ZESZYTY NAUKOWE UNIWERSYTETU JAGIELLOŃSKIEGO
CCCCLXXV PRACE MATEMATYCZNE, ZESZYT 19 1977

More general sufficient conditions for the convergence of the Bubnov-Galerkin method

by J. Bochenek

Introduction. In 1915 B. G. Galerkin published in paper [2] an approximative method of solution of certain equations. This method (we shall say the Bubnov-Galerkin method cf. [5]) had a wide application, but for some time had no basis. Many authors spoke of the convergence of the Bubnov-Galerkin method, but most of the general sufficient conditions were proposed by S. G. Michlin in 1948 and 1950 (cf. [3] and [4]).

The purpose of this paper is to give more general sufficient conditions for the convergence of this method.

§ 1. The Bubnov-Galerkin method. Let H be a Hilbert separable space and let A be a linear operator with domain $D(A) \subset H$. We shall consider the following equation

$$Au = f, \quad f \in H. \tag{1}$$

In order to solve the equation (1), we choose a sequence $\{\varphi_n\}$ such that $\varphi_n \in D(A)$ and the sequence $\{\varphi_n\}$ is a complete system in H. Let us denote by

$$u_n = \sum_{k=1}^{n} a_k \varphi_k \tag{2}$$

the n-th approximative of the solution of the equation (1) in the sense of Bubnov-Galerkin. The coefficients a_k $(k = 1, \ldots, n)$ are defined as solution of equations

$$(Au_n - f, \varphi_j) = 0 \quad (j = 1, \ldots, n). \tag{3}$$

The equations (3) may be written in the form

$$\sum_{k=1}^{n} a_k (A\varphi_k, \varphi_j) = (f, \varphi_j) \quad j = 1, \ldots, n. \tag{4}$$

The equations (4) are called the Bubnov-Galerkin system.

§ 2. A special case of the Bubnov-Galerkin method. Let the operator A in the equation (1) have the form

$$A = E + T, \tag{5}$$

where E is the identity operator, but T is a linear operator defined in H such that

$$\|T\| \leqslant a < 1 . \tag{6}$$

We shall prove the following

THEOREM 1. *If the operator A in the equation* (1) *satisfies conditions* (5) *and* (6) *then the Bubnov-Galerkin method for the equation* (1) *is convergent for any sequence $\{\varphi_n\}$ of choice satisfying the conditions laid down in* § 1 *and then*

$$\|u_0 - u_n\| \leqslant \frac{1}{1-a} \|u_0 - P_n u_0\| , \tag{7}$$

where u_0 denotes a solution of the equation (1) *), *but P_n is the projector of H into the close linear subspace L_n spanned by*

$$\{\varphi_1, \ldots, \varphi_n\} .$$

Proof. Under the assumptions of Theorem 1 the system (4) take the form

$$\sum_{k=1}^{n} a_k(\varphi_k + T\varphi_k, \varphi_j) = (f, \varphi_j) \qquad j = 1, \ldots, n , \tag{8}$$

or the other form

$$u_n + P_n T u_n = P_n f . \tag{9}$$

On the other hand by (1) we have

$$P_n u_0 + P_n T u_0 = P_n f . \tag{10}$$

From (9) and (10) we have

$$u_n - P_n u_0 = P_n T(u_0 - u_n) \tag{11}$$

From this follows

$$\|u_n - P_n u_0\| \leqslant a\|u_0 - u_n\| , \qquad (\|P_n\| = 1) \tag{12}$$

Let us observe that

$$\|u_0 - u_n\| = \|(u_0 - P_n u_0) + (P_n u_0 - u_n)\| \leqslant \|u_0 - P_n u_0\| + \\ + \|u_n - P_n u_0\| \leqslant \|u_0 - P_n u_0\| + a\|u_0 - u_n\| ,$$

From this follows inequality (7). Since $\{\varphi_n\}$ is a complete system in H, therefore we have the thesis of Theorem 1.

Remark 1. The inequality (7) gives the rate of convergence of sequence $\{u_n\}$ to u_0. From this it follows that the rate of convergence is as the rate con-

*) From (5) and (6) it follows that equation (1) has the unique solution u_0.

vergence $\{P_n u_0\}$ to u_0. As we know, the rate of convergence $\{P_n u_0\}$ to u_0 is dependent on the choice of the sequence $\{\varphi_n\}$.

§ 3. The sufficient conditions for the convergence of the Bubnov-Galerkin method. Let B be a linear self-adjoint positive definite operator such that $D(B)$ is dense in H. In $D(B)$ we define the scalar product $[\cdot, \cdot]$ and the norm $\|\cdot\|_0$ by the formulas $[u, v] = (Bu, v)$ and $\|u\|_0 = [u, u]^{\frac{1}{2}}$, $u, v \in D(B)$.

In the sequel we define the Hilbert space H_0 as the close the domain $D(B)$ in the norm $\|\cdot\|_0$. Let A be a linear operator defined in $D(A)$ dense in H.

First we shall prove the following two lemmas

LEMMA 1. *If G is a linear self-adjoint positive definite operator and C is a bounded linear operator in H such that $C(D(G)) \subset D(G)$, then GCG^{-1} is the bounded operator in H.*

Proof. Let us observe that $D(GCG^{-1}) = H$. Indeed, $D(G^{-1}) = R(G) = H$, because the operator G is the self-adjoint and positive definite in H. Let f be any element of H; then $GCG^{-1}f$ is defined. Now we shall prove that the operator GCG^{-1} is close. Let $f_n \to f$ and $GCG^{-1}f_n \to g$. Let us denote $h_n = CG^{-1}f_n$. Since CG^{-1} is the bounded operator, therefore $CG^{-1}f_n \to CG^{-1}f$, so $h_n \to h$. On the other hand $Gh_n \to g$. Since the operator G is a close operator, therefore $h \in D(G)$ and $Gh = g$, so GCG^{-1} is a close operator. Thus GCG^{-1} is defined in the whole Hilbert space H and the close operator, therefore as is known GCG^{-1} is a bounded operator (cf. [6], p. 427).

LEMMA 2. *If $B^{-1}A$ is a bounded operator in H, then $B^{-1}A$ is a bounded operator in H_0.*

Proof. Let us observe first that if $B^{-1}A$ is a bounded operator in H, then $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ is a bounded operator in H. Indeed, $B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = B^{\frac{1}{2}}B^{-1}AB^{-\frac{1}{2}} = GCG^{-1}$ where $G = B^{\frac{1}{2}}$, $C = B^{-1}A$, therefore by Lemma 1 $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ is a bounded operator in H.

Let u be any element in H_0. Since $H_0 \subset H$, so $u \in H$ and we have

$$\|B^{-1}Au\|_0^2 = [B^{-1}Au, B^{-1}Au] = (Au, B^{-1}Au) = (B^{-\frac{1}{2}}Au, B^{-\frac{1}{2}}Au)$$

$$= (B^{-\frac{1}{2}}AB^{-\frac{1}{2}}B^{\frac{1}{2}}u, B^{-\frac{1}{2}}AB^{-\frac{1}{2}}B^{\frac{1}{2}}u) \leqslant \|B^{-\frac{1}{2}}AB^{-\frac{1}{2}}\|^2\|B^{\frac{1}{2}}u\|^2 = \|B^{-\frac{1}{2}}AB^{-\frac{1}{2}}\|^2\|u\|_0^2 .$$

From this follows the thesis of Lemma 2 and the inequality

$$\|B^{-1}A\|_0 \leqslant \|B^{-\frac{1}{2}}AB^{-\frac{1}{2}}\| . \tag{13}$$

Now we shall consider the equation (1). We now assume additionally that $D(A) \subset D(A^*)$. Since $D(A^*)$ is a dense subspace in H, the operator A may expand to a close operator in H (cf. [6], p. 557). In the sequel by $\bar{A}$ we shall denote the smallest close expanding of A.

THEOREM 2. *If 1° $D(\bar{A}) = D(B)$ where A and B are the operators fulfilling the above mentioned conditions, 2° $(Bu, u) \leqslant c|(Au, u)|$, $u \in D(A)$, $c > 0$, 3° $\{\varphi_n\}$ is a complete system in H_0, $\varphi_n \in D(\bar{A})$, 4° equation $\bar{A}u = f$ has the unique solution u_0, then the sequence $\{u_n\}$ defined by (2) is convergent in the space H_0.*

Proof. The coefficients a_k, $k = 1, \ldots, n$ in formula (2) satisfy system (4). System (4) may be written in the form

$$\sum_{k=1}^{n} a_k[\psi_k, \varphi_j] = [g, \varphi_j] \quad j = 1, \ldots, n, \tag{14}$$

or in the other form

$$P_n z_n = P_n g, \tag{15}$$

where $z_n = \sum_{k=1}^{n} a_k B^{-1}\bar{A}\varphi_k = \sum_{k=1}^{n} a_k \psi_k$; $\psi_k = B^{-1}\bar{A}\varphi_k$, $g = B^{-1}f$.

Let $v_n \in M_n$ (M_n be a close linear subspace spanned by $\{\psi_1, \ldots, \psi_n\}$), therefore $v_n = \sum_{k=1}^{n} \beta_k \psi_k$ and $P_n v_n = \sum_{k=1}^{n} \alpha_k \varphi_k$ where $\alpha_1, \ldots, \alpha_n$ are defined by the condition

$$\|v_n - P_n v_n\|_0 = \min. \tag{16}$$

From (16) follows that the coefficients $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ satisfy the following system

$$\sum_{k=1}^{n} \beta_k[\psi_k, \varphi_i] = \alpha_i \quad i = 1, \ldots, n. \tag{17}$$

Let us observe that

$$\|v_n\|_0^2 = \|\sum_{k=1}^{n} \beta_k \psi_k\|_0^2 = \|B^{-1}\bar{A} \sum_{k=1}^{n} \beta_k \varphi_k\|_0^2 \leqslant \|B^{-\frac{1}{2}}\bar{A}B^{-\frac{1}{2}}\|^2 \|\sum_{k=1}^{n} \beta_k \varphi_k\|^2 = \tag{18}$$

$$= \|B^{-\frac{1}{2}}\bar{A}B^{-\frac{1}{2}}\|^2 \sum_{i=1}^{n} \beta_i^2.$$

On the other hand

$$\sum_{i=1}^{n} \beta_i^2 = \Big(B \sum_{k=1}^{n} \beta_k \varphi_k, \sum_{k=1}^{n} \beta_k \varphi_k\Big) \leqslant c \Big|\Big(\bar{A} \sum_{k=1}^{n} \beta_k \varphi_k, \sum_{k=1}^{n} \beta_k \varphi_k\Big)\Big| =$$

$$= c\Big|\Big[\sum_{k=1}^{n} \beta_k B^{-1}\bar{A}\varphi_k, \sum_{i=1}^{n} \beta_i \varphi_i\Big]\Big| \leqslant c \sum_{i=1}^{n} |\beta_i| \sum_{k=1}^{n} |\beta_k[\psi_k, \varphi_i]| =$$

$$= c \sum_{i=1}^{n} |\alpha_i \beta_i| \leqslant c\Big(\sum_{i=1}^{n} \alpha_i^2\Big)^{\frac{1}{2}}\Big(\sum_{i=1}^{n} \beta_i^2\Big)^{\frac{1}{2}}.$$

From this we obtain

$$\Big(\sum_{i=1}^{n} \beta_i^2\Big)^{\frac{1}{2}} \leqslant c\Big(\sum_{i=1}^{n} \alpha_i^2\Big)^{\frac{1}{2}}. \tag{19}$$

By (18) and (19) we get

(20) $$\|v_n\|_0 \leqslant c_1 \|P_n v_n\|_0 \quad \text{where } c_1 = c\|B^{-\frac{1}{2}}\bar{A}B^{-\frac{1}{2}}\| .$$

Since $z_n - \pi_n g \in M_n$ *) in virtue of (20) we have

(21) $$\|z_n - \pi_n g\|_0 \leqslant c_1 \|P_n(z_n - \pi_n g)\|_0 = c_1 \|P_n(g - \pi_n g)\|_0$$
$$\leqslant c_1 \|g - \pi_n g\|_0 .$$

From (21) by the completeness of $\{\varphi_n\}$ in H_0 it follows that

(22) $$\|z_n - \pi_n g\|_0 \to 0 \quad \text{when } n \to \infty .$$

Since $\|g - z_n\|_0 \leqslant \|g - \pi_n g\|_0 + \|z_n - \pi_n g\|_0$, so

(23) $$\|z_n - g\|_0 \to 0 \quad \text{when } n \to \infty .$$

By definition of the norm in H_0, sequence $\{z_n\}$ and g we have

$$\|z_n - g\|_0^2 = [B^{-1}\bar{A}u_n - B^{-1}f, B^{-1}\bar{A}u_n - B^{-1}f]$$
$$= (\bar{A}u_n - f, B^{-1}(\bar{A}u_n - f)) = \|B^{-\frac{1}{2}}(\bar{A}u_n - f)\|^2 =$$
$$= \|B^{-\frac{1}{2}}(\bar{A}u_n - \bar{A}u_0)\|^2 = \|B^{-\frac{1}{2}}\bar{A}(u_n - u_0)\|^2.$$

From the last inequality by (23) we get

(24) $$B^{-\frac{1}{2}}\bar{A}(u_n - u_0) \to 0 \quad \text{when } n \to \infty .$$

Since $B^{\frac{1}{2}}\bar{A}^{-1}B^{\frac{1}{2}}$ is a bounded operator, therefore

$$B^{\frac{1}{2}}\bar{A}^{-1}B^{\frac{1}{2}}\{B^{-\frac{1}{2}}\bar{A}(u_n - u_0)\} \to 0 \text{ **)}$$

or

$$B^{\frac{1}{2}}(u_n - u_0) \to 0$$

that is

$$\|u_n - u_0\|_0 \to 0 .$$

The proof of Theorem 2 is completed.

Remark 2. The result of Theorem 2 may be carried into an equation of the form

(25) $$Au + Ku = f ,$$

where A is a operator satisfying all the previous conditions and K is a linear operator such that $D(A) \subset D(K) \subset D(K^*)$, and $A^{-1}K$ and KA^{-1} are completely continuous operators in H. In the case of the equation (25) the reasoning is analogous to that in the paper [1], § 3.

---

*) π_n is the projector of H_0 into M_n.

**) Since $\bar{A}^{-1}B$ is a bounded operator; it follows from Lemma 1 that $B^{\frac{1}{2}}\bar{A}^{-1}B^{\frac{1}{2}}$ is a bounded operator.

References

[1] J. Bochenek, *Some remarks on the convergence of Ritz and of Galerkin methods*. Prace Matematyczne 17 (1973), 17-27.
[2] Б. Г. Галеркин, *Стержни и пластины*, Вестник инженеров, 19 (1915), 897—908.
[3] С. Г. Михлин, *О сходимости метода Галеркина*, ДАН СССР, 62 №2 (1948).
[4] —, *Прямые методы в математической физике*, Москва 1950.
[5] —, *Численная реализация вариационных методов*, Москва 1966.
[6] В. И. Смирнов, *Курс высшей математики*, Т. 5, Москва 1959.