Atti Accademia Nazionale dei Lincei

Classe Scienze Fisiche Matematiche Naturali

Rendiconti

Solomon G. Mikhlin

Some theorems on the stability of numerical processes

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Serie 8, Vol. **72** (1982), n.2, p. 71–76.

Accademia Nazionale dei Lincei

<http://www.bdim.eu/item?id=RLINA_1982_8_72_2_71_0>

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Accademia Nazionale dei Lincei, 1982.

Analisi numerica. — *Some theorems on the stability of numerical processes* (*). Nota (**) del Socio straniero SOLOMON G. MIKHLIN.

RIASSUNTO. — Nell'articolo si dimostrano alcuni teoremi sulla stabilità dei processi numerici di Ritz e della collocazione in rapporto agli errori di « distorsione ».

1. Let us consider a numerical process which consists in solving a sequence of independent equations

$$A_n x^{(n)} = f^{(n)} ; \qquad n = 1, 2, \cdots \tag{1}$$

Here $x^{(n)} \in X_n , f^{(n)} \in Y_n$; A_n is an operator acting from X_n into Y_n ; X_n , Y_n are metric spaces. In this paper we only consider the case when X_n , Y_n are separable Banach spaces (in the sections 2-4—Hilbert spaces) and A_n are linear operators. Processes (1) arise, for example, when one uses the Ritz method (particularly, the finite elements method) for solving linear equations. In these cases the operators A_n and the right-hand terms $f^{(n)}$ are not given a priori. How it is natural, they are calculated with some errors. As a result we have to solve equations of a certain "distorted" sequence

$$(A_n + \Gamma_n) z^{(n)} = f^{(n)} + \delta^{(n)} \tag{2}$$

instead of sequence (1).

We say that the process (1) is stable, with respect to the distortions errors, in the sequence of pairs of spaces (X_n , Y_n) if there exist positive numbers p , q , r, such that the inequality $\| \Gamma_n \|_{X_n \to Y_n} \leq r$ involves the estimate

$$\| z^{(n)} - x^{(n)} \|_{X_n} \leq p \| \Gamma_n \|_{X_n \to Y_n} + q \| \delta^{(n)} \|_{Y_n} . \tag{3}$$

Some other definitions of stability are also possible.

It is demonstrated in [1] that the process (1) is stable, according to the above definition if and only if the conditions

$$\| A_n^{-1} \|_{Y_n \to X_n} \leq c_1 , \| A_n^{-1} B_n x^{(n)} \|_{X_n} \leq c_2 \tag{4}$$

are fulfilled; here c_1 , c_2 do not depend on n , $x^{(n)}$ is the solution of (1) and B_n is an arbitrary operator with unit norm, acting from X_n into Y_n.

2. Let us consider the equation

$$Ax = f , \tag{5}$$

(*) Dedicated to Prof. G. Fichera on the occasion of his 60th birthday.
(**) Presentata nella seduta del 13 febbraio 1982.

where A is a positive definite [2] operator acting in a separable Hilbert space H; we designate by H_A the energy space of the operator A, for the definition see [2]. We choose a sequence of finite-dimensional subspaces $H_A^{(n)} \subset H_A$; let this sequence be complete in H_A. We put $\dim H_A^{(n)} = N(n) = N$. Further let $(\varphi_{n1}, \varphi_{n2}, \cdots, \varphi_{nN})$ be a basis in $H_A^{(n)}$. Following the Ritz method one constructs the approximate solution $x^{(n)}$ of (5) as an element of $H_A^{(n)}$

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}$$

with coefficients $a_n^{(k)}$ satisfying the system of equations

$$M_n a^{(n)} = f^{(n)}. \tag{6}$$

Here M_n is the matrix of elements $[\varphi_{nk}, \varphi_{nj}]$; $a^{(n)}$ and $f^{(n)}$ are vectors in R_N with components $(a_1^{(n)}, a_2^{(n)}, \cdots, a_N^{(n)})$ and $(f, \varphi_{n1}), (f, \varphi_{n2}), \cdots, (f, \varphi_{nN})$ respectively. The indices j, k change in the limits $1 \leq j, k \leq N$; the square and round brackets designate the inner product in H_A and H respectively.

Remark. We obtain the classical Ritz method, if $\forall n, H_A^{(n)} \subset H_A^{(n+1)}$ [3]. The idea of using subspaces $H_A^{(n)} \notin H_A^{(n+1)}$ is due to Courant [4]; this idea contains the basis of the finite elements method.

Let A_n be the operator acting in R_N and generated by the matrix M_n. If $a^{(n)}$ and $f^{(n)}$ are treated as elements of R_N, then one can write the equation (6) in the form

$$A_n a^{(n)} = f^{(n)}. \tag{7}$$

It is demonstrated in [5] (see also [6]) that the numerical process (7) for the classical Ritz process is stable in the sequence (R_N, R_N) if and only if the least eigenvalue $\lambda_1^{(n)}$ of the matrix M_n is bounded below by a positive constant. The proof can be transferred without change on the case of non-expanding subspaces H_A.

3. We investigate now the stability of the Ritz process in the general case $\inf \lambda_1^{(n)} \geq 0$. We introduce two N-dimensional Hilbert spaces X_N and Y_N with the norms

$$\forall b \in R_N ; \| b \|_{X_N} = \sqrt{\lambda_1^{(n)}} \, \| b \|_{R_N} , \| b \|_{X_N} = \frac{1}{\sqrt{\lambda_1^{(n)}}} \| b \|_{R_N} . \tag{8}$$

Let us designate here by A_n the operator generated by the matrix M_n and acting from X_N into Y_N; the vectors $a^{(n)}$ and $f^{(n)}$ are treated as elements of X_N and Y_N respectively.

THEOREM 1. *The process* (7) *is stable in the sequence* (X_N, Y_N).

It is sufficient to prove that the inequalities (4) are satisfied.

Let $v^{(n)} \in H_A^{(n)}$, then

$$v^{(n)} = \sum_{k=1}^{N} b_k^{(n)} \varphi_{nk} ; \tag{9}$$

if we put $b^{(n)} = (b_1^{(n)}, b_2^{(n)}, \cdots, b_N^{(n)})$, we obtain

$$|\!|\!| v^{(n)} |\!|\!|^2 = (M_n b^{(n)}, b^{(n)})_{R_N} \geq \lambda_1^{(n)} \| b^{(n)} \|_{R_N}^2 = \| b^{(n)} \|_{X_N}^2 . \tag{10}$$

$|\!|\!| \cdot |\!|\!|$ designates the norm in H_A. Now

$$\| A_n^{-1} \|_{Y_n \to X_n} = \sup_{b \in R_N} \frac{\| A_n^{-1} b \|_{X_N}}{\| b \|_{Y_N}} = \lambda_1^{(n)} \sup_{b \in R_N} \frac{\| M_n^{-1} b \|_{R_N}}{\| b \|_{R_N}} = 1 ; \tag{11}$$

hence the first inequality (4) is proved.

The Ritz method converges in H_A for the equation (5) because A is positive definite [2]. Consequently, $|\!|\!| x^{(n)} |\!|\!| \leq c_3 = \text{const}$; according to (10), $\| a^{(n)} \|_{X_N} \leq c_3$. Now $\| A_n^{-1} B_n a^{(n)} \| \leq c_3$, and the second inequality (4) is also proved.

4. Formula (9) defines an operator Π_n which transforms any vector $b^{(n)} \in R_N$ in an element $v^{(n)} \in H_A^{(n)}$, so that $v^{(n)} = \Pi_n a^{(n)}$. The operator Π_n is invertible: $b^{(n)} = \Pi_n^{-1} v^{(n)}$; particularly, $a^{(n)} = \Pi_n^{-1} x^{(n)}$. Substituting this in (7), we obtain the numerical process giving the approximate solution $x^{(n)}$:

$$A_n \Pi_n^{-1} x^{(n)} = f^{(n)} . \tag{12}$$

THEOREM 2. *The numerical process* (12) *is stable in the sequence* $(H_A^{(n)}, Y_N)$.

We use the method of [7] in order to prove Theorem 2.

Let Γ_n and $\delta^{(n)}$ be the distortions of A_n and $f^{(n)}$ respectively, and let $c^{(n)}$ be the solution of the distorted equation

$$(A_n + \Gamma_n) c^{(n)} = f^{(n)} + \delta^{(n)} . \tag{13}$$

The distorted approximate Ritz solution is $z^{(n)} = \Pi_n c^{(n)}$, and

$$|\!|\!| z^{(n)} - x^{(n)} |\!|\!|^2 = (M_n (c^{(n)} - a^{(n)}), c^{(n)} - a^{(n)})_{R_N} \leq$$
$$\leq \| A_n (c^{(n)} - a^{(n)}) \|_{Y_N} \cdot \| c^{(n)} - a^{(n)} \|_{X_N} .$$

According to Theorem 1 there exist numbers $p, q, r > 0$ with the following property: if $\| \Gamma_n \|_{X_N \to Y_N} \leq r$, then

$$\| c^{(n)} - a^{(n)} \|_{X_N} \leq p \| \Gamma_n \|_{X_N \to Y_N} + q \| \delta^{(n)} \|_{Y_N} .$$

It follows from (7) and (13) that

$$(A_n + \Gamma_n)(c^{(n)} - a^{(n)}) = (I_n + \Gamma_n A_n^{-1}) A_n (c^{(n)} - a^{(n)}) = \delta^{(n)} - \Gamma_n a^{(n)},$$

where I_n is the identical operator in Y_n. Let r' be a number in the interval $(0, 1)$, and let $\| A_n^{-1} \| \cdot \| \Gamma_n \|_{X_N \to Y_N} \leq r'$. Then

$$\| (I_n + \Gamma_n A_n^{-1})^{-1} \| \leq (1 - r')^{-1}$$

and

$$\| A_n (c^{(n)} - a^{(n)}) \| \leq \frac{1}{1 - r'} [c_3 \| \Gamma_n \|_{X_N \to Y_N} + \| \delta^{(n)} \|_{Y_N}] .$$

Now obviously

$$\| z^{(n)} - x^{(n)} \| \leq p' \| \Gamma_n \|_{X_N \to Y_N} + q' \| \delta^{(n)} \|_{Y_N} , \tag{14}$$

where p', q' are suitable constants. Theorem 2 is proved.

Remark. One can define the norms in X_N, Y_N as follows:

$$\forall b \in R_N ; \| b \|_{X_N} = \gamma(n) \| b \|_{R_N} , \| b \|_{Y_N} = \frac{1}{\gamma(n)} \| b \|_{R_N} . \tag{15}$$

Here $\gamma(n)$ is any positive function of n, satisfying the inequality

$$\forall b \in R_N , \| \Pi_n b \| \geq C\gamma(n) \| b \|_{R_N} ; C = \text{const} ,$$

or, what is the same,

$$\lambda_1^{(n)} \geq C\gamma^2(n) . \tag{16}$$

In particular, it is sufficient that $\mu_1^{(n)} \geq C\gamma^2(n)$, where $\mu_1^{(n)}$ is the least eigenvalue of the matrix of inner products $(\varphi_{nk}, \varphi_{nj})_H$; $j, k = 1, 2, \cdots, N$. Theorems 1 and 2 with their proofs still hold, only the relation (11) must be replaced by the inequality $\| A_n^{-1} \|_{Y_N \to X_N} \leq C^{-1}$, where C is the constant of (16).

The theorems on stability of the finite elements method given in [8] are particular cases of the Theorems 1 and 2. The function $\gamma(n)$ used in [8] is equal to $h^{m/2}$, where h is the step of the net and m is the dimension of the space of coordinates.

5. We consider now the problem of stability of the collocation method; this method was first formulated in [9]. The main points of the collocation method are the following. Let be given the problem of solving the equation (5) where A is a linear operator acting from a Banach space X into a Banach space Y, so that the domain D (A) and the range R (A) are dense in X and Y respectively. We suppose that Y consists only of functions which are continouus on a certain compact $K \subset R_m$. We choose a sequence of finite-dimensional

subspaces $X_n \subset D(A)$ which is complete in X and put $\dim X_n = N(n) = N$, $Y_n = AX_n$. If the inverse operator A^{-1} exists then $\dim Y_n = N$. Further, if $\{\varphi_{nk}\}$, $1 \leq k \leq N$, is a basis in X_n and $\psi_{nk} = A\varphi_{nk}$, then $\{\psi_{nk}\}$ is a basis in Y_n.

Let us choose some points $t_k^{(n)} \in K$, $1 \leq k \leq N$, the so-called collocation knots. One constructs the approximate solution of (5) as an element of X_n:

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk} ; \tag{17}$$

the coefficients $a_k^{(n)}$ are to be defined from the algebraic system

$$\sum_{k=1}^{N} a_k^{(n)} \psi_{kn}(t_j^{(n)}) = f(t_j^{(n)}) ; \qquad 1 \leq j \leq N. \tag{18}$$

6. Let $t_j^{(n)}$ be the vertices of a certain parallelepipedal net. Further let $h_k^{(n)}$ be the length of the edge of the parallelepiped which is parallel to the k-th coordinate axis. Suppose

$$c_1 h_k^{(n)} \leq h_n \leq c_2 h_k^{(n)} \quad ; \quad c_1, c_2 = \text{const}, \ h_n \xrightarrow[n\to\infty]{} 0 .$$

One can write down (18) in the form (6); the meaning of the notations is obvious. We consider $a^{(n)}$ as an element of R_N and $f^{(n)}$ as an element of the N-dimensional Hilbert space F_{Ns} with the norm $\| \cdot \| F_{Ns} = h^{s/2} \| \cdot \|_{R_N}$.

Let A_n be the operator generated by the matrix M_n and acting from R_N into F_{Ns}. One can write the system (18) in the form (7).

Let $s_n^{(1)}$ designate the least singular number of the matrix M_n, i.e., the least eigenvalue of the non-negative matrix $M_n^* M_n$.

THEOREM 3. *If* $\| a^{(n)} \| \leq c_3 = \text{const}$ *and* $s_1^{(n)} \geq c_4 h_n^{-s}$, *where* $c_3, c_4 = \text{const} > 0$, *then the process* (7) *for the collocation method is stable in the sequence* (R_N, F_{Ns}). *If* $s_1^{(n)} \leq \gamma(n) h_n^{-s}$, $\gamma(n) \xrightarrow[n\to\infty]{} 0$, *then the same process in unstable.*

Any numerical process of the kind (1) is stable if and only if the conditions (5) are fulfilled. It is easy to see that $\| A_n^{-1} \| = h_n^{-s/2} \| M_n^{-1} \|_{R_N \to R_N}$. The greatest singular number of M_n^{-1} is equal to $1/s_1^{(n)}$, hence $\| M_n^{-1} \|_{R_N \to R_N} = 1/\sqrt{s_1^{(n)}}$; consequently $\| A_n^{-1} \| = (h^s s_1^{(n)})^{-1/2}$. If $s_1^{(n)} \leq \gamma(n) h_n^{-s}$ then $\| A_n^{-1} \| \geq 1/\sqrt{\gamma(n)} \xrightarrow[n\to\infty]{} \infty$, and the process (7) is unstable. On the contrary, if $s_1^{(n)} \geq c_4 h_n^{-s}$ then $\| A_n^{-1} \| \leq 1/\sqrt{c_4}$ and the first condition (5) is satisfied. The second condition (5) is satisfied by assumption, and the collocation process is stable.

References

[1] Mikhlin S. G. (1964) - *On the stability of some numerical processes.* « Dokl. Akad. Nauk SSSR », *157*, N. 2, 271-273. English translation: « Soviet Math. Dokl. », *5*, 931-933.

[2] Mikhlin S. G. (1957-1964) - *Variational methods in mathematical Physics.* Moscow. English translation: Pergamon Press, Oxford.

[3] Ritz W. (1908) - *Über eine neue Methode zur Lösung gewisser Variations-probleme der mathmatischen Physik.* Journ. f. d. reine und angewandte Mathematik », *135*, H. 1.

[4] Courant R. (1943) - *Variational methods for the solution of problems of equilibrium and vibrations.* « Bull. Amer. Math. Soc. », *49*, 1-23.

[5] Mikhlin S. G. (1960-1961) - *On the stability of the method of Ritz.* « Dokl. Akad. Nauk SSSR », *135*, 16-19. English translation: « Soviet Math. Dokl. », *1*, 1230-1233.

[6] Mikhlin S. G. (1971) - *The numerical performance of variational methods*, Moscow, 1966. English translation: Wolters-Noordhoff Publ., Groningen.

[7] Yaskova G. N., Yakovlev M. N. (1962) - *Some condition for the stability of the method of Petrov - Galerkin.* « Trudy Matem. in-ta im. ». Steklova, *66*, 182-189.

[8] Mikhlin S. G. (1979) - *Approximation on a rectangular grid.* Sijthoff & Noordhoff, Alphen aan den Rijn.

[9] Kantorovich L. V. (1934) - *On a method for approximate solution of differential equations.* « Dokl. Akad. Nauk SSSR », *2*, 532-536.