ANNALES
POLONICI MATHEMATICI
XXXIV (1977)

Uniform convergence of a Galerkin-type method for a non-linear boundary value problem

by SATISH SHIRALI* (Trieste, Italy)

Abstract. A Galerkin-type approximation method for a fixed point problem $x = Fx$ is considered in the abstract setting of a Banach space with an auxiliary norm. Such a method consists in seeking an x_n in a subset of the n-th finite-dimensional subspace which minimizes the auxiliary norm of $x_n - P_n F x_n$, where P_n is a projection onto the subspace. A partial differential equation of the type $\Delta u(P) = f(P, u(P), \nabla u(P))$ is fitted into this setting and convergence theorems obtained for two Galerkin-type methods. One resembles the usual Galerkin (moment) method and the other resembles the least squares method. The latter is shown to give uniform convergence in some cases when the dimension is 3 or less.

Introduction. The convergence of Galerkin-type methods for linear equations has been studied by several authors, Keldysh [3], Kantorovich [2], Mikhlin [7] and others. In the non-linear case, Krasnosel'skii [6] proved convergence in an abstract Banach space setting. But, fitting a concrete differential equation to this setting presents difficulties if uniform convergence is desired. Vainikko [9] has fitted the case of a non-linear *ODE* and obtained uniform convergence for the interpolation method under suitable but broad assumptions. One of the crucial theorems used by him is not available in the case of *PDE*'s as indicated at the end of Section 2 below. Therefore, his methods do not readily extend to *PDE*'s and a certain modification of the abstract setting becomes necessary.

In this paper, we obtain, under some rather broad conditions, uniform convergence for the unknown function itself and mean square convergence for first partial derivatives but for dimension $\leqslant 3$. The 2-dimensional and simplest case of our *PDE* is

$$\Delta u = f\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$$

* On leave of absence from Department of Mathematics, Panjab University, Chandigarh, India.

on a bounded domain with sufficiently smooth boundary, the boundary condition being that u must vanish there. It is intended that u should have Hölder continuous second partial derivatives on the closed domain. The approximate method we consider is of the Galerkin-type in the abstract setting, but in actual computation it strongly resembles the method of least squares. Details are discussed in Section 4 below.

First we establish that the problem is equivalent to that of finding the fixed point of a completely continuous operator. This operator, which naturally involves the function f, cannot be thought of as operating in a Hilbert space unless some conditions besides continuous differentiability are imposed on f. Roughly speaking, this is because a Hilbert space is not a Banach algebra. We are therefore forced to look at a space of continuous functions, with an inner product providing an auxiliary (incomplete) norm. Consequently there is no explicit mention of Sobolev spaces even when they may be there.

1. Let D be a bounded domain in p-dimensional space and ∂D its boundary. We assume that the boundary of D is smooth enough to apply the Schauder estimate as in (5′) on p. 336 of Courant and Hilbert, Vol. II, [1]. Let $C^{k+\alpha}(D)$ be the usual Banach space (op. cit.) of real valued functions on $\bar{D}$ with continuous partial derivatives up to order k, the k-th order derivatives being α-Hölder continuous $(0<\alpha<1)$. Let

$$L = \sum \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right), \qquad a_{ij} = a_{ji},$$

be an elliptic partial differential operator in D with C^2 coefficients. We consider the partial differential equation,

$$\text{(A)} \qquad (Lu)(P) = f\left(P, u(P), \frac{\partial u}{\partial x_1}(P), \ldots, \frac{\partial u}{\partial x_p}(P)\right), \qquad u(\partial D) \equiv 0,$$

the (classical) solution being sought in $C^{2+\alpha}(D)$ for some α. Here $f\colon \bar{D}\times \times \boldsymbol{R}^{p+1}\to\boldsymbol{R}$ is assumed to have continuous first partial derivatives. In particular, f satisfies a Lipschitz condition on every bounded subset of $\bar{D}\times\boldsymbol{R}^{p+1}$, the constant being dependent on the bounded set.

Let K be the linear operator which maps each $\varphi\in C^{\alpha}(D)$ into the solution in $C^{2+\alpha}(D)$ of

$$Lu = \varphi, \qquad u(\partial D) \equiv 0.$$

Then, as an operator from $C^{\alpha}(D)$ to $C^2(D)$, K is completely continuous because of the Schauder estimates. We consider the equation

$$\text{(B)} \quad v(P) = f\left(P, (Kv)(P), \left(\frac{\partial}{\partial x_1}Kv\right)(P), \ldots, \left(\frac{\partial}{\partial x_p}Kv\right)(P)\right), \qquad v\in C^1(D).$$

It is clear that if v is a solution of (B), then Kv is a solution of (A).

Now suppose that u is a solution of (A) and set $v = Lu$. Then $u = Kv$ and v is a solution of (B), except that $v \in C^{\alpha}(D)$. However, as long as $v \in C^{\alpha}(D)$, the functions occurring "inside" f on the right-hand side in (B) are in $C^{1+\alpha}(D)$. Moreover, f has continuous first partial derivatives; therefore the entire right-hand side in (B) is in $C^1(D)$. Therefore, if $v \in C^{\alpha}(D)$ and satisfies the equation in (B), then v actually lies in $C^1(D)$. Therefore, for any solution u of (A), $v = Lu$ is a solution of (B).

This proves that (A) and (B) are equivalent in the sense that, for any solution u of (A), Lu is a solution of (B) and at the same time, for any solution u of (B), Kv is a solution of (A).

For any $v \in C^1(D)$, define Fv as follows:

$$(Fv)(P) = f\left(P, (Kv)(P), \left(\frac{\partial}{\partial x_1} Kv\right)(P), \ldots, \left(\frac{\partial}{\partial x_p} Kv\right)(P)\right).$$

It is a by-product of our argument above that F maps $C^1(D)$ into itself. Note that (B) requires v to be a fixed point of the operator F in $C^1(D)$. Therefore, instead of considering the original boundary value problem, we may consider the equivalent fixed point problem.

We shall now prove that F is a completely continuous operator from $C^1(D)$ into itself. Let v_n be a bounded sequence in $C^1(D)$. Then v_n is a bounded sequence in $C^{\alpha}(D)$ for every α between 0 and 1, and by the Schauder estimates Kv_n is a bounded sequence in $C^{2+\alpha}(D)$ for every α. Therefore there is a subsequence v_{n_k} such that the corresponding sequences Kv_{n_k} and $\frac{\partial}{\partial x_1} Kv_{n_k}, \ldots, \frac{\partial}{\partial x_p} Kv_{n_k}$ all converge in $C^1(D)$. In particular, all these functions remain bounded in $C(D)$. Therefore the Lipschitz property of f may be used, and we conclude that Fv_{n_k} and its first partial derivatives all converge uniformly; in other words Fv_{n_k} converges in $C^1(D)$. All this shows that F is compact. The continuity of F is easily proved. Therefore F is completely continuous.

2. In this section the abstract function space set-up is described and shown to be applicable to equation (B). As far as possible, the notation will agree with that of [6].

It is assumed that "co-ordinate functions" can be found whose linear span is dense in $C^1(D)$. See e.g. Chapter IV of [4].

E is a Banach space over the reals, with norm $\| \; \|$. There is also an auxiliary norm $\| \; \|_2$ in E, but it need not be complete. F is a non-linear operator in E. There is an increasing sequence $L_1 \subset L_2 \subset \ldots$ of finite-dimensional linear subspaces of E and corresponding projections $P_1, P_2, \ldots$ of E into these subspaces. Furthermore, the following conditions are satisfied:

1° For every $x \in E$, $\|x\| \geqslant \|x\|_2$;

2° F is completely continuous under the complete norm $\| \ \|$;

3° F is *bound continuous* in the sense that if $\|x_n - x\|_2 \to 0$ and $\|x_n\|$ is bounded, then $\|Fx_n - Fx\|_2 \to 0$;

4° For every $x \in E$, $\|P_n x - x\|_2 \to 0$;

5° $\bigcup Lu$ is dense in E in the complete norm.

For the application to equation (B), the role of E is played by $C^1(D)$ with its usual complete norm as $\| \ \|$, and the mean square (i.e., L^2) norm as $\| \ \|_2$. Clearly, condition 1° is satisfied except for a constant factor, which we may ignore.

The role of F is played by the operator which has already been denoted by "F" in Section 1. It was proved at the end of Section 1 that condition 2° is satisfied.

Denote the inner product in $L^2(D)$ by (,) as usual. As a preliminary to condition 3°, we wish to prove the following. Let γ be a constant such that $(-Lu, u) \geqslant \gamma(u, u)$ for those $u \in C^2(D)$ which vanish on the boundary of D. Then we have $\gamma(-Kv, v) \leqslant (v, v)$ for every $v \in L^2(D)$. Here K is regarded as having been extended to $L^2(D)$, where it is a bounded operator. First we restate the inequality $(-Lu, u) \geqslant \gamma(u, u)$ as $(-Kv, v) \geqslant \gamma(Kv, Kv)$ for all $v \in C^1(D)$. Since K is a bounded operator, this holds for all $v \in L^2(D)$. Now we argue as follows:

$$(v + \gamma Kv, v + \gamma Kv) \geqslant 0.$$

Therefore,

$$\begin{aligned}(v, v + \gamma Kv) &\geqslant -(\gamma Kv, v + \gamma Kv) \\ &\geqslant -\gamma[(Kv, v) + \gamma(Kv, Kv)] \\ &\geqslant 0.\end{aligned}$$

This is precisely the inequality we wished to prove. We shall use this inequality to prove condition 3°.

Let $\|x_n - x\|_2 \to 0$. By what we just proved, $\big(K(x_n - x), x_n - x\big) \to 0$. But it is well known that by the ellipticity of L we have

$$-(Kv, v) \geqslant \mu \int_D \sum_{j=1}^{p} \left(\frac{\partial}{\partial x_j} Kv\right)(P)^2 dP \geqslant \gamma \int_D (Kv)\ (P)^2 dP \qquad \text{for } v \in C^1(D),$$

where μ is a suitable constant. The latter inequality here is Friedrich's inequality [8], p. 146. Therefore $K(x_n - x)$ and its first partial derivatives converge to 0 in $L^2(D)$. From this we would like to conclude that $Fx_n - Fx$ converges to 0 in $L^2(D)$ by applying the Lipschitz property of the numerical function f. But this property can be applied only if Kx_n and its partial derivatives remain bounded. This does happen, however, because $\|x_n\|$ is assumed bounded. This proves that condition 3° holds.

Actually, we have almost proved that condition 3° would hold even if $\| \ \|_2$ were taken to be the so-called energy norm $\sqrt{-(Kv, v)}$. In the preceding paragraph we started by moving over to this norm in any case. At the end of that argument, we can apply the inequality $\gamma(-Kv, v) \leqslant \leqslant (v, v)$ once more and we have the desired conclusion regarding the use of the energy norm to replace the mean square norm. Such a replacement can also be made in condition 1°, because the energy norm is essentially smaller than the mean square norm.

Now we choose co-ordinate functions $\varphi_1, \varphi_2, \ldots$ in such a way that condition 5° holds. (The φ_j need not satisfy the boundary condition. More about this later.) In particular, the linear combinations of the φ_j are uniformly dense in $C(D)$. Let L_n be the subspace spanned by the first n of these functions. Let P_n be the orthogonal projection of $C^1(D)$ onto L_n. We shall get different P_n depending upon whether $\| \;\|_2$ is chosen to be the L^2 norm or the energy norm $\sqrt{(-Kv, v)}$. But in either case, condition 4° holds.

We have so far described a concrete example of the abstract set-up, with two possible choices of the auxiliary norm $\| \;\|_2$.

The need for the uniform boundedness of Kx_n and their partial derivatives in proving condition 3° prevents us from considering the entire question directly in Hilbert space as has been done by Vainikko [9]. In his one variable case, $\|x_n - x\|_2 \to 0$ implies that the derivative of Kx_n converges uniformly to the derivative of Kx. A corresponding result in the multivariable case is not available. Otherwise F could have been extended to $L^2(D)$.

3. We now return to the abstract situation of a Banach space E with an auxiliary norm $\| \;\|_2$, a non-linear operator F in E and subspaces L_k with projections P_k, satisfying the four conditions of Section 2. Let L be a subset of E, which is bounded with respect to the complete norm and whose intersection with any finite-dimensional subspace is a polyhedron. Consider the vector field Φ on L given by $\Phi x = x - Fx$. We also suppose that Φ has no null vectors in the $\| \;\|_2$ closure of L. We shall show that the rotation of Φ on L can be obtained by using the projections P_k.

First we prove that there is some $a > 0$ such that $\|\Phi x\|_2 > a$ for every $x \in L$. If not, then there exists a sequence $x_n \in L$ such that $\|x_n - - Fx_n\|_2 \to 0$. Since F is completely continuous (condition 2°) there is some subsequence, which we shall also denote by x_n, such that $\|Fx_n - y\| \to 0$ for some $y \in E$. By condition 1°, $\|Fx_n - y\|_2 \to 0$. Now,

$$\|x_n - y\|_2 \leqslant \|x_n - Fx_n\|_2 + \|Fx_n - y\|_2,$$

so that $\|x_n - y\|_2 \to 0$. By condition 3°, $\|Fx_n - Fy\|_2 \to 0$. Therefore $y = Fy$. Since Φ is assumed to have no null vectors in the $\| \;\|_2$ closure of L, this is a contradiction. This proves that such a positive number a exists.

From condition 4° and the uniform boundedness principle, we conclude that there is an upper bound, say $A - 1$, for the norms of P_n regarded as mapping the Banach space E into the normed linear space E with norm $\| \ \|_2$. By condition 2° we can choose an $(a/6A)$-net for the totally bounded image of L under F. Again by condition 4°, choose a positive integer N so large that for $n \geqslant N$ and any x for which Fx is in the $(a/6A)$-net, we have

$$\|P_n Fx - Fx\|_2 < a/6.$$

It follows that

$$\|QFx - Fx\|_2 < a/2 \quad \text{for all } x \in L, \tag{3.1}$$

if $Q = P_n$, where $n \geqslant N$. Now it is known that a projection Q into a finite-dimensional subspace can be used for obtaining the rotation of Φ on L if

$$\|QFx - Fx\| < \beta/2 \quad \text{for all } x \in L, \tag{3.2}$$

where β is any number such that $0 < \beta < \inf\|x - Fx\|$, $x \in L$; and that such Q with arbitrarily small β can be found [6], p. 106. By condition 1°, we may choose $\beta = a$. But then, again by condition 1°, (3.2) will imply that (3.1) holds for Q. Now we argue exactly as in the well-known case that if Q and Q' both satisfy (3.1), then the vector fields $x - QFx$ and $x - Q'Fx$ have no null vectors on the intersection of L with the subspace into which Q and Q' map, and that they are homotopic there. Then either field gives the rotation of $x - Fx$ on L. Since Q can be chosen to be P_n for sufficiently large n, $x - P_n Fx$ gives the rotation of $x - Fx$ on L, if n is large enough.

We are now ready to discuss the approximation of the fixed point of the non-linear operator F. Suppose x_0 is a fixed point which is *strongly isolated* in the sense that there is some closed $\| \ \|$-ball B containing x_0 in its interior such that the $\| \ \|_2$ closure of this ball contains no other fixed points of F. All approximate solutions will be required to lie in this ball. An *n-th Galerkin approximate solution* of the equation $x = Fx$ will be understood to be an element $x_n \in B \cap L_n$ such that

$$\|x_n - P_n Fx_n\|_2 = \min \|x - P_n Fx\|_2, \quad x \in B \cap L_n. \tag{3.3}$$

Condition 5° of Section 2 guarantees that $B \cap L_n$ is non-empty for sufficiently large n. Moreover, the compactness of $B \cap L_n$ guarantees the occurrence of the minimum in (3.3). Therefore, an n-th Galerkin approximate solution exists if n is large enough. The usual notion of a Galerkin approximate solution requires that the minimum in (3.3) should be zero. This possibility is covered by our definition. The only reason for this departure from usual practice is to guarantee existence.

We shall now prove that

$$\lim \|x_n - P_n Fx_n\|_2 = 0. \tag{3.4}$$

If this does not hold, then there is some subsequence x_{n_k} and some $\beta > 0$ such that

$$\|P_{n_k}Fx_{n_k} - x_{n_k}\|_2 > \beta. \tag{3.5}$$

As in (3.1), we can assert that for sufficiently large k

$$\|P_{n_k}Fx - Fx\|_2 < \beta/2 \quad \text{for all } x \in B. \tag{3.6}$$

Now $x_0 - Fx_0 = 0$ and F is continuous. Therefore, there is some $\| \ \|$-ball $B' \subset B$ about x_0 such that

$$\|x - Fx\|_2 < \beta/2 \quad \text{for all } x \in B'. \tag{3.7}$$

$B' \subset B$ and, therefore, for any $x \in B' \cap L_n$ we have $\|x - P_nFx\|_2 \geqslant \|x_n - P_nFx_n\|_2$. Now consider any $x \in B' \cap L_{n_k}$, which is non-empty if k is sufficiently large. For such an x we have

$$\begin{aligned}\|x - Fx\|_2 &\geqslant \|x - P_{n_k}Fx\|_2 - \|P_{n_k}Fx - Fx\|_2 \\ &\geqslant \|x_{n_k} - P_{n_k}Fx_{n_k}\|_2 - \|P_{n_k}Fx - Fx\|_2 \\ &> \beta - \beta/2 = \beta/2,\end{aligned}$$

where the last inequality follows from (3.5) and (3.6) and the fact $B' \subset B$. But the inequality just proved contradicts (3.7). This shows that (3.4) must hold.

Now we use (3.4) to prove that the Galerkin approximate solutions x_n converge to the exact solution x_0 in the auxiliary norm:

$$\lim \|x_n - x_0\|_2 = 0.$$

Consider any subsequence x_{n_k} of approximate solutions, and an arbitrary $\varepsilon > 0$. Then one can ensure as in (3.1) that for sufficiently large k we have

$$\|P_{n_k}Fx - Fx\|_2 < \varepsilon/2 \quad \text{for all } x \in B.$$

In particular, this holds for $x = x_{n_k}$. This fact combined with (3.4) shows that $\|x_{n_k} - Fx_{n_k}\|_2 < \varepsilon$ for sufficiently large k. But this means that $\lim \|x_{n_k} - Fx_{n_k}\|_2 = 0$. One can now proceed as in the second paragraph of this section to show that x_{n_k} has a sub-subsequence which approaches x_0 in the auxiliary norm. It is now an elementary argument that if every subsequence of a sequence has a sub-subsequence approaching a given element, then the sequence itself approaches that element. Thus our contention that $\|x_n - x_0\|_2 \to 0$ stands proved.

Special mention should be made of one case when the minimum occurring in (3.3) is 0. Suppose the rotation of the vector field Φ on the boundary L of B is non-zero, i.e., the index of the fixed point x_0 is not zero. We have already shown that this rotation is the same as that of

$x - P_n Fx$ on $L \cap L_n$. Therefore, it follows exactly as in the well-known case [6] that the minimum occurring in (3.3) is 0 for sufficiently large n.

4. So far we have shown that in the abstract situation of Section 2 the Galerkin approximations to a strongly isolated fixed point of F converge to that fixed with respect to the auxiliary norm. We shall now apply this to the concrete example described in Section 2. First we consider the case when the auxiliary norm is the mean square norm. Then the Galerkin approximate solutions v_n of equation (B) must converge to the exact solution v in mean square. Suppose that the inverse operator K can be represented as an integral operator whose kernel $G(P, Q)$, the Green's function, has the property that $\int_D G(P, Q)^2 dQ$ is bounded independently of P. Then, by Theorem 1 (2.X) on p. 356 of [3], we can conclude that K maps a mean square convergent sequence into an L^∞ convergent sequence. Such a G cannot be found if $p > 3$. Now the Kv_n are in C^1. Therefore Kv_n converges uniformly to Kv, which is the exact solution of (A). Since $\gamma(-Kv_0, v_0) \leqslant (v_0, v_0)$, the partial derivatives of the Kv_n converge to those of the exact solution Kv in mean square.

The Kv_n, however, are Galerkin approximate solutions for (A) in the sense of the following:

Definition. Let $\varphi_1, \varphi_2, \ldots$ be a sequence of linearly independent functions in $C^1(D)$ whose linear span is dense. Let P_n be the orthogonal projection (in the sense of $L^2(D)$) onto the linear span of the first n of these co-ordinate functions φ_k, and let K be the inverse of the operator L. Then an *n-th Galerkin approximate solution* of equation (A) relative to a given subset of $C^1(D)$ is a linear combination $u = \sum_{i=1}^{n} c_i K\varphi_i$ which achieves the minimum value of

$$\int_D \left| (Lu)(P) - P_n\left[f\left(P, u(P), \frac{\partial u}{\partial x_1}(P), \ldots, \frac{\partial u}{\partial x_p}(P)\right)\right]\right|^2 dP$$

among all such linear combinations for which $Lu = \sum_{i=1}^{n} c_i\varphi_i$ belongs to the given subset of $C^1(D)$.

If the φ_i are orthonormal, then without explicitly referring to the projections P_n, this may be described as a linear combination which achieves the minimum value of

$$\sum_{i=1}^{n}\left[c_i - \int_D \varphi_i(P) f\left(P, u(P), \frac{\partial u}{\partial x_1}(P), \ldots, \frac{\partial u}{\partial x_p}(P)\right) dP\right]^2$$

among all such linear combinations for which Lu belongs to the given subset of $C^1(D)$.

Two comments on this definition are in order here. The first is that a subset of $C^1(D)$ is supposed to be given. This is because in the abstract situation we seek approximate solutions only within the ball which has been denoted by B above. Without this, existence of approximate solutions could have been proved only when the index of the solution was assumed non-zero, and convergence could not have been proved at all, because more than one solution would become relevant.

The second comment is that there is an explicit reference to K. The approximate solution is sought as a linear combination of the $K\varphi_i$, but the expression to be minimized involves the φ_i. This is where the above notion of a Galerkin approximate solution differs from the usual one in an essential way. To put this difference in a more familiar perspective, let us suppose, as one usually does, that the minimum is equal to 0. Then the c_i are given by the equations

$$(\varphi_i, Lu) = c_i = \int_D \varphi_i(P) f\left(P, u(P), \frac{\partial u}{\partial x_1}(P), \ldots, \frac{\partial u}{\partial x_p}(P)\right) dP.$$

This is so even if the φ_i are not orthonormalized. Let us denote the right-hand side by (φ_i, f) for short. Then we are seeking solutions in the form $u = \sum_{i=1}^{n} c_i \psi_i$, where $\psi_i = K\varphi_i$, $\varphi_i = L\psi_i$ and the c_i are given by the equations $(\varphi_i, Lu) = (\varphi_i, f)$. In the more usual Galerkin method we set $(\psi_i, Lu) = (\psi_i, f)$, where the ψ_i rather than φ_i are called co-ordinate functions. This is why we need not require the φ_i to satisfy the boundary condition.

Before proceeding further, let us state the result which has been proved regarding equation (A).

THEOREM 1. *Let D be a bounded domain in $\boldsymbol{R}^p$, with a boundary which is sufficiently smooth for the Schauder estimates to hold, and let $f\colon \bar{D} \times \boldsymbol{R}^{p+1} \to \boldsymbol{R}$ have continuous first partial derivatives. Let L be an elliptic operator in D, of the form $Lu = \sum \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)$, $a_{ij} = a_{ji}$ the coefficients being in $C^2(D)$. Suppose that u_0 is an (exact) solution of equation* (A), *and B_0 is a closed ball in $C^1(D)$ containing Lu_0 in its interior and having no other solution of* (A) *in its mean square closure. Suppose also that $\varphi_1, \varphi_2, \ldots$ is a linearly independent sequence of co-ordinate functions in $C^1(D)$ whose linear span is dense.*

Then, for sufficiently large n, the n-th Galerkin approximate solution u_n of equation (A) *relative to B_0 exists. Moreover, $Lu_n, \frac{\partial}{\partial x_i}u_n$ and u_n approach $Lu_0, \frac{\partial}{\partial x_i}u_0$ and u_0 in mean square. If the inverse K of L can* *be represented as an integral operator with kernel $G(P,Q)$ for which $\int_D G(P,Q)^2 dQ$ is bounded independently of P, then u_n converges to u_0 uniformly* ($p \leqslant 3$).

If the index of the solution Lu_0 of equation (B) *is non-zero, then the minimum achieved by the n-th Galerkin approximation is zero for sufficiently large n.*

Having stated the theorem, let us resume the discussion about the φ_i and ψ_i. We already pointed out that the more usual Galerkin method consists in setting $(\psi_i, Lu) = (\psi_i, f)$. This may be written as $(-K\varphi_i, Lu) = (-K\varphi_i, f)$ or $[\varphi_i, Lu] = [\varphi_i, f]$, where $[\ ,\ ]$ denotes the inner product associated with the energy norm. These equations are the same as those of the modified Galerkin method above, except that the energy norm replaces the mean square norm. Therefore, our function space set-up covers this case when we choose the auxiliary norm to be the energy norm. The definition of an n-th Galerkin approximate solution of (A) must now be modified by replacing the mean square integral by the energy integral; in case the φ_i are orthonormal in energy, then the finite sum to be minimized must be replaced by a similar sum, where the ψ_i appear in place of φ_i. In case the minimum is expected to be zero, then the equations for the c_i reduce to the standard ones.

THEOREM 2. *Suppose the hypotheses of Theorem* 1 *hold, B_0 has no other solution of* (A) *in its energy norm closure, and the approximate solutions u_n are obtained* (*as is usual*) *by minimizing an energy integral instead of a mean square integral. Then the same conclusions hold as in Theorem* 1, *except for the convergence of Lu_n and the uniform convergence of u_n.*

It was noted earlier that in the standard Galerkin method it is the ψ_i which are called co-ordinate functions. Theorem 2 requires that the $L\psi_i$ span a dense subspace of $C^1(D)$.

The above theorems are a little awkward because of the elaborate assumption that the solution is strongly isolated and because the minimum given by the approximate solution may not be 0. All this can be avoided if we assume that the solution is unique and that the approximate solutions u_n obtained by taking the minimum to be 0 have the property that Lu_n are bounded in $C^1(D)$.

The energy norm is equivalent to the norm of Sobolev space H^1. So, Theorem 2 also tells us about H^1-convergence. Similarly, Theorem 1 tells us about H^2-convergence because $\|Lu\|_{L^2}$ is equivalent to $\|u\|_{H^2}$.

Acknowledgments. The author wishes to thank A. Bensoussan and B. V. Singbal for their help and encouragement during the preparation of this article.

He is also grateful to Professor Abdus Salam, the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

References

[1] R. Courant and D. Hilbert, *Methods of mathematical physics*, Vol. II, Interscience Publishers, New York 1962.

[2] L. V. Kantorovich, *Functional analysis and applied mathematics*, Usp. Mat. Nauk, Vol. 3, No. 28 (1948). (English translation by Nat. Bur. Standards, N.B.S. Rep. 1509 (1952).)

[3] — and G. P. Akilov, *Functional analysis in normed spaces*, Fizmatgiz, Moscow 1959. (English translation by Pergamon Press, Oxford 1964.)

[4] — and V. I. Krylov, *Approximate methods of higher analysis* (English translation by P. Noordhoff, Ltd., Groningen 1958).

[5] M. V. Keldysh, *On B. G. Galerkin's method for the solution of boundary problems*, Izv. Akad. Nauk USSR, Ser. Mat. 6 (1942).

[6] M. A. Krasnosel'skii, *Topological methods in the theory of non-linear integral equations*, Gostekhizdat, 1956. (English translation by Pergamon Press, Oxford 1964.)

[7] S. G. Mikhlin, *On the convergence of Galerkin's method*, Dokl. Akad. Nauk SSSR, Vol. 61, No. 2 (1948).

[8] — *Variational methods in mathematical physics*, Gostekhizdat, 1957. (English translation by Pergamon Press, Oxford 1964.)

[9] G. Vainikko, *The convergence of the collocation method for non-linear differential equations*, Zh. Vychisl. Mat. Mat. Fiz. Vol. 6, No. 1 (1966), p. 35–42. (English translation in USSR, Comp. Math. and Math. Phys. Vol 6, No. 1, (1966), p. 47–58.)

INTERNATIONAL ATOMIC ENERGY AGENCY
and
UNITED NATIONS EDUCATIONAL SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

Reçu par la Rédaction le 2. 12. 1974