class of Runge-Kutta methods that this procedure can be carried out and that the diagonal entries of T are positive. This means that the linear systems that are to be solved have a non-singular matrix.

CR Subject Classification (1991): G.1.7
Keywords & Phrases: numerical analysis, Runge-Kutta methods, parallelism
Note: The research reported in this paper was supported by STW (Dutch Foundation for Technical Sciences).

1. Introduction and motivation

For solving the stiff initial value problem

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad y, f \in \mathbf{R}^d, \quad t_0 \le t \le t_e,$$

one of the most powerful methods is an implicit Runge-Kutta (RK) method. In such a method we have to solve every time step a system of non-linear equations of the form

$$R(Y_n) = 0; \quad R(Y_n) := Y_n - (e \otimes I)y_{n-1} - h_n(A \otimes I)F(Y_n), \tag{1.1}$$

where A denotes the $s \times s$ matrix containing the parameters of the s-stage RK method, y_{n-1} the approximation to $y(t_{n-1})$, e is the s-dimensional vector with unit entries, I is the $d \times d$ identity matrix, h_n is the step size $t_n - t_{n-1}$ and $\otimes$ denotes the Kronecker product.

The s components $Y_{n,i}$ of the sd-dimensional solution vector Y_n represent s numerical approximations to the s exact solution vectors $y(et_{n-1} + ch_n)$; here, c denotes the abscissa vector. Furthermore, for any vector $X = (X_i)$, $F(X)$ contains the derivative values $(f(X_i))$. It is assumed that the components of c are distinct and positive.

Once we have solved (1.1), we obtain the step point value $y_n \approx y(t_n)$ by some step point formula

$$y_n = y_{n-1} + h_n(b^T \otimes I)F(Y_n),$$

where b is a vector of dimension s containing method parameters. For stiffly accurate RK methods, $c_s = 1$, so that the step point value equals the sth component of Y_n and the step point formula may be omitted.

To solve (1.1), in general one uses a Newton-type iteration scheme of the form

$$(I - B \otimes h_n J_n)\Delta Y_n^{(j+1)} = -R(Y_n^{(j)}); \quad Y_n^{(j+1)} = Y_n^{(j)} + \Delta Y_n^{(j+1)}, \tag{1.2}$$

where J_n is an approximation to the Jacobian of the right hand side function f at t_{n-1}, $Y_n^{(0)}$ is the initial iterate to be provided by some predictor formula and B is an $s \times s$ matrix that defines the type of Newton iteration. To get insight in the convergence behaviour of (1.2), we apply the scheme to the scalar test equation $y' = \lambda y$. Defining the iteration error $\epsilon_n^{(j)}$ by $Y_n^{(j)} - Y_n$, we see from (1.1) and (1.2) that these errors are amplified by the matrix Z defined by

$$Z(z) = z(I - zB)^{-1}(A - B); \quad z := \lambda h_n.$$

We introduce the *stiff* and *non-stiff amplification matrices* of scheme (1.2), notation $Z_\infty(B)$ and $Z_0(B)$, respectively, by:

$$Z_\infty(B) := \lim_{|z|\to\infty} Z(z) = I - B^{-1}A \quad \text{and} \quad Z_0(B) := \lim_{|z|\to 0} Z(z)/|z| = A - B.$$

Choosing $B = A$ would lead to the modified Newton process, for which $Z(z) = 0$ for all z. However, the computation of $Y_n^{(j)}$ now requires the solution of a linear system of dimension sd. For high-dimensional problems this requires a lot of computational effort. Several attempts have been made to reduce these costs by selecting matrices B different from A.

In [CB83], Cooper & Butcher propose the choice $B = P$, where P is a matrix that has a one-point spectrum. By performing a similarity transformation to (1.2) they arrive at the scheme

$$
\begin{aligned}
&P\,Q = Q\,L,\\
&(I - L \otimes h_n J_n)\Delta X_n^{(j+1)} = -(Q^{-1} \otimes I) R(Y_n^{(j)}),\\
&Y_n^{(j+1)} = Y_n^{(j)} + (Q \otimes I)\Delta X_n^{(j+1)},
\end{aligned}
\tag{1.3}
$$

where L and Q are lower triangular and orthogonal matrices, respectively, that define the Schur decomposition of P. Since the diagonal entries of L are equal, implementing (1.3) requires only one LU-decomposition of dimension d.

In [HSo91], the authors select $B = D$, where D is a diagonal matrix. Scheme (1.2) is now suitable for implementation on an s processor machine, since the s components of $Y_n^{(j)}$ can be computed independently. The matrix D is constructed such that $\rho(Z_\infty(D)) = 0$, where $\rho(\cdot)$ denotes the spectral radius function. This method was called PDIRK, Parallel Diagonal-implicit Iterated Runge-Kutta.

Recently, in [HSw95], a mixture of the two strategies described above was presented and given the name PTIRK, Parallel Triangularly-implicit Iterated Runge-Kutta. Here, the matrix B was identified with a lower triangular matrix T such that $A = T\,U$ is the Crout decomposition of A, i.e. U is unit upper triangular. One easily verifies that for this T the stiff amplification matrix $Z_\infty(T)$ is strictly upper triangular. Throughout this paper, T will always denote this special lower triangular matrix. This choice of B yields, just like in PDIRK, a stiff amplification matrix that has a zero spectral radius. However, the new strategy leads to an amplification matrix $Z(z)$ that has a much smaller departure from normality than the amplification matrix in PDIRK. Consequently, the amplification after several iterations, i.e. the norm of the powers of $Z(z)$ is now considerably smaller (see [HSw95], Table 3.1). Suppose that all diagonal entries of T are distinct and that the eigenvalue decomposition of T is given by $T\,Q = Q\,D$, where D is diagonal and Q non-singular. Applying a similarity transformation in an analogous way as in [CB83], we arrive at the scheme

$$
\begin{aligned}
&T\,Q = Q\,D,\\
&(I - D \otimes h_n J_n)\Delta X_n^{(j+1)} = -(Q^{-1} \otimes I) R(Y_n^{(j)}),\\
&Y_n^{(j+1)} = Y_n^{(j)} + (Q \otimes I)\Delta X_n^{(j+1)}.
\end{aligned}
\tag{1.4}
$$

It is clear that the s components of $Y_n^{(j)}$ can be computed in parallel. The only additional costs of (1.4) with respect to PDIRK are the appliance of the transformations $(Q \otimes I)$ and $(Q^{-1} \otimes I)$.

In order to ensure the non-singularity of the matrix $(I - D \otimes h_n J_n)$ in (1.4), the positiveness of the diagonal entries of D is required. In [HSw95] the positiveness of D was proved for $s \leq 5$ and conjectured for $s > 5$. The main scope of this paper is to prove this conjecture. This will be done in Section 3, using operator theory.

The outline of the rest of the paper is as follows. Section 2 gives some preliminaries to the conjecture. In Section 4 we prove for $s = 2$, that the choice $B = T$ made in PTIRK is in some sense optimal.

2. Preliminaries

The $s \times s$ matrix A belonging to the RK collocation method with abscissa vector c has the form ([HW91], p. 82)

$$A = C\,V R\,V^{-1},$$

where $C = \operatorname{diag}\{c_1, c_2, \ldots, c_s\}$, $R = \operatorname{diag}\{1, 1/2, \ldots, 1/s\}$ and V is the Vandermonde matrix generated by c, i.e.

$$V = \begin{pmatrix} 1 & c_1 & \ldots & c_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \ldots & c_s^{s-1} \end{pmatrix}.$$

Here, the abscissae c_i have to be distinct. In the sequel the abscissae are also supposed to be positive. Without loss of generality, we assume that the RK method is written such that $c_1 < c_2 < \ldots < c_s$. Let $A = T\,U$ denote the Crout decomposition of A. The diagonal entries t_{kk} of T satisfy the formula [HSw95]

$$t_{kk} = \frac{|A_k|}{|A_{k-1}|}, \tag{2.1}$$

where $|A_j|$ denotes the determinant of the jth principal sub-matrix of A and $|A_0| := 1$. From (2.1) we see that the existence of the Crout decomposition immediately follows from the positiveness of t_{kk}.

In [HSw95] the authors proved the positiveness of t_{kk}, $k \in \{1, 2, \ldots, s\}$, for $s \le 5$ in the following way: first they showed that $|A_1|$ and $|A_s|$ are positive (for general s); then the positiveness of the remaining $|A_2|, \ldots, |A_{s-1}|$ was demonstrated by computing them explicitly; this approach does not lead to a proof for general s.

Another idea is to investigate whether the matrix $V R\,V^{-1}$ is positive definite. By using the result that every positive definite matrix has an LU-decomposition with positive diagonal entries ([GL89], p. 140), the proof of the conjecture would then easily follow, realizing that $T = C\,L$, where L is the lower triangular matrix in the Crout decomposition of $V R\,V^{-1}$. However, the following example shows that $V R\,V^{-1}$ is not always positive definite: If $s = 3$, $c = (1/3, 1/2, 2/3)^T$ and $x = (1, -3, -7)^T$, then $x^T V R\,V^{-1} x = -11$.

In the following section the proof of the conjecture will be given by considering VRV^{-1} as the matrix of an operator on the space of polynomials of degree less than s with respect to a basis of Lagrange polynomials.

3. Proof of the conjecture

Theorem 1 *Let V be the $s \times s$ Vandermonde matrix generated by $c_1, c_2, \ldots, c_s$, where $0 < c_1 < c_2 < \ldots < c_s$, let R be the diagonal matrix* $\operatorname{diag}(1, 1/2, \ldots, 1/s)$. *There exist a lower triangular matrix L, and unit upper triangular matrix U, such that $LU = VRV^{-1}$. The diagonal entries of L are positive.*

Notice that from this theorem it immediately follows that for any $s \times s$ RK collocation matrix A with positive distinct abscissae, there exists a lower triangular matrix T with positive diagonal entries such that $Z_\infty(T)$ is strictly upper triangular, by setting $T = C\,L$.

Proof of Theorem 1: Let $\mathbf{P}_s$ be the s-dimensional linear space of polynomials of degree less than s with real coefficients, and $\mathcal{C}$ the canonical basis for $\mathbf{P}_s$, i.e.

$$\mathcal{C} = \{1, x, \ldots, x^{s-1}\}.$$

Define the operator $H : \mathbf{P}_s \to \mathbf{P}_s$ by $H(p) = q$ where q is defined by

$$q(x) = \frac{1}{x} \int_0^x p(t)\, dt.$$

We use the notation $\operatorname{mat}(H)_\mathcal{C}$ for the matrix of the operator H with respect to the basis $\mathcal{C}$. It can be easily verified that

$$\operatorname{mat}(H)_\mathcal{C} = R.$$

We denote the kth Lagrange polynomial with respect to $c_1, c_2, \ldots, c_s$ by l_k:

$$l_k(x) = \prod_{i \neq k} \frac{x - c_i}{c_k - c_i} \quad ; \quad k \in \{1, 2, \ldots, s\}.$$

Notice that l_k is of degree $s-1$ and thus element of $\mathbf{P}_s$. The Lagrange polynomials define also a basis for $\mathbf{P}_s$, which will be denoted by $\mathcal{L}$:

$$\mathcal{L} = \{l_1, l_2, \ldots, l_s\}.$$

We write $\mathcal{C}_\mathcal{L}$ for the matrix that expresses the canonical basis in the Lagrange basis. Since for every $m \in \{0, 1, \ldots, s-1\}$ the equality

$$x^m = {c_1}^m l_1 + {c_2}^m l_2 + \ldots + {c_s}^m l_s$$

should hold, it can be seen that $\mathcal{C}_\mathcal{L} = V$. Consequently, the matrix of the operator H with respect to the basis $\mathcal{L}$ is given by

$$\mathrm{mat}(H)_{\mathcal{L}} = \mathcal{C}_{\mathcal{L}} \cdot \mathrm{mat}(H)_{\mathcal{C}} \cdot {\mathcal{C}_{\mathcal{L}}}^{-1} = VRV^{-1} =: B.$$

If $\big(H(l_k)\big)_{\mathcal{L}}$ denotes the image under H of l_k with respect to the basis $\mathcal{L}$, then

$$\big(H(l_k)\big)_{\mathcal{L}} = Be_k = \begin{pmatrix} \beta_{1k} \\ \vdots \\ \beta_{nk} \end{pmatrix},$$

where e_k is the kth canonical basis vector of $\mathbf{R}^s$ and $(\beta_{ij}) = B$.
We claim that $\beta_{11} > 0$. To see this, notice that $H(l_1)$ is a polynomial with coefficient β_{11} in the direction of l_1. Since $l_k(c_1) = 0$ for $k > 1$, it is clear that

$$\left(H(l_1)\right)(c_1) = \beta_{11}.$$

With respect to the value of l_1 in zero, we observe that $l_1(c_1) = 1$, and that all its roots are to the right of c_1; therefore l_1 is positive on $[0, c_1]$, which implies

$$\left(H(l_1)\right)(c_1) = \frac{1}{c_1}\int_0^{c_1} l_1(t)\,dt > 0.$$

Consequently, $\beta_{11} > 0$.
It is now possible to define

$$v_{1k} := -\frac{\beta_{1k}}{\beta_{11}} \quad ; \quad k \in \{2, \dots, s\}.$$

From this definition it follows that, for $k > 1$,

$$\big(H(l_k + v_{1k}l_1)\big)_{\mathcal{L}} = B(e_k + v_{1k}e_1) = B \begin{pmatrix} v_{1k} \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \beta_{2k}^{(1)} \\ \vdots \\ \beta_{nk}^{(1)} \end{pmatrix}.$$

Assuming $\beta_{22}^{(1)} \neq 0$, we are able to define

$$v_{2k} := -\frac{\beta_{2k}^{(1)}}{\beta_{22}^{(1)}} \quad ; \quad k \in \{3, \dots, s\},$$

such that

$$\big(H(l_k + v_{2k}l_2 + v_{1k}l_1)\big)_{\mathcal{L}} = B \begin{pmatrix} v_{1k} \\ v_{2k} \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \beta_{3k}^{(2)} \\ \vdots \\ \beta_{nk}^{(2)} \end{pmatrix}.$$

Continuing this procedure, we finally arrive at

$$\big(H(\sum_{i=1}^{k} v_{ik}l_i)\big)_{\mathcal{L}} = Bu_k = r_k,$$

where

$$v_{ik} = \begin{cases} -\frac{\beta_{ik}^{(i-1)}}{\beta_{ii}^{(i-1)}} & \text{for} \quad i < k, \\ 1 & \text{for} \quad i = k, \end{cases}$$

(defining $\beta_{ij}^{(0)} = \beta_{ij}$) and u_k and r_k are vectors defined by

$$u_k = \begin{pmatrix} v_{1k} \\ \vdots \\ v_{k-1,k} \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad r_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta_{kk}^{(k-1)} \\ \vdots \\ \beta_{nk}^{(k-1)} \end{pmatrix}.$$

If we can show that $\beta_{kk}^{(k-1)} > 0$ for $k \in \{2, 3, \ldots, s\}$, we have demonstrated that the procedure outlined above can be carried out. By observing that u_k and r_k are columns of matrices $\widetilde{U}$ and L, respectively, for which the relation $B\widetilde{U} = L$ holds, we then have proved Theorem 1 using U for $\widetilde{U}^{-1}$.

The vectors u_k and r_k can be considered as polynomials in $\mathbf{P}_s$ with respect to the basis $\mathcal{L}$. Moreover, r_k is the image of u_k under the operator H:

$$H(u_k) = r_k.$$

Since $r_k(c_k) = \beta_{kk}^{(k-1)}$, we have to prove that $r_k(c_k) > 0$. We define the polynomial U_k of degree $s{+}1$ by

$$U_k(x) = \int_0^x u_k(t)\,dt.$$

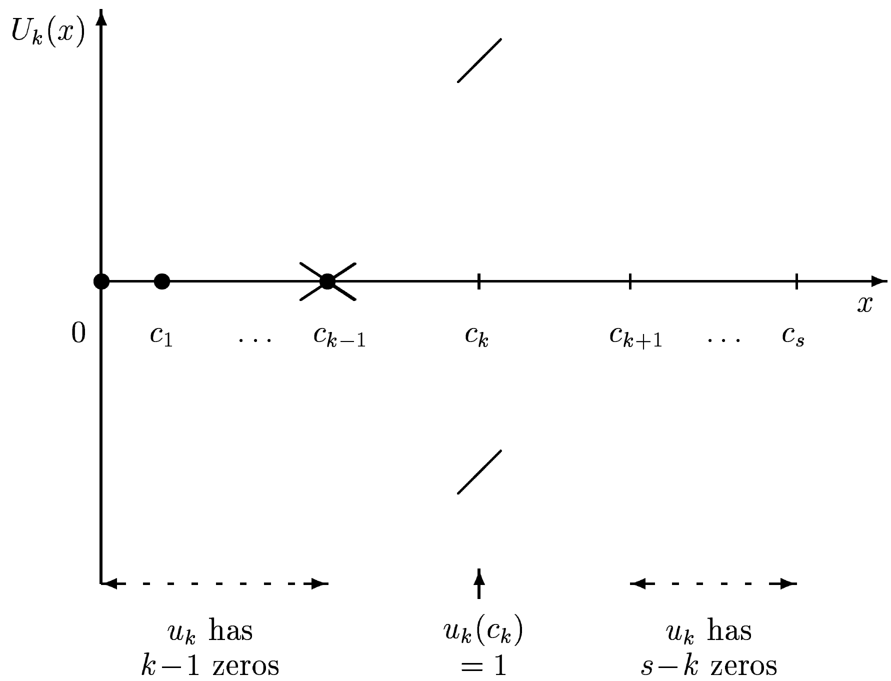

Figure 1: Sketch of $U_k(x)$

Notice that $U_k(0) = 0$ and, for $x > 0$, the sign of r_k equals the sign of U_k (the latter holds since $U_k = xr_k$). Since $l_k(c_i) = 0$ for $i < k$ and r_k has only components in the direction of l_j with $j \geq k$, we see that $r_k(c_i) = 0$ for $i < k$ and consequently

$$U_k(c_i) = 0 \quad \text{for} \quad i < k.$$

This means that u_k (being the derivative of U_k) has $k-1$ zeros in the interval $(0, c_{k-1})$. All components of u_k in the direction of the last $s-k$ Lagrange polynomials are zero. Consequently, $u_k(c_i) = 0$ for $i > k$, so that u_k has $s-k$ zeros in the interval $[c_{k+1}, c_k]$.

We now consider 2 cases (see also Figure 1):

$$u_k(c_{k-1}) > 0, \tag{3.1}$$
$$u_k(c_{k-1}) < 0. \tag{3.2}$$

Remark that, since all c_i are distinct, U_k has a single zero in c_{k-1}, so that the situation $u_k(c_{k-1}) = 0$ does not arise. Suppose that (3.2) holds. Since $u_k(c_k) = 1$, the polynomial u_k should have a zero in the interval (c_{k-1}, c_k). In that case, u_k has $(k-1) + (s-k) + 1 = s$

zeros. However, the degree of u_k is only $s-1$, proving that only situation (3.1) can occur, and $u_k > 0$ on (c_{k-1}, c_k).

From $U_k(c_{k-1}) = 0$, it now follows that $U_k(c_k) > 0$. Since r_k has the same sign as U_k, we have proved the theorem. □

4. Is PTIRK optimal?

In this section we investigate the optimality of the matrix T in PTIRK. Since the number of parameters becomes too large to handle conveniently for $s > 2$, we restrict ourselves here to methods with 2 implicit stages, i.e. $s = 2$.

In the class of lower triangular matrices, T is optimal in the sense that it leads to the smallest stiff amplification matrix measured in the infinity norm:

Theorem 2 *If L is a 2×2 lower triangular matrix, then*

$$\|Z_\infty(L)\|_\infty \geq \|Z_\infty(T)\|_\infty.$$

Proof: Write $L^{-1} = (l_{ij})$ with $l_{12} = 0$. Then

$$Z_\infty(L) = \begin{pmatrix} 1 + \frac{l_{1,1} c_1 (-2\, c_2 + c_1)}{2\,(c_2 - c_1)} & \frac{l_{1,1} {c_1}^2}{2\,(c_2 - c_1)} \\ * & * \end{pmatrix}.$$

Define for $x > 0$:

$$g(x) = \left| 1 + \frac{c_1\,(-2\,c_2 + c_1)}{2\,(c_2 - c_1)}\, x \right| + \frac{{c_1}^2}{2\,(c_2 - c_1)}\, x.$$

Then $g(x) \geq g(x_{\min}) = c_1/(2\,c_2 - c_1)$, where $x_{\min} = 2\,(c_2 - c_1)/(c_1\,(2\,c_2 - c_1))$. Since $\|Z_\infty(T)\|_\infty = g(x_{\min})$, it follows that $\|Z_\infty(L)\|_\infty \geq \|Z_\infty(T)\|_\infty$. □

For two well-known stiffly accurate RK methods with 2 implicit stages, it is possible to show that in the class of lower triangular matrices that lead to a 'small' stiff amplification matrix, T is optimal in the sense that it has the smallest non-stiff amplification matrix, again measured in the infinity norm:

Theorem 3 *If L is a 2×2 lower triangular matrix with the property that $\rho(Z_\infty(L)) = 0$, then, for the 2-stage RadauIIA, and the 3-stage Lobatto IIIA method,*

$$\|Z_0(L)\|_\infty \geq \|Z_0(T)\|_\infty.$$

Proof: Write $A = (a_{ij})$ and $L = (l_{ij})$ with $l_{12} = 0$. Then $\|Z_0(L)\|_\infty = \max(m_1, m_2)$, where m_1 and m_2 are given by

$$m_1 = |a_{11} - l_{11}| + |a_{12}| \quad \text{and} \quad m_2 = |a_{21} - l_{21}| + |a_{12} - l_{22}|\,.$$

Let J be the interval such that if $l_{11} \notin J$, then $m_1 > \|Z_0(T)\|_\infty$. Notice that J only depends on c. From $\sigma(Z_\infty(L)) = 0$ it follows that $\text{trace}(Z_\infty(L)) = \det(Z_\infty(L)) = 0$. Using these two equations, it is possible to express l_{21} and l_{22}, and thus m_2, in l_{11}. We have to proof that for $l_{11} \in J$, $m_2 \geq \|Z_0(T)\|_\infty$. We treat the two methods separately.

RadauIIA: $c = (1/3, 1)^T$, $\|Z_0(T)\|_\infty = 3/20$, $J = [7/20, 29/60]$, and

$$m_2(l_{11}) = \left|\frac{3}{4} + \frac{-24\, l_{1,1} + 5 + 18\, {l_{1,1}}^2}{6\, l_{1,1}}\right| + \left|\frac{1}{4} - \frac{1}{6\, l_{1,1}}\right|.$$

It can be verified that $\min\limits_{l_{11} \in J} (m_2(l_{11})) = m_2(t_{11}) = 3/20$.

LobattoIIIA: $c = (0, 1/2, 1)^T$, $\|Z_0(T)\|_\infty = 1/12$, $J = [7/24, 3/8]$, and

$$m_2(l_{11}) = \left|\frac{2}{3} + \frac{-12\, l_{1,1} + 2 + 12\, {l_{1,1}}^2}{3\, l_{1,1}}\right| + \left|\frac{1}{6} - \frac{1}{12\, l_{1,1}}\right|$$

The reader is invited to check that $\min\limits_{l_{11} \in J} (m_2(l_{11})) = m_2(t_{11}) = 1/12$.

□

ACKNOWLEDGEMENTS
With great pleasure we acknowledge the insight of Hendrik W. Lenstra who shared with the first author the idea to use the materialization of the operator H on different bases for the purpose of the proof of Theorem 1.

REFERENCES

[CB83] G.J. Cooper and J.C. Butcher. An iteration scheme for implicit Runge-Kutta methods. *IMA J. Numer. Anal.*, 3:127–140, 1983.

[GL89] G.H. Golub and C.F. van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore and London, second edition, 1989.

[HSo91] P.J. van der Houwen and B.P. Sommeijer. Iterated Runge-Kutta methods on parallel computers. *SIAM J. Sci. Stat. Comput.*, 12:1000–1028, 1991.

[HSw95] P.J. van der Houwen and J.J.B. de Swart. Triangularly implicit iteration methods for ODE-IVP solvers. Technical Report NM-R9510, CWI, Amsterdam, 1995. Submitted for publication.

[HW91] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-algebraic Problems.* Springer-Verlag, 1991.