Japan J. Appl. Math., 5 (1988), 313–332

A Numerical Approach to the Proof of Existence of Solutions for Elliptic Problems

Mitsuhiro T. NAKAO

Faculty of Science, Kyushu University 33, Hakozaki, Fukuoka 812, Japan

Received July 27, 1987

In this paper, we describe a method which proves by computers the existence of weak solutions for linear elliptic boundary value problems of second order. It is shown that we can constitute the computing procedures to verify the existence, uniqueness and inclusion set of a solution based on Schauder's fixed point theorem. Using the finite element approximations for some simple Poisson's equations and the results of error estimates, we generate iteratively a set sequence composed of functions and attempt to construct automatically the set including the exact solution. Further, the conditions of verifiability by this method are considered and some numerical examples of verification are presented.

Key words: boundary value problems, finite element method, error estimates, fixed point theorem

§ 1. Introduction

In recent years, several techniques have been developed to use computers in proving theorems in analysis such as existence and uniqueness of solutions for functional equations. Some of them have partially worked upon the integral equations, ordinary differential equations and the special functional equations ([5], [8]). It seems, however, that no such attempts have resulted in success for partial differential equations up to now. In this paper, we describe a method which automatically proves the existence of weak solutions for elliptic boundary value problems by computers. The main task consists of the computing procedures to verify the existence, uniqueness and inclusion set of a solution based on Schauder's fixed point theorem. Using the finite element approximations for some simple Poisson's equations and the error estimates, we generate iteratively a set sequence composed of functions and attempt to construct automatically the set including the exact solution. Further, the conditions of verifiability by this method are considered and some numerical examples of verification are presented.

This report is an initial investigation and only one case study, so there are many difficulties to overcome for construction of the general techniques which make applications possible to more broad ranges. However, the author believes that the study in this direction will open up a new approach by numerical analysis in the field of the existence theory of solutions for various partial differential equations appeared

in mathematical analysis.

The outline of this paper is as follows. In §2, we formulate the boundary value problem as the fixed point equation. And in §3, the concepts of rounding and rounding error are introduced. The former implies the operation which maps the functions in an infinite dimensional space into an appropriate finite element subspace by the orthogonal projection. The latter corresponds to the error followed by the rounding which is obtained according to the results of error estimates for the projection. Also using these concepts, we present the conditions for verification of the existence, uniqueness and the inclusion set of the solution. §4 contains the concrete and detailed algorithms to generate the set of functions in computers satisfying the verification conditions. In §5, we consider about the verifiability, that is, we prove some propositions which suggests, under certain assumptions, the normal completion of the verification process defined in the previous section. Furthermore, in §6, we extend our techniques to the simultaneous verification of a set of equations with interval variable coefficients instead of a single equation. Finally, some numerical results of verifications are presented for one and two dimensional cases in §7.

§2. Problem and the Fixed Point Formulation

Consider the following elliptic boundary value problem of second order:

$$\begin{cases} \Delta u + b \cdot \nabla u + cu = -f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where Ω is a bounded convex domain in R^n, $1 \le n \le 3$, with piecewise smooth boundary and $b=(b_i)$, $1 \le i \le n$. We assume that $b_i \in W^1_\infty(\Omega)$, $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$, where $W^1_\infty(\Omega)$ denotes the usual L^∞-Sobolev space of first order on Ω. Then (2.1) can be rewritten in weak form as: find $u \in H^1_0(\Omega)$ such that

$$(\nabla u, \nabla \phi) = (b \cdot \nabla u + cu, \phi) + (f, \phi), \qquad \phi \in H^1_0(\Omega), \tag{2.2}$$

where $(\cdot, \cdot)$ implies L^2-inner product on Ω and $H^1_0(\Omega)$ denotes the L^2-Sobolev space of order 1 but we adopt the inner product on $H^1_0(\Omega)$ by $\langle \phi, \psi \rangle \equiv (\nabla \phi, \nabla \psi)$ and the associated norm is denoted by $\|\phi\|^2_{H^1_0} = \langle \phi, \phi \rangle$. From now on, we will suppress the symbol Ω in $H^1_0(\Omega)$ and $L^\infty(\Omega)$ etc., and simply denote by H^1_0 and L^∞, respectively. As well known, we can rewrite (2.1) by the following fixed point form.

$$u = Au + F, \tag{2.3}$$

where the map $A: H^1_0 \to H^1_0$ is a compact operator satisfying

$$\langle Au, \phi \rangle = (b \cdot \nabla u + cu, \phi), \qquad \phi \in H^1_0,$$

and F is an element of H^1_0 such that $\langle F, \phi \rangle = (f, \phi)$ for all $\phi \in H^1_0$.

§ 3. Rounding and Verification Conditions

In order to handle the functions, which are in the infinite dimensional space H_0^1, on computers, we attempt to make round them into an appropriate finite dimensional subspace.

We now provide a finite element subspace S_h of $H_0^1(\Omega)$ with the following approximation property. For each $u \in H_0^1 \cap H^2$,

$$\inf_{\chi \in S_h} \|u-\chi\|_{H_0^1} \leq C_1 h |u|_{H^2}, \tag{3.1}$$

where h is a parameter with $0<h<1$, C_1 is a positive constant independent of u and h, and $|u|_{H^2}$ implies the semi-norm of u on $H^2(\Omega)$ defined by

$$|u|_{H^2}^2 \equiv \sum_{i,j=1}^{n} \left\| \frac{\partial^2 u}{\partial x_i \partial x_j} \right\|_{L^2(\Omega)}^2 .$$

It is well known that the estimates (3.1) is valid for many finite element spaces which consist of piecewise linear polynomials on each element T_h of Ω, where $\Omega = \bigcup T_h$, (see e.g. [1]).

Now, for each $u \in H_0^1 \cap H^2$, we define the rounding $R(u) \in S_h$ by

$$(\nabla(u-R(u)), \nabla v) = 0, \qquad v \in S_h . \tag{3.2}$$

$R(u)$ is the so-called H_0^1-projection of u into S_h, so we name it H_0^1-rounding, which can be regarded as a kind of problem dependent roundings by Kaucher and Miranker [5]. By the basic error estimates using the property (3.1), we have the following H^1-error.

$$\|u-R(u)\|_{H_0^1} \leq C_1 h |u|_{H^2} . \tag{3.3}$$

Further, it easily follows by the well-known Aubin-Nitsche technique, that we obtain L^2-error, using another constant C_1', as

$$\|u-R(u)\|_{L^2} \leq C_1' h^2 |u|_{H^2} . \tag{3.4}$$

Based upon the estimates (3.3) and (3.4) we define a subset $RE(u)$ of H_0^1 which is called the rounding error of u, corresponding to the usual round off error, by

$$RE(u) = \{\phi \in H_0^1;\ \|\phi\|_{H_0^1} \leq C_1 h |u|_{H^2} \text{ and } \|\phi\|_{L^2} \leq C_1' h^2 |u|_{H^2}\} . \tag{3.5}$$

From the fact that $\{\phi_n\} \to \phi$ in H_0^1 also implies $\{\phi_n\} \to \phi$ in L^2, $RE(u)$ is a bounded, convex and closed subset in H_0^1. And, clearly, we have $u \in R(u) + RE(u)$.

Next, for each family of functions $U \subset H_0^1 \cap H^2$ rounding $R(U)$ and rounding error $RE(U)$ are defined as

$$R(U) = \bigcup_{u \in U} R(u) \tag{3.6}$$

and

$$RE(U)=\bigcup_{u\in U} RE(u)\,, \tag{3.7}$$

respectively. Then, we have

$$U\subset R(U)+RE(U)\,. \tag{3.8}$$

Now we denote the closure and the interior of set S by $\bar{S}$ and $\mathring{S}$, respectively. Then let $S_1\mathring{\subset}S_2$ imply $\bar{S}_1\subset\mathring{S}_2$. The following lemma is the direct consequence of (3.8) and Schauder's fixed point theorem with some additional considerations on the uniqueness for linear case (e.g. [5]).

LEMMA 1. *Let U be a bounded convex and closed subset in H_0^1 such that*

$$R(AU+F)+RE(AU+F)\mathring{\subset}U\,. \tag{3.9}$$

Then, there exists a unique solution u for (2.3) *in U.*

§4. Verification Procedures by Computers

In the present section, we consider about the fully automatic generation of the set U satisfying (3.9) on the digital computer. Let $\{\phi_j\}$, $1\le j\le M$, be a basis for S_h. In the following arguments from now on, we will often use the same symbol to denote a set of functions as well as each function belonging to the set. The reason why we adopt such notations is to make the essentials of arguments more clearly owing to avoid the complicated use of various symbols. We believe that it causes no troubles by any confusions of symbols.

First, we generate a sequence of sets $\{U^{(i)}\}$, $i=0, 1, \cdots$, which consists of the subsets in H_0^1 as the following manner. Let R^+ denote the set of nonnegative real numbers. For $\alpha\in R^+$ we set

$$[\alpha]\equiv\{\phi\in H_0^1;\ \|\phi\|_{H_0^1}\le\alpha,\ \|\phi\|_{L^2}\le C_1'/C_1 h\alpha\}\,.$$

For $i=0$, we choose appropriate initial value $u_h^{(0)}\in S_h$ and $\alpha_0\in R^+$, and define $U^{(0)}\subseteq H_0^1$ by

$$U^{(0)}=u_h^{(0)}+[\alpha_0]\,. \tag{4.1}$$

Usually, $u_h^{(0)}$ will be determined as the following element in S_h which corresponds to the Galerkin approximation for (2.1).

$$(\nabla u_h^{(0)},\nabla\phi_j)=(b\cdot\nabla u_h^{(0)}+cu_h^{(0)}+f,\phi_j)\,,\qquad 1\le j\le M\,. \tag{4.2}$$

And the standard selection for α_0 will be $\alpha_0=0$. In order to define $U^{(i)}$ for $i\ge 1$, we need some additional definitions. Let $U=\sum_{j=1}^{M} A_j\phi_j$ be a linear combination of $\{\phi_j\}$ with interval coefficients A_j, $1\le j\le M$, which is defined as an element of the power set 2^{S_h} in the following sense:

$$\sum_{j=1}^{M} A_j\phi_j=\left\{\sum_{j=1}^{M} a_j\phi_j;\ a_j\in A_j,\ 1\le j\le M\right\}.$$

For $U=\sum_{j=1}^{M} A_j\phi_j$ and $\alpha\in R^+$, we choose $\tilde{U}=\sum_{j=1}^{M}\tilde{A}_j\phi_j$ and $\tilde{\alpha}\in R^+$ satisfying

$$(4.3)\quad (\nabla\tilde{U},\nabla\phi_j)=(b\cdot\nabla U+cU+f,\phi_j)+[-1,1]Ch\alpha\|-\nabla(b\phi_j)+c\phi_j\|_{L^2},\quad 1\le j\le M,$$

and

$$(4.4)\qquad \tilde{\alpha}=Ch(\|b\cdot\nabla U+cU+f\|_{L^2}+(\|b\|_{L^\infty}+Ch\|c\|_{L^\infty})\alpha),$$

respectively. Here, (4.3) implies that $\tilde{U}$ is determined by the solution, an interval vector, of a linear system of equations with the interval right hand side. Besides, $\|\cdot\|_{L^2}$ in (4.4) means the supremum of norms for all functions in U. Further the positive constant C implies that $C=C_1'/C_1=C_1C_2$, where C_1 is the same constant in (3.1) and C_2 is given by the regularity estimates (4.5) below for the following boundary value problem.

$$\begin{cases}-\Delta\phi=\psi & \text{in } \Omega,\\ \phi=0 & \text{on } \partial\Omega,\end{cases}$$

where $\psi\in L^2(\Omega)$. Then, we have

$$(4.5)\qquad |\phi|_{H^2}\le C_2\|\psi\|_{L^2}.$$

Now, let $\mathfrak{S}_I$ denote the set of all linear combinations of $\{\phi_j\}$ with interval coefficients. Using (4.3) and (4.4), we define a map T: $\mathfrak{S}_I\times R^+\to\mathfrak{S}_I\times R^+$ by

$$(4.6)\qquad T(U,\alpha)=(\tilde{U},\tilde{\alpha}).$$

We now define $(u_h^{(i)},\alpha_i)$ for $i\ge 1$

$$(4.7)\qquad (u_h^{(i)},\alpha_i)=T(u_h^{(i-1)},\alpha_{i-1})$$

and let

$$(4.8)\qquad U^{(i)}=u_h^{(i)}+[\alpha_i].$$

Notice that, when we choose $u_h^{(0)}$ satisfying (4.2) and $\alpha_0=0$, $\{U^{(i)}\}$ forms a monotone increasing sequence, i.e., $U^{(0)}\subsetneqq U^{(1)}\subsetneqq\cdots$. We have the following fundamental property for $\{U^{(i)}\}$.

LEMMA 2. *For the iterative sequences* $\{(u_h^{(i)},\alpha_i)\}$ and $\{U^{(i)}\}$ *defined by* (4.7) *and* (4.8), *respectively, with any initial values* $(u_h^{(0)},\alpha_0)$, *it is valid that*

$$R(AU^{(i-1)}+F)\subset u_h^{(i)},$$

$$RE(AU^{(i-1)}+F)\subset[\alpha_i]$$

and also

$$AU^{(i-1)}+F\subset U^{(i)}.$$

Proof. Consider the set of Poisson's equations

$$\begin{cases} -\Delta u^{(i)}=b\cdot\nabla U^{(i-1)}+cU^{(i-1)}+f & \text{in } \Omega, \\ u^{(i)}=0 & \text{on } \partial\Omega. \end{cases} \tag{4.9}$$

Here, as previously stated, we have used the same symbol $u^{(i)}$ for the set of functions as well as for each function in it. Let $\tilde{u}_h^{(i)}$ be Galerkin approximations for $u^{(i)}$. Then, for each ϕ_j, from integration by parts and the definition of $[\alpha_{i-1}]$, we have

$$\begin{aligned}(\nabla\tilde{u}_h^{(i)}, \nabla\phi_j)&\in(b\cdot\nabla u_h^{(i-1)}+cu_h^{(i-1)}+f, \phi_j)+(b\cdot\nabla[\alpha_{i-1}]+c[\alpha_{i-1}], \phi_j)\\ &=(b\cdot\nabla u_h^{(i-1)}+cu_h^{(i-1)}+f, \phi_j)+([\alpha_{i-1}], -\nabla(b\phi_j)+c\phi_j)\\ &\subset(b\cdot\nabla u_h^{(i-1)}+cu_h^{(i-1)}+f, \phi_j)+[-1, 1]Ch\alpha_{i-1}\|-\nabla(b\phi_j)+c\phi_j\|_{L^2}.\end{aligned}$$

Thus, noting that $u^{(i)}=AU^{(i-1)}+F$ and $\tilde{u}_h^{(i)}=R(u^{(i)})$, we obtain, by (4.3) and (4.7)

$$R(AU^{(i-1)}+F)=\tilde{u}_h^{(i)}\subset u_h^{(i)}.$$

Next, from (3.3), (3.4) and the estimates (4.5) for the problem (4.9), it follows that

$$\begin{aligned}\|u^{(i)}-R(u^{(i)})\|_{H_0^1}&\le C_1h\,|u^{(i)}|_{H^2}\\ &\le Ch(\|b\cdot\nabla U^{(i-1)}+cU^{(i-1)}+f\|_{L^2})\\ &\le Ch(\|b\cdot\nabla u_h^{(i-1)}+cu_h^{(i-1)}+f\|_{L^2}+\|b\cdot\nabla[\alpha_{i-1}]+c[\alpha_{i-1}]\|_{L^2})\\ &\le Ch(\|b\cdot\nabla u_h^{(i-1)}+cu_h^{(i-1)}+f\|_{L^2}+(\|b\|_{L^\infty}+Ch\|c\|_{L^\infty})\alpha_{i-1})\\ &=\alpha_i.\end{aligned}$$

Hence, we have

$$RE(AU^{(i-1)}+F)=RE(u^{(i)})\subset[\alpha_i].$$

Since the last inclusion can be directly derived by (3.8), we complete the proof.

Notice that if we can determine the constants C_1 and C_2, and also can estimate in computers the values of integrations and the maxima of functions occurred in (4.3) and (4.4), then the iterative sequence $U^{(i)}=u_h^{(i)}+[\alpha_i]$, or the sequence including it, can be automatically generated.

Next, we consider the problem of the verification condition based upon the iteration $\{U^{(i)}\}$. Since $(u_h^{(i)}, \alpha_i)\in\mathfrak{S}_I\times R^+$, the convergence problem for $\{U^{(i)}\}$ is reduced to the finite dimensional case. We now suppose that $\{U^{(i)}\}$ converges to $U^{(\infty)}$ as $i\to\infty$. Then by Lemma 2 and Schauder's fixed point theorem, the existence of a solution for (2.3) is verifed in $U^{(\infty)}$. Furthermore, it is also expected that the operator $A(\cdot)+F$ is contractive on a neighborhood $U_\varepsilon^{(\infty)}$ of $U^{(\infty)}$ in the sense of $A(U_\varepsilon^{(\infty)})+F\mathring{\subset}U_\varepsilon^{(\infty)}$. In that case, by Lemma 1 the unique solution can be found in $U_\varepsilon^{(\infty)}$.

We now provide a computer algorithm to decide the convergence of $\{U^{(i)}\}$ and validate the contractivity near $U^{(\infty)}$ as follows. For $u_h^{(i)}=\sum_{j=1}^{M} A_j^{(i)}\phi_j$, where $A_j^{(i)}=[\underline{A}_j^{(i)}, \bar{A}_j^{(i)}]$, $1\le j\le M$, let

$$\|u_h^{(i)}-u_h^{(k)}\|=\max_{1\le j\le M}\{|\underline{A}_j^{(i)}-\underline{A}_j^{(k)}|, |\bar{A}_j^{(i)}-\bar{A}_j^{(k)}|\}.$$

If, for sufficiently small $\varepsilon>0$ and a positive integer N, we attain estimates

$$\|u_h^{(N)}-u_h^{(N-1)}\|<\varepsilon \quad \text{and} \quad |\alpha_N-\alpha_{N-1}|<\varepsilon, \tag{4.10}$$

then stop the iteration, that is, we regard (4.10) as a suggestion that $\{U^{(i)}\}$ converges and that $U^{(N)}$ is close to the limit set $U^{(\infty)}$. Next, for an appropriate $\delta>0$, set

$$\tilde{u}_h^{(N)}=\sum_{j=1}^{M}\tilde{A}_j^{(N)}\phi_j \quad \text{and} \quad \tilde{\alpha}_N=\alpha_N+\delta, \tag{4.11}$$

where $\tilde{A}_j^{(N)}=[\underline{A}_j^{(N)}-\delta, \bar{A}_j^{(N)}+\delta]$, $1\le j\le M$. The procedure (4.11) is called δ-extension of $U^{(N)}$. Then, we compute (u_h, α) by

$$(u_h, \alpha)=T(\tilde{u}_h^{(N)}, \tilde{\alpha}_N). \tag{4.12}$$

From the arguments described up to now, we obtain the following conclusion.

THEOREM 1. *For (u_h, α) defined by* (4.12), *if*

$$u_h\subset\tilde{u}_h^{(N)} \quad \textit{and} \quad \alpha<\tilde{\alpha}_N, \tag{4.13}$$

then there exists a unique solution u for (2.3) *in $u_h+[\alpha]$. Here, $u_h\subset\tilde{u}_h^{(N)}$ implies that each coefficient interval in u_h is included in the corresponding interval in $\tilde{u}_h^{(N)}$.*

REMARK 1. While the determinations of $\{(u_h^{(i)}, \alpha_i)\}$ for $i=0, 1, \cdots, N$ may contain the indefinite round off errors, the calculation of (4.12) has to be done rigorously by the use of the strict interval arithmetic.

§5. Convergence of $\{U^{(i)}\}$ and Verifiability Conditions

In this section, we consider the convergence condition for $\{U^{(i)}\}$ defined in the previous section and the attainability of the verification condition (4.13) in Theorem 1.

Now, for each $u\in H_0^1$ we define $\tilde{A}u\in S_h$ by

$$\langle\tilde{A}u, \phi_j\rangle=(b\cdot\nabla u+cu, \phi_j), \qquad 1\le j\le M. \tag{5.1}$$

We also introduce an operator $\kappa: R^+\to 2^{S_h}$ such that, for $\alpha\in R^+$,

$$\langle\kappa\alpha, \phi_j\rangle\in[-1, 1]Ch\alpha\|-\nabla(b\phi_j)+c\phi_j\|_{L^2}, \qquad 1\le j\le M. \tag{5.2}$$

That is, $\psi\in\kappa\alpha$ is an element of S_h for which $\langle\psi, \phi_j\rangle$ is included in the right hand side of (5.2). Then, κ depends on the selection of the basis $\{\phi_j\}$. As the topology of 2^{S_h} we

define the following Hausdorff metric $D(\cdot, \cdot)$ based on H_0^1-norm. For $U_h, V_h \in 2^{S_h}$,

$$D(U_h, V_h) = \max\left\{\sup_{\phi \in U_h} d(\phi, V_h), \sup_{\psi \in V_h} d(\psi, U_h)\right\}, \tag{5.3}$$

where $d(\phi, V_h) = \inf_{\psi \in V_h} \|\phi - \psi\|_{H_0^1}$.

We now newly define an iterative sequence $\{(u_h^{(i)}, \alpha_i)\}$ which is essentially close to the one introduced in §4 but slightly different. Let $u_h^{(0)} \in S_h$ and $\alpha_0 \in R^+$ be appropriate initial values. For $i \geq 1$, $(u_h^{(i)}, \alpha_i)$ is iteratively defined as $u_h^{(i)} \in 2^{S_h}$ and $\alpha_i \in R^+$ such that, denoting the H_0^1-projection of F into S_h by $\tilde{F}$,

$$u_h^{(i)} = \tilde{A} u_h^{(i-1)} + \kappa \alpha_{i-1} + \tilde{F}, \tag{5.4}$$

$$\alpha_i = \tilde{C}_1 h \|u_h^{(i-1)}\|_1 + \tilde{C}_2 h \alpha_{i-1} + \tilde{C}_3 h \|f\|_{L^2}. \tag{5.5}$$

Here, $\tilde{C}_1$, $\tilde{C}_2$ and $\tilde{C}_3$ are constants independent of h, and

$$\|u_h^{(i-1)}\|_1 \equiv \sup_{\phi \in u_h^{(i-1)}} \|\phi\|_{H_0^1}. \tag{5.6}$$

Let $\tilde{T}: 2^{S_h} \times R^+ \to 2^{S_h} \times R^+$ be the map defined by (5.4), (5.5) which is similar to the map T in §4. Then, we can write

$$(u_h^{(i)}, \alpha_i) = \tilde{T}(u_h^{(i-1)}, \alpha_{i-1}).$$

Note that $u_h^{(i)}$ determined by (5.4), in general, cannot be represented as a linear combination of $\{\phi_j\}$ with interval coefficients. We can say that $u_h^{(i)}$ in the previous section is an extension to the minimum set in $\mathfrak{S}_I$ which includes the above $u_h^{(i)}$. It will be also easy to guess the meaning of each $\tilde{C}_i$ in (5.5) from the right hand side of (4.4). As shown below, when we denote the spectral radius of the operator A defined in (2.3) by $r(A)$, the sequence $(u_h^{(i)}, \alpha_i)$, $i = 0, 1, \cdots$, converges to the unique fixed point of $\tilde{T}$ provided $r(A) < 1$. We begin with the following lemma.

LEMMA 3. *Let $A: H_0^1 \to H_0^1$ and $\tilde{A}: H_0^1 \to S_h$ be operators defined by* (2.3) *and* (5.1), *respectively. If $r(A) < 1$ then, for sufficiently small h, $r(\tilde{A}) < 1$ also holds.*

Proof. As well known, for arbitrary $\varepsilon > 0$, there exists a norm $\|\cdot\|_\varepsilon$ on H_0^1 which is equivalent to the norm $\|\cdot\|_{H_0^1}$ and satisfies $\|A\|_\varepsilon < r(A) + \varepsilon$. Therefore, we can now take sufficiently small ε such that $\|A\|_\varepsilon < 1$. Notice that, for each $u \in H_0^1$, $\phi \equiv Au \in H_0^1 \cap H^2$ is a solution for the Poisson's equation of the weak form

$$\langle \phi, \psi \rangle = (b \cdot \nabla u + cu, \psi), \qquad \psi \in H_0^1.$$

On the other hand, $\phi_h = \tilde{A} u \in S_h$ satisfies

$$\langle \phi_h, v \rangle = (b \cdot \nabla u + cu, v), \qquad v \in S_h.$$

Hence, it follows that ϕ_h is the H_0^1-projection of ϕ. Therefore, using the constants defined previously, we obtain

$$\|(A-\tilde{A})u\|_{H_0^1} \leq Ch\|b\cdot \nabla u + cu\|_{L^2}.$$

Thus, by the equivalency of norms, for a constant $\tilde{C}$

$$\|A-\tilde{A}\|_\varepsilon \leq \tilde{C}h.$$

From this fact, for sufficiently small h, we have

$$\|\tilde{A}\|_\varepsilon \leq \|A-\tilde{A}\|_\varepsilon + \|A\|_\varepsilon < 1,$$

which concludes the proof.

The following theorem is the main result on the convergence of $\{(u_h^{(i)}, \alpha_i)\}$.

THEOREM 2. *Assume that* (i) $\{\phi_j\}$, $1 \leq j \leq M$ *is an orthogonal basis of* S_h, (ii) *there exists a positive constant* $\hat{C}$, *independent of* h, *such that* $Mh^n \leq \hat{C}$ *and that* (iii) $r(A) < 1$. *Then, for sufficiently small* h, $\{(u_h^{(i)}, \alpha_i)\}$ *defined by* (5.4), (5.5) *converges to a unique limit* (u_h, α) *in* $2^{S_h} \times R^+$, *with an arbitrary initial value* $(u_h^{(0)}, \alpha_0)$, *which is also a unique fixed point of* $\tilde{T}$.

Notice that property (ii) will be always satisfied if S_h is the usual piecewise linear finite element space with quasi-uniform partition i.e. the measure of each element is bounded below by Ch^n, where C is independent of h.

Proof. Let $\|\cdot\|_\varepsilon$ denote the same as in the proof of Lemma 3, and also denote the corresponding Hausdorff metric on 2^{S_h} by D_ε and the usual distance by d_ε, respectively. We show that $\{(u_h^{(i)}, \alpha_i)\}$ is the Cauchy sequence in $2^{S_h} \times R^+$.

First, observe that

$$D_\varepsilon(u_h^{(i)}, u_h^{(i-1)}) = \max\left\{\sup_{\phi \in u_h^{(i)}} d_\varepsilon(\phi, u_h^{(i-1)}), \sup_{\psi \in u_h^{(i-1)}} d_\varepsilon(\psi, u_h^{(i)})\right\}. \tag{5.7}$$

We now estimate the former term in $\{\ \}$ of (5.7). From (5.4), we have

$$\begin{aligned}\sup_{\phi \in u_h^{(i)}} d_\varepsilon(\phi, u_h^{(i-1)}) &= \sup_{(i-1)} \inf_{(i-2)} \|(\tilde{A}\hat{u}_h^{(i-1)} + \hat{\kappa}\alpha_{i-1}) - (\tilde{A}\hat{u}_h^{(i-2)} + \hat{\kappa}\alpha_{i-2})\|_\varepsilon \\ &\leq \sup_{(i-1)} \inf_{(i-2)} (\|\tilde{A}(\hat{u}_h^{(i-1)} - \hat{u}_h^{(i-2)})\|_\varepsilon + \|\hat{\kappa}\alpha_{i-1} - \hat{\kappa}\alpha_{i-2}\|_\varepsilon),\end{aligned}$$

where $(i-1)$ and $(i-2)$ imply that

$$(i-1) \equiv \begin{cases} \hat{u}_h^{(i-1)} \in u_h^{(i-1)} \\ \hat{\kappa}\alpha_{i-1} \in \kappa\alpha_{i-1} \end{cases} \quad \text{and} \quad (i-2) \equiv \begin{cases} \hat{u}_h^{(i-2)} \in u_h^{(i-2)} \\ \hat{\kappa}\alpha_{i-2} \in \kappa\alpha_{i-2} \end{cases}, \quad \text{respectively}.$$

By noting that we may take the infimum independently with each other in the last right hand side, we get

$$\begin{aligned}\sup_{\phi \in u_h^{(i)}} d_\varepsilon(\phi, u^{(i-1)}) &\leq \sup_{(i-1)} \left(\inf_{(i-2)} \|\tilde{A}(\hat{u}_h^{(i-1)} - \hat{u}_h^{(i-2)})\|_\varepsilon + \inf_{(i-2)} \|\hat{\kappa}\alpha_{i-1} - \hat{\kappa}\alpha_{i-2}\|_\varepsilon) \right) \\ &\leq \|\tilde{A}\|_\varepsilon D_\varepsilon(u_h^{(i-2)}, u_h^{(i-2)}) + D_\varepsilon(\kappa\alpha_{i-1}, \kappa\alpha_{i-2}).\end{aligned}$$

Since the estimates of the latter half in (5.7) are also similar, we obtain

$$D_\varepsilon(u_h^{(i)}, u_h^{(i-1)}) \le \|\tilde{A}\|_\varepsilon D_\varepsilon(u_h^{(i-1)}, u_h^{(i-2)}) + D_\varepsilon(\kappa\alpha_{i-1}, \kappa\alpha_{i-2}) . \tag{5.8}$$

Next, we estimate $D_\varepsilon(\kappa\alpha_{i-1}, \kappa\alpha_{i-2})$ in (5.8). We may assume $\alpha_{i-1} > \alpha_{i-2}$ without loss of generality. Then, by the fact that $\kappa\alpha_{i-2} \subset \kappa\alpha_{i-1}$

$$D_\varepsilon(\kappa\alpha_{i-1}, \kappa\alpha_{i-2}) = \sup_{(i-1)} \inf_{(i-2)} \|\hat{\kappa}\alpha_{i-1} - \hat{\kappa}\alpha_{i-2}\|_\varepsilon , \tag{5.9}$$

where $(i-1)$ and $(i-2)$ mean $\hat{\kappa}\alpha_{i-1} \in \kappa\alpha_{i-1} - \kappa\alpha_{i-2}$ and $\hat{\kappa}\alpha_{i-2} \in \kappa\alpha_{i-2}$, respectively. Now taking notice that $\{\phi_j\}$ is orthogonal, we can easily deduce the following estimates from the definition (5.2)

$$\begin{aligned} \sup_{(i-1)} \inf_{(i-2)} \|\hat{\kappa}\alpha_{i-1} - \hat{\kappa}\alpha_{i-2}\|_{H_0^1}^2 &= \sup_{(i-1)} \inf_{(i-2)} \sum_{j=1}^{M} \frac{|\langle \hat{\kappa}\alpha_{i-1} - \hat{\kappa}\alpha_{i-2}, \phi_j \rangle|^2}{\|\phi_j\|_{H_0^1}^2} \\ &\le C'^2 h^2 |\alpha_{i-1} - \alpha_{i-2}|^2 M , \end{aligned} \tag{5.10}$$

where C' is a positive constant. Therefore, using the equivalency of $\|\cdot\|_\varepsilon$ and $\|\cdot\|_{H_0^1}$, from (5.9) and (5.10) we obtain

$$D_\varepsilon(\kappa\alpha_{i-1}, \kappa\alpha_{i-2}) \le C'' h |\alpha_{i-1} - \alpha_{i-2}| \sqrt{M} .$$

Thus, we have by (5.8)

$$D_\varepsilon(u_h^{(i)}, u_h^{(i-1)}) \le \|\tilde{A}\|_\varepsilon D_\varepsilon(u_h^{(i-1)}, u_h^{(i-2)}) + C'' h \sqrt{M} |\alpha_{i-1} - \alpha_{i-2}| . \tag{5.11}$$

Now, as before, let $D(\cdot, \cdot)$ denote the Hausdorff metric based on the norm $\|\cdot\|_{H_0^1}$. Then, using the triangle inequality for $D(\cdot, \cdot)$ and the norm equivalency, we have by (5.5)

$$\begin{aligned} |\alpha_i - \alpha_{i-1}| &\le \tilde{C}_1 h |\|u_h^{(i-1)}\|_1 - \|u_h^{(i-2)}\|_1| + \tilde{C}_2 h |\alpha_{i-1} - \alpha_{i-2}| \\ &= \tilde{C}_1 h |D(u_h^{(i-1)}, \{0\}) - D(u_h^{(i-2)}, \{0\})| \\ &\quad + \tilde{C}_2 h |\alpha_{i-1} - \alpha_{i-2}| \\ &\le \tilde{C}_1 h D(u_h^{(i-1)}, u_h^{(i-2)}) + \tilde{C}_2 h |\alpha_{i-1} - \alpha_{i-2}| \\ &\le C'_1 h D_\varepsilon(u_h^{(i-1)}, u_h^{(i-2)}) + \tilde{C}_2 h |\alpha_{i-1} - \alpha_{i-2}| . \end{aligned} \tag{5.12}$$

(5.11) and (5.12) can be rewritten as the following matrix form.

$$\begin{bmatrix} D_\varepsilon(u_h^{(i)}, u_h^{(i-1)}) \\ |\alpha_i - \alpha_{i-1}| \end{bmatrix} \le \begin{bmatrix} \|\tilde{A}\|_\varepsilon & C'' h\sqrt{M} \\ C'_1 h & \tilde{C}_2 h \end{bmatrix} \begin{bmatrix} D_\varepsilon(u_h^{(i-1)}, u_h^{(i-2)}) \\ |\alpha_{i-1} - \alpha_{i-2}| \end{bmatrix} . \tag{5.13}$$

When P denotes the 2×2 matrix in the right hand side of (5.13), for the eigen values λ_1, λ_2 of P with $\lambda_1 \ge \lambda_2$, we have, for h such that $\|\tilde{A}\|_\varepsilon \ge \tilde{C}_2 h$,

$$0<\lambda_1=\frac{\|\tilde{A}\|_\varepsilon+\tilde{C}_2h+\sqrt{(\|\tilde{A}\|_\varepsilon-\tilde{C}_2h)^2+4C'_1C''h^2\sqrt{M}}}{2} \tag{5.14}$$
$$\leq\|\tilde{A}\|_\varepsilon+h\sqrt{C'_1C''}\sqrt[4]{M},$$

and

$$\lambda_2=\frac{\|\tilde{A}\|_\varepsilon+\tilde{C}_2h-\sqrt{(\|\tilde{A}\|_\varepsilon-\tilde{C}_2h)^2+4C'_1C''h^2\sqrt{M}}}{2} \tag{5.15}$$
$$\geq\frac{\|\tilde{A}\|_\varepsilon+\tilde{C}_2h-(\|\tilde{A}\|_\varepsilon-\tilde{C}_2h)-2\sqrt{C'_1C''h^2}\sqrt[4]{M}}{2}$$
$$=(\tilde{C}_2-\sqrt{C'_1C''}\sqrt[4]{M})h.$$

By the assumption (ii), when $1\leq n\leq 3$, for sufficiently small h, from (5.14), (5.15) and Lemma 3 we obtain $0<\lambda_1<1$ and $-1<\lambda_2<1$. Since it is easily seen that this fact also follows for $\|\tilde{A}\|_\varepsilon<\tilde{C}_2h$, consequently we have $r(P)<1$, where $r(P)$ is the spectral radius of the matrix P. Combining it with (5.13), by the use of standard techniques, it can be readily seen that $\{(u_h^{(i)}, \alpha_i)\}$ becomes the Cauchy sequence in $2^{S_h}\times R^+$ with the metric $D_\varepsilon(\cdot,\cdot)$ on 2^{S_h}. From the uniform boundedness of $u_h^{(i)}$ in S_h, $\{(u_h^{(i)}, \alpha_i)\}$ belongs to a compact subset in $2^{S_h}\times R^+$. Therefore, it follows that there exists a unique limit (u_h, α) of $\{(u_h^{(i)}, \alpha_i)\}$ which is also a limit under the metric $D(\cdot,\cdot)$. Further, by virtue of (5.13) and $r(A)<1$, (u_h, α) becomes a unique fixed point of $\tilde{T}$. Thus we can complete the proof of the theorem.

Next, we consider about the attainability of the verification conditions in Theorem 1. For $V_h\subset S_h$ and $\alpha\in R^+$, let $\tilde{T}(V_h, \alpha)=(\tilde{V}_h, \tilde{\alpha})$. When it holds that $\tilde{V}_h\subseteq V_h$ and $\tilde{\alpha}<\alpha$, T is said to be strictly inclusive at (V_h, α). We now assume that for $(u_h, \alpha)\in 2^{S_h}\times R^+$, $\tilde{T}(u_h, \alpha)=(u_h, \alpha)$, that is,

$$\begin{cases} u_h=\tilde{A}u_h+\kappa\alpha+\tilde{F} & \text{and} \\ \alpha=\tilde{C}_1h\|u_h\|_1+\tilde{C}_2h\alpha+\tilde{C}_3h\|f\|_{L^2}. \end{cases} \tag{5.16}$$

Let $\Delta u_h\subset S_h$ and $\delta_\alpha\in R^+$ be variations of u_h and α, respectively, such that $\sup_{\phi\in\Delta u_h}\|\phi\|_\varepsilon\leq\delta_u$, $\delta_u\in R^+$. Then, by (5.16)

$$\tilde{A}(u_h+\Delta u_h)+\kappa(\alpha+\delta_\alpha)+\tilde{F}=u_h+\tilde{A}\Delta u_h+\kappa\delta_\alpha \tag{5.17}$$

and

$$\tilde{C}_1h\|u_h+\Delta u_h\|_1+\tilde{C}_2h(\alpha+\delta_\alpha)+\tilde{C}_3h\|f\|_{L^2} \tag{5.18}$$
$$\leq\alpha+\tilde{C}_1h\|\Delta u_h\|_1+\tilde{C}_2h\delta_\alpha.$$

Taking into account $\|\kappa\delta_\alpha\|_\varepsilon\leq C''h\sqrt{M}\delta_\alpha$, we have

$$\begin{bmatrix} \|A\tilde{\Delta}u_h+\kappa\delta_\alpha\|_\varepsilon \\ \tilde{C}_1 h\|\Delta u_h\|_1+\tilde{C}_2 h\delta_\alpha \end{bmatrix} \leq Pd\,, \tag{5.19}$$

where $d=(\delta_u, \delta_\alpha)^T$ and P is the same 2×2 matrix defined as in the proof of Theorem 2. From (5.17)–(5.19), in order to show that $\tilde{T}$ is strictly inclusive at $(u_h+\Delta u_h, \alpha+\delta_\alpha)$, it is sufficient to prove

$$\|\tilde{A}\|_\varepsilon\delta_u+C''h\sqrt{M}\delta_\alpha<\delta_u \tag{5.20}$$

and

$$C_1'h\delta_u+\tilde{C}_2h\delta_\alpha<\delta_\alpha\,. \tag{5.21}$$

Now let γ be a positive number less than 1. For given δ_u, if we choose δ_α such that

$$\delta_\alpha=\frac{\gamma}{C''h\sqrt{M}}(1-\|\tilde{A}\|_\varepsilon)\delta_u\,,$$

then (5.20) is satisfied. Further, observe that

$$C_1'h\delta_u+\tilde{C}_2h\delta_\alpha=\frac{C_1'C''h\sqrt{M}}{(1-\|\tilde{A}\|_\varepsilon)\gamma}+\tilde{C}_2h\delta_\alpha\,.$$

Hence, under the same assumptions as in Theorem 2, for sufficiently small h, (5.21) also holds. When δ_α is given first, we can also choose corresponding δ_u to satisfy (5.20) and (5.21).

Thus, from the above arguments, we obtain the following conclusion.

THEOREM 3. *Assume the same conditions as in Theorem* 2 *and that* (u_h, α) *is a fixed point of* $\tilde{T}$. *Then, for sufficiently small* h, *there exist two positive constants* δ_u *and* δ_α *such that* $\tilde{T}$ *is strictly inclusive at* $(V_h(u_h;\delta_u), \alpha+\delta_\alpha)$, *where* $V_h(u_h;\delta_u)$ *is a* δ_u*-neighborhood of* u_h *in the sense of*

$$V_h(u_h;\delta_u)=\{\phi\in S_h;\ \phi=\phi_1+\phi_2,\ \phi_1\in u_h,\ \|\phi_2\|_{H_0^1}\leq\delta_u\}\,.$$

Theorem 3 suggests that we can complete the verification process under certain conditions, that is, (4.13) in Theorem 1 is attainable. However, the manner of δ-extension (4.11) does not coincide, in general, with $(\delta_u, \delta_\alpha)$-enlargement of the above. So, even if the stopping criteria (4.10) are satisfied, it is possible that we cannot establish the strictly inclusive relation (4.13). Besides, it is not necessary conditions that δ_u and δ_α must be sufficiently small. If a pair $(\delta_u, \delta_\alpha)$ satisfies (5.20), (5.21), then so for all $(K\delta_u, K\delta_\alpha)$ with arbitrary $K>0$.

§6. For Equations with Interval Coefficients

We attempt to extend the arguments of verification for the equations whose

coefficients are intervals. Here, each coefficient interval implies a bounded function such that the range is included in that interval.

Now consider the problem as before:

$$\begin{cases} \Delta u + b \cdot \nabla u + cu = -f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{6.1}$$

where $b = (b^{(i)})$ and

$$\begin{cases} b^{(i)} \in L^{\infty}(\Omega), & b^{(i)}(x) \in [b_1^{(i)}, b_2^{(i)}], \quad 1 \le i \le n, \\ c \in L^{\infty}(\Omega), & c(x) \in [c_1, c_2], \\ f \in L^{\infty}(\Omega), & f(x) \in [f_1, f_2], \quad \forall x \in \Omega, \end{cases} \tag{6.2}$$

where $b_1^{(i)}$, c_1, f_1 etc. are real constants. Note that each coefficient in (6.1) can be also regarded as the set of functions satisfying (6.2).

Now let A_j, $1 \le j \le M$, be constant intervals, i.e., usual sets of constants. And let ψ, g_j be Riemann integrable functions on Ω such that $\psi(x) \in [a_1, a_2]$ and $g_j(x) \ge 0$, $1 \le j \ge M$. Then the Riemann integral on Ω can be written in the form, under the appropriate subdivisions of Ω into N segments,

$$J \equiv \int_{\Omega} \sum_{j=1}^{M} \psi(x) A_j g_j(x) dx = \lim_{N \to \infty} \sum_{p=1}^{N} \sum_{j=1}^{M} A_j \psi(x_p) g_j(x_p) \Delta x_p .$$

Hence, taking notice of $g_j(x) \ge 0$, we have

$$\begin{aligned} J &= \sum_{j=1}^{M} A_j \lim_{N \to \infty} \sum_{p=1}^{N} \psi(x_p) g_j(x_p) \Delta x_p \\ &\subset \sum_{j=1}^{M} A_j \lim_{N \to \infty} \sum_{p=1}^{N} [a_1, a_2] g_j(x_p) \Delta x_p \\ &= \sum_{j=1}^{M} A_j \lim_{N \to \infty} [a_1, a_2] \sum_{p=1}^{N} g_j(x_p) \Delta x_p \\ &= \sum_{j=1}^{M} A_j [a_1, a_2] \int_{\Omega} g_j(x) dx . \end{aligned} \tag{6.3}$$

Next, let $\Omega = \Omega_j^{(1)} \cup \Omega_j^{(2)}$ such that

$$g_j(x) \begin{cases} \ge 0 & \text{on } \Omega_j^{(1)}, \\ < 0 & \text{on } \Omega_j^{(2)}, \end{cases} \quad 1 \le j \le M .$$

Then it can be easily seen that

$$\int_{\Omega} \sum_{j=1}^{M} \psi(x) A_j g_j(x) dx \subset \sum_{j=1}^{M} A_j \left([a_1, a_2] \int_{\Omega_j^{(1)}} g_j(x) dx + [a_1, a_2] \int_{\Omega_j^{(2)}} g_j(x) dx \right). \tag{6.4}$$

Now by the above consideration, we modify the iterative sequence $\{(u_h^{(i)}, \alpha_i)\}$ in §4 as follows. Suppose that the basis $\{\phi_j\}$, $1 \le j \le M$, of S_h can be taken to satisfy $\phi_j(x) \ge 0$ and that

$$\frac{\partial \phi_j}{\partial x_l} \begin{cases} \ge 0 & \text{on } \Omega_{jl}^{(1)}, \\ < 0 & \text{on } \Omega_{jl}^{(2)}, \end{cases}$$

where $\Omega = \Omega_{jl}^{(1)} \cup \Omega_{jl}^{(2)}$, $1 \le j \le M$, $1 \le l \le n$. We choose $(u_h^{(0)}, \alpha_0)$ appropriately, and for $i \ge 1$ we define $(u_h^{(i)}, \alpha_i) \in \mathfrak{S}_I \times R^+$, where let $u_h^{(i)} = \sum_{j=1}^{M} A_j^{(i)} \phi_j$, by

$$(6.5) \quad (\nabla u_h^{(i)}, \nabla \phi_k) = \sum_{j=1}^{M} A_j^{(i-1)} \{ b \cdot ((\nabla \phi_j, \phi_k)) + [c_1, c_2](\phi_j, \phi_k) \} + [f_1, f_2](\phi_k, 1) + [-1, 1](|b| + Ch|c|)\alpha_{i-1} \|\phi_k\|_{L^2}$$

and

$$(6.6) \quad \alpha_i = Ch(|b| \cdot \|\nabla u_h^{(i-1)}\|_{L^2} + |c| \cdot \|u_h^{(i-1)}\|_{L^2} + |f|(m(\Omega))^{1/2} + (|b| + Ch|c|)\alpha_{i-1}),$$

where

$$b \cdot ((\nabla \phi_j, \phi_k)) = \sum_{l=1}^{n} \left([b_1^{(l)}, b_2^{(l)}] \int_{\Omega_{jl}^{(1)}} \frac{\partial \phi_j}{\partial x_l} \phi_k dx + [b_1^{(l)}, b_2^{(l)}] \int_{\Omega_{jl}^{(2)}} \frac{\partial \phi_j}{\partial x_l} \phi_k dx \right),$$

and $|b| = \max(|b_1^{(i)}|, |b_2^{(i)}|, 1 \le i \le n)$, $|c| = \max(|c_1|, |c_2|)$, $|f| = \max(|f_1|, |f_2|)$, $m(\Omega) = [\text{measure of } \Omega]$ and norms in (6.6) mean as before. Then, the verification procedures are quite analogous to that in §4 and therefore the proposition similar to Theorem 1 holds.

REMARK 2. In order to obtain (6.3), we supposed that the functions are Riemann integrable, however, this assumption is removable by virtue of the fact that $C(\Omega)$ is dense in $L^\infty(\Omega)$ and $L^2(\Omega)$. Further, it would be readily deduced that it is also possible to formulate the verification procedures based on the intervals of values of the norm, e.g. $\|b\|_{H^1} \in [0, L]$ or $\|c\|_{L^2} \in [0, K]$ etc., instead of the ranges of coefficient functions.

§7. Numerical Examples of Verification

We now show some examples of equations which we actually carried out the verifications. In those results, we used the ACRITH which is the software for verified numerical computations developed by IBM corporation based upon the studies [6] and [7]. The ACRITH achieves the reliable computations with high accuracy by

utilizing the following techniques:

- Interval arithmetic
- Automatic verification of results according to Brouwer's fixed point theorem
- Iterative residual correction method.

It contains, for example, evaluations of arithmetic expressions, computations of standard mathematical functions and solutions of systems of linear equations, etc., see [4] for detail. Therefore, for each example below the verification condition (4.13) in Theorem 1 is strictly validated and each floating point number occurred is accurate up to the least significant mantissa.

(i) One dimensional case.

Let $\Omega=(0, 1)$, then (2.1) reduces to the following two point boundary value problem.

$$\begin{cases} u''+bu'+cu=-f, & x\in I=(0, 1), \\ u(0)=u(1)=0. \end{cases} \tag{7.1}$$

For simplicity, we use a uniform partition of I: $x_i=i/L$, $0\le i\le L$, and set $I_i=(x_{i-1}, x_i)$. Then we have $h=1/L$. When $P_1(I_i)$ denotes the set of linear polynomials on I_i, we take S_h as

$$S_h\equiv \mathcal{M}_0^1=\{v\in C(I);\ v|_{I_i}\in P_1(I_i),\ 1\le i\le L,\ v(0)=v(1)=0\}. \tag{7.2}$$

Then clearly $M=\dim S_h=L-1$. Further, it is easily seen, by the use of the interpolating polynomials at each mesh point, that we may take $C_1=1$ in (3.1). We can also choose C_2 in (4.5) as $C_2=1$ for the present case. Since the rounding $R(\phi)$ by (3.2) becomes the interpolation of ϕ at each node x_j (e.g. [2]), we may take $\alpha_i(x_j)=0$ for each iteration step. Therefore, when we have completed the verification, it holds that

$$u(x_j)\in \tilde{u}_h^{(N)}(x_j)=\sum_{k=1}^{L-1}\tilde{A}_k^{(N)}\phi_k(x_j). \tag{7.3}$$

We now choose the basis of S_h as the hat functions illustrated in Fig. 1. Then, by (7.3), we have $u(x_j)\in\tilde{A}_j^{(N)}$, $1\le j\le L-1$, for the solution u of (7.1).

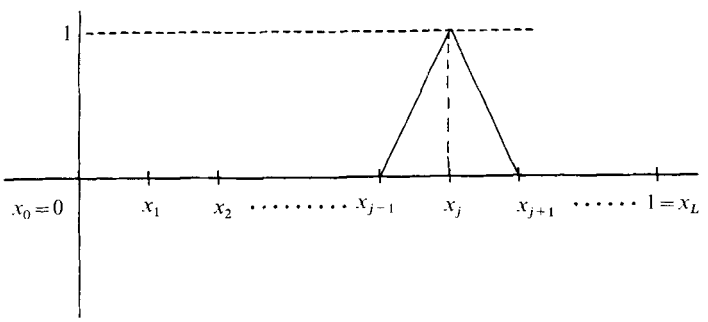

Fig. 1. Basis functions of S_h.

We show an example whose verification normally completed.

Example 1. (Two point boundary value problem)
Problem:

$$\begin{cases} -u'' - \pi u = (\pi - 1)\sin \pi x\,, & x \in I\,, \\ u(0) = u(1) = 0\,. \end{cases} \tag{7.4}$$

Exact solution: $u(x) = \dfrac{1}{\pi}\sin \pi x$

Conditions: Numbers of elements $= 20$, $\dim S_h = 19$
Initial values: $u_h^{(0)} =$ Galerkin approximation (4.2), $\alpha_0 = 0$
Stopping and Extension parameters: $\varepsilon = \delta = 10^{-8}$

Results: Iteration numbers: $N = 15$
L^2-error bound ($= h\tilde{\alpha}_N$): 0.0056608
Coefficient intervals: as in Table 1.

Table 1. $-u'' - \pi u = (\pi - 1)\sin \pi x$

j	$\tilde{A}_j^{(N)} = [\underline{\tilde{A}}_j^{(N)}, \bar{\tilde{A}}_j^{(N)}]$	Exact solution $((1/\pi)\sin \pi x_j)$
1	[0.0475887, 0.0519051]	(0.0497946)
2	[0.0941316, 0.1024062]	(0.0983632)
3	[0.1384494, 0.1502931]	(0.1445096)
4	[0.1794208, 0.1944164]	(0.1870978)
5	[0.2160106, 0.2337163]	(0.2250791)
6	[0.2472951, 0.2672477]	(0.2575181)
7	[0.2724851, 0.2942038]	(0.2836162)
8	[0.2909456, 0.3139358]	(0.3027307)
9	[0.3022112, 0.3259683]	(0.3143910)
10	[0.3059982, 0.3300116]	(0.3183098)
11	[0.3022112, 0.3259683]	(0.3143910)
12	[0.2909456, 0.3139358]	(0.3027307)
13	[0.2724851, 0.2942038]	(0.2836162)
14	[0.2472951, 0.2672477]	(0.2575181)
15	[0.2160106, 0.2337163]	(0.2250791)
16	[0.1794208, 0.1944164]	(0.1870978)
17	[0.1384494, 0.1502931]	(0.1445096)
18	[0.0941316, 0.1024062]	(0.0983632)
19	[0.0475887, 0.0519051]	(0.0497946)

Notice that $(1/\pi)\sin \pi x_j \in \tilde{A}_j^{(N)}$ which follows according to (7.3).

(ii) Two dimensional case.

Let Ω be a rectangular domain in R^2 such that $\Omega = (0, 1) \times (0, 1)$. Also let δ_x: $0 = x_0 < x_1 < \cdots < x_L = 1$ and $\delta_y = \delta_x$. We define the partition δ of Ω by $\delta = \delta_x \otimes \delta_y$. Further, let $S_h = \mathcal{M}_0^1(x) \otimes \mathcal{M}_0^1(y)$, where $\mathcal{M}_0^1(x)$ and $\mathcal{M}_0^1(y)$ are sets of piecewise linear polynomials on (0, 1) defined by (7.2) in the variable x and y, respectively.

We now estimate the constant C_1 in (3.1). For fixed $y \in (0, 1)$ we define a projection $P_x u(\cdot, y)$ of $u \in H_0^1$ into $\mathcal{M}_0^1(x)$ by

$$\left(\frac{\partial}{\partial x}(u(\cdot, y)-P_x u(\cdot, y)), v_x\right)_I = 0, \qquad v \in \mathcal{M}_0^1(x), \tag{7.5}$$

where $(\cdot, \cdot)_I$ implies the inner product on $L^2(I)$ for $I=(0, 1)$. $P_y: H_0^1 \to \mathcal{M}_0^1(y)$ is also defined similarly. Then, clearly $P_y P_x u \in S_h$ and by virtue of (7.5) we have

$$\left\|\frac{\partial}{\partial x}(u-P_y P_x u)\right\|_{L^2}^2 = \left\|\frac{\partial}{\partial x}(u-P_x u)\right\|_{L^2}^2 + \left\|\frac{\partial}{\partial x}(P_x u-P_y P_x u)\right\|_{L^2}^2. \tag{7.6}$$

It follows that, from the well known error estimates for (7.5),

$$\left\|\frac{\partial}{\partial x}(u-P_x u)\right\|_{L^2}^2 \leq h^2 \|u_{xx}\|_{L^2}^2 \tag{7.7}$$

and

$$\left\|\frac{\partial}{\partial x}(P_x u-P_y P_x u)\right\|_{L^2}^2 \leq h^2 \|u_{xy}\|_{L^2}^2. \tag{7.8}$$

(7.6)–(7.8) and the analogous estimates for y-derivative yield

$$\|u-P_y P_x u\|_{H_0^1}^2 \leq h^2 |u|_{H^2}^2, \tag{7.9}$$

which implies that we may take $C_1=1$.

Next, it is seen that from [3], in particular the proof of Theorem 4.3.1.4, we can choose the constant C_2 in (4.5) as $C_2=1$. Following are verified problems for two dimensional case.

Example 2. (Unknown solution)
Problem:

$$\begin{cases} -\Delta u - 20xy \cdot u = (2\pi-1)\sin \pi x \cdot \sin \pi y, & (x, y) \in \Omega \\ u = 0, & (x, y) \in \partial\Omega. \end{cases} \tag{7.10}$$

Exact solution: unknown

Conditions: Numbers of elements $=100$ ($h=0.1$), $\dim S_h = 81$
Initial values: $u_h^{(0)}$ = Galerkin approximation (4.2), $\alpha_0 = 0$
Stopping and Extension parameters: $\varepsilon = 10^{-4}$, $\delta = 10^{-3}$

Results: Iteration numbers: $N=20$
L^2 error bound ($=h\tilde{\alpha}_N$): 0.0789060
Coefficient intervals: as in Table 2.

Table 2. $-\Delta u-20xy\cdot u=(2\pi-1)\sin\pi x\cdot\sin\pi y$

j	$[\underline{\tilde{A}}_j^{(N)}, \bar{\tilde{A}}_j^{(N)}]$
1	[0.0044953, 0.0603775]
2	[0.0097055, 0.1147773]
3	[0.0124607, 0.1610751]
4	[0.0115127, 0.1956391]
5	[0.0066034, 0.2148457]
6	[−0.0013600, 0.2154861]
7	[−0.0102185, 0.1951729]
8	[−0.0165818, 0.1526802]
9	[−0.0158443, 0.0880232]
10	[0.0097055, 0.1147773]
11	[0.0194966, 0.2196898]
12	[0.0246207, 0.3093965]
13	[0.0227255, 0.3767712]
14	[0.0133514, 0.4145900]
15	[−0.0016884, 0.4163249]
16	[−0.0181372, 0.3769854]
17	[−0.0294121, 0.2939300]
18	[−0.0269411, 0.1674148]
19	[0.0124607, 0.1610751]
20	[0.0246207, 0.3093965]
21	[0.0303857, 0.4371331]

(Omitted for $j=22\sim81$)

Example 3. (Interval coefficients)
Problem:

$$(7.11)\qquad \begin{cases} -\Delta u+[-5,5]u=[-6,6], & (x,y)\in\Omega, \\ u=0, & (x,y)\in\partial\Omega. \end{cases}$$

Conditions: Numbers of elements $=100$ $(h=0.1)$, $\dim S_h=81$
Initial values: $u_h^{(0)}=0$, $\alpha_0=0$
Stopping and Extension parameters: $\varepsilon=10^{-4}$, $\delta=10^{-3}$

Results: Iteration numbers: $N=14$
L^2 error bound $(=\tilde{h}\alpha_N)$: 0.1349496
Coefficient intervals: as in Table 3.

Example 4. (For comparison with problem (7.11))
Problem:

$$(7.12)\qquad \begin{cases} -\Delta u-\pi u=(2\pi-1)\sin\pi x\cdot\sin\pi y, & (x,y)\in\Omega, \\ u=0, & (x,y)\in\partial\Omega. \end{cases}$$

Exact solution: $u(x,y)=\dfrac{1}{\pi}\sin\pi x\cdot\sin\pi y$

Conditions: Numbers of elements $=100$ $(h=0.1)$, $\dim S_h=81$
Initial values: $u_h^{(0)}=$Galerkin approximation (4.2), $\alpha_0=0$
Stopping and Extension parameters: $\varepsilon=\delta=10^{-8}$

Results: Iteration numbers: $N=16$
L^2 error bound ($=\tilde{h}\alpha_N$): 0.0487168
Coefficient intervals: as in Table 4.

Table 3. $-\Delta u+[-5, 5]u=[-6, 6]$

j	$\tilde{A}_j^{(N)}=[\underline{\tilde{A}}_j^{(N)}, \bar{\tilde{A}}_j^{(N)}]$
1	[−0.1715288, 0.1715288]
2	[−0.2785206, 0.2785206]
3	[−0.3476306, 0.3476306]
4	[−0.3868606, 0.3868606]
5	[−0.3996123, 0.3996123]
6	[−0.3868606, 0.3868606]
7	[−0.3476306, 0.3476306]
8	[−0.2785206, 0.2785206]
9	[−0.1715288, 0.1715288]
10	[−0.2785206, 0.2785206]
11	[−0.4727604, 0.4727604]
12	[−0.6012401, 0.6012401]
13	[−0.6748777, 0.6748777]
14	[−0.6988963, 0.6988963]
15	[−0.6748777, 0.6748777]
16	[−0.6012401, 0.6012401]
17	[−0.4727604, 0.4727604]
18	[−0.2785206, 0.2785206]
19	[−0.3476306, 0.3476306]
20	[−0.6012401, 0.6012401]
21	[−0.7734027, 0.7734027]

(Omitted for $j=22\sim81$)

Table 4. $-\Delta u-\pi u=(2\pi-1)\sin\pi x\cdot\sin\pi y$

j	$\tilde{A}_j^{(N)}=[\underline{\tilde{A}}_j^{(N)}, \bar{\tilde{A}}_j^{(N)}]$
1	[0.0152852, 0.0459134]
2	[0.0335920, 0.0828146]
3	[0.0496244, 0.1105956]
4	[0.0604027, 0.1279473]
5	[0.0641884, 0.1338544]
6	[0.0604027, 0.1279473]
7	[0.0496244, 0.1105956]
8	[0.0335920, 0.0828146]
9	[0.0152852, 0.0459134]
10	[0.0335920, 0.0828146]
11	[0.0692942, 0.1521244]
12	[0.1000692, 0.2046874]
13	[0.1206619, 0.2376009]
14	[0.1278847, 0.2488152]
15	[0.1206619, 0.2376009]
16	[0.1000692, 0.2046874]
17	[0.0692942, 0.1521244]
18	[0.0335920, 0.0828146]
19	[0.0496244, 0.1105956]
20	[0.1000692, 0.2046874]
21	[0.1428661, 0.2765954]

(Omitted for $j=22\sim81$)

Note that the problem (7.12) is clearly contained to (7.11) and that the numerical results also illustrate such a situation.

REMARK 3. All the arguments in the present paper are based on the assumption that the finite element mesh exactly coincides with the given domain Ω, and that the integrals in the calculations of the inner products are evaluated exactly. In many practical cases, however, these assumptions may be violated, and it becomes necessary to evaluate the effects of these errors. It seems that some interval evaluations as described in §6 will resolve such difficulties.

Acknowledgements. We obtained much cooperation from the IBM Japan, Ltd. concerning the use of ACRITH library. The author would like to greatly acknowledge the staff members of the same company who readily offered the facility.

References

[1] P. G. Ciaret, The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam, 1978.
[2] J. Douglas, Jr. and T. Dupont, Galerkin approximation for the two point boundary problem using continuous piecewise polynomial spaces. Numer. Math., **22** (1974), 99–109.
[3] P. Grisvard, Elliptic Problems in Nonsmooth Domains. Pitman, Boston, 1985.

[4] IBM manual, High-Accuracy Arithmetic Subroutine Library (ACRITH), Program Description and User's Guide, SC33-6164-02, 1986.
[5] E. W. Kaucher and W. L. Miranker, Self-Validating Numerics for Function Space Problems. Academic Press, New York, 1984.
[6] U. W. Kulisch and W. L. Miranker, Comoputer Arithmetic in Theory and Practice. Academic Press, New York, 1981.
[7] U. W. Kulisch and W. L. Miranker, A new approach to scientific computation. Proceedings of the IBM Symposium, Academic Press, New York, 1983.
[8] O. E. Lanford III, Computer-assisted proofs in analysis. Physica **124**A (1984), 465–470.