SELECTION OF THE ITERATIVE PARAMETERS IN RICHARDSON'S METHOD*

E. S. NIKOLAEV and A. A. SAMARSKII

Moscow

(*Received* 24 *March* 1972)

AN ordering of the set of iterative parameters of Richardson's method for which it becomes numerically stable is presented. The number of parameters is arbitrary.

In this paper we consider the question of the numerical stability of Richardson's iterative method of solving an operator equation of the first kind in Hilbert space. This method possesses a high rate of convergence; however, its numerical instability for problems with an ill-conditioned operator has been revealed in practice [1–3].

It was shown in [4] that the instability of the method is connected with the order of use of the iterative parameters, and that previously [1–3] proposed methods of ordering the set of parameters do not remove the numerical instability, but only decrease it.

Further investigations have shown that there exists an order and a set of iterative parameters for which the method becomes numerically stable. This approach was proposed in [4] and [5], chapter VIII, for the case where the number of parameters is $n = 2^p$.

In the present paper the method of ordering the parameters explained in [5] is generalized to the case of an arbitrary number of parameters n.

In Section 1 a description is given of Richardson's method and of the ordering of the set of parameters for the case $n = 2^p$. Theorems on numerical stability are also formulated there. A detailed proof of the theorem will be given separately. Section 2 is devoted to defining the order in the set of parameters for the case of arbitrary n. The results of an experimental study of the numerical stability of the method with a description of the set of parameters is presented in Section 3.

* *Zh. vȳchisl. Mat. mat. Fiz.*, 12, 4, 960–973, 1972.

1. Formulation of the problem

1. In the real Hilbert space H let there be given an operator equation of the first kind with self-conjugate operator ($A = A^* > 0$)

$$Au = f, \tag{1.1}$$

where f is a prescribed, and u an unknown, element of H.

For the approximate solution of problem (1.1) we consider the implicit two-level iterative scheme

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \qquad k = 0, 1, \ldots, n-1, \tag{1.2}$$

with the arbitrary initial approximation $y_0 \in H$. We distinguish the family of schemes (1.2) by the condition

$$B = B^* \geqslant \beta E,\ \beta > 0, \tag{1.3}$$

and we will suppose that the operators A and B are energetically equivalent to the constants γ_1 and γ_2 [5]:

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B,\ \gamma_2 > \gamma_1 > 0. \tag{1.4}$$

On the assumptions (1.3), (1.4) the solution of the problem of the optimal set of iterative parameters τ_h is of the form [5]

$$\begin{aligned} &\tau_k = \frac{\tau_0}{1 + \rho_0\mu_k}, \\ &\mu_k \in \mathfrak{M}_n = \left\{\cos\frac{2i-1}{2n}\pi,\quad i = 1, 2, .., n\right\}, \qquad k = 1, 2, \ldots, n, \end{aligned} \tag{1.5}$$

where $\mathfrak{M}_n$ is a set of n elements arranged in the order in which i increases

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \qquad \rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2}.$$

With this choice of parameters the following estimates hold:

$$\|y_n - u\|_D \leqslant q_n \|y_0 - u\|_D,\ D = A \text{ or } B.$$

Here the norm in the energy space H_D is defined as follows:

$$\|x\|_D = \sqrt{(Dx, x)} \quad \text{for} \quad D^* = D > 0, \qquad x \in H,$$

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \qquad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

In order to decrease the norm of the initial error in the space H_A (H_B) by the factor $1/\epsilon$ it is sufficient to take n iterations:

$$n = n(\varepsilon, \xi) \approx \frac{\ln 0.5\varepsilon}{\ln \rho_1} \approx \frac{|\ln 0.5\varepsilon|}{2\sqrt{\xi}}.$$

The scheme of (1.2) with the set of parameters (1.5) is called Richardson's implicit iterative method.

2. In the study of the convergence of the method of (1.2) we have assumed that the computational process is ideal, that is, the calculations are carried out to an infinite number of places. However, the process of rounding the results of the arithmetic operations introduces some errors into the solution y_n at each stage of the calculations. We will assume that the introduction of these errors is equivalent to a perturbation of the input data of the problem - the initial approximation, the right side and the operators A and B.

Then the actual solution $\tilde{y}_k$ may be regarded as the exact solution of the problem

$$(1.6) \qquad \tilde{B}\frac{\tilde{y}_{k+1} - \tilde{y}_k}{\tau_{k+1}} + \tilde{A}\tilde{y}_k = \tilde{f}_{k+1} + \frac{1}{\tau_{k+1}}\tilde{w}_{k+1}, \qquad k = 0, 1, \ldots, n,$$

$\tilde{y}_0$ given.

With this approach the problem of the computational error of the method reduces to an investigation of the stability of the scheme (1.6) with the perturbed operators $\tilde{A}$ and $\tilde{B}$ with respect to the initial data and right side.

The second question of the numerical stability of the method of (1.2) is the question of the increase of the intermediate solutions y_k for various orderings of the set $\mathfrak{M}_n$. The study of the stability of the scheme (1.6) makes it possible to obtain an estimate for the value of the intermediate solution.

Some examples showing the effect of the order of use of the parameters τ_k on the increase of the solution and the accuracy of the method are given in [5] and in Section 3 of this paper.

We will suppose that the scheme of (1.6) belongs to the original family of schemes, that is, the following conditions are satisfied:

$$\tilde{A}=\tilde{A}^*>0,\ \tilde{B}=\tilde{B}^*\geqslant\tilde{\beta}E,\ \tilde{\beta}>0. \tag{1.7}$$

As a measure of the perturbation of the operators A and B we will take the relative variation of their energy $(0\leqslant a_1, a_2<1)$

$$|((A-\tilde{A})x,x)|\leqslant a_1(Ax,x),\ |((B-\tilde{B})x,x)|\leqslant a_2(Bx,x). \tag{1.8}$$

In order to study the increase of the intermediate solutions $\tilde{y}_k$ we will change from the implicit scheme (1.6) to the equivalent explicit scheme

$$\begin{aligned} &x_{k+1}=\tilde{S}_{k+1}x_k+\tau_{k+1}\varphi_{k+1}+\psi_{k+1}, \qquad k=0,1,\dots,n-1,\\ &\tilde{S}_{k+1}=E-\tau_{k+1}\tilde{C}, \end{aligned} \tag{1.9}$$

where

$$x_k=\tilde{B}^{1/2}\tilde{y}_k, \qquad \tilde{C}=\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2}, \qquad \varphi_k=\tilde{B}^{-1/2}\tilde{f}_k,$$
$$\psi_k=\tilde{B}^{-1/2}\tilde{w}_k.$$

In order to study the numerical accuracy of the method we will consider the problem for the error $z_k=\tilde{y}_k-u$:

$$\tilde{B}\frac{z_{k+1}-z_k}{\tau_{k+1}}+\tilde{A}z_k=\tilde{f}_{k+1}-f+\frac{1}{\tau_{k+1}}\tilde{w}_{k+1}+(A-\tilde{A})u,$$
$$z_0=\tilde{y}_0-u.$$

The equivalent explicit scheme is of the form

$$\begin{aligned} &x_{k+1}-\varphi=\tilde{S}_{k+1}(x_k-\varphi)+\tau_{k+1}\varphi_{k+1}+\psi_{k+1}, \quad k=0,1,\dots,n-1,\\ &\tilde{S}_{k+1}=E-\tau_{k+1}\tilde{C}, \end{aligned} \tag{1.10}$$

where

$$x_k=\tilde{B}^{1/2}z_k, \qquad \tilde{C}=\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2}, \qquad \varphi_k=\tilde{B}^{-1/2}(\tilde{f}_k-f),$$
$$\psi_k=\tilde{B}^{-1/2}\tilde{w}_k, \qquad \varphi=\tilde{B}^{1/2}\tilde{A}^{-1}(A-\tilde{A})u.$$

It follows from (1.9), (1.10) that in order to obtain the estimates we require it is sufficient to study the stability of the scheme (1.9) with respect to the initial data and the right side.

Using the recurrence formula (1.9), we find

$$x_k = \tilde{T}_{k,0} x_0 + \sum_{j=1}^{k} \tau_j \tilde{T}_{k,j} \psi_j + \sum_{j=1}^{k} \tilde{T}_{k,j} \psi_j, \qquad k = 1, 2, \ldots, n$$

(1.11)

$$\tilde{T}_{k,j} = \prod_{i=j+1}^{k} \tilde{S}_i, \qquad \tilde{T}_{k,k} \equiv E.$$

The operator $\tilde{T}_{k,j}$ is called the resolving operator from the level j to the level k.

Definition. We call the collection of parameters $\{\tau_k\}$ stable, if positive C_1, C_2, C_3, perhaps depending on γ_1, γ_2, a_1, a_2, but independent of n, exist such that

$$\max_{1 \leqslant k \leqslant n} \|\tilde{T}_{k,0}\| \leqslant C_1, \qquad \max_{1 \leqslant k \leqslant n} \sum_{j=1}^{k} \tau_j \|\tilde{T}_{k,j}\| \leqslant C_2, \qquad \max_{1 \leqslant k \leqslant n} \sum_{j=1}^{k} \|\tilde{T}_{k,j}\| \leqslant C_3.$$

For a stable collection of parameters for any k the estimates

$$\|x_k\| \leqslant C_1 \|x_0\| + C_2 \max_{1 \leqslant j \leqslant k} \|\varphi_j\| + C_3 \max_{1 \leqslant j \leqslant k} \|\psi_j\|$$

hold for problem (1.9) and

$$\|x_k\| \leqslant C_1 \|x_0\| + C_2 \max_{1 \leqslant j \leqslant k} \|\varphi_j\| + C_3 \max_{1 \leqslant j \leqslant k} \|\psi_j\| + (1 + C_1) \|\varphi\|$$

for problem (1.10), from which it follows that the schemes (1.9), (1.10) are stable.

3. We will now specify the order of the elements in the set $\mathfrak{M}_n$ which will generate a stable collection of iterative parameters $\{\tau_k\}$ of Richardson's method Here we consider the case where $n = 2^p$, $p > 0$.

In the construction of the sequence of parameters $\{\tau_k\}$, following [5], we will begin from minimal $\beta = \beta_1 = \pi/2n$ and construct recursively the sets

$$M_1(\beta) = \{-\cos \beta\},$$

(1.12)

$$M_{2^k}(\beta) = M_{2^{k-1}}(\beta) \cup M_{2^{k-1}}(\pi / 2^{k-1} - \beta), \quad k = 1, 2, \ldots, p.$$

Then

(1.13) $$\mathfrak{M}_{2^p} = M_{2^p}(\beta_1).$$

The following theorem states that the set $\mathfrak{M}_{2p}$ ordered in this way generates a stable sequence $\{r_k\}$ (in formula (1.5) μ_k is the k-th element of the set $\mathfrak{M}_{2p}$).

Theorem 1

If the conditions of (1.3) and (1.4) are satisfied and the set $\mathfrak{M}_{2p}$ is ordered in accordance with (1.13), then if $\alpha_1 = \alpha_2 = 0$, the following estimates hold independently of n ($m = j \cdot 2^k$, j is an odd number, $k \geq 0$):

$$\|T_{m,0}\| \leqslant 1/\xi, \qquad m = 1, 2, \dots, 2^p,$$

$$\sum_{i=1}^{m} \tau_i \|T_{m,i}\| \leqslant \frac{1}{\gamma_1}\left[1 + \left(1 - \delta_{1,j}\ {}^{\backprime} \frac{1}{\xi^2}\right],\right.$$

$$\sum_{i=1}^{m} \|T_{m,i}\| \leqslant \frac{1}{\xi}\left[1 + (1 - \delta_{1,j})\frac{1}{\xi^2}\right],$$

where $\delta_{1,j}$ is the Kronecker delta. For $m = 2^p$ the following more precise estimates hold:

$$(1.14) \qquad \|T_{2^p,0}\| \leqslant q_{2^p} < 1, \qquad \sum_{j=1}^{2^p} \|T_{2^p,j}\| \tau_j \leqslant \frac{1 - q_{2^p}}{\gamma_1} < \frac{1}{\gamma_1},$$

$$(1.15) \qquad \sum_{j=1}^{2^p} \|T_{2^p,j}\| \leqslant \frac{4}{3\sqrt{\xi}}.$$

Theorem 1 expresses the fact that if in the scheme (1.6) the perturbations of the operators A and B is neglected, the intermediate solutions are bounded in norm ($m = j \cdot 2^k$).

$$\|\tilde{y}_m\|_B \leqslant \frac{1}{\xi}\|\tilde{y}_0\|_B + \left[1 + (1 - \delta_{1,j})\frac{1}{\xi^2}\right]\left(\frac{1}{\gamma_1}\max_{1\leqslant i\leqslant m}\|f_i\|_{B^{-1}} + \frac{1}{\xi}\max_{1\leqslant i\leqslant m}\|\tilde{w}_i\|_{B^{-1}}\right)$$

and for the error of the solution after n iterations we have the estimate

$$(1.16) \qquad \|\tilde{y}_n - u\|_B \leqslant q_n\|\tilde{y}_0 - u\|_B + \frac{1 - q_n}{\gamma_1}\max_{1\leqslant i\leqslant n}\|\tilde{f}_i - f\|_{B^{-1}} + \frac{4}{3\sqrt{\xi}}\max_{1\leqslant i\leqslant n}\|\tilde{w}_i\|_{B^{-1}}.$$

Theorem 2

If the conditions (1.3), (1.4), (1.7), (1.8) are satisfied and the set $\mathfrak{M}_2^P$ is ordered in accordance with (1.13), then subject to the condition

$$\alpha=\frac{\alpha_1+\alpha_2}{1-\alpha_2}\leqslant\frac{\xi}{2}$$

we have for the error of the solution of problem (1.6) after n iterations the estimate

$$\|\tilde{y}_n-u\|_{\tilde{B}}\leqslant\frac{q_n}{\tilde{q}_n}\|\tilde{y}_0-u\|_{\tilde{B}}+\frac{1}{\tilde{\gamma}_1}\left(1-\frac{q_n}{\tilde{q}_n}\right)\max_{1\leqslant i\leqslant n}\|\tilde{f}_i-f\|_{\tilde{B}^{-1}}+ \tag{1.17}$$
$$\frac{4}{3}\frac{1+\sqrt{\alpha}}{\sqrt{\xi}-\sqrt{\alpha}}\max_{1\leqslant i\leqslant n}\|\tilde{w}_i\|_{\tilde{B}^{-1}}+\frac{\alpha_1}{\tilde{\gamma}_1}\left(1+\frac{q_n}{\tilde{q}_n}\right)\|f\|_{\tilde{B}^{-1}}.$$

Here

$$\tilde{\gamma}_1=\gamma_1-\alpha\gamma_2\geqslant 0.5\,\gamma_1,$$
$$\tilde{q}_n=\frac{2\tilde{\rho}_1^{\,n}}{1+\tilde{\rho}_1^{\,2n}}>q_n,\qquad \tilde{\rho}_1=\frac{1-\sqrt{\xi_0}}{1+\sqrt{\xi_0}},\qquad \xi_0=\frac{\alpha}{1+\alpha-\xi}.$$

Theorem 2 expresses the numerical stability of Richardson's iterative method for the case where $n=2^p$. As $\alpha_1,\alpha_2\to 0$ the estimate (1.17) becomes the previously obtained estimate (1.6). The estimates (1.14) given by Theorem 1 cannot be improved for any ordering of the set $\mathfrak{M}_n$, and (1.15) is exact in respect of the order of smallness of ξ for large n (compare with Lemma 6 of [4]).

Note. If it is assumed that ψ_j in (1.11) is of the form $\psi_j=T_{j,0}v_j$, this perturbation is equivalent to a perturbation of x_0 by an amount $\sum_{j-1}^{k}v_j$.

2. Construction of a sequence of parameters for arbitrary n

1. The idea of the construction of an order in the set $\mathfrak{M}_n$ is based on two considerations: the passage from the ordered set $\mathfrak{M}_{2k}$ to the specification of an order in the set $\mathfrak{M}_{2k+1}$, and the passage from $\mathfrak{M}_k$ to $\mathfrak{M}_{2k}$, where k is an arbitrary integer. (We have been informed that in [6, 7] V. I. Lebedev gave other sequences of parameters for any n.)

We assume that the set $\mathfrak{M}_{2k}$ has already been ordered in the way we require. We then represent it as the following sum:

$$\mathfrak{M}_{2k}=\bigcup_{i=1}^{2k}M_1(\beta_{2k,i})=M_1(\beta_{2k,1})\cup M_1(\beta_{2k,2})\cup\ldots\cup M_1(\beta_{2k,2k}),$$

where for any i the element $\beta_{2k,\,i}$ belongs to the set

$$\left\{\frac{2j-1}{4k}\pi,\ \ j=1,2,\ldots,2k\right\}.$$

We then order the set $\mathfrak{M}_{2k+1}$ as follows:

$$\mathfrak{M}_{2k+1}=\bigcup_{i=1}^{2k} M_1(\beta_{2k+1,i})\cup M_1\left(\frac{\pi}{2}\right),$$

where $\beta_{2k+1,\,i}$ is a number close to $\beta_{2k,i}$ of the set

$$\left\{\frac{2j-1}{2(2k+1)}\pi,\ j=1,2,\ldots,2k+1,\ j\neq k+1\right\}.$$

The passage from $\mathfrak{M}_{2k}$ to $\mathfrak{M}_{2k+1}$ has been accomplished.

Also, let the set $\mathfrak{M}_k$ be ordered:

$$\mathfrak{M}_k=\bigcup_{i=1}^{k} M_1(\beta_{k,i})=M_1(\beta_{k,1})\cup\ldots\cup M_1(\beta_{k,k}),$$

$$\beta_{k,i}\in\left\{\frac{2j-1}{2k}\pi,\ \ j=1,2,\ldots,k\right\}.$$

Then, using formulas (1.12), we specify the order in the set $\mathfrak{M}_{2k}$ as follows:

$$\mathfrak{M}_{2k}=\bigcup_{i=1}^{k} M_2(\beta_{2k,i})=\bigcup_{i=1}^{k}\left(M_1(\beta_{2k,i})\cup M_1(\pi-\beta_{2k,i})\right),$$

$$\beta_{2k,\,i}=0.5\beta_{k,\,i},\qquad i=1,2,\ldots,k.$$

These considerations enable us to pass from the ordered set $\mathfrak{M}_1=M_1(\pi/2)$, consisting of one element, to the set $\mathfrak{M}_n$ with arbitrary n, alternating as necessary transitions from a set with an even number of elements to a set with an odd number and from a set of k elements to a set of $2k$ elements.

This procedure for ordering the set $\mathfrak{M}_n$ for arbitrary n can be formalized as follows.

We represent n as an expansion in the sum of powers of 2 with the integral exponents k_i:

$$n=2^{k_1}+2^{k_2}+\ldots+2^{k_t},\ k_j\leqslant k_{j-1}-1,\ k_t\geqslant 0.$$

Here t is an integral subscript. We introduce the quantities

$$(2.1)\qquad n_j = \sum_{i=1}^{j} 2^{k_i - k_j}, \qquad \delta_j = \frac{n_j}{n} 2^{k_j}, \qquad j = 1, 2, \ldots, t,$$

and put $k_{t+1} = -1$. We notice that all the n_j are odd numbers.

In order to construct the sequence of parameters we will begin with the least

$$\beta = \beta_1 = \pi/2n.$$

We form the ordered sum of sets

$$(2.2)\qquad M_n(\beta) = \bigcup_{i=1}^{t} M_{2^{k_i}}(n_i\beta) = M_{2^{k_1}}(n_1\beta) \cup \ldots \cup M_{2^{k_t}}(n_t\beta),$$

where $M_{2^k}(\beta)$ is defined recursively ($j = 1, 2, \ldots, t$):

$$(2.3)\qquad \begin{aligned} &M_1(\beta) = \{-\cos\beta\}, \\ &M_{2^k}(\beta) = M_{2^{k-1}}(\beta) \cup M_{2^{k-1}}\left(\frac{\pi}{2^{k-1}}\delta_j - \beta\right), \end{aligned}$$

if $k_{j+1} + 2 \leqslant k \leqslant k_j + 1$.

Then

$$(2.4)\qquad \mathfrak{M}_n = M_n(\beta_1).$$

We notice that if $n = 2^p$, we have $t = 1$, $n_1 = 1$, $\delta_1 = 1$, $k_1 = p$. Then formula (2.2) becomes

$$M_n(\beta) = M_{2p}(\beta)$$

and the recurrence relations (2.3) defining $M_{2p}(\beta)$ become formulas (1.12) for $n = 2^p$, given above. Consequently, by (2.4) the ordering of the set $\mathfrak{M}_n$ is a generalization of the constructions for the case $n = 2^p$.

We now explain an algorithm which enables us to order the set $\mathfrak{M}_n$ in accordance with (2.4). It is difficult to construct the set $\mathfrak{M}_n$ directly from formulas (2.2), (2.3); we use them to describe the ordering of $\mathfrak{M}_n$ in Theorems 1, 2.

Let θ_m be a set of m integer-valued elements

$$\theta_m = \{\theta_m(1), \theta_m(2), \dots, \theta_m(m)\}.$$

We put $n_{t+1} = 2n + 1$. Let $j = 1$. We construct the sets

$$(2.5) \qquad \theta_{n_j} = \{\theta_{n_j}(i) = \theta_{n_j - 1}(i), \quad i = 1, 2, \dots, n_j - 1; \ \theta_{n_j}(n_j) = n_j\},$$

$$(2.6) \qquad \theta_{2m} = \{\theta_{2m}(2i) = 4m - \theta_m(i), \ \theta_{2m}(2i-1) = \theta_m(i), \ i = 1, 2, \dots, m\},$$
$$m = n_j, 2n_j, 4n_j, \dots, 0.25(n_{j+1} - 1).$$

If $j = t$, the required set θ_n has already been constructed, otherwise we construct the set

$$(2.7) \qquad \theta_{n_{j+1}-1} = \{\theta_{n_{j+1}-1}(2i) = 2n_{j+1} - \theta_{0.5(n_{j+1}-1)}(i), \quad \theta_{n_{j+1}-1}(2i-1) =$$
$$\theta_{0.5(n_{j+1}-1)}(i), \quad i = 1, 2, \dots, 0.5(n_{j+1} - 1)\}.$$

Then j is increased by 1 and the process is repeated, beginning with (2.5). As a result the set θ_n will be constructed.

Then

$$(2.8) \qquad \mathfrak{M}_n = \left\{-\cos\beta_i, \ i = 1, 2, \dots, n, \ \beta_i = \theta_n(i)\frac{\pi}{2n}\right\}$$

and μ_k in formula (1.5) is the k-th element of the set $\mathfrak{M}_n$.

For the case $n = 2^p$ the algorithm (2.5)–(2.7) is simplified:

$$\theta_1 = \{\theta_1(1) = 1\},$$
$$\theta_{2m} = \{\theta_{2m}(2i) = 4m - \theta_m(i), \ \theta_{2m}(2i-1) = \theta_m(i), \ i = 1, 2, \dots, m\},$$
$$m = 1, 2, 4, \dots, 2^{p-1},$$

and after finding θ_{2p} the set $\mathfrak{M}_{2p}$ is constructed by (2.8). We give some examples:

$$\theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\},$$
$$\theta_9 = \{1, 17, 7, 11, 3, 15, 5, 13, 9\},$$
$$\theta_{12} = \{1, 23, 11, 13, 5, 19, 7, 17, 3, 21, 9, 15\},$$
$$\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27. 11, 21\},$$
$$\theta_{18} = \{1, 35, 17, 19, 7, 29, 11, 25, 3, 33, 15, 21, 5, 31, 13, 23, 9, 27\},$$

3. Numerical stability of the method

1. A numerical experiment is used in order to study the stability of the sequence of iteration parameters $\{\tau_k\}$ constructed in accordance with the ordering of the set $\mathfrak{M}_n$ by formula (2.4).

The nature of the iterative process can obviously be investigated on the simplest model, since the nature of the process is determined by the specific form of the operator A and its fundamental functional properties as an operator in Hilbert space.

An experiment on a simulated problem enables us to compare the theoretical estimates obtained for the case $n = 2^p$ in Theorems 1, 2, with the numerical results.

The simulated problem chosen is the difference approximation of the boundary value problem

$$\begin{aligned} &\frac{d^4v}{dx^4} = f(x), \qquad 0 < x < 1, \qquad v(0) = \nu_1, \\ &\frac{d^2v}{dx^2}(0) = \nu_2, \qquad v(1) = \nu_3, \qquad \frac{d^2v}{dx^2}(1) = \nu_4. \end{aligned} \tag{3.1}$$

On a uniform grid with step $h = 1/N$ we construct a difference scheme approximating problem (3.1) with an error of order $O(h^2)$:

$$\begin{aligned} &u_{\bar{x}x\bar{x}x} = f, \qquad 2h \leqslant x \leqslant 1 - 2h, \\ &u(0) = \nu_1, \qquad u_{\bar{x}x} - hu_{\bar{x}xx} = \nu_2, \qquad x = h, \\ &u(1) = \nu_3, \qquad u_{\bar{x}x} + hu_{\bar{x}x\bar{x}} = \nu_4, \qquad x = 1 - h. \end{aligned} \tag{3.2}$$

The operator A corresponding to problem (3.2) is selfconjugate in the space H of grid functions defined at the internal nodes of the grid. The scalar product in H is defined as follows:

$$(u, v) = \sum_{x=h}^{1-h} u(x)v(x)h.$$

Here the eigenfunctions of the operator A will be $\mu_k(x) = \sin k\pi x$ and the corresponding eigenvalues will be

$$\lambda_k = \frac{16}{h^4}\sin^4\frac{k\pi h}{2}, \qquad k = 1, 2, \ldots, N-1.$$

We note that $\|A\| = \lambda_{N-1} \approx 16/h^4 = 1.6 \times 10^5$ for $N = 10$, and $\|A\| \approx 1.6 \times 10^9$ for $N = 100$.

The choice of problem (3.2) as the object of experiment enables us to simulate ill-conditioned operators A on a coarse grid.

The explicit iterative process (1.2) ($B \equiv E$) is considered. Then with conditions (1.4) the energy equivalence constants γ_1 and γ_2 are $\gamma_1 = \lambda_1$, $\gamma_2 = \lambda_{N-1}$. Then

$$\xi = \operatorname{tg}^4 \frac{\pi h}{2}.$$

2. Several series of experiments were performed. In the first series the effect of the ordering of the set $\mathfrak{M}_n$ on the increase of the intermediate solutions and the accuracy attained after n iterations were studied.

In the simulated problem (3.2) the values $\nu_1 = 1$, $\nu_2 = \nu_3 = \nu_4 = 0$, $f \equiv 0$ were taken. With these initial data the exact solution of the problem is $u(x) = 1 - x$.

Let n be a multiple of 8. The following orderings of the set $\mathfrak{M}_n$ were considered:

(1)
$$\mathfrak{M}_n = \bigcup_{k=1}^{n} M_1\left(\frac{2k-1}{2n}\pi\right) =$$
$$M_1\left(\frac{\pi}{2n}\right) \cup M_1\left(\frac{3\pi}{2n}\right) \cup \ldots \cup M_1\left(\frac{2n-1}{2n}\pi\right);$$

To this ordering there corresponds the usual "inverse" sequence $\{r_k\}$:

(2)
$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \qquad \mu_k = -\cos\frac{2k-1}{2n}\pi, \qquad k = 1, 2, \ldots, n;$$
$$\mathfrak{M}_n = \bigcup_{k=1}^{n} M_1\left(\pi - \frac{2k-1}{2n}\pi\right), \quad \text{here} \quad \mu_k = \cos\frac{2k-1}{2n}\pi,$$

which corresponds to the usual "direct" sequence of parameters $\{r_k\}$;

(3)
$$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{2k-1}{2n}\pi\right) =$$
$$\bigcup_{k=1}^{n/2}\left(M_1\left(\frac{2k-1}{2n}\pi\right) \cup M_1\left(\pi - \frac{2k-1}{2n}\pi\right)\right);$$

this ordering corresponds to the partition of the sequence $\{r_k\}$ into blocks of two elements ($\mu_{2k-1} = -\cos[(2k-1)/2n]\pi$, $\mu_{2k} = \cos[(2k-1)/2n]\pi$); here and below we use the recurrence formulas (1.12);

(4)
$$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{\pi}{2} - \frac{2k-1}{2n}\pi\right) =$$
$$\bigcup_{k=1}^{n/2}\left(M_1\left(\frac{\pi}{2} - \frac{2k-1}{2n}\pi\right) \cup M_1\left(\frac{\pi}{2} + \frac{2k-1}{2n}\pi\right)\right).$$

The ordering recommended in [1] assumes, in our notation, the form

(5) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{\pi}{2}+\frac{2k-1}{2n}\pi\right);$$

$$\mathfrak{M}_n = \bigcup_{k=1}^{n/4} M_4\left(\frac{2k-1}{2n}\pi\right) =$$

$$\bigcup_{k=1}^{n/4}\left(M_1\left(\frac{2k-1}{2n}\pi\right)\cup M_1\left(\pi-\frac{2k-1}{2n}\pi\right)\cup\right.$$

$$\left. M_1\left(\frac{\pi}{2}-\frac{2k-1}{2n}\pi\right)\cup M_1\left(\frac{\pi}{2}+\frac{2k-1}{2n}\pi\right)\right).$$

Here the blocks have been organized to consist of four elements:

$$\left(-\cos\beta_k,\ \cos\beta_k,\ -\sin\beta_k,\ \sin\beta_k;\quad \beta_k=\frac{2k-1}{2n}\pi\right);$$

(6) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/8} M_8\left(\frac{2k-1}{2n}\pi\right) =$$

$$\bigcup_{k=1}^{n/8}\left(M_4\left(\frac{2k-1}{2n}\pi\right)\cup M_4\left(\frac{\pi}{4}-\frac{2k-1}{2n}\pi\right)\right);$$

here a block consists of 8 elements;

(7) the ordering of the set $\mathfrak{M}_n$ in accordance with (2.4).

In order to exhibit the effect of the conditionality of the operator A on the computational stability of the method, the calculations were carried out on the sequence of grids $h = {}^1\!/_{10},\ {}^1\!/_{12},\ {}^1\!/_{14}$.

The setting of n used was from 8 to 512 with step 8. Iterations in accordance with the scheme (1.2) were carried out for every n and each of the orderings $\mathfrak{M}_n$ indicated. The critical value n_{acc} for which the actual relative accuracy did not yet exceed the theoretical accuracy was determined, that is, the inequality

$$\frac{\|y_n-u\|}{\|y_0-u\|}=\varepsilon_{\mathrm{real}}\leqslant q_n.$$

was satisfied. In the same way a determination was made of n_{stop} for which at some intermediate iteration the solution exceeded the maximum possible number (10^{19}) for representation in the computer.

The calculations showed that for the orderings (1)–(6) as n increased, there was first a loss of accuracy, when n becomes greater than n_{acc}, and then an

TABLE 1

	$h=1/10,\quad \xi=6.29\cdot 10^{-4}$				$h=1/12,\quad \xi=3.004\cdot 10^{-4}$				$h=1/14,\quad \xi=1.61\cdot 10^{-4}$			
	n_{acc}	q_n	$\varepsilon_{\text{real}}$	n_{stop}	n_{acc}	q_n	$\varepsilon_{\text{real}}$	n_{stop}	n_{acc}	q_n	$\varepsilon_{\text{real}}$	n_{stop}
1	24	0.55	0.495	48	24	0.732	0.729	48	24	0.839	0.773	40
	24	0.55	0.548	48	24	0.732	0.731	48	24	0.839	0.837	48
2	24	0.55	0.481	72	24	0.732	0.713	64	24	0.839	0.765	64
	16	0.746	0.746	64	16	0.864	0.86	64	16	0.923	0.918	64
3	48	0.178	0.17	88	48	0.366	0.364	80	48	0.544	0.376	80
	48	0.178	0.177	96	48	0.366	0.366	88	48	0.544	0.542	88
4	48	0.178	0.171	136	40	0.47	0.42	136	48	0.544	0.482	128
	40	0.264	0.263	136	40	0.47	0.469	128	40	0.64	0.639	128
5	88	$2.4\cdot 10^{-2}$	$2.2\cdot 10^{-2}$	176	96	$7.2\cdot 10^{-2}$	$5.8\cdot 10^{-2}$	168	104	0.142	0.141	160
	104	$1.08\cdot 10^{-2}$	$8.68\cdot 10^{-3}$	192	104	$5.4\cdot 10^{-2}$	$4.9\cdot 10^{-2}$	184	104	0.142	0.105	176
6	184	$1.95\cdot 10^{-4}$	$1.24\cdot 10^{-4}$	416	176	$4.48\cdot 10^{-3}$	$3.74\cdot 10^{-3}$	408	176	$2.29\cdot 10^{-2}$	$1.98\cdot 10^{-2}$	336
	184	$1.95\cdot 10^{-4}$	$1.9\cdot 10^{-4}$	456	176	$4.48\cdot 10^{-3}$	$4.44\cdot 10^{-3}$	448	200	$1.25\cdot 10^{-2}$	$9\cdot 10^{-3}$	376

emergency stop due to an increase in the intermediate solutions.

The use of blocks of more complex structure leads to a decrease of the numerical instability of the method, and the greater the size of a block the greater n_{stop}.

When the ordering (7) is used the real relative accuracy ϵ_{real} for all n did not exceed the theoretical accuracy, which for $n = 512$ was given by $q_n = 1.4 \times 10^{-11}$, 3.9×10^{-8}, 4.5×10^{-6} for $h = 1/10, 1/12, 1/14$. The intermediate results were bounded and the quantity

$$R_n = \max_{\substack{0 \leqslant x \leqslant 1 \\ 1 \leqslant k \leqslant n}} |y_k(x)|$$

was a monotonically increasing function of n and for large n was unchanged as n increased, that is, it attained its asymptotic value.

The calculations showed that the ordering of $\mathfrak{M}_n$ in accordance with (2.4) for an n differing from a power of 2, preserved the same characteristics of computational stability of the method as in the case $n = 2^p$.

The results are shown in Table 1. The first rows correspond to the initial approximation

$$y_0(x)=\begin{cases}1, & x=0,\\ 0, & x\neq 0,\end{cases}$$

and the second correspond to $y_0(x) = \cos(\pi x/2)$. For these approximations with the ordering (7)

$$\max_{1\leqslant n\leqslant 512} R_n = 208,\ 427,\ 784 \text{ и } \max_{1\leqslant n\leqslant 512} R_n = 1.63,\ 2.73,\ 4.00$$

for $h = {}^1/_{10}, {}^1/_{12}, {}^1/_{14}$ respectively.

3. In the second series estimates were found for

$$\|T_{n,0}\|, \qquad \sum_{j=1}^{n} \tau_j \|T_{n,j}\|, \qquad \sum_{j=1}^{n} \|T_{n,j}\|,$$

where $\mathfrak{M}_n$ is ordered by (2.4). For this the following quantities were computed:

(3.3) $$I_1 = \|T_{n,0}y\| \,/\, \|y\| \leqslant \|T_{n,0}\|,$$

(3.4) $$I_2 = \sum_{j=1}^{n} \tau_j \frac{\|T_{n,j}y\|}{\|y\|} \leqslant \sum_{j=1}^{n} \tau_j \|T_{n,j}\|,$$

(3.5) $$I_3 = \sum_{j=1}^{n} \frac{\|T_{n,j}y\|}{\|y\|} \leqslant \sum_{j=1}^{n} \|T_{n,j}\|.$$

For the case $n = 2^p$ in Theorem 1 it was proved that the equations in (3.3)–(3.5) are attained if y is an eigenfunction corresponding to the minimum eigenvalue of the problem

$$Ay - \lambda By = 0.$$

Here the equations

(3.6) $$I_1 = \|T_{2^p,0}\| = q_{2^p}, \qquad I_2 = \sum_{j=1}^{2^p} \tau_j \|T_{2^p,j}\| = \frac{1-q_{2^p}}{\gamma_1}$$

and the estimate

(3.7) $$I_3 = \sum_{j=1}^{2^p} \|T_{2^p,j}\| \leqslant \frac{4}{3\sqrt{\xi}}.$$

TABLE 2

$h = 1/10,\ \xi = 6.29 \times 10^{-4}$

n	I_1	q_n	I_2	$(1-q_n)/\gamma_1$	I_3	$I_3\sqrt{\xi}$
64	$8.0451\cdot10^{-2}$	$8.0451\cdot10^{-2}$	$9.5968\cdot10^{-3}$ $3.5085\cdot10^{-4}$	$9.5968\cdot10^{-3}$	42.726 27.171	1.072 0.6816
96	$1.6174\cdot10^{-2}$	$1.6174\cdot10^{-2}$	$1.0268\cdot10^{-2}$ $3.6973\cdot10^{-4}$	$1.0268\cdot10^{-2}$	45.034 28.641	1.1297 0.718
128	$3.2467\cdot10^{-3}$	$3.2467\cdot10^{-3}$	$1.0403\cdot10^{-2}$ $3.8662\cdot10^{-4}$	$1.0403\cdot10^{-2}$	47.072 29.933	1.181 0.7509
192	$1.308\cdot10^{-4}$	$1.308\cdot10^{-4}$	$1.0435\cdot10^{-2}$ $3.8184\cdot10^{-4}$	$1.0435\cdot10^{-2}$	46.5 29.57	1.167 0.7418
256	$5.27\cdot10^{-6}$	$5.27\cdot10^{-6}$	$1.044\cdot10^{-2}$ $3.868\cdot10^{-4}$	$1.044\cdot10^{-2}$	47.098 29.95	1.182 0.7513
344	$6.37\cdot10^{-8}$	$6.37\cdot10^{-8}$	$1.044\cdot10^{-2}$ $4.3697\cdot10^{-4}$	$1.044\cdot10^{-2}$	53.143 33.768	1.333 0.8471
384	$8.55\cdot10^{-9}$	$8.55\cdot10^{-9}$	$1.044\cdot10^{-2}$ $3.8787\cdot10^{-4}$	$1.044\cdot10^{-2}$	47.225 30.03	1.185 0.753

TABLE 3

$h = 1/20,\ \xi = 3.84 \times 10^{-5}$

n	I_1	q_n	I_2	$(1-q_n)/\gamma_1$	I_3	$I_3\sqrt{\xi}$
64	0.75125	0.75125	$2.5641\cdot10^{-3}$ $3.115\cdot10^{-5}$	$2.5641\cdot10^{-3}$	62.066 39.506	0.384 0.245
96	0.55725	0.55725	$4.564\cdot10^{-3}$ $4.48\cdot10^{-5}$	$4.564\cdot10^{-3}$	89.331 56.863	0.553 0.352
128	0.39313	0.39313	$6.2558\cdot10^{-3}$ $5.708\cdot10^{-5}$	$6.2558\cdot10^{-3}$	113.86 72.474	0.705 0.449
192	0.1838	0.1838	$8.4137\cdot10^{-3}$ $7.42\cdot10^{-5}$	$8.4137\cdot10^{-3}$	148.04 94.234	0.917 0.584
256	$8.3747\cdot10^{-2}$	$8.3747\cdot10^{-2}$	$9.445\cdot10^{-3}$ $8.64\cdot10^{-5}$	$9.445\cdot10^{-3}$	172.26 109.65	1.067 0.679
344	$2.8197\cdot10^{-2}$	$2.8197\cdot10^{-2}$	$1.002\cdot10^{-2}$ $9.88\cdot10^{-5}$	$1.002\cdot10^{-2}$	197.03 125.4	1.22 0.777
384	$1.7181\cdot10^{-2}$	$1.7181\cdot10^{-2}$	$1.013\cdot10^{-2}$ $9.136\cdot10^{-5}$	$1.013\cdot10^{-2}$	182.23 116.0	1.129 0.718
512	$3.5191\cdot10^{-3}$	$3.5191\cdot10^{-3}$	$1.027\cdot10^{-2}$ $9.56\cdot10^{-5}$	$1.027\cdot10^{-2}$	190.66 121.37	1.181 0.752
768	$1.4762\cdot10^{-4}$	$1.4762\cdot10^{-4}$	$1.0307\cdot10^{-2}$ $9.43\cdot10^{-5}$	$1.0307\cdot10^{-2}$	188.18 119.78	1.166 0.742
1024	$6.192\cdot10^{-6}$	$6.192\cdot10^{-6}$	$1.0308\cdot10^{-2}$ $9.56\cdot10^{-5}$	$1.0308\cdot10^{-2}$	190.72 121.4	1.181 0.752

hold. For the simulated problem of (3.2) we have $y = \sin \pi x$.

It was also proved in Theorem 1 that the estimate

$$\sum_{j=1}^{n} \tau_j \| T_{n,j} \| \geqslant \frac{1 - q_n}{\gamma_1}.$$

holds for arbitrary n for any ordering of the set $\mathfrak{M}_n$.

In order to investigate how all the three norms I_1, I_2, I_3 change in the neighbourhood of $n = 2^p$, a numerical experiment was used. To calculate these norms it is sufficient to find

$$\max_{y} I_1, \qquad \max_{y} I_2, \qquad \max_{y} I_3,$$

when y runs through the set of eigenfunctions of the operator A.

Calculations were carried out on the sequence of grids The number of iterations n was specified as in the first series. It turned out that in (3.3)–(3.5) equality was attained at the first eigenfunction of the operator A, as for the case $n = 2^p$, and equations (3.6) were valid.

In the process we verified that the theoretical estimates (3.6), (3.7) for $n = 2^p$ were attained.

Some results are shown in Tables 2, 3. In the first rows the values of I_1, I_2, I_3 are given for $y = \sin \pi x$, and in the second rows for comparison the values of I_2, I_3 are given for $y = \sin (N - 1)\pi x$.

The experiments on the simulated problem considered above show that for an n differing from a power of 2, in an ordering of the set $\mathfrak{M}_n$ in accordance with (2.4), the same characteristics of numerical stability of the method are preserved as for the case $n = 2^p$, for which the theoretical estimates were obtained in Theorems 1, 2.

The use of the ordering of $\mathfrak{M}_n$ by (2.4) is necessary for the solution of problems with an ill-conditioned operator, for example, difference problem arising from the approximation of elliptic equations of high order. At the same time, as is shown by example 1, if n is not too great and the problem is not very ill-conditioned, a simpler ordering with shorter blocks can be used.

Translated by J. Berry

REFERENCES

1. YOUNG, D. On Richardson's method for solving linear system with positive definite matrices, *J. Math. and Phys.*, **32**, 4, 243–255, 1954.

2. FORSYTHE, G. E. and WASOW, W. R. *Finite-Difference Methods for Solving Partial Differential Equations* (Raznostnye metody resheniya differentsial'nykh uravnenii v chastnykh proizvodnykh). Izd-vo in. lit., Moscow, 1963.

3. FADDEEV, D. K. and FADEEVA, V. N. *Numerical Methods of Linear Algebra* (Vychislitel'nye metody lineinoi algebry). Fizmatgiz, Moscow, 1963.

4. LEBEDEV, V. N. and FINOGENOV, S. A. The order of selection of the iterative parameters in the Chebyshev cyclic iterative method. *Zh. vychisl. Mat. mat. Fiz.*, **11**, 2, 425–438, 1971.

5. SAMARSKII, A. A. *Introduction to the Theory of Difference Schemes* (Vvedenie v teoriyu raznostnykh skhem). "Nauka", Moscow, 1971.

6. MARCHUK, G. I. and LEBEDEV, V. I. *Numerical Methods in Neutron Transport Theory* (Chislennye metody v teoriyu perenos neitronov), Atomizdat, Moscow, 1971.

7. LEBEDEV, V. I. and FINOGENOV, S. A. An algorithm for the selection of the parameters in the Chebyshev cyclic methods. In: *Numerical Methods of Linear Algebra* (Vychisl. metody lineinoi algebry), 21–27, VTs SO Akad. Nauk SSSR, Novosibirsk, 1972.