NUMER. FUNCT. ANAL. AND OPTIMIZ., 4(3), 211-239 (1981-1982)

SOME REMARKS ON THE METHOD OF MINIMAL RESIDUES

V.S. Kozjakin
Cardiology Research Center of the USSR
Petroverigskiy Lane 10
Moscow 101837, USSR

M.A. Krasnosel'skiĭ
Institute for Control Problems
Profsojuznaja Street 65
Moscow 117342, USSR

ABSTRACT

Two procedures of approximate solution of systems of linear equations with non-symmetric matrices, the method of minimal residues and that of simple iterations, are considered. A comparison of the methods under consideration is made. According to the comparison of the rates of convergence preference should be given to the method of simple iterations. Some arguments are presented nevertheless, in virtue of which, the method of minimal residues should be preferred.

In this paper we consider systems of linear equations

$$Ax = b \qquad (x, b \in R^N) \qquad (1)$$

and discuss two procedures of approximate construction of their solutions, the method of minimal residues and that of simple iterations, respectively.

The method of minimal residues was introduced and investigated by KRASNOSEL´SKIǏ & KREIN [1] for systems with positive definite symmetric matrices A. It seems that the transition to the non-symmetric case was carried out for the first time by KUZNETSOV [2] who found out that the convergence follows only from a positive definitness of the matrix A. KUZNETSOV & MARCHUK [3, 4] investigated the method of minimal residues for the case of singular matrices. SAMARSKIǏ [5] established an estimate of the rate of convergence of the method of minimal residues for equations (1) with non-symmetric matrices A.

At the beginning of this paper the method of minimal residues is described and Samarskiǐ´s estimate is quoted (Theorem 1). Then the method of simple iterations is described and an explicit expression for the rate of convergence is deduced (Theorems 2 and 3). After this the comparison of the methods under consideration is made (Theorem 4). The paper is completed by proofs. In particular, a new proof of Samarskiǐ´s estimate is presented for the case of a general Hilbert space.

As the comparison of the rates of convergence shows, the method of simple iterations should be preferred. Some arguments are given here with which we nevertheless choose the method of minimal residues.

1. THE METHOD OF MINIMAL RESIDUES

We shall assume that in R^N a scalar product is introduced. For an approximate solution of the system (1) one can construct the iterations

$$x_{n+1} = x_n - p_n(Ax_n - b), \tag{2}$$

defining p_n so that the norm of the residue $Ax_{n+1} - b$ be minimal. Therefore p_n is defined by the formula

$$p_n = \frac{(Ay_n, y_n)}{(Ay_n, Ay_n)} , \tag{3}$$

where $y_n = Ax_n - b$. The iteration process (2)-(3) is called the method of minimal residues.

It is convenient for us to put further tools into effect at once for the equation (1) in a real Hilbert space H.

Denote by $S(m,M,k,H)$ $(0 < m \leq M < \infty , k \geq 0)$ the set of linear continuous operators in H for which

$$m(x,x) \leq (Ax,x) \leq M(x,x), \tag{4}$$

$$\| A - A^* \| \leq 2k. \tag{5}$$

Let be $B = (A+A^*)/2$, $C = (A-A^*)/2$. Then the conditions (4)-(5) are equivalent to the fact that the spectrum of the symmetric operator B lies in the interval $[m , M]$, and the norm of antisymmetric operator C is less than k.

For an investigation of the method (2)-(3) we shall study the behaviour of the residues $y_n = Ax_n - b$. Evidently,

$$y_{n+1} = y_n - \frac{(Ay_n, y_n)}{(Ay_n, Ay_n)} Ay_n .$$

Let us set

$$q(m,M,k,H) = \sup_{A \in S} \sup_{\|x\|=1} \| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \| . \qquad (6)$$

Then in case of $A \in S(m,M,k,H)$ the estimation

$$\|y_{n+1}\| \leq q(m,M,k,H) \|y_n\| \qquad (7)$$

is fulfilled. It is clear that this estimation cannot be improved if all equations are considered with operators A from $S(m,M,k,H)$.

Theorem 1. If $\dim H \geq 2$, then the following equality is true:

$$q(m,M,k,H) = \frac{M^2 - m^2 + 4k(k^2 + Mm)^{1/2}}{(M+m)^2 + 4k^2} .$$

2. THE METHOD OF SIMPLE ITERATIONS

For an approximate solution of the system (1) one can construct the following iterations

$$x_{n+1} = x_n - p(Ax_n - b), \tag{8}$$

choosing p such that the spectral radius $r(I-pA)$ of the matrix $I-pA$ is as small as possible. As is known, in order to guarantee the convergence of the approximations (8), under any initial approximation, to the exact solution x^* of the system (1), it is necessary and sufficient that the inequality $r(I-pA) < 1$ should be satisfied. If $r(I-pA) < 1$, then the approximations given above converge faster than any geometric progression $\{c^n\}$, where $c = \varepsilon + r(I-pA)$ and ε is an arbitrary small positive number.

Complicated questions arise concerning a possibility of selection of the coefficient p and a practical choice of it. For example, if the matrix A has eigenvalues both with positive real parts and negative ones, then it is impossible to choose required coefficient p.

Denote by $T(m,M,k,R^N)$ $(0 < m \leq M < \infty, \quad k \geq 0)$ the family of square matrices of size N, whose spectra lie in the rectangular region: $m \leq \operatorname{Re}\lambda \leq M$, $|\operatorname{Im}\lambda| \leq k$. A coefficient $p = p_0(m,M,k,R^N)$ we call an optimum one with respect to the class $T(m,M,k,R^N)$ if the following equality holds:

$$\sup_{A\in T} r(I-pA) = \sup_{A\in T} \min_{-\infty<t<\infty} r(I-tA). \tag{9}$$

In accordance with this the iteration process (8) with the coefficient $p = p_0(m,M,k,R^N)$ is called an optimum one with respect to the class $T(m,M,k,R^N)$.

Let us set

$$Q(m,M,k,R^N) = \sup_{A\in T} r(I-p_0A). \tag{10}$$

Theorem 2. If $\dim R^N \geq 4$, then the optimum coefficient $p = p_0(m,M,k,R^N)$ with respect to the class $T(m,M,k,R^N)$ exists and is defined by the formula

$$p_0(m,M,k,R^N) = \begin{cases} \dfrac{m}{m^2+k^2}, & \text{if} \quad 2k^2 \geq Mm-m^2, \\[2ex] \dfrac{2}{M+m}, & \text{if} \quad 2k^2 < Mm-m^2. \end{cases} \tag{11}$$

Besides

$$Q(m,M,k,R^N) = \begin{cases} \dfrac{k}{(m^2+k^2)^{1/2}}, & \text{if} \quad 2k^2 \geq Mm-m^2, \\[2ex] \dfrac{\{4k^2+(M-m)^2\}^{1/2}}{M+m}, & \text{if} \quad 2k^2 < Mm-m^2. \end{cases} \tag{12}$$

Let matrix A belong to the class $T(m,M,k,R^N)$. Then, evidently, the iteration procedure of the type (8) can be used only in case when

$$Q(m,M,k,R^N) < 1 \tag{13}$$

In this instance one can guarantee that the successive approximations (8) converge faster than any geometric progression $\{c^n\}$, where $c = \varepsilon + Q(m,M,k,R^N)$, $\varepsilon > 0$.

In order that $A \in T(m,M,k,R^N)$ it is sufficient that $A \in S(m,M,k,R^N)$. In this instance one can say a little more about the rate of convergence of successive approximations. We recall that the number $c < 1$ is called the rate of convergence of the sequence $\{x_n\}$ if $\|x_n\| = O(c^n)$.

Theorem 3. If $A \in S(m,M,k,R^N)$, then the successive approximations (8) with $p = p_0(m,M,k,R^N)$ converge with the rate $Q(m,M,k,R^N)$, i.e. the following inequalities hold

$$\| x_n - x^* \| \leq LQ^n(m,M,k,R^N) \| x_0 - x^* \|$$

with a certain $L \in (0,\infty)$.

On all occasions, as it follows from (12), the condition (13) is valid. Therefore the possibility of applying the iteration procedure (8) to solution of systems (1) with matrices A from $T(m,M,k,R^N)$ follows from Theorem 2. The formula (11) provides the rule of selection of coefficient p.

3. COMPARISON OF THE METHODS

The method (8) is applicable in essentially more general conditions than the method (2)-(3). However, in the conditions when both the methods are applicable preference should be given to one of them. One can compare the method (8) of simple iterations and the method (2)-(3) of minimal residues on the basis of different characteristics. The much more usual way is the comparison of optimum estimations of the decrease of the norm of residues on one step of a process. In our case this corresponds to the comparison of the values $Q(m,M,k,R^N)$ and $q(m,M,k,R^N)$.

Theorem 4. If $m = M$ or $k = 0$, then

$$Q(m,M,k,R^N) = q(m,M,k,R^N).$$

If $m \neq M$ and $k \neq 0$, then

$$Q(m,M,k,R^N) < q(m,M,k,R^N).$$

By this theorem we should prefer the method (8) of simple iterations. We shall give some reasons why, to our mind, the method of minimal residues should be preferred.

a). Suppose, for the beginning, that $k = 0$. In this case the optimum estimation of the rate of convergence of the method (8) and that of (2)-(3) are the same (more details about this are in [1]). However, for the method (8) an unpleasant fact occurs. Namely, under almost all initial approximations the asymptotic rate of convergence of successive approximations is worst possible and equals to $Q(m,M,k,R^N)$. In case of the method of minmal residues, as it follows from the results by FORSYTHE [6] and EMELIN [7] (and as I.V. Emelin noticed), under almost all initial approximations the rate of convergence is strictly less than $Q(m,M,k,R^N)$.

It would be interesting and important to obtain estimations of an average rate of convergence under various natural distributions of initial approximations.

b). Now let $m \neq M$ and $k \neq 0$. From (7), for each n, the estimations

$$\|y_n\| \leq q^n(m,M,k,R^N)\|y_0\|. \tag{14}$$

follow. In spite of exactness of the estimations (7), the estimations (14) are no longer exact.

One should find or estimate as exactly as possible the following values:

$$q_n(m,M,k,R^N) = \sup_{\|y_0\|=1} \|y_n\|^{1/n}. \tag{15}$$

The authors failed to estimate the value (15) exactly enough and to prove any assertions for a general case, which are analogous to the theorems from [6] and [7]. In accordance with this an extensive numerical test was carried out with low-dimensional matrices. I.V. Emelin took part actively in carrying out this test. It turned out that the method of minimal residues is better than the method of simple iterations in the sense of magnitudes of residues with large indices in an absolute plurality of cases. As a rule, directions $z_n = y_n/\|y_n\|$ and coefficients $d_n = \|y_n\|/\|y_{n-1}\|$ with even indices (analogously - with odd ones) behave as follows: for a significant number of indices the directions change little and the quantities d_n are close to the worst one. However, then the directions z_n change suddenly and the quantities d_n become small. After this a new cycle of slow change of z_n and of "bad" d_n comes, and so on.

Explanations and applications of the phenomenon under consideration would be iteresting and important.

c). Applications of the method of minimal residues do not call for knowledge of values $m, M, k,$ in contrast with the method of simple iterations, if che latter is implemented in accordance with Theorem 2.

Of course, cases of ill-conditioned matrices A present an interest. Let, for example, m be small and k be large. Let us set $\varepsilon = m/k$, then, in virtue of Theorem 2, the method (8) under $p = p_0(m,M,k,R^N)$ converges with rate $Q = 1 - \varepsilon^2/2$. In virtue of Theorem 1, the rate of convergence of the method of minimal residues is the same. However, the values m and k in some problems are not known exactly. Therefore the method (8) will be implemented with a not necessarily optimum p. Hence it will converge essentially much slower.

4. PROOF OF THEOREM 1

The following lemma is evident.

Lemma 1. *Let P be an orthogonal projection operator in a Hilbert space H onto a subspace PH. Let be $A \in S(m,M,k,H)$, and let $(PA)_0$ be a restriction of the operator PA onto a subspace PH. Then $(PA)_0 \in S(m,M,k,PH)$.*

Now, let be $x \in H$, $\|x\| = 1$ and $Ax \neq \lambda x$. And let P be an orthogonal projection operator onto two-dimensional plane $H_0 = H_0(x)$ with the basis x, Ax. Since $PAx = Ax$, then

$$x - \frac{(Ax,x)}{(Ax,Ax)} Ax = x - \frac{(PAx,x)}{(PAx,PAx)} PAx .$$

Therefore

$$\| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \| \leq \sup_{z \in H, \|z\|=1} \| z - \frac{(PAz,z)}{(PAz,PAz)} PAz \| .$$

Hence, by virtue of Lemma 1,

$$\| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \| \leq q(m,M,k,H_0),$$

where H_0 is a two-dimensional plane. But then

$$q(m,M,k,H) \leq q(m,M,k,H_0). \qquad (16)$$

Evidently, under fixed m, M, k the function

$q(m,M,k,H)$ depends only on the dimension $\dim H$ of the space H. In case of increasing of $\dim H$ the value $q(m,M,k,H)$ does not decrease. Therefore Lemma 2 follows from (16).

Lemma 2. If $\dim H \geq 2$, then

$$q(m,M,k,H) = q(m,M,k,H_0) \qquad (\dim H_0 = 2)$$

Let be $A \in S(m,M,k,H_0)$. Let g, h be an orthogonal basis in H_0 consisting of eigenvectors of the symmetric operator $B = (A+A^*)/2$. Then in this basis the operator A has matrix:

$$A = \begin{Vmatrix} u & -w \\ w & v \end{Vmatrix}, \tag{17}$$

where $u = (Ag,g)$, $v = (Ah,h)$, $w = (Ag,h) = -(Ah,g)$.

One can assume, without loss in generality, that $u \leq v$. Let us set

$$r(u,v,w) = \max_{\|x\|=1} \left\| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \right\|$$

and show that

$$r(u,v,w) = \frac{v^2 - u^2 + 4|w|(w^2 + uv)^{1/2}}{(u+v)^2 + 4w^2} . \tag{18}$$

Before proving the above equality, we notice that the assertion of Theorem 1 follows from it. Indeed the function in the right-hand side increases in the variables v and $|w|$ and decreases in the variable u. But since $A \in S(m,M,k,H_0)$, then $m \leq u \leq v \leq M$, $|w| \leq k$. Hence

$$\max_{u,v,w} r(u,v,w) = \frac{M^2 - m^2 + 4k(k^2 + Mm)^{1/2}}{(M+m)^2 + 4k^2} .$$

By definition, the left-hand side of this equality is equal to $q(m,M,k,H_0)$. From here and from Lemma 2 we conclude:

$$q(m,M,k,H) = \frac{M^2 - m^2 + 4k(k^2 + Mm)^{1/2}}{(M+m)^2 + 4k^2} . \qquad (19)$$

Thus, we shall only prove the equality (18).

By definition,

$$r(u,v,w) = \max_{\|x\|=1} f(x) ,$$

where

$$f(x) = 1 - \frac{(Ax,x)^2}{(Ax,Ax)} . \qquad (20)$$

Let $z = sg + th$ ($\|z\| = 1$) be the point attaining an extremum to the function $f(x)$. Then the gradient of the function $f(x)$ at the point z is parallel with z, i.e.,

$$(Az,Az)(A+A^*)z - (Az,z)A^*Az = (Az,Az)(Az,z)z.$$

Scalar multiplying this equality by $A^{-1}z$, we obtain

$$1 - \frac{(Az,z)^2}{(Az,Az)} = -(Az - (Az,z)z , A^{-1}z).$$

Since

$$A^{-1} = \frac{u+v}{w^2 + uv} I - \frac{1}{w^2 + uv} A,$$

we get

$$f(z) = \frac{(Az,Az) - (Az,z)^2}{w^2 + uv} .$$

Hence

$$f(z) = \frac{\{(v-u)st - w\}^2}{w^2 + uv} \qquad (21)$$

and $(Az,z) = w^2 + uv$, i.e.

$$u^2s^2 + v^2t^2 = uv - 2w(v-u)st. \qquad (22)$$

Adding the equality $s^2 + t^2 = 1$ to (22) and then expressing s^2 and t^2 by st, we obtain:

$$(u+v)s^2 = v - 2wst, \qquad (u+v)t^2 = u + 2wst.$$

Hence the product st satisfies the equality

$$[4w^2 + (u+v)^2](st)^2 - 2w(v-u)st - uv = 0.$$

Thus, the product st takes one of two values:

$$st = \frac{w(v-u) \pm (v+u)(w^2 + uv)^{1/2}}{(u+v)^2 + 4w^2} .$$

Then from (21) follows that only two values

$$f(z) = \{ \frac{v^2 - u^2 \pm 4|w|(w^2 + uv)^{1/2}}{(u+v)^2 + 4w^2} \}^2$$

may be extremal for the function (20). From here follows the equality (18).

Theorem 1 is proved.

If the fact that the estimation (7) can not be improved is not interesting for us, and we want only to derive inequality

$$\| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \| \le \frac{M^2 - m^2 + 4k(k^2 + Mm)^{1/2}}{(M + m)^2 + 4k^2} \|x\| ,$$

then the proof can essentially be simplified.

Indeed, by Lemma 2, it is sufficient to prove the inequality

$$\| x - \frac{(Ax,x)}{(Ax,Ax)} Ax \| \le \frac{v^2 - u^2 + 4|w|(w^2 + uv)^{1/2}}{(u+v)^2 + 4w^2} \|x\| ,$$

where x is an arbitrary point of a two-dimensional space and A is the matrix (17).

From the construction of the method of minimal residues follows relations:

$$\|x-\frac{(Ax,x)}{(Ax,Ax)}Ax\| = \min_{-\infty<t<\infty}\|x-tAx\| \le \min_{-\infty<t<\infty}\|I-tA\|\|x\|.$$

Consequently, it is sufficient to prove the equality

$$\min_{-\infty<t<\infty}\|I-tA\| = \frac{v^2-u^2+4|w|(w^2+uv)^{1/2}}{(u+v)^2+4w^2}. \qquad (23)$$

We note that the norm of the matrix $I-tA$ is equal to the norm of the following symmetric matrix

$$\left\|\begin{matrix}1 & 0\\ 0 & -1\end{matrix}\right\|(I-tA) = \left\|\begin{matrix}1-tu & -tw\\ -tw & -1+tv\end{matrix}\right\|. \qquad (24)$$

For the calculation of the norm of the latter matrix it is sufficient to find the roots $\lambda_1(t)$ and $\lambda_2(t)$ of its characteristic equation because, as is known, this norm equals to the maximum in $\lambda_1(t)$ and $\lambda_2(t)$. A simple calculation shows that

$$\min_{-\infty<t<\infty}\max\{\lambda_1(t),\lambda_2(t)\} = \frac{v^2-u^2+4|w|(w^2+uv)^{1/2}}{(u+v)^2+4w^2}.$$

This equality is equivalent to (23).

5. PROOF OF THEOREM 2

For the first time, we prove an existence of the optimum coefficient $p_0 = p_0(m,M,k,R^N)$. By definition, the spectrum of each matrix A from $T(m,M,k,R^N)$ lies inside the rectangle with vertices at $m \pm ik$

and $M \pm ik$. Hence it is easy to obtain the following expression for the value $\sup r(I-tA)$:

$$\sup_{A\in T} r(I-tA) = \max\{|1-t(m+ik)|, |1-t(M+ik)|\} \qquad (25)$$

In virtue of this equality,

$$\sup_{A\in T} r(I-tA) = r(I-tK), \qquad (26)$$

where

$$K = \left\| \begin{matrix} m & -k & 0 & 0 \\ k & m & 0 & 0 \\ 0 & 0 & M & -k \\ 0 & 0 & k & M \end{matrix} \right\| . \qquad (27)$$

Denote by $p_0(m,M,k,R^N)$ such value of t under which $\sup r(I-tA)$ attains minimum. In virtue of (25), $\sup r(I-tA) \longrightarrow \infty$ when $|t| \longrightarrow \infty$, therefore the number $p_0 = p_0(m,M,k,R^N)$ is defined correctly.

Write the following relations:

$$\begin{array}{ll}
\text{(i)} & \sup\limits_{A\in T} \min\limits_{t} r(I-tA) \geq \min\limits_{t} r(I-tK) \\[2ex]
\text{(ii)} & \sup\limits_{A\in T} \min\limits_{t} r(I-tA) \leq \min\limits_{t} \sup\limits_{A\in T} r(I-tA) \\[2ex]
\text{(iii)} & \min\limits_{t} r(I-tK) = \min\limits_{t} \sup\limits_{A\in T} r(I-tA) \\[2ex]
\text{(iv)} & \min\limits_{t} \sup\limits_{A\in T} r(I-tA) = \sup\limits_{A\in T} r(I-p_0A)
\end{array} \qquad (28)$$

Here, the inequality (i) is evident, (ii) is the usual minimax estimation, the equality (iii) follows from (26), the equality (iv) is fulfilled by definition of the value p_0. The equality (9) follows from relations (28). The existence of the optimum p_0 is proved.

Calculate the value $Q(m,M,k,R^N)$. By definition of the value $Q(m,M,k,R^N)$, see (10), from the equality (iv) of the relations (28) and from the equality (25) we obtain:

$$Q(m,M,k,R^N) = \min_t F(t),$$

where $F(t) = \max\{|1-t(m+ik)|,|1-t(M+ik)|\}$.

We notice that at the point $t = 0$ and at the point $t = 2/(M+m)$ the following equality

$$|1 - t(m+ik)| = |1 - t(M+ik)|$$

holds. Outside the interval $[0,2/(M+m)]$ the pointwise maximum in the functions $|1 - t(m+ik)|$ and $|1 - t(M+ik)|$ increases monotonously when t moves away from the interval $[0,2/(M+m)]$. Hence $\min F(t)$ is attained at the point $t = p_0 \in [0,2/(M+m)]$. Now, since the following inequality

$$|1 - t(m+ik)| \geq |1 - t(M+ik)|$$

is fulfilled at any point $t \in [0,2/(M+m)]$, then

$$\min_t F(t) = \min_{0\leq t\leq 2/(M+m)} |1 - t(m+ik)|.$$

If the function $| 1 - t(m+ik) |$ attains minimum at some point lying outside the interval $[0,2/(M+m)]$ (this occurs, if $2k^2 < Mm - m^2$), then the minimum of $F(t)$ is attained at a boundary point of the interval $[0,2/(M+m)]$ and it equals to the minimal of the values

$$F[0] = 1 \quad \text{and} \quad F[2/(M+m)] = \frac{\{4k^2+(M-m)^2\}^{1/2}}{M+m}.$$

Because of $2k^2 < Mm - m^2$, the inequality $F[2/(M+m)] < < F[0]$ holds, then the parts of formulae (12), (13), which are corresponding to the case $2k^2 < Mm - m^2$, follow from here.

If the function $| 1 - t(m+ik) |$ attains minimum at some point lying inside the interval $[0,2/(M+m)]$ (this occurs, if $2k^2 \geq Mm - m^2$), then its minimum $k/(m^2+k^2)^{1/2}$ is attained at the point $t = p_0 = = m/(m^2+k^2)$ and it equals to the minimum of the function $F(t)$. Parts of the formulae (12), (13), which are corresponding to the case $2k^2 \geq Mm - m^2$, follow from here.

Theorem 2 is proved.

6. PROOF OF THEOREM 3

Since in R^N all norms are equivalent to one another, then for proving our theorem it is sufficient tc estaolish existence of such a norm $\|\cdot\|^*$ that

$$\|I-p_0A\|^* \leq Q(m,M,k,R^N), \tag{29}$$

in so far as from (29) under some $L \in (0,\infty)$ the following inequalities

$$\| x_n-x^* \| \leq L\| x_n-x^* \|^* \leq LQ^n(m,M,k,R^N)\| x_0-x^* \|^*$$

follow.

Let us denote $r_0 = r(I-p_0A)$ and observe two cases. One of them is in which $r_0 < Q(m,M,k,R^N)$, and the other is in which $r_0 = Q(m,M,k,R^N)$.

In the case when $r_0 < Q(m,M,k,R^N)$ we choose $\varepsilon > 0$ such that $r_0 + \varepsilon$ is less than $Q(m,M,k,R^N)$ and determine the norm $\|\cdot\|^*$ with the help of equality, cf. KRASNOSEL´SKIǏ et al. [8],

$$\|x\|^* = \sup_{n\geq 0} \frac{\|A_0^n x\|}{(r_0+\varepsilon)^n} \qquad (A_0=I-p_0A)$$

In this norm the inequality $\|I-p_0A\|^* \leq r_0 + \varepsilon$ will be fulfilled, and hence the inequality (29) will be fulfilled too.

Now, observe the case in which $r_0 = Q(m,M,k,R^N)$. We shall call an eigenvalue a semisimple eigenvalue, see e.g. KATO [9], if it has no adjoint vectors. The following lemma is evident.

Lemma 3. Let any eigenvalue of the matrix $B : R^N \longrightarrow R^N$, which is equal to its spectral radius

$r(B)$ by absolute value, be semisimple. Then there exists a norm $\|\cdot\|^*$ such that $\|B\|^* = r(B)$.

In virtue of Lemma 3, for proving the existence of a norm under requirement in the estimation (29) it is sufficient to show that all eigenvalues of the matrix $I-p_0A$, which are equal to $r(I-p_0A)$ by absolute value, are semisimple. For a matrix A from the class $S(m,M,k,R^N)$, under condition $r(I-p_0A) = Q(m,M,k,R^N)$, all above mentioned eigenvalues lie on the boundary of the rectangle $I-p_0\Pi$, where

$$\Pi = \{ z\in C^1 : m \leq \operatorname{Re} z \leq M, |\operatorname{Im} z| \leq k \}. \qquad (30)$$

Then it is suffucient for us to show that all eigenvalues of A, which are lying on the boundary of Π, are semisimple. It will be proved in Lemma 5. But now, we shall prove a statement complementing Lemma 3.

Lemma 4. Let be $\|B\| = r(B)$. Then any eigenvalue of the matrix B, which is equal to $r(B)$ by absolute value, is semisimple.

We shall prove Lemma 4 for complex matrices B because only some unessential details distinguish the real case from complex one.

Proof. Let λ be an eigenvalue of the matrix B, $|\lambda| = r(B)$. Let λ has an eigenvector z ($z \neq 0$) and adjoint vector w ($\|w\| = 1$). We denote by t_0 the greatest real non-negative t, for which $\| w + \lambda^{-1}tz \| \leq 1$. Then, by condition of lemma, we get $\| w + \lambda^{-1}(1+t_0)z \| = \| \lambda^{-1}B(w + \lambda^{-1}t_0 z) \| \leq 1$. That contradicts the definition of t_0. This contradiction shows that λ can not have any adjoint vectors. Lemma 4 is proved.

Lemma 5. *Let λ be a point lying on the boundary of the rectangular Π, see (30). If λ is an eigenvalue of a matrix A from the class $S(m,M,k,R^N)$, then λ is a semisimple eigenvalue.*

Proof. Denote by $A^c : C^N \longrightarrow C^N$ a complexification of the matrix A defined by the equality

$$A^c(x+iy) = Ax + iAy \qquad (x, y \in R^N).$$

We define a scalar product $\langle\cdot,\cdot\rangle$ in C^N by the equality

$$\langle x+iy, u+iv\rangle = (x,u)+(y,v)-i(x,v)+i(y,u) \qquad (x,y,u,v \in R^N).$$

Conditions of the lemma mean that at least one of the following relations is fulfilled

$$m(x,x) \leq (Ax,x), \qquad \mathrm{Re}\,\lambda = m.$$

$$(Ax,x) \leq M(x,x), \qquad \mathrm{Re}\,\lambda = M.$$

$$\| A-A^* \| \leq 2k, \qquad |\,\mathrm{Im}\,\lambda\,| = k.$$

Then a simple examination shows that each pair of relations written above may be rewritten as follows

$$\mathrm{Re}\,\langle \bar{c}Bz,z\rangle \leq |c|^2\langle z,z\rangle, \qquad \mathrm{Re}\,\lambda\bar{c} = |c|^2 \qquad (31)$$

if we put $c = m$, $B = -A^c$ in the first case, $c = M$, $B = A^c$ in the second case, $c = -ik\ \mathrm{sign}(\mathrm{Im}\ \lambda)$, $B = A^c$ in the third case.

Lemma will be proved if we can show that a semisimplness of the eigenvalue λ follows from the relations (31). For proving this, let us look at the matrix

$$D = B(2cI-B)^{-1}.$$

It is easy to examine that, in virtue of (31), $\|D\|_{C^N} \leq 1$ and that the eigenvalue $d = \lambda(2c-\lambda)^{-1}$ of the matrix D lies on the unit circle in the complex plane. By Lemma 4, a semisimplness of the eigenvalue d of the matrix D follows from here. Then the eigenvalue λ of the matrix A^c, and also of the matrix A, is semisimple. Lemma 3, and Theorem 3 with it, is proved.

Assertions, close to Lemmas 3-5, are used in the theory of numerical range of a linear maps, see e.g. GLAZMAN & LJUBICH [10], MARCUS & MINC [11].

7. PROOF OF THEOREM 4

Lemma 6. If K is the matrix defined by the equality (27), then

$$Q(m,M,k,R^N) = \max_{\|x\|=1} \left\| x - \frac{(Kx,x)}{(Kx,Kx)} Kx \right\|.$$

Proof. We denote

$$f(x) = \| x - \frac{(Kx,x)}{(Kx,Kx)} Kx \|^2 .$$

Evidently, if $\|x\| = 1$, then

$$f(x) = 1 - \frac{(Kx,x)^2}{(Kx,Kx)} .$$

We denote the components of a vector $x \in R^4$ by x_1, x_2, x_3, x_4 and set $x_1^2+x_2^2 = t$, $x_3^2+x_4^2 = s$. If $\|x\| = 1$, then $s = 1-t$ $(0 \leq t \leq 1)$, and the value f may be written in the form

$$f = 1 - \frac{\{mt + M(1-t)\}^2}{(m^2+k^2)t + (M^2+k^2)(1-t)} .$$

Derivative of the function f on t equals to

$$f' = (M-m)[mt+M(1-t)] \frac{2k^2+(M-m)M-(M^2-m^2)t}{\{(m^2+k^2)t+(M^2+k^2)(1-t)\}^2} .$$

As is easy to see, if $2k^2 \geq Mm - m^2$, then the function f' is positive over $0 \leq t \leq 1$. If $2k^2 <$ $< Mm - m^2$, then f' has exactly one zero.

$$t_0 = \frac{2k^2 + (M-m)M}{M^2 - m^2},$$

which is lying in the interval $[0,1]$. From here, with the help of simple computations we obtain that

$$\max_{\|x\|=1} f(x) = 1 - \frac{m^2}{m^2+k^2} = \frac{k^2}{m^2+k^2}$$

under $2k^2 \geq Mm - m^2$, and

$$\max_{\|x\|=1} f(x) = 1 - \frac{\{mt_o + M(1-t_o)\}^2}{(m^2+k^2)t_o+(M^2+k^2)(1-t_o)} = \frac{4k^2+(M-m)^2}{(M+m)^2}$$

under $2k^2 < Mm - m^2$.

Thus, by Theorem 2, $\max_{\|x\|=1} \{f(x)\}^{1/2} = Q(m,M,k,R^N)$. Lemma 6 is proved.

Now, denote by z such x ($\|x\|=1$), under which the function $f(x)$ attains maximum.

Since the assertion of theorem is evident in the case when $M = m$ or $k = 0$, then we pass following considerations under assumption that $M \neq m$, $k \neq 0$. In this case the matrix K, see (27), has no real eigenvalues. Hence the vector Kz is not parallel with the vector z.

Denote by H the two-dimensional subspace of the space R^4, which is linear hull of the vectors z, Kz. And denote by P an orthogonal projection operator onto H. Then

$$\|z - \frac{(Kz,z)}{(Kz,Kz)} Kz\| = \|z - \frac{(Bz,z)}{(Bz,Bz)} Bz\| \leq$$

$$\leq \max_{z\in H,\|z\|=1} \|z - \frac{(Bz,z)}{(Bz,Bz)} Bz\| = \max_{z\in H,\|z\|=1} \min_{t} \|(I-tB)z\| \leq$$

$$\leq \min_{t} \| I-tB \| , \qquad (32)$$

where $B : H \longrightarrow H$ is a restriction of the operator PK onto H.

We choose an orthogonal basis g, h in H consisting of an eigenvectors of the symmetric operator $(B+B^*)/2$. In this basis the operator B has the following matrix

$$\left\| \begin{matrix} m^* & k^* \\ -k^* & M^* \end{matrix} \right\| ,$$

where

$$m^* = (Bg,g), \; M^* = (Bh,h), \; k^* = (Bg,h) = -(Bh,g). \qquad (33)$$

Without loss in generality one can suppose that $m^* \leq$ $\leq M^*$, $k^* \geq 0$. Then, by definition of the operator B and in virtue of Lemma 1, the following inequalities

$$m^* \geq m, \quad M^* \leq M, \quad k^* \leq k \qquad (34)$$

are valid.

For completing the proof of theorem it is sufficient for us to establish that at any rate one of the inequalities (34) is strict. Indeed, in virtue of (23), the inequality

$$\min_t \| I-tB \| \le \frac{M^{*2} - m^{*2} + 4k^*(k^{*2} + M^*m^*)^{1/2}}{(M^* + m^*)^2 + 4k^{*2}}$$

holds. The right-hand side of this inequality is increasing on M^* and k^* and is decreasing on m^*. Hence, if one of the inequalities (34) is strict, then the strict inequality

$$\min_t \| I-tB \| < \frac{M^2 - m^2 + 4k(k^2 + Mm)^{1/2}}{(M+m)^2 + 4k^2}$$

is fulfilled. In virtue of Lemma 6 and inequality (32) the assertion of the theorem follows from here.

For proving that at least one of the inequalities (34) is strict, suppose contrary, i.e.,

$$m^* = m, \quad M^* = M, \quad k^* = k,$$

and write the vectors g and h in a coordinate form

$$g = \{g_1,g_2,g_3,g_4\}, \qquad h = \{h_1,h_2,h_3,h_4\}.$$

Now, notice that $(Bg,g) = (Kg,g)$; therefore the first of the inequalities (34) can be given in the form

$$m(g_1^2+g_2^2) + M(g_3^2+g_4^2) \ge m$$

or, that is the same, in the form

$$m(g_1^2+g_2^2) + M(g_3^2+g_4^2) \ge m(g_1^2+g_2^2+g_3^2+g_4^2).$$

Since $M > m$ by assumption, then $g_3 = g_4 = 0$.

Analogously, from the second of the inequalities (34) we have $h_1 = h_2 = 0$. Then, as is easy to see, from the third of the equalities (33) we obtain $k^* = (Bg,h) = (Kg,h) = 0$. Hence $k^* < k$ because of the assumption $0 < k$. We have come to the contradiction with the assumption that $k^* = k$.

So, we establish that at any rate one of the inequalities (34) is strict. Theorem 4 is proved.

ACKNOWLEDGEMENTS

We are grateful to I.V.Emelin for a number of useful discussions. We also thank L.M.Mogilevkina, who helped us willingly, whenever we had linguistical problems.

REFERENCES

1. M.A.Krasnosel'skiĭ, and S.G.Krein, An iteration process with minimal residues. Mat.Sb. 31, 4 (1952), 315-334 (in Russian).

2. Ju.A.Kuznetsov, A contribution to the theory of iteration processes. Dokl.Akad.Nauk SSSR 184, 2 (1969), 274-277 (in Russian).

3. G.I.Marchuk, and Ju.A.Kuznetsov, On solution of linear equations by iteration techniques. In: Problems of an accuracy and efficiency of computation algorithms, Proc. Symposium vol.1, Kiev 1969, 60-74 (in Russian).

4. G.I.Marchuk, and Ju.A.Kuznetsov, Iteration methods and quadratic functionals. Nauka, Novosibirsk 1972 (in Russian).

5. A.A.Samarskiĭ, Two-layer iteration schemes for non-selfadjoint equations. Dokl.Akad.Nauk SSSR 186, 1 (1969), 35-38 (in Russian).

6. G.E.Forsythe, On the asymptotic directions of the s-dimensional optimum gradient method. Numer. Math 11, 1 (1968), 57-76.

7. I.V.Emelin, On the rate of convergence of the method of steepest descend. Uspehi Mat. Nauk 32, 1 (1977), 163-164 (in Russian).

8. M.A.Krasnosel′skiĭ et al., Approximate solution of operator equations. Nauka, Moscow 1969 (in Russian).

9. T.Kato, Pertrubation theory for linear operators. Springer-Verlag, Berlin-Heidelberg-New-York 1966.

10. I.M.Glazman, and Ju.I.Ljubich, Finite-dimensional linear spaces in problems. Nauka, Moscow 1969 (in Russian).

11. M.Marcus, and H.Minc, A survey of matrix theory and matrix inequalities, Allyn and Bacon, Inc., Boston 1964.

Received: September 1981

Accepted: January 1982