NOVI SAD J. MATH.
VOL. 26, NO. 2, 1996, 31–47

ON THE CONVERGENCE OF A MODIFIED BLOCK SOR ALGORITHM

Boško S. Jovanović [1]
University of Belgrade, Faculty of Mathematics
Studentski trg 16, P.O.Box 550, 11001 Belgrade, Yugoslavia
e–mail: xpmfm13@yubgss21.bg.ac.yu

Abstract

In this paper we consider a vector iterative alternating directions difference scheme for solving multidimensional Poisson equation. The scheme reduce to a modified block successive overrelaxation (SOR) algorithm. We investigate the stability and the convergence of the scheme, determine the optimal iterative parameters, and estimate the error.

AMS Mathematics Subject Classification (1991): 65 N 22
Key words and phrases: linear system, successive overrelaxation

1. Introduction

By discretisation of boundary value problems for linear partial differential equations one obtains large linear systems. Their matrices are sparse and have distinctive structure. Iterative alternating direction methods are often used to solve such systems (see [6], [3], [7], [5]).

[1] Supported by Ministry of Science and Technology of Serbia, grant number 04M03 / C

A new class of alternating direction difference schemes was proposed recently for solving multidimensional evolutive problems, so called multicomponent schemes (see [1], [2], [9]). The main idea lies in vectorization of the problem, i.e. the unknown solution is approximated by vector mesh–function. Convergence of these schemes was proved and their numerical stability was checked on numerous test and real problems.

It is well known that every difference scheme for solving initial boundary value problems of parabolic type can be interpreted as an iterative method for solving the corresponding stationary problem. The aim of this paper is investigation of one class of iterative multicomponent methods.

As a model problem we consider the Dirichlet boundary value problem for the Poisson equation in the domain $\Omega = (0, 1)^n$

$$(1) \qquad \begin{aligned} -\Delta u &= f, \qquad x \in \Omega, \\ u(x) &= 0, \qquad x \in \Gamma = \partial\Omega. \end{aligned}$$

Let $\overline{\omega}$ be a uniform mesh in $\overline{\Omega}$, with the step size $h = 1/(N+1)$. Let us use the notation $\omega = \overline{\omega} \cap \Omega$ and $\gamma = \overline{\omega} \setminus \omega$. For a function v defined on the mesh $\overline{\Omega}$ we introduce the finite differences $v_{x_i} = \big(v(x + h\, r_i) - v(x)\big)/h$ and $v_{\bar{x}_i} = \big(v(x) - v(x - h\, r_i)\big)/h$, where r_i is the unit vector of the x_i axis [8].

Let H_h denote the set of discrete functions defined on the mesh $\overline{\omega}$, which vanish on γ. We introduce the difference operators

$$\Lambda_i v = \begin{cases} -v_{x_i \bar{x}_i}, & x \in \omega \\ 0, & x \in \gamma \end{cases} \qquad \text{and} \qquad \Lambda v = \sum_{i=1}^{n} \Lambda_i v\,.$$

Let I denote the unit operator on H_h. We also define the discrete inner product

$$(v, w) = h^n \sum_{x \in \omega} v(x)\, w(x) \qquad \text{and the norm} \qquad \|v\| = (v, v)^{1/2}.$$

We approximate the boundary value problem (1) with the standard $(2\,n + 1)$–point difference scheme

$$(2) \qquad \Lambda v = \widetilde{f}, \qquad v, \widetilde{f} \in H_h,$$

where $\widetilde{f}$ is some approximation of f.

For solving the problem (2) we use the following multicomponent alternating directions scheme

$$(3) \qquad (I + \frac{\tau}{\sigma}\Lambda_i)\frac{v_k^i - v_{k-1}^i}{\tau} + \sum_{j=1}^{i-1}\Lambda_j v_k^j + \sum_{j=i}^{n}\Lambda_j v_{k-1}^j = \tilde{f},$$
$$i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots$$

where k is the iteration number, while σ and τ are free parameters. The scheme (3) is a system with n unknown mesh functions v_k^i. To determine v_k^i we must solve a linear system whose matrix can be represented in a tridiagonal form and whose order is N^n (= number of nodes of the mesh ω). For $\sigma = 1$ the scheme (3) reduces to the algorithm used in [1] for the solution of a initial–boundary value problem for the heat conduction equation.

In the paper, we shall often use the same notation for linear operators in H_h and their matrix representation.

2. Convergence Result

To investigate the convergence of the method (3) let us represent the equation (3) in the matrix form (see [4] and [9])

$$(4) \qquad \frac{v_k - v_{k-1}}{\tau} + \left(L + \frac{1}{\sigma}I\right)\Lambda v_k + \left[\left(1 - \frac{1}{\sigma}\right)I + U\right]\Lambda v_{k-1} = f,$$

where

$$v_k = (v_k^1, v_k^2, \ldots, v_k^n)^T, \qquad f = (\tilde{f}, \tilde{f}, \ldots, \tilde{f})^T,$$
$$\Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_n), \qquad I = \operatorname{diag}(I, I, \ldots, I),$$
$$L = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ I & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I & I & \ldots & 0 & 0 \\ I & I & \ldots & I & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & I & \ldots & I & I \\ 0 & 0 & \ldots & I & I \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & I \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

Let us also define the inner product and the norm of vector–functions

$$(v, w) = \sum_{i=1}^{n}(v^i, w^i) \qquad \text{and} \qquad \|v\| = (v, v)^{1/2}.$$

Expressing v_k from (4) we get the canonical form of the method

$$v_k = B\, v_{k-1} + c\,, \tag{5}$$

where $B = \left[I + \tau\,\left(L + \frac{1}{\sigma} I\right)\,\Lambda\right]^{-1}\,\left\{I - \tau\,\left[\left(1 - \frac{1}{\sigma}\right)\,I + U\right]\,\Lambda\right\}$ and $c = \tau\,\left[I + \tau\,\left(L + \frac{1}{\sigma} I\right)\,\Lambda\right]^{-1}\,f$. In such a way, the method (4) reduces to a modified block SOR algorithm (see [10]), while for $\sigma = 1$ it reduces to a modified block Gauss–Seidel algorithm.

It is well known (see [10]), that method (5) converges for arbitrary vectors v_0 and c from H_h^n if and only if the moduli of all eigenvalues of the matrix B are lower than 1. Such eigenvalues are the roots of the equation

$$\det\,(B - \lambda\, I) = 0\,. \tag{6}$$

Lemma 2.1. *Eigenvalues of the matrix* B *are* $\lambda = 1$ *and*

$$\lambda = \lambda_{k_1, k_2, \ldots, k_n} = \prod_{j=1}^{n} \frac{1 + \tau\,\left(\frac{1}{\sigma} - 1\right)\,\lambda_{k_j}}{1 + \frac{\tau}{\sigma}\,\lambda_{k_j}}\,, \qquad k_1, k_2, \ldots, k_n = 1, 2, \ldots, N$$

where $\lambda_{k_j} = \frac{4}{h^2}\,\sin^2\,\frac{k_j\,\pi\,h}{2}$.

Proof. From (6), after simple transformations, we obtain the equation

$$D = \det\,(M + \lambda\, L + I + U) = 0\,,$$

where $M = M(\lambda) = \frac{\lambda - 1}{\tau}\,\left(\Lambda^{-1} + \frac{\tau}{\sigma} I\right) = \mathrm{diag}\,(M_1, \ldots, M_n)$ and $M_i = \frac{\lambda - 1}{\tau}\,\left(\Lambda_i^{-1} + \frac{\tau}{\sigma} I\right)$. Further

$$0 = D = \begin{vmatrix} M_1 + I & I & I & \ldots & I \\ \lambda\, I & M_2 + I & I & \ldots & I \\ \lambda\, I & \lambda\, I & M_3 + I & \ldots & I \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda\, I & \lambda\, I & \lambda\, I & \ldots & M_n + I \end{vmatrix}$$

$$= \begin{vmatrix} M_1 + I & I & I & \ldots & I & I \\ -P_1 & M_2 & 0 & \ldots & 0 & 0 \\ 0 & -P_2 & M_3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & M_{n-1} & 0 \\ 0 & 0 & 0 & \ldots & -P_{n-1} & M_n \end{vmatrix}$$

$$= \begin{vmatrix} 0 & 0 & 0 & \dots & 0 & I \\ -P_1 & M_2 & 0 & \dots & 0 & 0 \\ 0 & -P_2 & M_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & M_{n-1} & 0 \\ D_1 & -M_n & -M_n & \dots & -P_{n-1} - M_n & M_n \end{vmatrix},$$

where $P_i = M_i - (\lambda - 1) I = \frac{\lambda - 1}{\tau} Q_i$, $Q_i = \Lambda_i^{-1} + \tau \left(\frac{1}{\sigma} - 1\right) I$ and $D_1 = -M_n (M_1 + I)$.

From here we obtain

$$0 = \begin{vmatrix} -P_1 & M_2 & 0 & \dots & 0 \\ 0 & -P_2 & M_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & M_{n-1} \\ D_1 & -M_n & -M_n & \dots & -P_{n-1} - M_n \end{vmatrix}$$

$$= \begin{vmatrix} -P_1 & 0 & 0 & \dots & 0 \\ 0 & -P_2 & M_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & M_{n-1} \\ D_1 & D_2 & -M_n & \dots & -P_{n-1} - M_n \end{vmatrix}$$

$$= \det(-P_1) \cdot \begin{vmatrix} -P_2 & M_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & M_{n-1} \\ D_2 & -M_n & \dots & -P_{n-1} - M_n \end{vmatrix},$$

where $D_2 = D_1 P_1^{-1} M_2 - M_n$. Continuing in the same manner, we get

$$\det(-P_1) \cdot \det(-P_2) \cdots \det(-P_{n-2}) \cdot \det D_{n-1} = 0, \tag{7}$$

where $D_i = -D_{i-1} P_{i-1}^{-1} M_i - M_n$, for $i = 2, 3, \dots, n-2$, and $D_{n-1} = -P_{n-1} - D_{n-2} P_{n-2}^{-1} M_{n-1} - M_n$.

Using the relation $M_i = \frac{\lambda - 1}{\tau} Q_i (I + \tau Q_i^{-1})$ and the commutativity of operators Λ_i, by recursion we obtain

$$D_{n-1} = \frac{\lambda - 1}{\tau} Q_{n-1} \cdot \frac{1}{\tau} Q_n \cdot (I + \tau Q_1^{-1}) \cdots (I + \tau Q_n^{-1}) \times$$
$$\times \left[(I + \tau Q_1^{-1})^{-1} \cdots (I + \tau Q_n^{-1})^{-1} - \lambda I \right],$$

which together with (7) yields

$$
\begin{aligned}
(8)\qquad & (-1)^{(n-2)N^n}\left(\frac{\lambda-1}{\tau}\right)^{(n-1)N^n}\left(\frac{1}{\tau}\right)^{N^n}\det Q_1\cdot\det Q_2\cdots\det Q_n \\
& \times\det\left(I+\tau Q_1^{-1}\right)\cdot\det\left(I+\tau Q_2^{-1}\right)\cdots\det\left(I+\tau Q_n^{-1}\right) \\
& \times\det\left[\left(I+\tau Q_1^{-1}\right)^{-1}\cdots\left(I+\tau Q_n^{-1}\right)^{-1}-\lambda I\right]=0\,.
\end{aligned}
$$

From (8) it follows that $\lambda=1$ is the eigenvalue of the matrix B of the multiplicity $(n-1)N^n$. The other eigenvalues can be obtained as eigenvalues of the matrix $(I+\tau Q_1^{-1})^{-1}\cdots(I+\tau Q_n^{-1})^{-1}$. Because $Q_j=\Lambda_j^{-1}+\tau(\frac{1}{\sigma}-1)I=\Lambda_j^{-1}\left[I+\tau(\frac{1}{\sigma}-1)\Lambda_j\right]$ and the eigenvalues of Λ_j equal to λ_{k_j}, $k_j=1,2,\ldots,N$, we get

$$
\lambda=\lambda_{k_1,k_2,\ldots,k_n}=\prod_{j=1}^{n}\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\lambda_{k_j}}{1+\frac{\tau}{\sigma}\lambda_{k_j}},\quad k_1,k_2,\ldots,k_n=1,2,\ldots,N\,.\square
$$

One can directly verify the following result.

Lemma 2.2. *The eigenvectors of the matrix B corresponding to eigenvalue $\lambda=1$ are*

$$
\begin{aligned}
\phi^{1;k_1,k_2,\ldots,k_n}&=\left(\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_1}},-\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_2}},0,\ldots,0\right)^T,\\
\phi^{2;k_1,k_2,\ldots,k_n}&=\left(\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_1}},0,-\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_3}},0,\ldots,0\right)^T,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\phi^{n-1;k_1,k_2,\ldots,k_n}&=\left(\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_1}},0,\ldots,0,-\frac{\varphi^{k_1,k_2,\ldots,k_n}}{\lambda_{k_n}}\right)^T,
\end{aligned}
$$

where

$$
\varphi^{k_1,k_2,\ldots,k_n}=\sin x_1\cdot\sin x_2\cdots\sin x_n\,,\qquad (x_1,x_2,\ldots,x_n)\in\omega\,.
$$

For the eigenvalue $\lambda_{k_1,k_2,\ldots,k_n}$ *corresponding eigenvector is*

$$\phi^{k_1,k_2,\ldots,k_n} = \phi^{n;k_1,k_2,\ldots,k_n} = \begin{pmatrix} \frac{\varphi^{k_1,k_2,\ldots,k_n}}{1+\frac{\tau}{\sigma}\lambda_{k_1}} \\ \\ \frac{\left[1+\tau\left(\frac{1}{\sigma}-1\right)\lambda_{k_1}\right]\varphi^{k_1,k_2,\ldots,k_n}}{\left(1+\frac{\tau}{\sigma}\lambda_{k_1}\right)\left(1+\frac{\tau}{\sigma}\lambda_{k_2}\right)} \\ \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \\ \frac{\left[1+\tau\left(\frac{1}{\sigma}-1\right)\lambda_{k_1}\right]\cdots\left[1+\tau\left(\frac{1}{\sigma}-1\right)\lambda_{k_{n-1}}\right]\varphi^{k_1,k_2,\ldots,k_n}}{\left(1+\frac{\tau}{\sigma}\lambda_{k_1}\right)\cdots\left(1+\frac{\tau}{\sigma}\lambda_{k_{n-1}}\right)\left(1+\frac{\tau}{\sigma}\lambda_{k_n}\right)} \end{pmatrix}.$$

The vectors $\varphi^{k_1,k_2,\ldots,k_n}$ represent an orthogonal basis of H_h. From the representation

$$\tilde{f} = \sum_{k_1,k_2,\ldots,k_n} \alpha_{k_1,k_2,\ldots,k_n}\,\varphi^{k_1,k_2,\ldots,k_n}$$

we immediately obtain

$$c = \tau \sum_{k_1,k_2,\ldots,k_n} \alpha_{k_1,k_2,\ldots,k_n}\,\phi^{k_1,k_2,\ldots,k_n}.$$

Choosing $v_0 = 0$, from (5) it follows

$$(9)\qquad v_k = (B^{k-1} + B^{k-2} + \ldots + I)\,c$$
$$= \tau \sum_{k_1,k_2,\ldots,k_n} \alpha_{k_1,k_2,\ldots,k_n}(\lambda^{k-1}_{k_1,k_2,\ldots,k_n} + \lambda^{k-2}_{k_1,k_2,\ldots,k_n} + \ldots + 1)\phi^{k_1,k_2,\ldots,k_n},$$

and this iterative process converges if

$$(10)\qquad \max_{k_1,k_2,\ldots,k_n} |\lambda_{k_1,k_2,\ldots,k_n}| < 1.$$

If (10) is satisfied, for $k \to \infty$ from (9) follows

$$(11)\qquad v_k \to v = \tau \sum_{k_1,k_2,\ldots,k_n} \alpha_{k_1,k_2,\ldots,k_n} \frac{1}{1-\lambda_{k_1,k_2,\ldots,k_n}}\,\phi^{k_1,k_2,\ldots,k_n}.$$

Using the orthogonality of vectors $\varphi^{k_1,k_2,\ldots,k_n}$ and the Parseval equation, from (9) and (11) follows

$$(12)\qquad \|v_k - v\| \le \left(\max_{k_1,k_2,\ldots,k_n} |\lambda_{k_1,k_2,\ldots,k_n}|\right)^k \|v_0 - v\|, \qquad v_0 \in \tilde{H}_h.$$

The same conclusion holds if v_0 is an arbitrary linear combination of vectors $\phi^{k_1,k_2,\ldots,k_n}$. The set of such vectors is a subspace of H_h^n, which will be denoted by $\widetilde{H}_h$. One can directly check that the condition (10) is satisfied for $\tau > 0$ and $0 < \sigma \leq 2$. Thus the following result holds true.

Lemma 2.3. *If $\tau > 0$, $0 < \sigma \leq 2$ and $v_0 \in \widetilde{H}_h$ then the iterative process (5) converges and the error estimate (12) holds.*

Let us estimate $\max_{k_1,k_2,\ldots,k_n} |\lambda_{k_1,k_2,\ldots,k_n}|$. The values λ_{k_j} belong to the interval $[m, M]$, where

$$m = \frac{4}{h^2}\sin^2\frac{\pi h}{2} \geq 8, \qquad M = \frac{4}{h^2}\cos^2\frac{\pi h}{2} < \frac{4}{h^2}.$$

Further

$$\max_{k_1,k_2,\ldots,k_n} |\lambda_{k_1,k_2,\ldots,k_n}| = \max_{k_1,k_2,\ldots,k_n} \prod_{j=1}^{n} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\lambda_{k_j}}{1+\frac{\tau}{\sigma}\lambda_{k_j}}\right|$$

$$\leq \sup_{\mu_j\in[m,M]} \prod_{j=1}^{n} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu_j}{1+\frac{\tau}{\sigma}\mu_j}\right| \leq \prod_{j=1}^{n} \sup_{\mu_j\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu_j}{1+\frac{\tau}{\sigma}\mu_j}\right|$$

$$= \left[\sup_{\mu\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right|\right]^n.$$

If the condition (10) is satisfied, the convergence rate of the sequence (9) is optimal when $\max_{k_1,k_2,\ldots,k_n} |\lambda_{k_1,k_2,\ldots,k_n}|$ is minimal. Supposing that σ and τ are nonnegative, in a natural way we obtain the following inf–sup problem: find

$$q = q(\sigma) = \inf_{\tau\geq 0} \sup_{\mu\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right|.$$

Let us first consider the case $0 < \sigma \leq 1$. Denote

$$\psi(t) = \left|\frac{1+\left(\frac{1}{\sigma}-1\right)t}{1+\frac{t}{\sigma}}\right|.$$

Let $\tau > 0$ be fixed. The function

$$\psi_0(\mu) = \psi(\tau\mu) = \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right|$$

is monotonically decreasing for $\mu > 0$, and

$$\sup_{\mu\in[m,M]} \psi_0(\mu) = \psi_0(m) = \psi(m\tau) = \frac{1+\tau\left(\frac{1}{\sigma}-1\right) m}{1+\frac{\tau}{\sigma} m}.$$

The function $\psi_1(\tau) = \psi(m\tau)$ is also decreasing for $\tau \geq 0$, so

$$(13) \qquad \begin{aligned} q &= q(\sigma) = \inf_{\tau\geq 0} \sup_{\mu\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right| \\ &= \inf_{\tau\geq 0} \psi_1(\tau) = \lim_{\tau\to\infty} \psi_1(\tau) = 1-\sigma. \end{aligned}$$

For $1 < \sigma \leq 2$, the function $\psi_0(\mu)$ is decreasing on the interval $[0, \frac{\sigma}{\tau(\sigma-1)}]$, increasing on the interval $[\frac{\sigma}{\tau(\sigma-1)}, +\infty]$ and we have $\psi_0(0) = 1$, $\psi_0(\sigma/\tau(\sigma-1)) = 0$ and $\lim_{\mu\to+\infty} \psi_0(\mu) = \sigma - 1$ (fig. 1). It follows that

$$\sup_{\mu\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right| = \sup_{\mu\in[m,M]} \psi_0(\mu) = \max\{\psi_0(m), \psi_0(M)\}.$$

Treated as functions of τ, $\psi_1(\tau) = \psi_0(m) = \psi(m\tau)$ and $\psi_2(\tau) = \psi_0(M) = \psi(M\tau)$ have an analogous behaviour as $\psi_0(\mu)$ (fig. 2). Because $0 < \frac{\sigma}{M(\sigma-1)} < \frac{\sigma}{m(\sigma-1)}$ there exists a point $\tau_0 \in (\frac{\sigma}{M(\sigma-1)}, \frac{\sigma}{m(\sigma-1)})$ such that $\psi_1(\tau_0) = \psi_2(\tau_0)$. Furthermore, $\psi_1(\tau) > \psi_2(\tau)$ for $\tau \in (0, \tau_0)$ and $\psi_1(\tau) < \psi_2(\tau)$ for $\tau \in (\tau_0, +\infty)$. In such a way

$$\max\{\psi_0(m), \psi_0(M)\} = \max\{\psi_1(\tau), \psi_2(\tau)\} = \begin{cases} \psi_1(\tau), & \tau \in (0, \tau_0] \\ \psi_2(\tau), & \tau \in [\tau_0, +\infty) \end{cases}$$

and

$$\inf_{\tau\geq 0} \sup_{\mu\in[m,M]} \left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right| = \psi_1(\tau_0) = \psi_2(\tau_0) < 1.$$

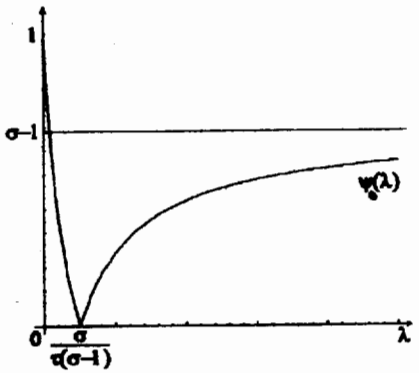

Fig. 1

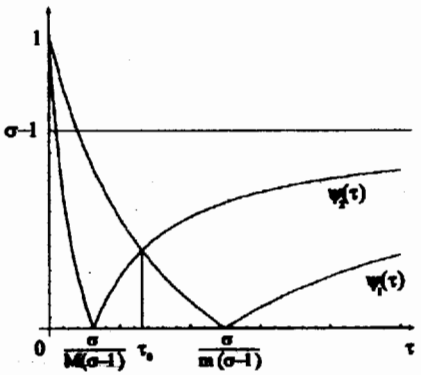

Fig. 2

The value τ_0 can be easily obtained as

$$\tau_0=\tau_0(\sigma)=\sigma\frac{(2-\sigma)(M+m)+\sqrt{(2-\sigma)^2(M+m)^2+16(\sigma-1)Mm}}{4(\sigma-1)Mm}, \tag{14}$$

and from where

$$q=q(\sigma)=\inf_{\tau\geq 0}\sup_{\mu\in[m,M]}\left|\frac{1+\tau\left(\frac{1}{\sigma}-1\right)\mu}{1+\frac{\tau}{\sigma}\mu}\right| \tag{15}$$

$$=(\sigma-1)\frac{(2+\sigma)M-(2-\sigma)m-\sqrt{(2-\sigma)^2(M+m)^2+16(\sigma-1)Mm}}{(3\sigma-2)M+(2-\sigma)m+\sqrt{(2-\sigma)^2(M+m)^2+16(\sigma-1)Mm}}.$$

Notice that the function $\tau_0(\sigma)$ is decreasing on the interval $(1, 2]$ (fig. 3). In particular

$$\lim_{\sigma\to 1+0}\tau_0(\sigma)=+\infty\,;\qquad \tau_0(2)=\frac{2}{\sqrt{Mm}}=\frac{h^2}{\sin\pi h}\asymp\frac{h}{\pi},\quad h\to 0\,.$$

The function $q(\sigma)$ is increasing on $(1, 2]$ (fig. 4), and

$$q(1)=0\,;\qquad q(2)=\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}=\frac{1-\sin\pi h}{\cos\pi h}\asymp 1-\pi h,\quad h\to 0\,.$$

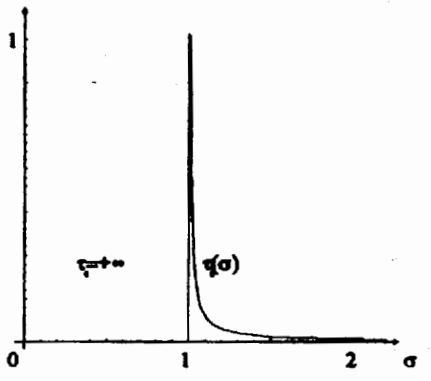

Fig. 3

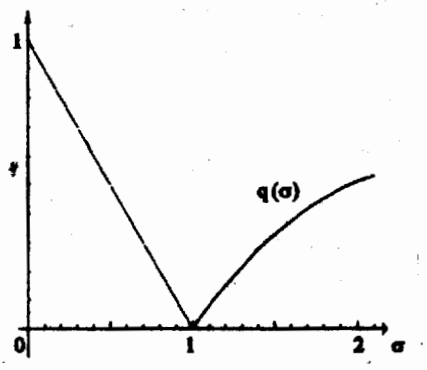

Fig. 4

In such a manner we have proved the following proposition:

Theorem 2.1. *If the initial vector* v_0 *of the sequence (4) belongs to the subspace* $\widetilde{H}_h$ *then the optimal parameter* τ, *by the maximal convergence rate criterion is* $\tau = \infty$, *for* $\sigma \in (0, 1]$, *or* $\tau = \tau_0(\sigma)$, *for* $\sigma \in (1, 2]$. *In this case, the convergence rate estimate holds*

$$\|v_k - v\| \leq q^{nk} \|v_0 - v\|, \qquad v_0 \in \widetilde{H}_h,$$

where $q = q(\sigma)$ *is defined by (13) or (15).*

Remark. In the case when $\sigma = 1$, $\tau = \infty$ the method (4) becomes

$$(L + I)\,\Lambda v_k + U\,\Lambda v_{k-1} = f$$

and converges in a single step if $v_0 \in \widetilde{H}_h$. For example, if $v_0 = 0$ then

$$v_1 = v_2 = \ldots = v = (\Lambda_1^{-1} \tilde{f}, 0, \ldots, 0)^T.$$

Note that in this case the limit vector does not determine the solution of the starting problem (2).

3. Internal Error of the Method

In the previous paragraph we have proved that for a suitable choice of iterative parameter τ the iterative process (4) converges to the limit vector (11). On the other hand, the exact solution of the problem (2) is

$$(16)\quad v^* = (\Lambda_1 + \Lambda_2 + \ldots + \Lambda_n)^{-1}\tilde{f} = \sum_{k_1,k_2,\ldots,k_n} \frac{\alpha_{k_1,k_2,\ldots,k_n}}{\lambda_{k_1} + \ldots + \lambda_{k_n}} \varphi^{k_1,k_2,\ldots,k_n}.$$

Let us estimate the distance from the exact solution v^* to the symmetrized first component

$$(17)\quad \overline{v} = (I + \frac{\tau}{\sigma}\Lambda_1)\, v^1 = \tau \sum_{k_1,k_2,\ldots,k_n} \frac{\alpha_{k_1,k_2,\ldots,k_n}}{1 - \lambda_{k_1,k_2,\ldots,k_n}} \varphi^{k_1,k_2,\ldots,k_n}$$

of the limit vector v. From (16) and (17) after some algebraic manipulation we obtain

$$(18)\quad v^* - \overline{v} = \sum_{k_1,k_2,\ldots,k_n} \alpha_{k_1,k_2,\ldots,k_n} \varphi^{k_1,k_2,\ldots,k_n} \left\{ \frac{1}{\lambda_{k_1} + \ldots + \lambda_{k_n}} - \tau \frac{\left(1 + \frac{\tau}{\sigma}\lambda_{k_1}\right) \cdots \left(1 + \frac{\tau}{\sigma}\lambda_{k_n}\right)}{\left(1 + \frac{\tau}{\sigma}\lambda_{k_1}\right) \cdots \left(1 + \frac{\tau}{\sigma}\lambda_{k_n}\right) - \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\lambda_{k_1}\right] \cdots \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\lambda_{k_n}\right]} \right\}.$$

In such a way, the error estimation $v^* - \overline{v}$ reduces to estimation of the maximum of modulus of the function

$$\chi(\mu_1, \mu_2, \ldots, \mu_n) = \frac{1}{\mu_1 + \ldots + \mu_n} - \tau \frac{\left(1 + \frac{\tau}{\sigma}\mu_1\right) \cdots \left(1 + \frac{\tau}{\sigma}\mu_n\right)}{\left(1 + \frac{\tau}{\sigma}\mu_1\right) \cdots \left(1 + \frac{\tau}{\sigma}\mu_n\right) - \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\mu_1\right] \cdots \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\mu_n\right]}$$

in the domain $\mu_1, \ldots, \mu_n \in [m, M]$.

The function $\chi(\mu_1, \mu_2, \ldots, \mu_n)$ can be represented in the form

$$\chi(\mu_1, \mu_2, \ldots, \mu_n) = \frac{\chi_2(\mu_1, \mu_2, \ldots, \mu_n)}{(\mu_1 + \ldots + \mu_n)\, \chi_1(\mu_1, \mu_2, \ldots, \mu_n)},$$

where

$$\chi_1(\mu_1, \mu_2, \ldots, \mu_n) = \left(1 + \frac{\tau}{\sigma}\mu_1\right) \cdots \left(1 + \frac{\tau}{\sigma}\mu_n\right) - \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\mu_1\right] \cdots \left[1 + \tau\left(\frac{1}{\sigma} - 1\right)\mu_n\right],$$

and

$$\chi_2(\mu_1, \mu_2, \ldots, \mu_n) = \chi_1(\mu_1, \mu_2, \ldots, \mu_n) - \tau(\mu_1 + \ldots + \mu_n)\left(1 + \frac{\tau}{\sigma}\mu_1\right)\cdots\left(1 + \frac{\tau}{\sigma}\mu_n\right).$$

Rearanging the previous expression we get

$$\begin{aligned}(19)\quad \chi_1(\mu_1, \mu_2, \ldots, \mu_n) &= \tau(\mu_1 + \ldots + \mu_n) \\ &+\tau^2\left(\frac{2}{\sigma} - 1\right)(\mu_1\mu_2 + \ldots + \mu_{n-1}\mu_n) + \ldots \\ &+\tau^i\left[\frac{1}{\sigma^i} - \left(\frac{1}{\sigma} - 1\right)^i\right](\mu_1\cdots\mu_i + \ldots + \mu_{n-i+1}\cdots\mu_n) + \ldots \\ &+\tau^n\left[\frac{1}{\sigma^n} - \left(\frac{1}{\sigma} - 1\right)^n\right]\mu_1\cdots\mu_n .\end{aligned}$$

For $0 < \sigma < 2$

$$\frac{1}{\sigma} > \left|\frac{1}{\sigma} - 1\right|, \qquad \text{and consequently} \qquad \frac{1}{\sigma^i} - \left(\frac{1}{\sigma} - 1\right)^i > 0 .$$

In such a way, all the coefficients in (19) are positive. Similarly, the function χ_2 can be represented in the form

$$\begin{aligned}(20)\quad -\chi_2(\mu_1, \mu_2, \ldots, \mu_n) &= \\ &= \tau^2\left\{\frac{1}{\sigma}(\mu_1 + \ldots + \mu_n)^2 - \left(\frac{2}{\sigma} - 1\right)(\mu_1\mu_2 + \ldots + \mu_{n-1}\mu_n)\right\} \\ &+\tau^3\left\{\frac{1}{\sigma^2}(\mu_1 + \ldots + \mu_n)(\mu_1\mu_2 + \ldots + \mu_{n-1}\mu_n)\right. \\ &\quad \left.-\left[\frac{1}{\sigma^3} - \left(\frac{1}{\sigma} - 1\right)^3\right](\mu_1\mu_2\mu_3 + \ldots + \mu_{n-2}\mu_{n-1}\mu_n)\right\} + \ldots \\ &+\tau^i\left\{\frac{1}{\sigma^{i-1}}(\mu_1 + \ldots + \mu_n)(\mu_1\cdots\mu_{i-1} + \ldots + \mu_{n-i+2}\cdots\mu_n)\right. \\ &\quad \left.-\left[\frac{1}{\sigma^i} - \left(\frac{1}{\sigma} - 1\right)^i\right](\mu_1\cdots\mu_i + \ldots + \mu_{n-i+1}\cdots\mu_n)\right\} + \ldots \\ &+\tau^n\left\{\frac{1}{\sigma^{n-1}}(\mu_1 + \ldots + \mu_n)(\mu_1\cdots\mu_{n-1} + \ldots + \mu_2\cdots\mu_n)\right. \\ &\quad \left.-\left[\frac{1}{\sigma^n} - \left(\frac{1}{\sigma} - 1\right)^n\right]\mu_1\cdots\mu_n\right\} \\ &+\tau^{n+1}\left\{\frac{1}{\sigma^n}(\mu_1 + \ldots + \mu_n)\mu_1\cdots\mu_n\right\}.\end{aligned}$$

Notice, that all the coefficients multiplying τ^i in (20) are positive

$$\begin{aligned} & \frac{1}{\sigma^{i-1}}(\mu_1 + \ldots + \mu_n)(\mu_1 \cdots \mu_{i-1} + \ldots + \mu_{n-i+2} \cdots \mu_n) \\ - & \left[\frac{1}{\sigma^i} - \left(\frac{1}{\sigma} - 1\right)^i\right](\mu_1 \cdots \mu_i + \ldots + \mu_{n-i+1} \cdots \mu_n) \\ \geq & \left\{\frac{i}{\sigma^{i-1}} - \left[\frac{1}{\sigma^i} - \left(\frac{1}{\sigma} - 1\right)^i\right]\right\}(\mu_1 \cdots \mu_i + \ldots + \mu_{n-i+1} \cdots \mu_n). \end{aligned}$$

For positive $\mu_1, \ldots, \mu_n$ the last inequality is equivalent to

$$i\,\sigma \geq 1 - (1-\sigma)^i = \sigma\left[1 + (1-\sigma) + \ldots + (1-\sigma)^{i-1}\right],$$

which is satisfied for $0 < \sigma < 2$.

Comparing the coefficients multiplying τ^i in (19) and (20) we get

$$|\chi_2(\mu_1, \mu_2, \ldots, \mu_n)| \leq \frac{\tau}{\sigma(2-\sigma)}(\mu_1 + \mu_2 + \ldots + \mu_n)\,\chi_1(\mu_1, \mu_2, \ldots, \mu_n),$$

for $0 < \sigma < 2$ and $\mu_j > 0$. From here we finally obtain the desired estimate

$$(21) \qquad \max_{\mu_j \in [m, M]} |\chi(\mu_1, \mu_2, \ldots, \mu_n)| \leq \frac{\tau}{\sigma(2-\sigma)}, \qquad \text{for} \qquad 0 < \sigma < 2.$$

Notice that the estimate (21) is of optimal order. Really, for $n = 4$

$$\lim_{h \to 0} |\chi(M, M, \ldots, M)| \geq \frac{\tau}{4}\left(\frac{1}{2-\sigma} + \frac{1}{\sigma}\right) = \frac{1}{2} \cdot \frac{\tau}{\sigma(2-\sigma)}.$$

From (18) and (21) we immediately obtain the following proposition

Theorem 3.1. *For the iterative process (4), in the case of convergence, the error estimate*

$$(22) \qquad \|v^* - \overline{v}\| \leq \frac{\tau}{\sigma(2-\sigma)}\|\widetilde{f}\|, \qquad \textit{holds for} \qquad 0 < \sigma < 2.$$

Remark 1. Estimate (22) is sufficient to describe the internal error of the method. Let $v = (v^1, v^2, \ldots, v^n)^T$ be the limit vector. Then

$$v^{i+1} = \left(I + \frac{\tau}{\sigma}\Lambda_{i+1}\right)^{-1}\left[I + \tau\left(\frac{1}{\sigma} - 1\right)\Lambda_i\right]v^i, \qquad i = 1, 2, \ldots, n-1.$$

From here follows

$$v^{i+1} - v^i = \left(I + \frac{\tau}{\sigma}\Lambda_{i+1}\right)^{-1}\left[\tau\left(\frac{1}{\sigma} - 1\right)\Lambda_i - \frac{\tau}{\sigma}\Lambda_{i+1}\right]v^i, \qquad \text{and}$$

$$\|v^{i+1} - v^i\| \leq \tau\left(\left|\frac{1}{\sigma} - 1\right|\|\Lambda_i v^i\| + \frac{1}{\sigma}\|\Lambda_{i+1} v^i\|\right),$$

i.e. the mutual distance of the limit vector components is also of the order $O(\tau)$. Consequently, for $i = 1, 2, \ldots, n$

$$\|v^* - v^i\| \leq \|v^* - \overline{v}\| + \|\overline{v} - v^i\|$$

$$\leq \|v^* - \overline{v}\| + \|\overline{v} - v^1\| + \sum_{j=2}^{i}\|v^{j-1} - v^j\| = O(\tau).$$

Remark 2. From the previous it follows that the free parameter σ must be determined by two criteria: to maximize the convergence rate (minimize $q(\sigma)$) and to minimize the error $v^* - \overline{v}$ (in practice it is sufficient to set $\tau/\sigma(2-\sigma) = O(h^2)$). Unfortunately, these conditions are contradictory. Choosing parameter τ by the maximal convergence rate criterion for the sequence v_k we conclude that

$$\frac{\tau_0(\sigma)}{\sigma(2-\sigma)} = +\infty, \qquad \text{for} \qquad 0 < \sigma \leq 1,$$

and

$$\frac{\tau_0(\sigma)}{\sigma(2-\sigma)} = \frac{M + m + \sqrt{(M+m)^2 + 16\,\frac{\sigma-1}{(2-\sigma)^2}\,Mm}}{4(\sigma-1)Mm}$$

$$\geq \frac{M}{4(\sigma-1)Mm} = \frac{1}{4(\sigma-1)m} \geq \frac{1}{4m} \geq \frac{1}{4\pi^2}, \qquad \text{for} \qquad 1 < \sigma < 2.$$

In such a way, in the case of maximal convergence rate of v_k, the error $v^* - \overline{v}$ doesn't converge to zero when $h \to 0$.

Taking suboptimal $\tau \in (0, \tau_0(\sigma))$, $\max_{\mu\in[m,M]}\psi_0(\mu)$ is reached for $\mu = m$. Seting

$$\tau = C\,h^\alpha,$$

when $h \to 0$, we get

$$\widetilde{q} = \psi_0(m) = 1 - O(h^\alpha), \qquad \frac{\tau}{\sigma(2-\sigma)} = O(h^\alpha).$$

In such a manner, in the case of "fast" convergence – the error is "large" ($0 < \alpha \leq 1$), while in the case of "small" error – the convergence is "slow" ($\alpha = 2$).

The method can be, for instance, used for a rough approximation of the solution, which can be improved later with some other method.

Remark 3. The observed discrepancy is not in disharmony with the good convergence properties of the method in parabolic case [1]. Namely, the parameter τ is then a time step size, and is controlled by the natural convergence condition $\tau = O(h^2)$.

References

[1] Abrashin, V.N., A variant of the method of variable direction for the solution of multidimensional problems in mathematical physics, Differentsial'nye Uravneniya 26 (1990), 314–323 (Russian).

[2] Abrashin, V.N., Mukha, V.A., A class of efficient difference schemes for solving multidimensional problems in mathematical physics, Differentsial'nye Uravneniya 28 (1992), 1786–1799 (Russian).

[3] Douglas J., On numerical integration of $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = \partial u/\partial t$ by implicit methods, J. Soc. Industr. Appl. Math. 3 (1955), 42–65.

[4] Jovanović, B.S., On the convergence of a multicomponent alternating direction difference scheme, Publ. Inst. Math. 55 (69) (1994).

[5] Marchuk, G.I., Splitting Methods, Nauka, Moscow 1988. (Russian).

[6] Peaceman, D.W., Rachford, H.H., The numerical solution of parabolic and elliptic differential equations, J. Soc. Industr. Appl. Math. 3 (1955), 28–42.

[7] Samarskiĭ, A.A., Additive difference schemes, International congress of mathematics, Moscow (1966), Section 14, 46–47. (Russian).

[8] Samarskiĭ, A.A., Theory of difference schemes, Nauka, Moscow 1983. (Russian).

[9] Vabishchevich, P.N., Vector additive difference schemes for first order evolution equations, Zh. Vychisl. Mat. Mat. Fiz. 36 (1996), 44–51 (Russian).

[10] Young, D.M., Iterative Solution of Large Linear Systems, Academic Press, New York – London 1971.

Received by the editors November 11, 1996.