ANNALES SOCIETATIS MATHEMATICAE POLONAE
Series I: COMMENTATIONES MATHEMATICAE XVII (1973)
ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria I: PRACE MATEMATYCZNE XVII (1973)

JAN BOCHENEK (Kraków)

On the reasonable choice of the coordinate functions in the Bubnov–Galerkin method

In the Mikhlin's monograph [5] conditions of the reasonable choice of coordinate functions for the approximate solution of the operator equation

$$Au = f \tag{1}$$

were given with the use of the Ritz method. It is shown in [5] that on the choice of these coordinate functions the following properties depend:

1° convergence of the approximate solution and, possibly, convergence of the "residuum $Au_n - f$" to zero,

2° stability of the method,

3° rate of convergence of an approximating sequence.

It is known, however (see [3], p 23), that for the application of the Ritz method to equation (1) it is needed that the operator A in (1) was selfadjoint and positive-definite. If A is not such, then for the solution of (1) with some additional assumptions, more general Bubnov–Galerkin method can be applied.

Let H be a separable Hilbert space. We assume that the domain of A, $D(A)$ is a dense subset of H, A is a linear operator. The Bubnov–Galerkin method as applied to (1) goes as follows: we choose the sequence of elements

$$\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots \tag{2}$$

satisfying conditions:

1° $\varphi_n \in D(A)$ $(n = 1, 2, \ldots)$,

2° for every n the elements $\varphi_1, \varphi_2, \ldots, \varphi_n$ are linearly independent,

3° the sequence (2) is a complete set in H.

We want to obtain the approximate solution of equation (1) in the form

$$u_n = \sum_{k=1}^{n} a_k \varphi_k, \tag{3}$$

where the a_k's $(k = 1, 2, \ldots, n)$ satisfy the system of equations:

$$\sum_{k=1}^{n} a_k(A\varphi_k, \varphi_j) = (f, \varphi_j), \quad j = 1, 2, \ldots, n. \tag{4}$$

The a_k's in (3) depend on n and should be denoted $a_k^{(n)}$, but we feel free to neglect such pedantry and shall simply write them a_k.

The problem of convergence of the Bubnov–Galerkin method was dealt by many authors, but the most general sufficient condition of its convergence were given by Mikhlin (see [1], [2] or [3], chapter V).

The aim of this paper is to give a certain fashion of the reasonable choice of the sequence (2). This system (2) will be hereinafter called the *system of coordinate functions*. We shall show that, under some restrictions imposed on the operator A in (1), the system of coordinate functions can be chosen in such a manner that conditions 1° and 2° mentioned earlier will be satisfied. In this paper we shall not consider the rate of convergence of an approximating sequence, although it will be the subject of another paper.

First we shall prove some lemmas.

LEMMA 1. *If B is selfadjoint, positive-definite operator, A — a linear operator such that $D(A) = D(B) \subset D(A^*)$, then the operator $B^{-1}A$ is bounded.*

Proof. We shall show first that A^*B^{-1} is bounded. We have $D(A^*B^{-1}) = H$. Indeed, $D(B^{-1}) = R(B) = H$, since B is selfadjoint and positive-definite. On the other hand $R(B^{-1}) = D(B) \subset D(A^*)$, hence, if f is an arbitrary element, then the equality $A^*B^{-1}f = A^*(B^{-1}f)$ makes sense. We shall now show that A^*B^{-1} is closed. Let $f_n \to f$ and $A^*B^{-1}f_n \to g$. We designate $B^{-1}f_n = h_n$. B^{-1} is bounded, hence $B^{-1}f_n \to B^{-1}f$. Let $B^{-1}f = h$, therefore $h_n \to h$ and $A^*h_n \to g$. It is known ([7], p. 557) that A^* is closed, thus $h \in D(A)$ and $A^*h = g$, that is $A^*B^{-1}f = g$, what means that A^*B^{-1} is closed, since operator A^*B^{-1} is defined in the whole space H and closed it is then bounded (see [7], p. 560). Boundedness of $B^{-1}A$ follows from the equality $B^{-1}A = (A^*B^{-1})^*$.

LEMMA 2. *If*: 1° *operators A and B are such as in Lemma* 1, 2° *A^{-1} exists*, 3° *the inequality*

$$|(Au, Bu)| \geqslant c\|Bu\|^2, \quad c > 0, \; u \in D(B), \; \textit{holds}, \tag{5}$$

then the operator BA^{-1} is bounded.

Proof. Let us put in (5) $v = Bu$. Then $u = B^{-1}v$ and so we have

$$c\|v\|^2 \leqslant |(AB^{-1}v, v)| \leqslant \|AB^{-1}v\| \cdot \|v\|$$

or, in other words

$$\|AB^{-1}v\| \geqslant c\|v\|. \tag{6}$$

From (6) it follows the existence and boundedness of the reciprocal operator for the operator AB^{-1}. In view of the assumptions made on A and B it is evident that $(AB^{-1})^{-1} = BA^{-1}$, and hence, Lemma 2 is proved.

LEMMA 3. *If operators A and B satisfy conditions* 1° *and* 3° *of Lemma* 2 *and if $\varphi_1, \varphi_2, \ldots, \varphi_n$ is an orthonormal sequence of elements belonging to the domain of A^*, then the matrix*

$$\psi_n = \|(\varphi_j, B^{-1}A^*\varphi_k)\|_{j,k=1}^n \tag{7}$$

posesses the reciprocal matrix ψ_n^{-1} and, moreover, $\|\psi_n^{-1}\| \leqslant c_2$, where c_2 is a constant which does not depend on n.

Proof. Let t denote any vector of the form $t = (t_1, t_2, \ldots, t_n)$. Let us denote, for brevity $B^{-1}A^*\varphi_m = \psi_m$, we have

$$\|\psi_n t\|^2 = \sum_{k=1}^n \Big| \sum_{m=1}^n (\varphi_k, \psi_m) t_m \Big|^2 = \sum_{k=1}^n |(\psi, \varphi_k)|^2, \tag{8}$$

where $\psi = \sum_{m=1}^n \bar{t}_m \psi_m = B^{-1}A^* \sum_{m=1}^n \bar{t}_m \varphi_m$.

We now apply inequality (5), putting there $Bu = v$. We obtain

$$|(AB^{-1}v, v)| \geqslant c\|v\|^2,$$

where v is an arbitrary element of H. Further, we have

$$|(AB^{-1}v, v)| = |(v, B^{-1}A^*v)| = |(B^{-1}A^*v, v)| \geqslant c\,\|v\|^2.$$

In this last inequality we now replace v by the sum $\sum_{m=1}^n \bar{t}_m \varphi_m$ and this yields

$$|(\psi, v)| \geqslant c\|v\|^2 = c\|t\|^2.$$

On the other hand

$$|(\psi, v)|^2 \leqslant \Big(\sum_{m=1}^n |t_m|\,|(\psi, \varphi_m)| \Big)^2 \leqslant \|t\|^2 \sum_{m=1}^n |(\psi, \varphi_m)|^2.$$

From (8) and this above inequality we obtain

$$\|\psi_n t\|^2 = \sum_{m=1}^n |(\psi, \varphi_m)|^2 \geqslant \|t\|^{-2} |(\psi, v)|^2 \geqslant c\|t\|^2. \tag{9}$$

Now, from (9) it follows that the matrix ψ_n^{-1} exists and also the inequality:

$$\|\psi_n^{-1}\| \leqslant c_2 = c^{-\frac{1}{2}}. \tag{10}$$

Now we shall prove the following theorem, whose proof is based on Lemmas 1, 2 and 3.

THEOREM 1. *If*:

1° *operators* A *and* B *satisfy hypotheses* 1° *and* 3° *of Lemma* 2,

2° *equation* (1) *has the unique solution* u_0,

3° *operator* B *posesses the discrete spectrum*

4° *the term*

$$u_n = \sum_{k=1}^{n} a_k \varphi_k \tag{11}$$

is a n-th successive approximation in the sense of Bubnov–Galerkin of the solution of equation (1); *where* $\{\varphi_n\}$ *is an orthonormal sequence of eigenvectors of* B *corresponding to the eigenvalues* $\{\lambda_n\}$, *then* $u_n \to u_0$ *and* $Au_n - f \to 0$ *when* $n \to \infty$, *in the metric of the space* H.

Proof. Let us denote $w_0 = Bu_0$ and $w_n = Bu_n$; then we can write (1) in the form

$$w_0 = BA^{-1}f \tag{12}$$

and

$$w_n = \sum_{k=1}^{n} c_k \varphi_k, \qquad c_k = \lambda_k a_k. \tag{13}$$

Coordinates a_k $(k = 1, 2, \ldots, n)$ in (11) satisfy the system of equation (4). This system can be transformed in the following manner:

$$(A\varphi_k, \varphi_j) = (\varphi_k, A^*\varphi_j) = \lambda_k(B^{-1}\varphi_k, A^*\varphi_j) = \lambda_k(\varphi_k, B^{-1}A^*\varphi_j),$$

$$(f, \varphi_j) = (AB^{-1}BA^{-1}f, \varphi_j) = (BA^{-1}f, B^{-1}A^*\varphi_j) = (w_0, B^{-1}A^*\varphi_j).$$

Then the system (4) takes the form

$$\sum_{k=1}^{n} c_k(\varphi_k, B^{-1}A^*\varphi_j) = (w_0, B^{-1}A^*\varphi_j); \quad j = 1, 2, \ldots, n. \tag{14}$$

But the system (14) may we write in the following form

$$P_n w_n = P_n w_0, \tag{15}$$

where P_n is a projection operator on the space K_n spanned by the vectors $\psi_j = B^{-1}A^*\varphi_j$, $j = 1, 2, \ldots, n$.

From a theorem of N. I. Polskiĭ we know that for the convergence $w_n \to w_0$ it suffices that the inequality

$$\|v\| \leqslant C\|P_n v\|, \qquad v \in L_n, \tag{16}$$

is satisfied. In this inequality L_n denotes the space spanned by the vectors $\varphi_1, \varphi_2, \ldots, \varphi_n$, where $\{\varphi_n\}$ is a complete system in H, and the constant C does not depend on n (see [6] or [5], p. 122).

Operator P_n can be defined as follows: we find such constants μ_{jk}, that

$$\|\varphi_j - \sum_{k=1}^{n} \mu_{jk} B^{-1} A^* \varphi_k\|^2 = \min, \quad \text{for } j = 1, 2, \ldots, n, \tag{17}$$

then for arbitrary $v \in L_n$ we have

$$v = \sum_{k=1}^{n} \gamma_k \varphi_k, \quad P_n v = \sum_{j,k=1}^{n} \gamma_j \mu_{jk} B^{-1} A^* \varphi_k. \tag{18}$$

Inequality (16) takes then the following form

$$\Big\| \sum_{k=1}^{n} \gamma_k \varphi_k \Big\|^2 \leqslant C^2 \Big\| \sum_{j,k=1}^{n} \gamma_j \mu_{jk} B^{-1} A^* \varphi_k \Big\|^2,$$

or it can be written:

$$\sum_{k=1}^{n} |\gamma_k|^2 \leqslant C^2 \sum_{j,k=1}^{n} \gamma_j \bar{\gamma}_k \sum_{r,s=1}^{n} \mu_{jr} \mu_{ks} (B^{-1} A^* \varphi_r, B^{-1} A^* \varphi_s). \tag{19}$$

From (19) it follows, that to show that inequality (16) is true it suffices to show that the minimal value of the quadratic form

$$\Gamma = \sum_{j,k=1}^{n} \gamma_j \bar{\gamma}_k \sum_{r,s=1}^{n} \mu_{jr} \bar{\mu}_{ks} (B^{-1} A^* \varphi_r, B^{-1} A^* \varphi_s), \tag{20}$$

is bounded from below by the non-negative constant independent of n. We denote

$$\delta_r = \sum_{j=1}^{n} \gamma_j \mu_{jr}, \tag{21}$$

then

$$\Gamma = \sum_{r,s=1}^{n} \delta_r \bar{\delta}_s (B^{-1} A^* \varphi_r, B^{-1} A^* \varphi_s) = \Big\| B^{-1} A^* \sum_{r=1}^{n} \delta_r \varphi_r \Big\|^2.$$

Let

$$\sum_{r=1}^{n} \delta_r \varphi_r = \xi, \quad B^{-1} A^* \xi = \eta,$$

hence

$$\xi = (A^*)^{-1} B \eta \quad \text{and} \quad \|\xi\| \leqslant \|(A^*)^{-1} B\| \, \|\eta\| = \|B A^{-1}\| \, \|\eta\|.$$

We deduce from Lemma 2 that the operator BA^{-1} is bounded, th us

$$\|\eta\| \geqslant \|B A^{-1}\|^{-1} \|\xi\|,$$

and so we have

$$\Gamma \geqslant \|B A^{-1}\|^{-2} \|\sum_{r=1}^{n} \delta_r \varphi_r\|^2 = \|B A^{-1}\|^{-2} \cdot \sum_{r=1}^{n} |\delta_r|^2. \tag{22}$$

Let, further, M_n denote the matrix of transformation (21)

$$M_n = \|\mu_{jr}\|_{j,r=1}^{n}.$$

We shall show that there exists the reciprocal matrix M_n^{-1} and $\|M_n^{-1}\| \leqslant C_1$, where C_1 is a constant which does not depend on n. Indeed, from (17) follows, that the constants μ_{jk} satisfy the system of equations:

$$\sum_{k=1}^{n} (B^{-1}A^*\varphi_k, B^{-1}A^*\varphi_m)\mu_{jk} = (\varphi_j, B^{-1}A^*\varphi_m), \quad j, m = 1, \ldots, n. \tag{23}$$

Denote now by Φ_n the matrix, whose elements are $(B^{-1}A^*\varphi_k, B^{-1}A^*\varphi_m)$. Then the system (23) can be written in the form:

$$M_n\Phi_n = \psi_n, \tag{24}$$

where the matrix ψ_n is defined in (7).

Let us observe, that the matrix Φ_n in (24) is bounded.
In fact, similarly as in Lemma 3, let $t = (t_1, t_2, \ldots, t_n)$ be an arbitrary vector, hence

$$\|\Phi_n t\|^2 = \sum_{k,m=1}^{n} (B^{-1}A^*\varphi_k, B^{-1}A^*\varphi_m) t_k \bar{t}_m = \|B^{-1}A^* \sum_{k=1}^{n} t_k\varphi_k\|^2$$

$$\leqslant \|B^{-1}A^*\|^2 \| \sum_{k=1}^{n} t_k\varphi_k\|^2 = \|AB^{-1}\|^2 \sum_{k=1}^{n} |t_k|^2 = \|AB^{-1}\|^2\|t\|^2.$$

From Lemma 1 we deduce that the operator AB^{-1} is bounded, thus, from the latter inequality is follows that

$$\|\Phi_n\| \leqslant \|AB^{-1}\|. \tag{25}$$

From (24) in view of (10) and (25) we obtain

$$\|M_n^{-1}\| \leqslant C_3 = \|AB^{-1}\| c^{-\frac{1}{2}}.$$

Observe, that the system (21) can we write in the form $M_n^*\gamma = \delta$, where $\gamma = (\gamma_1, \ldots, \gamma_n)$ and $\delta = (\delta_1, \ldots, \delta_n)$. Hence $\|\delta\| \geqslant C_3^{-1}\|\gamma\|$, and so

$$\Gamma \geqslant \|BA^{-1}\|^{-2} C_3^{-2} \|\gamma\|^2. \tag{26}$$

We have already mentioned that inequality (26) implies inequality (16). From this, in view of the quoted theorem of Polskiĭ, follows that $w_n \to w_0$, or $Bu_n \to Bu_0$ in the metric of space H. But the operators B^{-1} and AB^{-1} are both bounded and so

$$B^{-1}(Bu_n) = u_n \to B^{-1}(Bu_0) = u_0$$

as well as

$$Au_n - f = AB^{-1}(Bu_n - Bu_0) \to 0,$$

and this is what had to be proved.

Remark 1. Operator B in (5) can be replaced by the operator $B+kE$, where E is the identity operator and k some non-negative constant (see [5], p. 129).

Stability of the Bubnov–Galerkin method.

Suppose, that in a certain computational process we deal with the system of equations:

$$A_n x^{(n)} = y^{(n)}, \quad n = 1, 2, 3, \ldots, \tag{27}$$

where A_n is an operator from one Banach space X_n to another Banach space Y_n. We assume also that, for every n, A_n^{-1} exists and is defined in the whole space Y_n. Simultanously with (27) we consider the sequence of equations

$$(A_n+\Gamma_n)z^{(n)} = y^{(n)} + \delta^{(n)}. \tag{28}$$

According to the definition, given by Mikhlin in [4] or [5], p. 70, we say that this computational process is stable, if there exist constants p, q, r independent on n and such that for $\|\Gamma_n\| \leqslant r$ and arbitrary $\delta^{(n)}$ equations (28) have solutions and there holds the inequality

$$\|z^{(n)} - x^{(n)}\| \leqslant p\|\Gamma_n\| + q\|\delta^{(n)}\|. \tag{29}$$

We say that the computational process (27) is convergent if there exists a limit $x_0 = \lim_{n\to\infty} x^{(n)}$ in the norm of a space X, where X_n are the subspaces of X.

We shall now prove the following

THEOREM 2. *If the sequence $\{\varphi_n\}$ of coordinate functions is chosen according to Theorem* 1, *then the Bubnov–Galerkin method for equation* (1) *is stable.*

Proof. Evidently, it suffices to show that the computational process for the solution of the sequence of equations

$$A_n a^{(n)} = f^{(n)}, \quad n = 1, 2, \ldots, \tag{30}$$

where $A_n = \|(A\varphi_k, \varphi_j)\|_{k,j=1}^n$, $a^{(n)} = (a_1, \ldots, a_n)$ and $f^{(n)} = \big((f,\varphi_1), \ldots, (f,\varphi_n)\big)$ is stable.

Indeed, in the situation we are considering $X_n = Y_n$ are both n-dimensional euclidean spaces, and X is a l_2 space. Since the sequence $\{\varphi_n\}$ is orthonormal, we have

$$\|u_n\|^2 = \Big(\sum_{k=1}^n a_k\varphi_k, \sum_{k=1}^n a_k\varphi_k\Big) = \sum_{k=1}^n a_k^2 = \|a^{(n)}\|_n^2. \tag{31}$$

From (31), in view of Theorem 1, it follows the convergence of the process (30). This in turn implies, by Theorem 13.3 of [5], p. 74, that the process

(30) is stable if and only if $\|A_n^{-1}\| \leqslant C$, where C is a constant not depending on n. We observe that

$$A_n = \|(A\varphi_k, \varphi_j)\|_{k,j=1}^n = \|\lambda_k(B^{-1}\varphi_k, A^*\varphi_j)\|_{k,j=1}^n \tag{32}$$
$$= \|\lambda_k(\varphi_k, B^{-1}A^*\varphi_j)\|_{k,j=1}^n = \Lambda_n\psi_n,$$

where Λ_n is a diagonal matrix

$$\Lambda_n = (\lambda_1, \ldots, \lambda_n),$$

and matrix ψ_n is defined in (7).

From (32), by Lemma 3 we conclude that A_n^{-1} exists.

Let $t = (t_1, t_2, \ldots, t_n)$ be an arbitrary vector. We have

$$\|A_n t\|^2 = \sum_{k=1}^n \Big| \sum_{j=1}^n \lambda_k(\varphi_k, B^{-1}A^*\varphi_j) t_j\|^2 \geqslant \sum_{k=1}^n \lambda_1^2 \Big| \sum_{j=1}^n (\varphi_k, B^{-1}A^*\varphi_j) t_j|^2$$
$$= \lambda_1^2 \|\psi_n t\|^2.$$

From this, by inequality (9), we obtain

$$\|A_n t\|^2 \geqslant \lambda_1^2 c^2 \|t\|^2. \tag{33}$$

So the proof of Theorem 2 is completed.

Remark 2. If we assume that the operator A in (1) is self-adjoint and positive-definite then, as is well-known the Bubnov–Galerkin method is equivalent to Ritz method. In this case, obviously, Theorem 1 of this paper is the same as Theorem 23.1 of [5], p. 124.

Remark 3. It was mentioned in the introduction that S. G. Mikhlin has given the sufficient conditions for the convergence of Bubnov–Galerkin method, when appropriate assumptions about operator A in (1) were made. We consider important to emphasize that these conditions imposed on A by Mikhlin do not overlap with the conditions of Theorem 1, hence Mikhlin's theorem does not imply the convergence of Bubnov–Galerkin method for equation (1) with the hypotheses of Theorem 1, neither the convergence to zero of "residuum $Au_n - f$".

References

[1] С. Г. Михлин, *О сходимости метода Галеркина*, ДАН, т. 61, № 2 (1948).

[2] — *Некоторые достаточные условия сходимости метода Галеркина*, Уч. зап. ЛГУ, № 135, сер. матем. наук, № 18 (1950).

[3] — *Прямые методы в математической физике*, Москва 1950.

[4] — *Об устойчивости некоторых вычислительных процессов*, ДАН 157, № 2 (1964).

[5] — *Численная реализация вариационных методов*, Москва 1966.

[6] Н. И. Польский, *О сходимости некоторых приближенных метод анализа*, Укр. Мат. Журн. № 1 (1955), p. 56–70.

[7] В. И. Смирнов, *Курс высшей математики*, т. 5, Москва 1957.