Applications of Mathematics

Jan Zítko

Using successive approximations for improving the convergence of GMRES method

Applications of Mathematics, Vol. 43 (1998), No. 5, 321–350

Persistent URL: http://dml.cz/dmlcz/134392

Terms of use:

© Institute of Mathematics AS CR, 1998

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://dml.cz

43 (1998) APPLICATIONS OF MATHEMATICS No. 5, 321–350

USING SUCCESSIVE APPROXIMATIONS FOR IMPROVING THE CONVERGENCE OF GMRES METHOD

Jan Zítko, Praha

(Received February 16, 1996)

Abstract. In this paper, our attention is concentrated on the GMRES method for the solution of the system $(I - T)x = b$ of linear algebraic equations with a nonsymmetric matrix. We perform m pre-iterations $y_{l+1} = Ty_l + b$ before starting GMRES and put y_m for the initial approximation in GMRES. We derive an upper estimate for the norm of the error vector in dependence on the mth powers of eigenvalues of the matrix T. Further we study under what eigenvalues lay-out this upper estimate is the best one. The estimate shows and numerical experiments verify that it is advisable to perform pre-iterations before starting GMRES as they require fewer arithmetic operations than GMRES. Towards the end of the paper we present a numerical experiment for a system obtained by the finite difference approximation of convection-diffusion equations.

Keywords: GMRES, iterative method, numerical experiments, solution of discretized equations

MSC 2000: 65F10, 65N22

1. Introduction

One of the basic problems of numerical computing is the ability to solve linear systems $Ax = f$ arising from finite difference or finite element approximations of partial differential equations or as intermediate steps in computing the solution of nonlinear problems. The matrices of such systems are usually large and sparse.

The iterative or semi-iterative methods generate a sequence of approximate solutions $\{x_k\}$ and the evaluationof an iterative method invariably focuses on how quickly the iterates x_k converge. A difficulty associated with SOR, ADI and other

This paper was supported by the Grant Agency of the Czech republic under Grant No 201/96/0918.

accelerated iterative methods or Chebyshev semiiterative methods is that they depend upon parameters that are sometimes hard to choose properly. How to avoid this difficulty? In 1952, Hestenes and Stiefel introduced the conjugate gradient method which is an algorithm for solving symmetric positive definite linear systems. The CG-type methods need no accelerated parameters and have been developed later by many authors.

For nonsymmetric systems a lot of methods were proposed which are based upon a projection onto a Krylov subspace. In 1986, Saad and Schultz ([S-S 86]) introduced an iterative method, theoretically equivalent to GCR (see [E 82]), which has the property of minimizing at every step the norm of the residual over a Krylov subspace. The GMRES method has become very popular for the solution of nonsingular and nonsymmetric linear systems.

Another way to accelerate the convergence of basic linear iterative methods has been the use of extrapolation procedures. However, it was shown in the paper [Si-88] that the extrapolation methods MPE, RRE and the topological epsilon algorithm when applied to linearly generated vector sequences are Krylov subspace methods.

On the basis of the paper [Zi 83], where an improvement of the convergence of iterative processes is described, the GMRES method is introduced here for theoretical investigations as an extrapolation procedure for accelerating the convergence of successive approximations $x_{k+1} = Tx_k + b$ for solving the linear system $(I-T)x = b$. Let $x^\star$ denote the solution of this system. By practical use of GMRES the storage requirements per iteration grow linearly and a number of multiplications grows quadratically. Therefore restarts are necessary. But successive approximations need only one iteration vector and every iteration needs the same number of operations. Therefore, a question occurs, what happens if we first proceed m successive approximations and then take the resulting iteration x_m as a starting vector for GMRES. Will the convergence behaviour of such a modified GMRES be better? Numerical tests confirm this hypothesis. The theory presented in this paper explains this observation.

In Sections 2 and 3 we present some auxiliary considerations which enable us for a given k to calculate $x_k - x^*$ in the case that the matrix T possesses a general Jordan canonical form. In Section 4 we show that $\|x_k - x^*\| \leqslant L(m)m^\delta|\lambda_\mu|^m(u_k(m) + w_k(m))$, where $L(m) \in \langle 1, \|I-T\|\|(I-T)^{-1}\|\rangle$, δ is an integer, λ_μ is an eigenvalue, $\{u_k(m)\}_{m=0}^\infty$ and $\{w_k(m)\}_{m=0}^\infty$ are sequences of nonnegative numbers such that $\limsup_{m\to\infty} u_k(m) < \infty$ and $\lim\limits_{k\to\infty} w_k(m) = 0$. We obtain the index $\mu = \mu(k)$ easily from the Jordan canonical form of the matrix T. Complete formulas and estimates are given in precisely formulated propositions. This estimate inspired us to use successive approximations before starting GMRES. Numerical examples show that this approach could be very advantageous because one successive approximation costs less work. In Section 5 we study under what eigenvalues lay-out the upper

estimate for $\|x_k - x^*\|$ is the best one. The numerical examples are focused on the systems which we obtain by finite difference approximation of convection-diffusion equations. We have tested our generalization of GMRES, i.e. the use of the so called "pre-iteration" on systems with ten and twenty thousand unknowns and we have found it very effective especially when SOR pre-iterations were used. Numerical examples and graphs conclude the paper.

2. Auxiliary considerations

Let us consider the system

$$(I - T)x = b, \tag{2.1}$$

where $T \in L(\mathbb{R}^n)$, $1 \notin \sigma(T)$, $x, b \in \mathbb{R}^n$. Here $\sigma(T)$ denotes the spectrum of the matrix T. Recall that the symbol $e_i^{(s)}$ denotes the ith column of the identity matrix $I_s \in L(\mathbb{R}^s)$ and $e^{(s)} = \sum_{i=1}^{s} e_i^{(s)}$. If $u_i \in \mathbb{R}^n$ or $u_i \in \mathbb{C}^n$, $i = 1, \ldots, r$, then the symbol $(u_1, u_2, \ldots, u_r)$ denotes a rectangular $n \times r$ matrix with columns $u_1, u_2, \ldots, u_r$. The symbol $(U_1, U_2, \ldots, U_r)$, where $U_i \in L(\mathbb{R}^{k_i}, \mathbb{R}^n)$ or $U_i \in L(\mathbb{C}^{k_i}, \mathbb{C}^n)$, denotes, analogously, a rectangular matrix with block columns $U_1, U_2, \ldots, U_r$. The symbol J_s denotes the square matrix $(\Theta, e_1^{(s)}, e_2^{(s)}, \ldots, e_{s-1}^{(s)})$.

Let $x_0 \in \mathbb{R}^n$, $r_0 = b - (I - T)x_0 \neq 0$. Let $k \leqslant n$ be an integer and

$$\mathscr{K}_k(r_0, I - T) = \operatorname{span}\{r_0, (I - T)r_0, \ldots, (I - T)^{k-1} r_0\} \tag{2.2}$$

the Krylov subspace generated by the matrix of the system (2.1) and the residual r_0. It is well known that the GMRES-algorithm computes for any starting vector $x_0 \in \mathbb{R}^n$ the kth approximation x_k to the solution x^* of the equation (2.1) in the form $x_k = x_0 + u_k$, where $u_k \in \mathscr{K}_k(r_0, I - T)$ with the minimization property

$$\|r_0 - (I - T)u_k\| = \min_{u \in \mathscr{K}_k(r_0, I-T)} \|r_0 - (I - T)u\|. \tag{2.3}$$

The symbol $\|x\|$ denotes the l_2-norm of the vector x. In practice the minimum in (2.3) is usually being looked for so that in the space $\mathscr{K}_k(r_0, I - T)$ an orthonormal base is constructed according to the Arnoldi process. For the sake of clarity let us present here the whole GMRES algorithm. The following algorithm combines the procedures given in [V-V 93] and in [S-S 86].

Algorithm 2.1.

1) Choose $x_0 \in \mathbb{R}^n$ and put $r_0 = b - (I - T)x_0$ and $v_1 = r_0/\|r_0\|$.

2) For $j = 1, 2 \ldots, k$ do
$\widehat{v}_{j+1} = (I - T)v_j$; $v_{j+1} = \widehat{v}_{j+1}$;
For $i = 1, 2 \ldots, j$ do
$h_{ij} = v_i^T \widehat{v}_{j+1}$;
$v_{j+1} = v_{j+1} - h_{ij} v_i$
End for i.
Put $h_{j+1,j} = \|v_{j+1}\|$; $\quad v_{j+1} = v_{j+1}/h_{j+1,j}$
End for j.
3) Calculate $x_k = x_0 + u_k$ on the condition (2.3).
End of Algorithm 2.1.

It is easy to show that if for $k < n$ the equality $\dim \mathscr{K}_{k+1}(r_0, I - T) = k + 1$ holds then it is possible to construct nonzero and mutually orthonormal vectors $v_1, v_2, \ldots, v_{k+1}$.

Assumption 1. *Let for $k < n$ $\dim \mathscr{K}_{k+1}(r_0, I - T) = k + 1$.*

If it would be

$$\dim \mathscr{K}_{\hat{k}}(r_0, I - T) = \hat{k} \quad \text{and} \quad \dim \mathscr{K}_{\hat{k}+1}(r_0, I - T) = \hat{k}$$

for some $\hat{k} < n$ then $x^* = x_{\hat{k}}$. We define a matrix $H_k = (h_{ij})_{i,j=1,\ldots,k}$ with $h_{ij} = 0$ for $i \geqslant j + 2$. Let

$$\widehat{H}_k = \begin{pmatrix} H_k \\ \widehat{h} \end{pmatrix}, \quad \text{where} \quad \widehat{h} = (0, 0, \ldots, 0, h_{k+1,k}). \tag{2.4}$$

The matrix $\widehat{H}_k$ is a rectangular $(k + 1) \times k$ matrix.

Algorithm 2.1 immediately reveals that

$$(I - T)V_k = V_{k+1}\widehat{H}_k, \tag{2.5}$$

where

$$V_k = (v_1, \ldots, v_k) \quad \text{and} \quad V_{k+1} = (v_1, \ldots, v_{k+1}). \tag{2.6}$$

If we multiply (2.5) from the left by the matrix V_k^T, we obtain

$$V_k^T (I - T)V_k = H_k. \tag{2.7}$$

Let us now put $u = V_k z$ in (2.3). The problem of finding u_k is usually formulated so that we look for a vector $z_k \in \mathbb{R}^k$ in such a way that

$$\|r_0 - (I - T)V_k z_k\| = \min_{z \in \mathbb{R}^k} \|r_0 - (I - T)V_k z\|. \tag{2.8}$$

The expression on the right hand side is modified using (2.5). Let us denote $\beta = \|r_0\|$. Then

$$\begin{aligned}\|r_0 - (I-T)V_k z\| &= \|\beta v_1 - V_{k+1}\widehat{H}_k z\| = \|\beta V_{k+1} e_1^{(k+1)} - V_{k+1}\widehat{H}_k z\| \\ &= \|V_{k+1}(\beta e_1^{(k+1)} - \widehat{H}_k z)\| = \|\beta e_1^{(k+1)} - \widehat{H}_k z\|.\end{aligned}$$

In practice the minimalization of the functional $\|\beta e_1^{(k+1)} - \widehat{H}_k z\|$ is done by using the QR-decomposition, and it is described in [S-S 86] in great detail. Computer programs are prepared according to the just presented procedure.

Now we will start by studying the behaviour of $\|x_k - x^*\|$, analogously as in the papers [Zi 83], [Zi 84] and [Si 88]. First, we will generalize the GMRES method in the sense of the following algorithm.

Algorithm 2.2.

1) Choose $y_0 \in \mathbb{R}^n$ and an integer $m \geqslant 0$.
2) Calculate the m-th iteration y_m by the following iterative process:

$$y_{l+1} = Ty_l + b\,, \quad l = 0, 1, 2, \ldots, m. \tag{2.9}$$

3) Put $x_0 = y_m$ and carry out k steps of the GMRES (i.e. calculate x_k according to Algorithm 2.1).
End of Algorithm 2.2.

Throughout the paper we assume exact arithmetic. It is easy to see that

$$\mathscr{K}_k(r_0, I-T) = \mathscr{K}_k(r_0, T) = \operatorname{span}\{r_0, Tr_0, T^2 r_0, \ldots, T^{k-1} r_0\}. \tag{2.10}$$

(See [F-F] or [Zi 96].)

On the basis of (2.10) the approximation x_k can theoretically be expressed in the form

$$x_k = x_0 + \nu_0 r_0 + \nu_1 T r_0 + \nu_2 T^2 r_0 + \ldots + \nu_{k-1} T^{k-1} r_0, \tag{2.11}$$

while in the sense of Algorithm 2.1, the numbers ν_i, $i = 0, \ldots k-1$, are constructed so that

$$\|r_k\| = \min_{(\mu_0, \mu_1, \ldots, \mu_{k-1})^T \in \mathbb{R}^k} = \|r_0 - (I-T)(\mu_0 r_0 + \mu_1 T r_0 + \ldots + \mu_{k-1} T^{k-1} r_0)\|.$$

Let us define numbers $\alpha_0^{(k)}, \alpha_1^{(k)}, \ldots, \alpha_k^{(k)}$ as the solution of the following nonsingular system with a lower triangular matrix,

$$\sum_{t=s}^{k} \alpha_t^{(k)} = \nu_{s-1} \quad \text{for } s = k, k-1, \ldots, 0, \tag{2.12}$$

putting $\nu_{-1} = 1$. If we define further iterations $y_{m+1}, y_{m+2}, \dots$ according to (2.9), then

$$r_0 = b - (I - T)x_0 = Tx_0 + b - x_0 = Ty_m + b - y_m = y_{m+1} - y_m \tag{2.13}$$

and thus

$$T^t r_0 = y_{m+t+1} - y_{m+t} \tag{2.14}$$

for any nonnegative integer t. If we substitute (2.12) and (2.14) into (2.11), we obtain

$$x_k = \sum_{t=0}^{k} \alpha_t^{(k)} y_{m+t}, \tag{2.15}$$

and analogously for r_k we obtain the expression

$$\begin{aligned} r_k &= b - (I - T)x_k = \sum_{t=0}^{k} \alpha_t^{(k)} (b - (I - T)y_{m+t}) \\ &= \sum_{t=0}^{k} \alpha_t^{(k)} (y_{m+t+1} - y_{m+t}) = \sum_{t=0}^{k} \alpha_t^{(k)} T^t r_0, \end{aligned} \tag{2.16}$$

where we have used the relation (2.14) in the last equation. Let us calculate

$$\begin{aligned} \|r_k\|^2 = r_k^T r_k &= \left(\sum_{t=0}^{k} \alpha_t^{(k)} T^t r_0 \right)^T \left(\sum_{t=0}^{k} \alpha_t^{(k)} T^t r_0 \right) \\ &= (\alpha_0^{(k)}, \alpha_1^{(k)}, \dots, \alpha_k^{(k)})^T A_{k+1} (\alpha_0^{(k)}, \alpha_1^{(k)}, \dots, \alpha_k^{(k)}), \end{aligned}$$

where $A_{k+1} \in L(\mathbb{R}^{k+1})$ is a matrix in the form

$$A_{k+1} = \begin{pmatrix} r_0^T r_0, & r_0^T T r_0, & r_0^T T^2 r_0, & \dots, & r_0^T T^k r_0 \\ (Tr_0)^T r_0, & (Tr_0)^T T r_0, & (Tr_0)^T T^2 r_0, & \dots, & (Tr_0)^T T^k r_0 \\ \dots & \dots & \dots & \dots & \dots \\ (T^k r_0)^T r_0, & (T^k r_0)^T T r_0, & (T^k r_0)^T T^2 r_0, & \dots, & (T^k r_0)^T T^k r_0 \end{pmatrix}.$$

We can see that

$$(A_{k+1})_{ij} = (T^{i-1} r_0)^T T^{j-1} r_0. \tag{2.17}$$

If we denote

$$\vec{\alpha}_{k+1} = (\alpha_0^{(k)}, \alpha_1^{(k)}, \dots, \alpha_k^{(k)})^T, \tag{2.18}$$

then

$$\|r_k\|^2 = \vec{\alpha}_{k+1}^T A_{k+1} \vec{\alpha}_{k+1}. \tag{2.19}$$

N o t e. The upper indices of the coefficients $\alpha_i^{(k)}$ mean that k relates to x_k or r_k. The lower index of $\vec{\alpha}_{k+1}$ or A_{k+1} denotes the dimension.

The matrix A_{k+1} is symmetric and according to Assumption 1 it is positive definite. Thus the condition of the minimum can be formulated as follows. Let us denote by the symbol $\mathscr{M}$ the set of all vectors from $\mathbb{R}^{k+1}$ for which the sum of their components equals 1. Then

$$\vec{\alpha}_{k+1} = \arg\min_{w \in \mathscr{M}} w^T A_{k+1} w. \tag{2.20}$$

Theorem 2.1. *There exists only one vector $\vec{\alpha}_{k+1}$ which solves the problem (2.20). If $B_{k+1} \in L(\mathbb{R}^{k+1})$ is a matrix whose elements have the form*

$$\begin{aligned} (B_{k+1})_{t,s} &= (T^{t-1}(T-I)r_0)^T T^{s-1} r_0 \quad \text{for } t = 1, \dots, k; s = 1, \dots, k+1, \\ (B_{k+1})_{k+1,s} &= 1 \quad \text{for } s = 1, \dots, k+1, \end{aligned} \tag{2.21}$$

then the vector $\vec{\alpha}_{k+1}$ is the only solution of the system

$$B_{k+1} z = e_{k+1}^{(k+1)}, \quad z \in \mathbb{R}^{k+1}. \tag{2.22}$$

P r o o f. For $w \in \mathbb{R}^{k+1}$, $w = (w_1, w_2, \dots, w_{k+1})^T$ let us put

$$F(w) = w^T A_{k+1} w, \quad G(w) = \sum_{t=0}^{k+1} w_t - 1,$$

in order to shorten the notation. If the function F possesses on the set $\mathscr{M}$ its minimum at $\vec{\alpha}_{k+1}$, then there exists $\lambda \in \mathbb{R}$ such that the following equations hold:

$$\begin{aligned} F'(\vec{\alpha}_{k+1}) + \lambda G'(\vec{\alpha}_{k+1}) &= 0, \\ G(\vec{\alpha}_{k+1}) &= 0, \end{aligned}$$

where $F'(\vec{\alpha}_{k+1})$ and $G'(\vec{\alpha}_{k+1})$ denote the Gâteaux derivative of the functionals F and G respectively at $\vec{\alpha}_{k+1}$. In our case the Gâteaux derivative equals the Fréchet derivative. If we calculate both derivatives and make a transposition, we obtain a system of $k+2$ linear algebraic equations

$$A_{k+1}\vec{\alpha}_{k+1} + \tfrac{\lambda}{2} e^{(k+1)} = \Theta, \tag{2.23}$$

$$(e^{(k+1)})^T \vec{\alpha}_{k+1} = 1. \tag{2.24}$$

From (2.23) and (2.24) we easily obtain that $\vec{\alpha}_{k+1}$ is a solution of (2.22). As the system (2.23), (2.24) is nonsingular, the matrix B_{k+1} is nonsingular as well. □

Let us remark that the matrix B_{k+1} has the form

$$\begin{pmatrix} ((T-I)r_0)^T)r_0, & ((T-I)r_0)^T T r_0, & \ldots, & ((T-I)r_0)^T T^k r_0 \\ (T(T-I)r_0)^T r_0, & (T(T-I)r_0)^T T r_0, & \ldots, & (T(T-I)r_0)^T T^k r_0 \\ \cdots & \cdots & \cdots & \cdots \\ (T^{k-1}(T-I)r_0)^T r_0, & (T^{k-1}(T-I)r_0)^T T r_0, & \ldots, & (T^{k-1}(T-I)r_0)^T T^k r_0 \\ 1 & 1 & ,\ldots, & 1 \end{pmatrix}.$$

The relations (2.15) and (2.24) imply

$$x_k - x^* = \sum_{t=0}^{k} \alpha_t^{(k)} (y_{m+t} - x^*). \tag{2.25}$$

For calculating $x_k - x^*$ we need to express $y_{m+t} - x^*$ and estimate $\alpha_t^{(k)}$. For this purpose we first express vectors

$$T^{m+t}(y_0 - x^*), \quad T^{t-1}(T-I)r_0 = T^t r_0 - T^{t-1} r_0, \quad T^s r_0$$

for integers $t \geqslant 1$, $s \geqslant 0$ and on the basis of the obtained formulas we will modify the elements of the matrix B_{k+1}.

3. Modification of the auxiliary elements of the matrix B_{k+1}

Let the matrix U transform the matrix T into its Jordan canonical form, i.e.,

$$U^{-1}TU = \operatorname{diag}(\lambda_1 I_{i_1} + J_{i_1}, \lambda_2 I_{i_2} + J_{i_2}, \lambda_3 I_{i_3} + J_{i_3}, \ldots, \lambda_p I_{i_p} + J_{i_p}). \tag{3.1}$$

The matrix U is in general a complex matrix, i.e., it is an element of $L(\mathbb{C}^n)$.

Assumption 2. *Let*

$$|\lambda_1| \geqslant |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_p| > 0, \tag{3.2}$$

and let all eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$ be mutually different. If $|\lambda_j| = |\lambda_{j+1}|$ let $i_j \geqslant i_{j+1}$.

Let us denote the columns of the transformation matrix U successively by $u_1, u_2, \ldots, u_n$ and let us put

$$\begin{aligned} t_l &= \sum_{s=1}^{l} i_s \quad \text{for } l = 0, 1, \ldots, p, \\ U_{i+1} &= (u_{t_i+1}, u_{t_i+2}, \ldots, u_{t_{i+1}}) \quad \text{for } i = 0, 1, \ldots, p-1. \end{aligned} \tag{3.3}$$

Accordingly,

$$U = (U_1, U_2, \ldots, U_p). \tag{3.4}$$

It follows from the relation (3.1) that

$$T^l U = U \operatorname{diag}((\lambda_1 I_{i_1} + J_{i_1})^l, (\lambda_2 I_{i_2} + J_{i_2})^l, (\lambda_3 I_{i_3} + J_{i_3})^l, \ldots, (\lambda_p I_{i_p} + J_{i_p})^l) \tag{3.5}$$

for any positive integer l. Let P_i denote the projection of the space $\mathbb{C}^n$ into the subspace generated by the columns of the matrix U_i for $i = 1, 2, \ldots, p$. Let

$$y_1 - y_0 = a_1 + a_2 + \ldots + a_p, \tag{3.6}$$

$$y_0 - x^* = \widehat{a}_1 + \widehat{a}_2 + \ldots + \widehat{a}_p, \tag{3.7}$$

where $a_j, \widehat{a}_j \in P_j \mathbb{C}^n \ \ \forall j$. Let us define the vectors $b_j, \widehat{b}_j \in \mathbb{C}^{i_j} \ \ \forall j$ by the relations

$$a_j = U_j b_j \quad \text{and} \quad \widehat{a}_j = U_j \widehat{b}_j. \tag{3.8}$$

For a positive integer l let us calculate $T^l r_0$ and $T^l(y_0 - x^*)$. We perform the calculation only for $T^l r_0$, the procedure for $T^l(y_0 - x^*)$ is identical. Since according to (3.5) $T^l a_j = U_j(\lambda_j I_{i_j} + J_{i_j})^l b_j$, we have

$$T^l(y_1 - y_0) = \sum_{j=1}^{l} U_j(\lambda_j I_{i_j} + J_{i_j})^l b_j. \tag{3.9}$$

If we denote $b_j = (\beta_1^{(j)}, \ldots, \beta_{i_j}^{(j)})^T$, then

$$U_j(\lambda_j I_{i_j} + J_{i_j})^l b_j = \sum_{i=1}^{i_j} \binom{l}{i-1} \lambda_j^l v_{ji} \tag{3.10}$$

where we have put

$$v_{ji} = \left(\sum_{s=i}^{i_j} \beta_s^{(j)} u_{t_{j-1}+s-i+1} \right) \Big/ \lambda_j^{i-1} \quad \text{and} \quad \binom{l}{s} = 0 \quad \text{for } s > l. \tag{3.11}$$

Consequently,

$$T^l(y_1 - y_0) = y_{l+1} - y_l = \sum_{j=1}^{p} \sum_{i=1}^{i_j} \binom{l}{i-1} \lambda_j^l v_{ji}. \tag{3.12}$$

By analogy, if we put

$$\widehat{v}_{ji} = \left(\sum_{s=i}^{i_j} \widehat{\beta}_s^{(j)} u_{t_{j-1}+s-i+1}\right) \Big/ \lambda_j^{i-1} \text{ where } \widehat{b}_j = (\widehat{\beta}_1^{(j)}, \ldots, \widehat{\beta}_{i_j}^{(j)})^T, \tag{3.11'}$$

we obtain

$$T^l(y_0 - x^*) = y_l - x^* = \sum_{j=1}^{p} \sum_{i=1}^{i_j} \binom{l}{i-1} \lambda_j^l \widehat{v}_{ji}. \tag{3.12'}$$

Without any loss of generality let us suppose

Assumption 3. *Let* $\beta_{i_j}^{(j)} \neq 0$ *and* $\widehat{\beta}_{i_j}^{(j)} \neq 0 \quad \forall j = 1, 2 \ldots p.$

We will further simplify the notation (3.12) and (3.12′).
Let $\mathscr{N}$ denote the set of the following pairs of positive integers:

$$(1, i_1), (1, i_1 - 1), \ldots, (1, 1), (2, i_2), \ldots, (2, 1), \ldots, (p, i_p), \ldots, (p, 1). \tag{3.13}$$

Let c_l be a vector from $\mathbb{R}^{t_p}$ whose q-th component equals $\binom{l}{i-1}\lambda_j^l$, where the pair (j, i) occupies the q-th position in the sequence (3.13).The number t_p was defined by the relation (3.3), i.e., $t_p = \sum\limits_{j=1}^{p} i_j$. Let us introduce matrices

$$V = (v_{1i_1}, \ldots, v_{11}, v_{2i_2}, \ldots, v_{21}, \ldots, v_{pi_p}, \ldots, v_{p1}), \tag{3.14}$$

$$\widehat{V} = (\widehat{v}_{1i_1}, \ldots, \widehat{v}_{11}, \widehat{v}_{2i_2}, \ldots, \widehat{v}_{21}, \ldots, \widehat{v}_{pi_p}, \ldots, \widehat{v}_{p1}), \tag{3.14'}$$

where the indices of the column vectors of these matrices are arranged in accordance with (3.13). On the basis of the above introduced notation it follows from the relations (3.12) and (3.12') that

$$y_{l+1} - y_l = Vc_l; \quad y_l - x^* = \widehat{V}c_l. \tag{3.15}$$

According to (2.14) and (3.15)

$$T^t r_0 = y_{m+t+1} - y_{m+t} = Vc_{m+t} \quad \text{and} \tag{3.16}$$

$$T^{t-1}(T - I)r_0 = T^t r_0 - T^{t-1} r_0 = \Delta T^{t-1} r_0 = V \Delta c_{m+t-1}. \tag{3.17}$$

Let the inequality $m \geqslant \max\limits_{j=1,\ldots,p} i_j - 1$ hold for the number m introduced in Algorithm 2.2. Let $q \leqslant t_p$ and let the pair (j, i) occupy the q-th position in the sequence

(3.13). Then

$$
\begin{aligned}
(e_q^{(t_p)})^T(c_{m+t}-c_{m+t-1}) &= \binom{m+t}{i-1}\lambda_j^{m+t} - \binom{m+t-1}{i-1}\lambda_j^{m+t-1} \\
&= \lambda_j^m m^{i-1}\Big[\Big(\frac{m+t}{m}\Big)\Big(\frac{m+t-1}{m}\Big)\dots\Big(\frac{m+t-i+2}{m}\Big)\lambda_j \\
&-\Big(\frac{m+t-1}{m}\Big)\Big(\frac{m+t-2}{m}\Big)\dots\Big(\frac{m+t-i+1}{m}\Big)\Big]\frac{\lambda_j^{t-1}}{(i-1)!} \\
&= \lambda_j^m m^{i-1}\Big[\Big(1+\frac{t}{m}\Big)\Big(1+\frac{t-1}{m}\Big)\dots\Big(1+\frac{t-i+2}{m}\Big)\lambda_j \\
&-\Big(1+\frac{t-1}{m}\Big)\Big(1+\frac{t-2}{m}\Big)\dots\Big(1+\frac{t-i+1}{m}\Big)\Big]\frac{\lambda_j^{t-1}}{(i-1)!} \\
&= \lambda_j^m m^{i-1}\sum_{l=0}^{\infty}\frac{\varphi_l^{(t)}(q)}{m^l},
\end{aligned}
$$

where the series has only a finite number of nonzero members, however, as we will use it further, it is of advantage of to consider it as an absolutely convergent series. Obviously $\varphi_0^{(t)}(q)\neq 0$. Thus

$$
(e_q^{(t_p)})^T(c_{m+t}-c_{m+t-1}) = \lambda_j^m m^{i-1}\sum_{l=0}^{\infty}\frac{\varphi_l^{(t)}(q)}{m^l}, \tag{3.18}
$$

and by analogy

$$
(e_q^{(t_p)})^T c_{m+s} = \lambda_j^m m^{i-1}\sum_{l=0}^{\infty}\frac{\psi_l^{(s)}(q)}{m^l}, \tag{3.19}
$$

where $\psi_0^{(t)}(q)\neq 0$. According to (2.21)

$$
(B_{k+1})_{t,s} = (T^t r_0 - T^{t-1}r_0)^T T^{s-1}r_0 \quad \text{where } t=1,\dots,k; \quad s=1,\dots,k+1.
$$

On the basis of the formulae (3.18), (3.19) let us define sequences of vectors $\{\Phi_l^{(t)}\}_{l=0}^{\infty}\subset\mathbb{C}^{t_p}$ and $\{\Psi_l^{(s)}\}_{l=0}^{\infty}\subset\mathbb{C}^{t_p}$ such that the q-th component od the vector $\Phi_l^{(t)}$ and $\Psi_l^{(s)}$ equals $\varphi_l^{(t)}(q)$ and $\psi_l^{(s)}(q)$, respectively. Moreover, if we define vectors

$$
g_l = (\underbrace{\lambda_1^l,\dots,\lambda_1^l}_{i_1-\text{times}},\underbrace{\lambda_2^l,\dots,\lambda_2^l}_{i_2-\text{times}},\dots,\underbrace{\lambda_p^l,\dots,\lambda_p^l}_{i_p-\text{times}})^T, \tag{3.20}
$$

$$
h_l = (l^{i_1-1},l^{i_1-2},\dots,1,l^{i_2-1},l^{i_2-2},\dots,1,\dots,l^{i_p-1},l^{i_p-2}\dots,1)^T \tag{3.21}
$$

for any positive integer l then according to (3.15), (3.16), (3.17), (3.18) and (3.19)

$$T^t r_0 - T^{t-1} r_0 = V \operatorname{diag}(g_m) \operatorname{diag}(h_m) \sum_{l=0}^{\infty} \frac{\Phi_l^{(t)}}{m^l}, \tag{3.22}$$

$$T^{s-1} r_0 = V \operatorname{diag}(g_m) \operatorname{diag}(h_m) \sum_{l=0}^{\infty} \frac{\Psi_l^{(s)}}{m^l}. \tag{3.23}$$

Let us summarize the result of the above consideration in the following theorem.

Theorem 3.1. *Let Assumptions* 1–3 *be valid. Let the inequality* $m \geqslant \max\limits_{j=1,\dots,p} i_j - 1$ *hold for the number* m *introduced in Algorithm* 2.2. *Then there exist sequences of vectors* $\{\Phi_l^{(t)}\}_{l=0}^{\infty} \subset \mathbb{C}^{t_p}$ *and* $\{\Psi_l^{(s)}\}_{l=0}^{\infty} \subset \mathbb{C}^{t_p}$ *for* $t = 1, \dots, k$ *and* $s = 1, \dots, k+1$ *such that all components of the vectors* $\Phi_0^{(t)}$ *and* $\Psi_0^{(s)}$ *are nonzero, the series (all components of the series)* $\sum\limits_{l=0}^{\infty} \frac{\Phi_l^{(t)}}{m^l}$ *and* $\sum\limits_{l=0}^{\infty} \frac{\Psi_l^{(s)}}{m^l}$ *are absolutely convergent and the equalities* (3.22), (3.23) *hold. Consequently, the elements of the matrix* B_{k+1} *are in the form*

$$\begin{gathered} (B_{k+1})_{k+1,s} = 1, \\ (B_{k+1})_{t,s} = \left(V_1 \sum_{l=0}^{\infty} \frac{\Phi_l^{(t)}}{m^l} \right)^T V_1 \left(\sum_{l=0}^{\infty} \frac{\Psi_l^{(s)}}{m^l} \right) \\ \text{for} \quad t = 1, \dots, k, \quad s = 1, \dots, k+1, \quad \text{where} \quad V_1 = V \operatorname{diag}(g_m) \operatorname{diag}(h_m). \end{gathered} \tag{3.24}$$

4. The estimate for $\|x_k - x^*\|$

If $m \geqslant 0$ is an integer then according to (3.12) we have

$$y_{m+t} - x^* = \widehat{V} c_{m+t} = \sum_{j=1}^{p} \sum_{i=1}^{i_j} \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji} \tag{4.1}$$

for any nonnegative integer t. The relation (2.15) implies that

$$x_k - x^* = \sum_{t=0}^{k} \alpha_t^{(k)} (y_{m+t} - x^*). \tag{4.2}$$

For the purpose of simplifying the formulas let us make the following assumption. We will describe in the discussion later what would happen if the assumption were not met.

Assumption 4. *If k is a positive integer for which we calculate x_k then let there exist a positive integer $\tau \in [1, p]$ such that*

$$k = \sum_{j=1}^{\tau} i_j. \tag{4.3}$$

Let us introduce the polynomial

$$\widehat{P}_k(z) = \widehat{\sigma}_k z^k + \widehat{\sigma}_{k-1} z^{k-1} + \ldots + \widehat{\sigma}_0 = (z - \lambda_1)^{i_1} (z - \lambda_2)^{i_2} \ldots (z - \lambda_\tau)^{i_\tau}.$$

N o t e. In case Assumption 4 were not fulfilled, we would have to construct the polynomial $\widehat{P}_k(z)$ in a different way.

Let us put

$$P_k(z) = \widehat{P}_k(z)/\widehat{P}_k(1) \equiv \sigma_k z^k + \sigma_{k-1} z^{k-1} + \ldots + \sigma_0. \tag{4.4}$$

If we now substitute in the relation (4.2) σ_j for $\alpha_j^{(k)}\ \forall j$ and for $y_{m+t} - x^*$ from the relation (4.1), we obtain

$$\begin{aligned} \sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*) &= \sum_{t=0}^{k} \sigma_t \sum_{j=1}^{p} \sum_{i=1}^{i_j} \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji} \\ &= \sum_{j=1}^{\tau} \sum_{i=1}^{i_j} \sum_{t=0}^{k} \sigma_t \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji} + \sum_{j=\tau+1}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} \sigma_t \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji}. \end{aligned} \tag{4.5}$$

Let us denote $Q_{m+k}(z) = z^m P_k(z)$. The polynomial $Q_{m+k}(z)$ has the root zero of multiplicity m while λ_j is the root of multiplicity i_j for $j = 1, 2, \ldots, \tau$. Thus the polynomial

$$Q_{m+k}(z) = \sigma_k z^{m+k} + \sigma_{k-1} z^{m+k-1} + \ldots + \sigma_0 z^m$$

satisfies

$$\left. \frac{d^{(i-1)} Q_{m+k}(z)}{dz^{(i-1)}} \right|_{z=\lambda_j} = 0 \quad \text{for } j = 1, 2, \ldots, \tau \quad \text{and} \quad i = 1, 2, \ldots, i_j. \tag{4.7}$$

If we multiply the equation (4.7) by the number $z^{i-1}/(i-1)!$, we obtain

$$\begin{aligned} &\left[\frac{(m+k)(m+k-1)\ldots(m+k-i+1)}{(i-1)!} \sigma_k z^{m+k} \right. \\ &\quad + \frac{(m+k-1)(m+k-2)\ldots(m+k-i)}{(i-1)!} \sigma_{k-1} z^{m+k-1} \\ &\quad \left. + \ldots + \frac{(m)(m-1)\ldots(m-i+1)}{(i-1)!} \sigma_0 z^m \right]_{z=\lambda_j} = 0 \\ &\quad \text{for } j = 1, 2, \ldots, \tau \quad \text{and} \quad i = 1, 2, \ldots, i_j. \end{aligned} \tag{4.8}$$

If we rewrite the relation (4.8) using binomial coefficients and reverse the order of addition, we obtain a system of equations

$$\sum_{t=0}^{k} \sigma_t \binom{m+t}{i-1} \lambda_j^{m+t} = 0 \quad \text{for} \quad j = 1, 2, \ldots, \tau \quad \text{and} \quad i = 1, 2, \ldots, i_j. \tag{4.9}$$

If we substitute (4.9) into (4.5), we obtain

$$\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*) = \sum_{j=\tau+1}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} \sigma_t \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji}. \tag{4.10}$$

On the basis of the above we can formulate a proposition.

Theorem 4.1. *Let Assumptions* 1–4 *be met. Let* $x_k(m)$ *be a vector obtained by applying Algorithm* 2.2 *for a non-negative integer* m. *Then there exist numbers* $L(m)$,

$$1 \leqslant L(m) \leqslant \kappa = \|(I - T)^{-1}\| \|(I - T)\|$$

independent of k *and sequences* $\{u_k(m)\}_{m=0}^{\infty}, \{w_k(m)\}_{m=0}^{\infty}$, $u_k(m) > 0$, $w_k(m) \geqslant 0 \quad \forall m$, *such that*

$$\|x_k(m) - x^*\| \leqslant L(m) m^{i_{\tau+1}-1} |\lambda_{\tau+1}|^m (u_k(m) + w_k(m)) \tag{4.11}$$

and

$$\lim_{m \to \infty} w_k(m) = \Theta. \tag{4.12}$$

If $i_{\tau+1} = 1$ *then*

$$u_k(m) \leqslant \sum_{\substack{j=\tau+1 \\ |\lambda_j| = |\lambda_{\tau+1}|}}^{p} \left| \sum_{t=0}^{k} \sigma_t \lambda_j^t \right| \|\widehat{v}_{j1}\| \quad \forall m. \tag{4.13}$$

If $i_{\tau+1} = 1$ *and* $|\lambda_{\tau+1}| > |\lambda_{\tau+2}|$ *then*

$$\overline{\lim_{m \to \infty}} \frac{1}{m^{i_{\tau+2}-1}} \left| \left(\frac{\lambda_{\tau+1}}{\lambda_{\tau+2}} \right)^m w_k(m) \right| < \infty. \tag{4.12'}$$

If $i_{\tau+1} > 1$ *then*

$$u_k(m) \leqslant \sum_{\substack{j=\tau+1 \\ |\lambda_j| = |\lambda_{\tau+1}|,\ i_j = i_{\tau+1}}}^{p} \left| \sum_{t=0}^{k} \sigma_t \lambda_j^t \right| \|\widehat{v}_{ji_j}\| / (i_j - 1)!. \tag{4.14}$$

In this case $w_k(m) \sim O(\frac{1}{m})$.

R e m a r k. An analogous theorem could be formulated for the behaviour of $\|r_k\|$. The upper bound for the constant $L(m)$ is $\|I - T\|$ and the other estimates remain unchanged in this case. Therefore, it is sufficient to investigate only the behaviour of the norm $\|x_k - x^*\|$ in what follows.

P r o o f. In the sequel we will write x_k instead of $x_k(m)$. The linear combination $\sum_{t=0}^{k} \sigma_t y_{m+t} \in x_0 + \mathscr{K}_k(r_0, T)$ in view of $\sum_{t=0}^{k} \sigma_t = 1$. From the minimization property (2.3) it follows that

$$\|(I-T)(x_k - x^*)\| \leqslant \|(I-T)\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*)\|$$

and hence

$$\|(x_k - x^*)\| \leqslant \|(I-T)^{-1}\|\|(I-T)\|\left\|\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*)\right\|.$$

If $\alpha_t^{(k)} = \sigma_t$, then $\|x_k - x^*\| = \|\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*)\|$ would hold according to (4.2). Therefore there exists a number $L(m) \in \langle 1, \kappa \rangle$ such that the inequality

$$\|(x_k - x^*)\| \leqslant L(m)\left\|\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*)\right\| \tag{4.15}$$

holds. For the sake of brevity let us denote

$$d(t, m, i, j) = \sigma_t \binom{m+t}{i-1} \lambda_j^{m+t} \widehat{v}_{ji}. \tag{4.16}$$

According to (4.10) we have

$$\sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*) = \sum_{j=\tau+1}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} d(t, m, i, j) = z_1(m) + z_2(m)$$

where

$$z_1(m) = \sum_{\substack{j=\tau+1 \\ |\lambda_j| = |\lambda_{\tau+1}|}}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} d(t, m, i, j), \tag{4.17}$$

$$z_2(m) = \sum_{\substack{j=\tau+1 \\ |\lambda_j| < |\lambda_{\tau+1}|}}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} d(t, m, i, j). \tag{4.18}$$

If $i_{\tau+1}=1$, then according to Assumption 2

$$z_1(m)=\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|}}^{p}\sum_{t=0}^{k}d(t,m,1,j).$$

If we put $u_k(m)=\|z_1(m)/\lambda_{\tau+1}^m\|$ and $w_k(m)=\|z_2(m)/\lambda_{\tau+1}^m\|$, then

$$\left\|\sum_{t=0}^{k}\sigma_t(y_{m+t}-x^*)\right\|\leqslant|\lambda_{\tau+1}|^m(u_k(m)+w_k(m)), \tag{4.19}$$

where the numbers $u_k(m), w_k(m)$ conform to the relations (4.13) and (4.12). The equality (4.12′) is evident from the transcription of $\sum\limits_{t=0}^{k}\sigma_t(y_{m+t}-x^*)$.

If $i_{\tau+1}>1$ then according to (4.17)

$$\begin{aligned}z_1(m)&=\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|,\ i_j=i_{\tau+1}}}^{p}\sum_{i=1}^{i_j}\sum_{t=0}^{k}d(t,m,i,j)+\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|,\ i_j<i_{\tau+1}}}^{p}\sum_{i=1}^{i_j}\sum_{t=0}^{k}d(t,m,i,j)\\ &=\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|,\ i_j=i_{\tau+1}}}^{p}\left\{\sum_{t=0}^{k}d(t,m,i_j,j)+\sum_{i=1}^{i_j-1}\sum_{t=0}^{k}d(t,m,i,j)\right\}\\ &\quad+\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|,\ i_j<i_{\tau+1}}}^{p}\sum_{i=1}^{i_j}\sum_{t=0}^{k}d(t,m,i,j).\end{aligned}$$

Taking into account that

$$\begin{aligned}\binom{m+t}{i-1}&=\frac{m^{i-1}}{(i-1)!}\Big(1+\frac{t}{m}\Big)\Big(1+\frac{t-1}{m}\Big)\dots\Big(1+\frac{t-i+2}{m}\Big)\\ &=\frac{m^{i-1}}{(i-1)!}\Big(1+\frac{1}{m}R_{i-2}\Big(\frac{1}{m}\Big)\Big),\end{aligned}$$

where R_{i-2} is polynomial of the degree $i-2$, and putting

$$\widehat{z}_1(m)=\sum_{\substack{j=\tau+1\\ |\lambda_j|=|\lambda_{\tau+1}|,\ i_j=i_{\tau+1}}}^{p}\sum_{t=0}^{k}\sigma_t m^{i_{\tau+1}-1}\lambda_j^{m+t}v_{ji_j}/(i_{\tau+1}-1)! \tag{4.20}$$

and

$$\widehat{z}_2(m)=z_1(m)-\widehat{z}_1(m)+z_2(m),$$

then we conclude

(4.21)

$$\widehat{z}_2(m) = \sum_{\substack{j=\tau+1 \\ |\lambda_j|=|\lambda_{\tau+1}|,\, i_j=i_{\tau+1}}}^{p} \sum_{t=0}^{k} \sigma_t m^{i_{\tau+1}-2} \lambda_j^{m+t} R_{i_{\tau+1}-2}(1/m) v_{ji_j}/(i_{\tau+1}-1)!$$
$$+ \sum_{\substack{j=\tau+1 \\ |\lambda_j|=|\lambda_{\tau+1}|,\, i_j=i_{\tau+1}}}^{p} \sum_{i=1}^{i_j-1} \sum_{t=0}^{k} d(t,m,i,j) + \sum_{\substack{j=\tau+1 \\ |\lambda_j|=|\lambda_{\tau+1}|,\, i_j<i_{\tau+1}}}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} d(t,m,i,j)$$
$$+ \sum_{\substack{j=\tau+1 \\ |\lambda_j|<|\lambda_{\tau+1}|}}^{p} \sum_{i=1}^{i_j} \sum_{t=0}^{k} d(t,m,i,j),$$

and also putting

$$u_k(m) = \|\widehat{z}_1(m)/(m^{i_{\tau+1}-1}\lambda_{\tau+1}^m)\|,$$
$$w_k(m) = \|\widehat{z}_2(m)/(m^{i_{\tau+1}-1}|\lambda_{\tau+1}|^m)\|$$

we obtain

$$(4.19') \qquad \left\| \sum_{t=0}^{k} \sigma_t (y_{m+t} - x^*) \right\| \leqslant m^{i_{\tau+1}-1} |\lambda_{\tau+1}|^m (u_k(m) + w_k(m)),$$

where $u_k(m)$ fulfils (4.14) and for $w_k(m)$ the relation (4.12) holds. Moreover, $w_k(m) = O(\frac{1}{m})$ for $m \to \infty$ in this case, i.e., $i_{\tau+1} > 1$. □

The formula (4.11) inspired us to use Algorithm 2.2 because successive approximations (2.9) (we have called them pre-iterations) cost less work then GMRES (see Algorithm 2.1).

For demonstration. let us consider the system obtained by the finite difference approximation of the convection-diffusion equation on the unit square (see (6.1). We took $n = 900$ and tested GMRES without any restart for such a small system. From Algorithm 2.1 it is easy to obtain that for the calculation of x_k with $m = 0$ we need

$$4nk + k(k+1)n + 2nk + 6n = nk^2 + 7nk + 6n = nk(k+7) + 6n$$

multiplications if we consider the system (6.3). An analogous formula holds for (6.4). The m successive approximations (2.9) need $4nm$ multiplications.

For $m = 600$ we found $\|(I - T)(x_9 - x^*)\| < 10^{-8}$ where x_9 is the first iteration for which the norm of the residual is less then 10^{-8}. In this case we needed $2\,550n$ multiplications and stored 9 vectors.

For $m = 0$ we found $\|(I-T)(x_{107}-x^*)\| < 10^{-8}$, we needed $12\,204n$ multiplications and stored 109 vectors.

Now it remains to specify what would happen if Assumption 4 were not fulfilled. Let

$$\sum_{j=1}^{\tau-1} i_j < k < \sum_{j=1}^{\tau} i_j.$$

Then only the construction of the polynomial P_k would be different. All the rest would remain the same. With regard to the above extensive description it is easy to imagine that the estimate obtained would be

$$\|x_k - x^*\| \leqslant L(m)m^{\delta}|\lambda_\tau|^m(u_k(m) + w_k(m)), \tag{4.22}$$

where again $\overline{\lim}_{m\to\infty} |u_k(m)| < \infty$ and $|w_k(m)|$ meets the relation (4.12). The integer $\delta \leqslant i_\tau - 1$ in general case. From the point of view of convergence the best k's are those for which Assumption 4 is fulfilled. □

Our exposition reveals that Algorithm 2.2 is theoretically derived for all nonnegative integers m while in practice we consider only one m. Thus from the theoretical viewpoint the components of the vector $\vec{\alpha}_{k+1}$ are functions of m although we have written only $\alpha_i^{(k)}$ so far. Further on we will respect this dependence, i.e., instead of $\alpha_i^{(k)}$ and $\vec{\alpha}_{k+1}$ we will write $\alpha_i^{(k)}(m)$ and $\vec{\alpha}_{k+1}(m)$, respectively.

Now the question is what relation occurs between the components of the vector $\vec{\alpha}_{k+1}(m)$ and the coefficients of the polynomial P_k, because if the sequence $\{\vec{\alpha}_{k+1}(m)\}$ converged to the coefficients of the polynomial P_k for $m \to \infty$ then the number $L(m)$ in the estimate (4.11) which is for all m bounded by the condition number of the matrix $I - T$ would be close to 1. That is why it will be useful to study the behaviour of $\alpha_i^{(k)}(m)$ in dependence on m.

5. The behaviour of the functions $\alpha_i^{(k)}(m)$ in dependence on m

Let us put

$$\sigma(k+1) = (\sigma_0, \sigma_1, \ldots, \sigma_k)^T, \tag{5.1}$$

$$\vartheta(k+1) = B_{k+1}\sigma(k+1) - e_{k+1}^{(k+1)}. \tag{5.2}$$

From the latter relation it follows that

$$\vec{\alpha}_{k+1}(m) = B_{k+1}^{-1}e_{k+1}^{(k+1)} = \sigma(k+1) - B_{k+1}^{-1}\vartheta(k+1). \tag{5.3}$$

As the last component of the vector $\vartheta(k+1)$ is zero, we omit the calculation of the last column of the matrix B_{k+1}^{-1}. Now we use the well-known formula for the calculation of the inverse matrix B_{k+1}^{-1}, namely

$$(B_{k+1}^{-1})_{ij} = \frac{B_{k+1}^{A}(i,j)}{\det B_{k+1}},$$

where B_{k+1}^{A} denotes the adjoint of B_{k+1}. To make things clear let us take the matrix B_4 (for $k=3$) and let us show what $B_4^A(3,2)$ looks like:

(5.4)
$$B_4^A(3,2) = -\det \begin{pmatrix} ((T-I)r_0)^T r_0, & ((T-I)r_0)^T T r_0, & ((T-I)r_0)^T T^3 r_0 \\ (T^2(T-I)r_0)^T r_0, & (T^2(T-I)r_0)^T T r_0, & (T^2(T-I)r_0)^T T^3 r_0) \\ 1, & 1, & 1 \end{pmatrix}.$$

We subtract the second column from the third and the first from the second obtaining

(5.5)
$$B_4^A(3,2) = -\det \begin{pmatrix} ((T-I)r_0)^T (T-I)r_0, & ((T-I)r_0)^T T(T^2-I)r_0 \\ (T^2(T-I)r_0)^T (T-I)r_0, & (T^2(T-I)r_0)^T T(T^2-I)r_0 \end{pmatrix}.$$

If we modify $\det B_4$ in a similar way, we obtain

(5.5′)
$$\det B_4 = \det \begin{pmatrix} ((T-I)r_0)^T (T-I)r_0, & ((T-I)r_0)^T T(T-I)r_0, & ((T-I)r_0)^T T^2(T-I)r_0 \\ (T(T-I)r_0)^T (T-I)r_0, & (T(T-I)r_0)^T T(T-I)r_0, & (T(T-I)r_0)^T T^2(T-I)r_0 \\ (T^2(T-I)r_0)^T (T-I)r_0, & (T^2(T-I)r_0)^T T(T-I)r_0, & (T^2(T-I)r_0)^T T^2(T-I)r_0 \end{pmatrix}.$$

According to Assumption 1 the matrix in (5.5′) is strongly nonsingular. Each element of this matrix is a scalar product of vectors in the form (3.22). The same holds for each element of the matrix (5.4), only the series are different.

Our concern here is to express the determinant of the matrix B_{k+1} and the determinant of the corresponding adjoints with one formula. Let us now go back to the general formulation. The matrices whose determinant we will calculate have in general the dimension k or $k-1$. Let us consider a matrix from $L(\mathbb{R}^k)$, the calculation for matrices from $L(\mathbb{R}^{k-1})$ would be analogous.

Let us now make the following general consideration. Let $\{\Lambda_{q,l}^{(s)}\}_{l=0}^{\infty}$, be sequences of vectors from $L(\mathbb{C}^{t_p})$ for $q = 1, \ldots, k$, $s = 1, 2$ such that the series

$$\sum_{l=0}^{\infty} \frac{\Lambda_{q,l}^{(s)}}{m^l} \tag{5.6}$$

absolutely converge for all q, s under consideration. We suppose that $k \leqslant t_p = n$. (See (3.3).) Let $K_q^{(s)}(m)$ denote the sum of the series and let us define also the vectors $L_q^{(s)}(m)$,

$$L_q^{(s)}(m) = \operatorname{diag}(g_m) \operatorname{diag}(h_m) K_q^{(s)}(m), \tag{5.7}$$

where g_m and h_m were respectively defined by (3.20) and (3.21). Let us consider the matrix $C_k \in L(\mathbb{C}^k)$ constructed as follows:

(5.7′)
$$C_k = \begin{pmatrix} (VL_1^{(1)}(m))^H(VL_1^{(2)}(m)),\ (VL_1^{(1)}(m))^H(VL_2^{(2)}(m)),\ \ldots,(VL_1^{(1)}(m))^H(VL_k^{(2)}(m)) \\ (VL_2^{(1)}(m))^H(VL_1^{(2)}(m)),\ (VL_2^{(1)}(m))^H(VL_2^{(2)}(m)),\ \ldots,(VL_2^{(1)}(m))^H(VL_k^{(2)}(m)) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (VL_k^{(1)}(m))^H(VL_1^{(2)}(m)),\ (VL_k^{(1)}(m))^H(VL_2^{(2)}(m)),\ \ldots,\ (VL_k^{(1)}(m))^H(VL_k^{(2)}(m)) \end{pmatrix},$$

where V was given by (3.14).

Let us suppose that the matrix C_k is regular. Then the matrices

$$\begin{aligned} L^{(1)}(m) &= (L_1^{(1)}(m), L_2^{(1)}(m), \ldots, L_k^{(1)}(m)), \\ L^{(2)}(m) &= (L_1^{(2)}(m), L_2^{(2)}(m), \ldots, L_k^{(2)}(m)) \end{aligned} \tag{5.8}$$

are of the rank k and therefore the matrices

$$\begin{aligned} K^{(1)}(m) &= (K_1^{(1)}(m), K_2^{(1)}(m), \ldots, K_k^{(1)}(m)), \\ K^{(2)}(m) &= (K_1^{(2)}(m), K_2^{(2)}(m), \ldots, K_k^{(2)}(m)) \end{aligned} \tag{5.9}$$

are of the rank k as well. According to (5.7)

$$L^{(s)}(m) = \operatorname{diag}(g_m)\operatorname{diag}(h_m)K^{(s)}(m) \quad \text{for } s = 1, 2. \tag{5.10}$$

In the following considerations we will write the index s in the sense that everything holds for $s = 1, 2$.

Assumption 5. *Let matrices $H^{(s)}(m)$ constructed from the first k rows of the matrix $K^{(s)}(m)$ be strongly nonsingular for $s = 1, 2$.*

Therefore there exists an upper triangular matrix $R_k^{(s)}(m)$ with a unit diagonal such that the matrix $H^{(s)}(m)R_k^{(s)}(m)$ is lower triangular. Furthermore there exists a lower triangular matrix $W_k^{(s)}(m)$ with a unit diagonal such that the matrix $H^{(s)}(m)R_k^{(s)}(m)W_k^{(s)}(m)$ is diagonal. Therefore the matrix $K^{(s)}(m)R_k^{(s)}(m)W_k^{(s)}(m)$ has the following form: It is a rectangular matrix $t_p \times k$ whose first k rows form a diagonal matrix. We expect that the elements of the matrix $K^{(s)}(m)R_k^{(s)}(m)W_k^{(s)}(m)$ have the form $\sum_{t=\nu}^{\infty} \delta_t^{(s)}/m^t$, where ν is an integer and the series is absolutely convergent. After all this is clear from the following auxiliary proposition and assumption if we bear in mind that by triangular decomposition we perform Gauss elimination.

Lemma 5.1. *Let the series $\sum_{t=0}^{\infty} \frac{\gamma_t}{m^t}$ be absolutely convergent and $\gamma_0 \neq 0$. Then there exists a positive integer m_0 and a sequence $\{\widehat{\gamma}_t\}_{t=0}^{\infty}$ such that the series $\sum_{t=0}^{\infty} \frac{\widehat{\gamma}_t}{m^t}$*

is absolutely convergent and

$$\left(\sum_{t=0}^{\infty}\frac{\gamma_t}{m^t}\right)^{-1}=\sum_{t=0}^{\infty}\frac{\widehat{\gamma}_t}{m^t}\quad\forall m\geqslant m_0 \tag{5.11}$$

holds.

This lemma is proved in [Zi 84].

Note. If we consider the series $\sum\limits_{t=\nu}^{\infty}\frac{\gamma_t}{m^t}$ where $\nu\neq 0$, then instead of (5.11) we obtain

$$\left(\sum_{t=\nu}^{\infty}\frac{\gamma_t}{m^t}\right)^{-1}=\sum_{t=-\nu}^{\infty}\frac{\widehat{\gamma}_t}{m^t}. \tag{5.11$'$}$$

Assumption 6. *Let* $|\lambda_\tau|>|\lambda_{\tau+1}|\quad$ *and* $\quad m\geqslant m_U$, *where all series are invertible in the sense of Lemma* 5.1 *by the decompositions* $K^{(s)}(m)R_k^{(s)}(m)W_k^{(s)}(m)$ *for* $m\geqslant m_U$.

Assumption 6 introduces the positive integer m_U. We have seen in the previous sections that polynomials or quotients of polynomials with an argument $\frac{1}{m}$ have been represented by absolute convergent series. Therefore we can expect that the second part of Assumption 6 will be fullfiled for all m considered.

Let us put $M_k^{(s)}(m)=R_k^{(s)}(m)W_k^{(s)}(m)$ and calculate $(M_k^{(1)}(m))^HC_kM_k^{(2)}(m)$. Taking into account what has been said about the form of the matrix $K^{(s)}(m)M_k^{(s)}(m)$ and using the formula (5.10) we can immediately write the following proposition.

Theorem 5.1. *Let Assumptions* 1–6 *be fulfilled. Let* $i\leqslant k$ *and* $j\leqslant k$, *let* (p_1,q_1) *and* (p_2,q_2) *be the pairs occupying the* i*-th and* j*-th position respectively in the sequence* $\mathscr{N}$ *(see* 3.13*). Then*

$$\begin{aligned}((M_k^{(1)}(m))^HC_kM_k^{(2)}(m))_{ij}&\\ =\left(m^{q_1-1}\lambda_{p_1}^m\sum_{t=\nu_{q_1}}^{\infty}\frac{\delta_t^{(1)}(p_1,q_1)}{m^t}v_{p_1q_1}+d_m^{(1)}(p_1,q_1)\right)^H&\\ \times\left(m^{q_2-1}\lambda_{p_2}^m\sum_{t=\nu_{q_2}}^{\infty}\frac{\delta_t^{(2)}(p_2,q_2)}{m^t}v_{p_2q_2}+d_m^{(2)}(p_2,q_2)\right),&\end{aligned} \tag{5.12}$$

where ν_{q_s} *are integers, both series absolutely converge,* $\delta_{\nu_{q_s}}^{(s)}(p_s,q_s)\neq 0$ *and there exists an integer* l_{ij} *such that the relations*

$$\overline{\lim_{m\to\infty}}\,\|d_m^{(s)}(p_s,q_s)/(m^{l_{ij}}\lambda_{\tau+1}^m)\|<\infty\quad\text{and}\quad\lim_{m\to\infty}\|d_m^{(s)}(p_s,q_s)/(m^l\lambda_\tau^m)\|=0 \tag{5.13}$$

hold for $s = 1, 2$ *and every integer* l.

If we carry out the multiplication in the formula (5.12), we obtain

$$((M_k^{(1)}(m))^H C_k M_k^{(2)}(m))_{ij} = m^{r_{ij}} (\overline{\lambda}_{p_1})^m \lambda_{p_2}^m \sum_{t=0}^{\infty} \frac{\gamma_t(i,j)}{m^t}(1 + f_m(i,j)), \tag{5.14}$$

where r_{ij} is an integer, $\lim_{m\to\infty} f_m(i,j) = 0$, $\gamma_0(i,j) = \delta^{(1)}_{\nu_{q_1}}(p_1,q_1)\delta^{(2)}_{\nu_{q_2}}(p_1,q_2) \neq 0$ and the implication $\nu_{q_1} = \nu_{q_2} = 0 \Rightarrow r_{ij} = q_1 + q_2 - 2$ holds. Moreover, there exists an integer $l_{ij}^{(1)}$ such that

$$\overline{\lim_{m\to\infty}} \left| m^{l_{ij}^{(1)}} \left(\frac{\lambda_\tau}{\lambda_{\tau+1}} \right)^m f_m(i,j) \right| < \infty. \tag{5.15}$$

Let us remark that if all sums start from zero then the number $l_{ij}^{(1)} = \min(q_1, q_2) - i_{\tau+1}$.

Based on the above, we can express the determinant of the matrix C_k.

Theorem 5.2. *Let Assumptions* 1–6 *be fulfilled. Then there exists an integer* $l^{(1)}$ *such that*

$$\det C_k = m^r (|\lambda_1|^{i_1} |\lambda_2|^{i_2} \dots |\lambda_{\tau-1}|^{i_{\tau-1}} |\lambda_\tau|^{i_\tau})^{2m} \left(\sum_{t=\nu}^{\infty} \frac{\xi_t}{m^t} \right)(1 + \zeta_m), \tag{5.16}$$

where

$$\lim_{m\to\infty} \zeta_m = 0, \quad \overline{\lim_{m\to\infty}} \left| m^{l^{(1)}} \left(\frac{\lambda_\tau}{\lambda_{\tau+1}} \right)^m \zeta_m \right| < \infty, \tag{5.17}$$

the series is absolutely convergent and $\xi_\nu \neq 0$. *If* $\nu = 0$ *then*

$$l^{(1)} = 1 - i_{\tau+1} \quad \textit{and} \quad r = \sum_{j=0}^{\tau} i_j^2 - k. \tag{5.17'}$$

P r o o f. We use the previous theorem which describes the element in the (i,j)th position in the matrix $(M_k^{(1)}(m))^H C_k M_k^{(2)}(m)$. The determinant of the matrix is the sum of $(k!)$ products with each term resulting in such a way that from each row and each column we take exactly one element and multiply it by the sign of the corresponding permutation. We can see from the forms of the elements that all summands have a common factor $m^r(|\lambda_1|^{i_1}|\lambda_2|^{i_2}\dots|\lambda_{\tau-1}|^{i_{\tau-1}}|\lambda_\tau|^{i_\tau})^{2m}$ which may be factored out and the remaining absolutely convergent series may be multiplied and summed up. After a simple modification we obtain what is given in (5.16) in

view of $\det((M_k^{(1)}(m))^H C_k M_k^{(2)}(m)) = \det(C_k)$. If (5.17') holds, then the exponent r equals the number

$$2\sum_{j=1}^{\tau}\frac{1}{2}i_j(i_j-1) = \sum_{j=1}^{\tau} i_j^2 - k. \tag{5.19}$$

The above expression for the exponent equals the sum of exponents of the dominant diagonal entries of the matrix $(M_k^{(1)}(m))^H C_k M_k^{(2)}(m)$. In the other products in the sum for $\det(M_k^{(1)}(m))^H C_k M_k^{(2)}(m)$ the exponents are added up in a different order (according to the appropriate permutation) but their sum equals (5.19) in all summands. □

For the case of the adjoints of the matrix B_{k+1} we can use exactly the same argument with the difference that instead of a matrix from $L(\mathbb{R}^k)$ we would consider matrices from $L(\mathbb{R}^{k-1})$. Let us take a matrix of the form (5.7′) from $L(\mathbb{R}^{k-1})$ and denote it by C_{k-1}. The assumption $k = \sum_{j=1}^{\tau} i_j$ is of course valid. Then the formula (5.16) remains the same with the difference that instead of the factor $(|\lambda_1|^{i_1}|\lambda_2|^{i_2}\dots|\lambda_{\tau-1}|^{i_{\tau-1}}|\lambda_\tau|^{i_\tau})^{2m}$ we have $(|\lambda_1|^{i_1}|\lambda_2|^{i_2}\dots|\lambda_{\tau-1}|^{i_{\tau-1}}|\lambda_\tau|^{i_\tau-1})^{2m}$. Moreover, we have assumed the following

Assumption 7. *Let $|\lambda_{\tau-1}| > |\lambda_\tau|$ if $\tau > 1$.*

The relation (5.17) changes in the following way: there exists an integer $\tilde{l}^{(1)}$ such that

$$\overline{\lim_{m\to\infty}}\left|m^{\tilde{l}^{(1)}}\left(\frac{\lambda_{\tau-1}}{\lambda_\tau}\right)^m\zeta_m\right| < \infty.$$

And now everything is ready for demonstrating what the elements of the inverse matrix B_{k+1}^{-1} look like.

Theorem 5.3. *Let Assumptions 1–7 be fulfilled. Then the element in the position (i,j) in the matrix B_{k+1}^{-1} for $i = 1,\dots,k, j = 1,\dots,k+1$ has the form*

$$\frac{m^{s_{ij}}}{|\lambda_\tau|^{2m}}\left(\sum_{t=0}^{\infty}\frac{\omega_t(i,j)}{m^t}\right)(1+\varepsilon_m(i,j)) \tag{5.20}$$

where s_{ij} is an integer, the series is absolutely convergent, $\omega_0(i,j) \neq 0$ and there exists an integer $l_{ij}^{(2)}$ such that

$$\lim_{m\to\infty}\varepsilon_m(i,j) = 0 \quad \textit{and} \quad \overline{\lim_{m\to\infty}}\left|m^{l_{ij}^{(2)}}q^m\varepsilon_m(i,j)\right| < \infty, \tag{5.21}$$

where $q = \min\left(\left|\frac{\lambda_\tau}{\lambda_{\tau+1}}\right|, \left|\frac{\lambda_{\tau-1}}{\lambda_\tau}\right|\right)$.

The proof follows immediately from (5.16) if instead of the matrix C_k we consider the matrix B_{k+1} modified for calculating the determinant as shown in the special case in (5.5'), and instead of the matrix C_{k-1} we consider the corresponding adjoint again modified as in the special case the determinant (5.5). The scalar product by which the elements of the matrices C_k and C_{k-1} respectively are expressed is substituted from (3.24).

Now we return to the beginning of the section. We will calculate the $1, \dots, k$-th component of the vector $\vartheta(k+1)$. The last component equals identically zero. According to (5.2) and (2.21) it suffices to calculate $\sum_{t=0}^{k} \sigma_t T^t r_0$. Using the relation (4.9), we have

$$\sum_{t=0}^{k} \sigma_t T^t r_0 = \sum_{t=0}^{k} V(c_{m+t}\sigma_t) = m^{i_{\tau+1}-1}\lambda_{\tau+1}^m y(m) \tag{5.22}$$

$$\text{where} \quad \overline{\lim_{m\to\infty}} \|y(m)\| < \infty.$$

N o t e. By analogy the vector $y(m)$ could be re-written using an absolutely convergent series. Since this is clear, we omit it. Here we have assumed the following

Assumption 8. *Let* $|\lambda_1| > |\lambda_2|$ *and* $|\lambda_{\tau+1}| > |\lambda_{\tau+2}|$.

Let us substitute the expression (3.22) instead of $T^{t-1}(T-I)r_0$ in B_{k+1} We immediately obtain the following theorem.

Theorem 5.4. *Let Assumptions* 1–4 *and Assumption* 8 *be fulfilled. Then for the components of the vector* $\vartheta(k+1)$ *the equalities*

$$(\vartheta(k+1))_{k+1} = 0,$$

$$(\vartheta(k+1))_i = m^{i_1+i_{\tau+1}-2}(\lambda_1\lambda_{\tau+1})^m \left(\sum_{t=0}^{\infty} \frac{\eta_t^{(i)}}{m^t}\right)(1+\varepsilon_m^{(i)}), \quad \textit{for } i = 1, \dots, k \tag{5.23}$$

hold, where the series are absolutely convergent, $\eta_0^{(i)} \neq 0\ \forall i$ *and there exist integers* l_i *such that*

$$\lim_{m\to\infty} \varepsilon_m^{(i)} = 0 \quad \textit{and} \quad \overline{\lim_{m\to\infty}} \left|m^{l_i} q_1^m \varepsilon_m^{(i)}\right| < \infty \quad \forall i, \tag{5.24}$$

where $q_1 = \min\left(\left|\frac{\lambda_1}{\lambda_2}\right|, \left|\frac{\lambda_{\tau+1}}{\lambda_{\tau+2}}\right|\right)$.

From the above mentioned theorems we finally have the main theorem of this section.

Theorem 5.5. *Let Assumptions* 1–8 *be fulfilled. Then for* $i = 0, \dots, k$

$$\alpha_i^{(k)}(m) - \sigma_i = m^{\chi_i}(\lambda_1\lambda_{\tau+1}/|\lambda_\tau|^2)^m \left(\sum_{t=0}^{\infty} \frac{\omega_t^{(i)}}{m^t}\right)(1 + z_m^{(i)}), \tag{5.25}$$

where the series absolutely converge, χ_i *are integers,* $\omega_0^{(i)} \neq 0\,\forall i$ *and there exist integers* l_i *such that*

$$\lim_{m\to\infty} z_m^{(i)} = 0 \quad \textit{and} \quad \overline{\lim_{m\to\infty}} \left|m^{l_i} q_2^m z_m^{(i)}\right| < \infty \quad \forall i, \tag{5.26}$$

where $q_2 = \min(q, q_1)$. *If* $|\lambda_1\lambda_{\tau+1}/\lambda_\tau^2| < 1$ *then*

$$\lim_{m\to\infty} \vec{\alpha}_{k+1}(m) = \sigma(k+1). \tag{5.27}$$

P r o o f. The formula (5.25) immediately follows from (5.3) and Theorems 5.3 and 5.4. The rest is obvious. □

The formula (5.25) reveals that the coefficients $\alpha_i^{(k)}(m)$ do not alvays converge to the corresponding numbers σ_i.

Now we focus our attention back on the estimate (4.11). Let us remark that $k = \sum\limits_{j=1}^{\tau} i_j$.

Let $|\lambda_{\tau+1}| > |\lambda_{\tau+2}|$. Then

$$u_k(m) \leqslant C_{\tau+1}\left|\sum_{t=0}^{k} \sigma_t \lambda_{\tau+1}^t\right|, \tag{5.28}$$

where $C_{\tau+1}$ are constants independent of k and m.

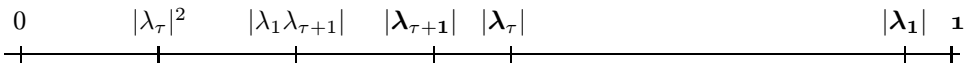
If $i_{\tau+1} = 1$ then

$$\overline{\lim_{m\to\infty}} \frac{1}{m^{i_{\tau+2}-1}} \left|\left(\frac{\lambda_{\tau+1}}{\lambda_{\tau+2}}\right)^m w_k(m)\right| < \infty,$$

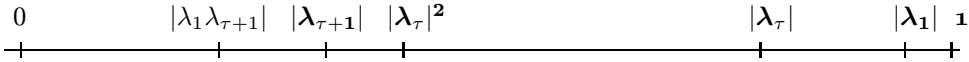
but if $i_{\tau+1} > 1$ then $w_k(m) \sim O(\frac{1}{m})$.

Moreover, let us suppose for the discussion that $\varrho(T) < 1$. We have two typical lay-outs of the numbers $|\lambda_\tau|, |\lambda_{\tau+1}|$ and $|\lambda_{\tau+2}|$.

Case 1



Case 2



In the case 1 we drew the situation when $|\lambda_1\lambda_{\tau+1}/\lambda_\tau^2| > 1$. In this case, $L(m)$ could increase to κ but the upper estimate for $u_k(m)$ is small (see (5.28)) in view of the continuous dependence of roots and coefficients of the polynomial P_k. In the case 2 we drew the situation when $|\lambda_1\lambda_{\tau+1}/\lambda_\tau^2| < 1$. In this case we can put $L(m) \doteq 1$ for $m \gg 1$ but $u_k(m)$ could be a larger number in view of the distance between λ_τ and $\lambda_{\tau+1}$. The dominant term in the estimate (4.11) is $|\lambda_{\tau+1}|^m$ and therefore if $|\lambda_{\tau+1}| < 1$ (or more generally if T is convergent), then for a large m we can obtain a very good estimate in (4.11).

R e m a r k. We have used the coefficients $\alpha_i^{(k)}(m)$ only for theoretical investigations. In practice we proceed according to Algorithm 2.1 and *do not calculate any* $\alpha_i^{(k)}(m)$.

6. Numerical experiments

Let us consider a discrete approximation of the partial differential equation

$$-\Delta\mathbf{x}(s,t) + 2s^2\mathbf{x}_s + 2s^2\mathbf{x}_t = \mathbf{f}(s,t) \tag{6.1}$$

on the square $\Omega = (0,1) \times (0,1)$ with homogeneous Dirichlet boundary conditions, where $\mathbf{x} = \mathbf{x}(s,t)$. We discretize (6.1) on a uniform $N \times N$ grid using Green's Theorem (see [V] Chapter 6). Let us suppose that the grid points are ordered using the rowwise natural ordering. Then the coefficient matrix A has the form

$$A = \mathrm{tri}[A_{j,j-1}, A_{j,j}, A_{j,j+1}],$$

where $A_{j,j-1}$, $A_{j,j+1}$ are diagonal matrices and $A_{j,j}$ tridiagonal matrices in $L(\mathbb{R}^N)$ for $j = 1, 2, \ldots, N$. We write the system of linear algebraic equations obtained as

$$Ax = f, \tag{6.2}$$

where $A \in L(\mathbb{R}^n)$ and $n = N \times N$.

Let us decompose $A = D - C_L - C_U$, where D is a diagonal matrix, C_L a strictly lower and C_U a strictly upper triangular matrix. Putting $T = D^{-1}(C_L + C_U)$ and $b = D^{-1}f$, we rewrite the system (6.2) in the form

$$(I - T)x = b. \tag{6.3}$$

We obtain the system to which Algorithm 2.2 has been applied. When the iteration index k increases, the number of vectors requiring storage in GMRES increases like k. The problems with memory occur for a large n. To avoid these difficulties we can use the GMRES iteratively (the so called restarted GMRES) or we can perform a large number of successive approximations (2.9) or (for large systems) we can restart GMRES with pre-iterations. A more detailed analysis of restarted GMRES will be the subject of a separate paper. We have put $T = D^{-1}(C_L + C_U)$, i.e., the system (6.2) is preconditioned by the diagonal of the matrix A. Another preconditioning strategy involving incomplete LU-decomposition is studied in the papers [Zi2 96] or [Zi 97]. We tested the computer time needed for the norm of error vector to be less than 10^{-5}. The pictures in Graph 1 show the dependence of time on the number of pre-iterations (m) for various values of the restart. The behaviour of GMRES with restart 100 is close to the behaviour of GMRES without restart. The matrix $T \in L(\mathbb{R}^n)$ for $n = 10000$. We took $y_0 = (1, 1, \ldots, 1)^T$ and $f = 0$, i.e., the error vector is identical with the iteration. It is seen from the figures that we have obtained the best result (the smallest time) for relatively small restart (for the restart = 20 (time = 84 s)). Moreover, in a large neighbourhood of 1500 pre-iterations, the decrease and increase of curves is very slow. Practically, we can choose the number of pre-iterations in the interval [1000, 2500] for the restart = 20. In other examples we observed an analogous phenomenon. Therefore, we propose (for diagonal preconditioning) to take the restart in the interval [20, 60] and $m = 1000$ and, if the convergence is slow, to add further pre-iterations.

We have further applied Algorithm 2.2 to the nonsymmetric linear systems

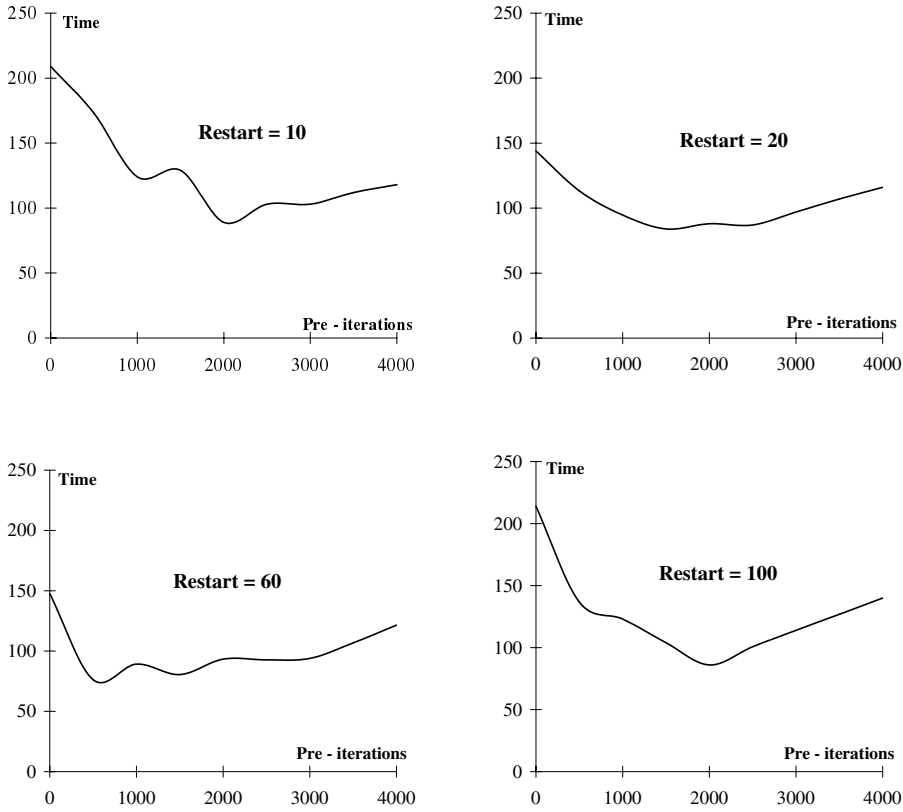
$$(I - \mathscr{L}_\omega)x = c_\omega, \tag{6.4}$$

where

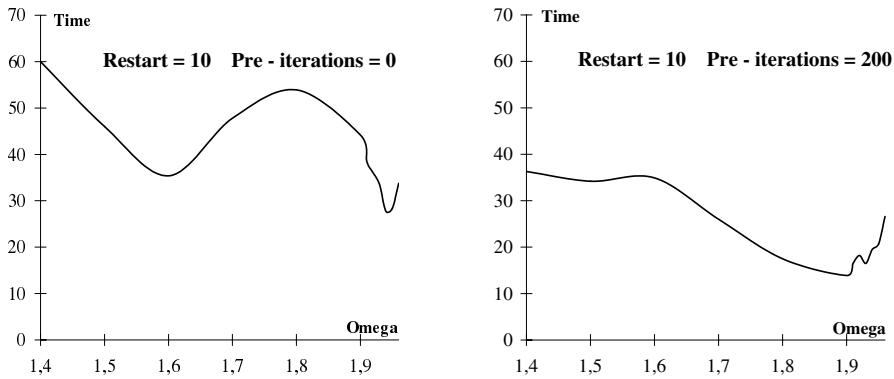
$$\mathscr{L}_\omega = (D - \omega C_L)^{-1}(\omega C_U + (1 - \omega)D),$$
$$c_\omega = (D - \omega C_L)^{-1}\omega b.$$

The systems (6.4), (6.2) and (6.3) have the same solution. For the test we took the same system with 10 000 unknowns and we tested the computer time needed for the norm of the error vector to be less than 10^{-5}. In both the last tests we took $y_0 = (1, 1, \ldots, 1)^T$. The following graphs show the dependence of time on ω for 0 and 200 pre-iterations.

We tested this example for various values of restarts and numerical results showed that for the use of GMRES on the modified system it is convenient to take a small number for the restart. We stored few vectors and as we saw in this case the convergence was, for $\omega = 1.9$, more than 5 times faster than in the case of Jacobi pre-iterations. In the optimum case the time was 14 s. The relaxation factor ω has been found experimentally.



Graph 1



Graph 2

The program for GMRES in FORTRAN 77 was prepared by my student Miroslav Folprecht.

References

[A 87] *O. Axelson*: A generalized conjugate gradient, least square methods. Numer. Math. *51* (1987), 209–227.

[G-L] *C. Brezinski, M.R. Zaglia*: Extrapolation Methods—Theory and Practice. North Holland, 1991.

[E 82] *H.C. Elman*: Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations. Ph. D. thesis, Computer Science Dept., Yale Univ., New Haven, CT, 1982.

[G-L] *G.H. Golub, Ch.F. Van Loan*: Matrix Computation. The John Hopkins University Press, Baltimore, 1984.

[F-F] *D.K.Fadeev, V.N.Fadeeva*:: Computational Methods of Linear Algebra. San Francisco: Freeman 1963.

[F-G-N 91] *R.W. Freund, G.H. Golub, N.M. Nachtigal*: Iterative solution of linear systems. Acta Numerica (1991), 57–100.

[H-Y] *L.A. Hageman, D.M. Young*: Applied Iterative Method. New York, Academic Press, 1981.

[Ho] *A.S. Householder*: The Theory of Matrices in Numerical Analysis. Blaisdell Publishing Company, 1964.

[J-Y] *K.C. Jea, D.M. Young*: Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods. Linear Algebra Appl. *34* (1980), 159–194.

[K-H] Iterative Methods for Large Linear Systems. Papers from a conference held Oct. 19–21, 1988 at the Center for Numerical Analysis of the University of Texas at Austin (Edited by D.R. Kincaid, L.J. Hayes, eds.). Academic Press, 1989.

[L 52] *C. Lanczos*: Solution of systems of linear equations by minimized iterations. J. Res. Nat. Bur. Stand. *49* (1952), 33–53.

[O-R] *J. M. Ortega, W. C. Rheinboldt*: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, 1970.

[Sa 81] *Y. Saad*: Krylov subspace methods for solving large unsymmetric linear systems. Math Comput. *37* (1981), 105–126.

[Sa 84] *Y. Saad*: Practical use of some Krylov subspace methods for solving indefinite and nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. *5* (1984), 203–227.

[S-S 86] *Y. Saad, M.H. Schultz*: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. *7* (1986), 856–869.

[Si-F-Sm] *A. Sidi, W.F. Ford, D.A. Smith*: Acceleration of convergence of vector sequences. SIAM J. Numer. Anal. *23* (1986), 178–196.

[Si 86] *A. Sidi*: Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms. SIAM J. Numer. Anal. *23* (1986), 197–209.

[Si 88] *A. Sidi*: Extrapolation vs. projection methods for linear systems of equations. J. Comput. Appl. Math. *22* (1988), 71–88.

[St 83] *J. Stoer*: Solution of large linear systems of equations by conjugate gradient type methods. Mathematical Programming—The State of the Art (A. Bachem, M. Grötschel and B. Korte, eds.). Springer (Berlin), 1983, pp. 540–565.

[V] *R.L. Varga*: Matrix Iterative Analysis. Prentice-Hall Englewood Clifs, New Jersey, 1962.

[V-V 93] *H.A. Van der Vorst, C. Vuik*: The superlinear convergence behaviour of GMRES. J. Comput. Appl. Math. *48* (1993), 327–341.

[Y] *D.M. Young*: Iterative Solution of Large Linear Systems. Academic Press, New York-London, 1971.

[W 81] *J. Wimp*: Sequence Transformations and their Aapplications. Academic Press, 1981.

[Zi 83] *J. Zítko*: Improving the convergence of iterative methods. Apl. Mat. *28* (1983), 215–229.

[Zi 84] *J. Zítko*: Convergence of extrapolation coefficients. Apl. Mat. *29* (1984), 114–133.

[Zi 94] *J. Zítko*: The behaviour of the error vector using the GMRES method. Technical report No 4/94, Prague. 1994, pp. 1–27.

[Zi 96] *J. Zítko*: Combining the preconditioned conjugate gradient method and a matrix iterative method. Appl. Math. *41* (1996), 19–39.

[Zi1 96] *J. Zítko*: Combining the GMRES and a matrix iterative method. ZAMM (Proceedings of ICIAM/GAMM 95) Vol. 76. 1996, pp. 595–596.

[Zi2 96] *J. Zítko*: Improving the convergence of GMRES using preconditioning and pre-iterations. Proceedings of the conference "Prague Mathematical Conference 1996". 1996, pp. 377–382.

[Zi 97] *J. Zítko*: Behaviour of GMRES iterations using preconditioning and pre-iterations. ZAMM (Proceedings of GAMM 96) Vol. 77. 1997, pp. 693–694.

Author's address: *Jan Zítko*, Katedra numerické matematiky MFF UK, Malostranské náměstí 25, 118 00 Praha 1, Czech Republic, e-mail `zitko@ms.mff.cuni.cz`.