数理解析研究所講究録
第673巻 1988年 107-126

A numerical approach to the proof of existence of solutions for elliptic problems
Part Ⅱ: for the case of large spectral radius

by

Mitsuhiro T. Nakao
(中尾 充宏)

§1. Introduction

In the preceding paper [2], we described a method which verifies, automatically using computers, existence of weak solutions for Dirichlet problems of second order based upon finite element approximations and Schauder's fixed point theorem. It was, however, difficult to apply the method to the problem of which associated spectral radius is greater than 1. In this paper, we propose an another approach, to overcome such a difficulty, utilizing Sadovskii's fixed point theorem instead of Schauder's theorem. This method can be applicable, at least theoretically, to general linear elliptic problems without any limitation of the spectral radius, if certain appropriate approximation spaces are provided.

In the following section, we formulate the Dirichlet problem as the fixed point equation associated with the condensing map by the use of an approximate Green's operator and a small positive parameter. This is done with a view to obtaining the equation with small spectral radius. In §3, as in [2], we define the

rounding for the set of functions using the orthogonal projection to the finite element subspace. We also clarify the computer oriented verification condition based on Sadovskii's fixed point theorem for the condensing map. The concrete algorithms of verification are presented in §4. We describe an iterative method to obtain the invariant set of functions satisfying the condition of the fixed point theorem. Further, we prove a theorem which suggests the conditions for verifiability by the algorithm. Finally, in §5, we show some numerical examples which confirm us that the present method is really applicable to problems having large spectral radius.

§2. Formulation of the problem

We consider the following linear elliptic boundary value problem :

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

where Ω is a bounded convex domain in R^n, $1 \le n \le 3$, with piecewise smooth boundary $\partial\Omega$ and $b = (b^i)$, $1 \le i \le n$. Assume that $b^i \in W^1_\infty(\Omega)$, $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$, where $W^1_\infty(\Omega)$ implies the usual L^∞-Sobolev space of first order on Ω. The weak solution $u \in H^1_0(\Omega)$ of (2.1) satisfies

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

where $(\cdot,\cdot)$ denotes L^2 inner product on Ω and $H^1_0(\Omega)$ means the

L^2-Sobolev space of first order whose element vanishes on $\partial\Omega$. The inner product on $H_0^1(\Omega)$ is defined as $\langle\phi,\psi\rangle \equiv (\nabla\phi,\nabla\psi)$ and the associated norm is denoted by $\|\phi\|_{H_0^1}^2 = \langle\phi,\phi\rangle$. Hereafter, we will usually suppress the symbol Ω in $H_0^1(\Omega)$ and $L^\infty(\Omega)$ etc., and simply denote by H_0^1 and L^∞, respectively. Notice that (2.1) is represented as the following operator form :

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

where the compact operator $A : H_0^1 \longrightarrow H_0^1$ is defined by

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

and $F \in H_0^1$ satisfies $\langle F,\phi\rangle = (f,\phi)$ for arbitrary $\phi \in H_0^1$.

We now take an appropriate finite element subspace S_h of H_0^1 for $0 < h < 1$. Let P_h be the orthogonal projection, i.e. H_0^1-projection, from H_0^1 into S_h. We denote $A_h \equiv P_hA$ and $I_h \equiv P_hI$, where I implies the identity operator on H_0^1. Let us suppose that the following is valid for a fixed h.

A1. The restriction of I_h-A_h to S_h has an inverse $[I-A]_h^{-1}$ on S_h.

This assumption means that there exists a unique Galerkin finite element solution $u_h \in S_h$ to (2.2) for each f in the sense that

$$(\nabla u_h,\nabla v) = (b\cdot\nabla u_h + cu_h + f,v), \qquad v \in S_h. \tag{2.4}$$

Next, for a fixed constant ε, $0 < \varepsilon < 1$, we define an operator $T : H_0^1 \longrightarrow H_0^1$ by

$$Tu \equiv (I - ([I-A]_h^{-1}P_h + \varepsilon)(I-A))u + ([I-A]_h^{-1}P_h + \varepsilon)F, \tag{2.5}$$

where $\varepsilon \equiv \varepsilon I$.

Then we have the following result which is the starting point of arguments in this paper.

Theorem 1. Let U be a bounded convex and closed subset in H_0^1. If $TU \overset{\circ}{\subset} U$ then there exists a unique solution u for (2.3) in U. Here, in general, $M_1 \overset{\circ}{\subset} M_2$ implies $\bar{M}_1 \subset \mathring{M}_2$ for any sets M_1, M_2.

Proof. We can rewrite (2.5) as

$$Tu = (1-\varepsilon)Iu + \{(\varepsilon A - [I-A]_h^{-1}P_h(I-A))u - ([I-A]_h^{-1}P_h + \varepsilon)F\}.$$

Here, the operator $(1-\varepsilon)I$ is obviously $(1-\varepsilon)$-contractive and $\varepsilon A-[I-A]_h^{-1}P_h(I-A)$ is compact. Therefore, T becomes a condensing map and, by Sadovskii's fixed point theorem (e.g.[6]), $TU \overset{\circ}{\subset} U$ implies that T has a fixed point in U. Further we can easily prove, by quite similar techniques in [5], particularly in the proof of Theorem 2.1, that both operators I-A and $[I-A]_h^{-1}P_h + \varepsilon I$ are invertible. Since, for a fixed point $u \in U$ of T, we have by (2.5)

$$([I-A]_h^{-1}P_h + \varepsilon)((I-A)u - F) = 0,$$

we obtain $u = Au + F$. The uniqueness result is straightforward by the non-singularity of I-A. Thus we have the theorem.

It is expected, from the appearance of (2.5), that the spectral radius of the linear part of the operator T becomes sufficiently small, particularly less than one, when ε is small enough and $[I-A]_h^{-1}$ is a good approximation of $(I-A)^{-1}$.

§3. Rounding and verification conditions

We introduce the concept of rounding which is similar to that in [2]. First, as one of the approximation properties of S_h, assume that

A2. For each $u \in H_0^1 \cap H^2$, there exists a positive constant C_1, independent of h, such that

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

where $|u|_{H^2}$ implies the semi-norm of u on $H^2(\Omega)$ defined by

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

Next, for each $\psi \in L^2(\Omega)$, let ϕ be a solution of the following problem :

$$\begin{cases} -\Delta \phi = \psi & \text{in } \Omega, \\ \phi = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.2}$$

Then, by the well-known result, $\phi \in H_0^1 \cap H^2$ and there exists a positive constant C_2 such that

$$|\phi|_{H^2} \le C_2 \|\psi\|_{L^2}. \tag{3.3}$$

Now we define for each subset $U \subset H_0^1$ the rounding $R(TU) \subset S_h$ as

$$R(TU) = \{u_h \in S_h \ ; \ u_h = T_h u, \ u \in U\}, \tag{3.4}$$

where $T_h = P_h T$. Further, for $u \in H_0^1$, set

$$e(u) \equiv (1-\varepsilon)\|u - P_h u\|_{H_0^1} + \tilde{C}\varepsilon h \|b \cdot \nabla u + cu + f\|_{L^2},$$

where $\tilde{C} = C_1 C_2$.

Then the rounding error $RE(TU) \subset S_h^{\perp}$ is defined by

$$RE(TU) = \{\phi \in S_h^{\perp} ; \|\phi\|_{H_0^1} \le e(U) \text{ and } \|\phi\|_{L^2} \le \tilde{C}he(U)\}, \tag{3.5}$$

where $S_h^{\perp}$ is the orthogonal complement of S_h and $e(U) = \sup_{u \in U} e(u)$.

Then we have following lemma from Theorem 1.

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

$$R(TU) + RE(TU) \overset{\circ}{\subset} U. \tag{3.6}$$

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

Proof. By virtue of Theorem 1 it is sufficient to show the inclusion $TU \subset R(TU) + RE(TU)$. Since, for each $u \in U$, Tu is uniquely decomposed as $Tu = T_h u + (Tu - T_h u)$, it is sufficient to prove $Tu - T_h u \in RE(TU)$.

First, observe that

$$Tu - T_h u = (1-\varepsilon)(I-I_h)u + \varepsilon((Au+F) - (A_h u+F_h)), \tag{3.7}$$

where $F_h = P_h F$.

Next, we estimate the second term in the right hand side of (3.7). We now notice that $Au + F$ is a solution of (3.2) for $\psi = b\cdot\nabla u + cu + f$ and $A_h u + F_h$ is its H_0^1-projection. Hence, by the use of estimates (3.1) and (3.3), we obtain

$$\|Au+F - (A_h u+F_h)\|_{H_0^1} \le \tilde{C}h\|b\cdot\nabla u + cu + f\|_{L^2}. \tag{3.8}$$

Thus (3.7), (3.8) yield that

$$\|Tu - T_h u\|_{H_0^1} \le e(u). \tag{3.9}$$

Furthermore, by the usual duality argument for the error estimates of the H_0^1-projection, we can easily get

$$\|Tu - T_h u\|_{L^2} \le \tilde{C}h\|Tu - T_h u\|_{H_0^1} \le \tilde{C}he(u). \tag{3.10}$$

(3.9) and (3.10) imply $Tu - T_h u \in RE(TU)$ which proves the lemma.

§4. Computing algorithm for verification

In order to construct the set U which satisfies the verification condition (3.6), we use an iterative method. Let R^+ denote the set of nonnegative real numbers and define for $\alpha \in R^+$

$$[\alpha] \equiv \{\phi \in S_h^{\perp} \ ; \ \|\phi\|_{H_0^1} \le \alpha \text{ and } \|\phi\|_{L^2} \le \tilde{C}h\alpha\}.$$

Also let $\{\phi_j\}_{j=1,\cdots,M}$ be a basis of S_h and denote the set of linear combinations of $\{\phi_j\}$ with interval coefficients by $\mathcal{G}_I$. Here, we interpret each $U \in \mathcal{G}_I$ as the same meaning as in [2], that is,

$$U \equiv \sum_{j=1}^{M} A_j\phi_j = \{ \sum_{j=1}^{M} a_j\phi_j \ ; \ a_j \in A_j, \ 1 \le j \le M\},$$

where A_j are real intervals.

Now, for $u_h = \sum_{j=1}^{M} A_j\phi_j \in \mathcal{G}_I$ and $\alpha \in R^+$, choose $\hat{\phi}^{(1)} = \sum_{j=1}^{M} B_j^{(1)}\phi_j$ and $\hat{\phi}^{(2)} = \sum_{j=1}^{M} B_j^{(2)}\phi_j$ satisfying

$$\sum_{j=1}^{M} (\nabla\phi_j, \nabla\phi_k)B_j^{(1)} = \sum_{j=1}^{M} A_j(b\cdot\nabla\phi_j + c\phi_j, \phi_k) + (f, \phi_k) + [-1,1]\tilde{C}h\alpha\|-\nabla\cdot(b\phi_k) + c\phi_k\|_{L^2} \tag{4.1}$$

and

$$(4.2)\qquad \sum_{j=1}^{M}\{(\nabla\phi_j,\nabla\phi_k) - (b\cdot\nabla\phi_j + c\phi_j,\phi_k)\}B_j^{(2)} =$$

$$[-1,1]\tilde{C}h\alpha\|-\nabla\cdot(b\phi_k) + c\phi_k\|_{L^2} + (f,\phi_k),$$

for $1 \le k \le M$, respectively. (4.1) and (4.2) imply that $\hat{\phi}^{(i)}$, $i = 1, 2$, are determined as the solutions for linear systems of equations with interval right hand side.

Then we set

$$(4.3)\qquad \tilde{u}_h = \varepsilon(\hat{\phi}^{(1)} - u_h) + \hat{\phi}^{(2)}.$$

Further let $\tilde{\alpha} \in R^+$ be taken as

$$(4.4)\qquad \tilde{\alpha} = \tilde{C}\varepsilon h \sup_{v\in u_h}\|b\cdot\nabla v + cv + f\|_{L^2}$$

$$+ \{(1 - \varepsilon) + \tilde{C}\varepsilon h(\|b\|_{L^\infty} + \tilde{C}h\|c\|_{L^\infty})\}\alpha,$$

where $\|b\|_{L^\infty} \equiv \max\{\|b^i\|_{L^\infty},\ 1 \le i \le n\}$.

Using (4.1) - (4.4), we define a map Φ from $\mathcal{G}_I \times R^+$ into itself by

$$(4.5)\qquad \Phi(u_h,\alpha) = (\tilde{u}_h,\tilde{\alpha}).$$

Now for appropriately chosen initial value $u_h^{(0)} \in S_h$ and $\alpha_0 \in R^+$, we generate an iterative sequence $\{(u_h^{(i)},\alpha_i)\}$, for $i \ge 1$, by

$$(4.6)\qquad (u_h^{(i)},\alpha_i) = \Phi(u_h^{(i-1)},\alpha_{i-1}).$$

Then the following property holds.

Lemma 2. For the sequence $\{(u_h^{(i)},\alpha_i)\}$ defined by (4.6)

$$(4.7)\qquad R(T(u_h^{(i-1)} + [\alpha_{i-1}])) \subset u_h^{(i)}$$

and

(4.8) $$RE(T(u_h^{(i-1)} + [\alpha_{i-1}])) \subset [\alpha_i], \quad \text{for } i \geq 1.$$

Proof. We fix $v \in u_h^{(i-1)}$ and $\phi \in [\alpha_{i-1}]$. By some simple calculations taking account of the assumption A1 and $I_h\phi = 0$, we have

(4.9)
$$\begin{aligned} T_h(v+\phi) &= (I_h - ([I-A]_h^{-1}P_h + \varepsilon)(I_h - A_h))(v+\phi) \\ &\qquad + ([I-A]_h^{-1} + \varepsilon)F_h \\ &= (I_h - [I-A]_h^{-1}(I_h - A_h))\phi - \varepsilon(I_h - A_h)(v+\phi) \\ &\qquad + ([I-A]_h^{-1} + \varepsilon)F_h \\ &= \varepsilon((A_h(v+\phi) + F_h) - v) + [I-A]_h^{-1}(A_h\phi + F_h). \end{aligned}$$

By the similar argument to that in [2], the proof of Lemma 2, we can show

(4.10) $$A_h(v+\phi) + F_h \in \hat{\phi}^{(1)},$$

where $\hat{\phi}^{(1)}$ is defined by (4.1) for $u_h = u_h^{(i-1)}$ and $\alpha = \alpha_{i-1}$.
Next, integrating by part we have

(4.11)
$$\begin{aligned} \langle A_h\phi + F_h, \phi_k \rangle &= (\phi, -\nabla\cdot(b\phi_k) + c\phi_k) + (f, \phi_k) \\ &\in [-1,1]\tilde{C}h\alpha_{i-1}\|-\nabla\cdot(b\phi_k) + c\phi_k\|_{L^2} + (f,\phi_k), \end{aligned}$$

for $1 \leq k \leq M$, where we have used $\|\phi\|_{L^2} \leq \tilde{C}h\alpha_{i-1}$.
Combining (4.11) with (4.2) we get

(4.12) $$[I-A]_h^{-1}(A_h\phi + F_h) \in \hat{\phi}^{(2)},$$

where $\hat{\phi}^{(2)}$ is defined by (4.2) for $\alpha = \alpha_{i-1}$.
Thus (4.9), (4.10), (4.12) and (4.3) imply (4.7).
Now we observe that, from the definition of $e(\cdot)$ in §3,

$$e(v+\phi) = (1-\varepsilon)\|\phi\|_{H_0^1} + \tilde{C}\varepsilon h\|b\cdot\nabla(v+\phi) + c(v+\phi) + f\|_{L^2}$$

$$\leq (1-\varepsilon)\|\phi\|_{H_0^1} + \tilde{C}\varepsilon h\|b\cdot\nabla v + cv + f\|_{L^2}$$

$$+ \tilde{C}\varepsilon h(\|b\|_{L^\infty}\|\phi\|_{H_0^1} + \|c\|_{L^\infty}\|\phi\|_{L^2})$$

$$\leq \tilde{C}\varepsilon h\|b\cdot\nabla v + cv + f\|_{L^2}$$

$$+ \{(1-\varepsilon) + \tilde{C}\varepsilon h(\|b\|_{L^\infty} + \tilde{C}h\|c\|_{L^\infty})\}\alpha_{i-1}.$$

Hence it holds that $e(v+\phi) \leq \alpha_i$, for arbitrary $v \in u_h^{(i-1)}$ and $\phi \in [\alpha_{i-1}]$ which yields (4.8), and we complete the proof.

From the Lemma 2 we also have $T(u_h^{(i-1)} + [\alpha_{i-1}]) \subset u_h^{(i)} + [\alpha_i]$. Thus we can say that the iterative sequence (4.6) is a computational sequence including the iteration $u_i = Tu_{i-1}$ with $u_0 = u_h^{(0)} + \phi_0$, where $\phi_0 \in [\alpha_0]$.

Now we are ready to present an algorithm for computing an inclusion of the solution of (2.3) which automatically verifies the correctness of the computed bounds.

First, we define a stopping criterion for the iteration (4.6). For

$$u_h^{(i)} = \sum_{j=1}^{M} A_j^{(i)}\phi_j \in \mathfrak{S}_I, \text{ where } A_j^{(i)} = [\underline{A}_j^{(i)}, \bar{A}_j^{(i)}],\ 1 \leq j \leq M,$$

define

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

If, for a given $\delta_1 > 0$, we have attained to the state at a number N such that

$$(4.13) \qquad \|u_h^{(N)} - u_h^{(N-1)}\| < \delta_1 \text{ and } |\alpha_N - \alpha_{N-1}| < \delta_1,$$

then we extend $u_h^{(N)}$ and α_N as follows :

for a given $\delta_2 > 0$, set

$$\hat{u}_h = u_h^{(N)} + \sum_{j=1}^{M} [-1,1]\delta_2\phi_j \tag{4.14}$$

and

$$\hat{\alpha} = \alpha_N + \delta_2. \tag{4.15}$$

Further we calculate $(u_h,\alpha) \in \mathbb{G}_I \times R^+$ as

$$(u_h,\alpha) = \Phi(\hat{u}_h,\hat{\alpha}). \tag{4.16}$$

Then we obtain the following result as the direct conclusion of the above arguments and it implies the completion of automatic verification.

Theorem 2. If (u_h,α) and $(\hat{u}_h,\hat{\alpha})$ satisfies

$$u_h \overset{\circ}{\subset} \hat{u}_h \quad \text{and} \quad \alpha < \hat{\alpha}, \tag{4.17}$$

where inclusion means that each coefficient interval in u_h is strictly covered by the corresponding interval in $\hat{u}_h$. Then there exists a unique solution u to (2.3) in $u_h + [\alpha]$.

Remark 1. The procedures described here to obtain the strict inclusion relation (4.17) are the same as in [2]. Although this is a technique to get such a inclusion, there might be another and more efficient methods. For example, if we adopt the iteration, instead of (4.6), such as $(u_h^{(i)},\alpha_i) = \Phi(u_h^{(i-1)} + \delta, \alpha_{i-1} + \delta)$ for appropriate $\delta > 0$, where $u_h^{(i-1)} + \delta$ means δ-extension of each coefficient interval, the strict inclusion

$$u_h^{(N)} \overset{\circ}{\subset} u_h^{(N-1)} \quad \text{and} \quad \alpha_N < \alpha_{N-1} \tag{4.18}$$

might be attained at some N step. Then it will be expected to save the computing time for verification.

Next, we shall provide a condition which enables the iteration (4.6) to converge independently of the spectral radius of the operator A and suggests that the verification process is normally completed. We assume the following additional hypotheses.

A3. $\{\phi_j\}$, $1 \le j \le M$, is an orthogonal basis of S_h.

A4. There exists a positive constant $\hat{C}$, independent of h, such that $Mh^n \le \hat{C}$.

Here, A4 will be always valid when S_h is the usual piecewise linear finite element space on quasi-uniform mesh (see e.g. [4]). We now make arguments similar to that in [2] on the assumptions A1 - A4.

First, we redefine the iterative sequence $\{(u_h^{(i)},\alpha_i)\}$. Let 2^{S_h} be the power set of S_h and define its topology by the following Hausdorff metric $D(\cdot,\cdot)$ based on H_0^1-norm on Ω. For U_h, $V_h \in 2^{S_h}$,

$$D(U_h,V_h) \equiv \max\{\sup_{\phi\in U_h} d(\phi,V_h),\ \sup_{\psi\in V_h} d(\psi,U_h)\}, \tag{4.19}$$

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

We now difine a map $T_1 : 2^{S_h} \times R^+ \longrightarrow 2^{S_h}$ by

$$T_1(u_h,\alpha) = -\varepsilon(I_h - A_h)u_h + ([I-A]_h^{-1} + \varepsilon)A_h[\alpha] \tag{4.20}$$

$$+ ([I-A]_h^{-1} + \varepsilon)F_h$$

for each $(u_h,\alpha) \in 2^{S_h} \times R^+$.

And $T_2 : 2^{S_h} \times R^+ \longrightarrow R^+$ is defined by

$$T_2(u_h,\alpha) = \tilde{C}_1 \varepsilon h \|u_h\|_1 + ((1 - \varepsilon) + \tilde{C}_2 \varepsilon h)\alpha + \tilde{C}_3 \varepsilon h \|f\|_{I,2}, \tag{4.21}$$

where $\tilde{C}_i$, $1 \le i \le 3$, are positive constants independent of h, and

$$\|u_h\|_1 \equiv \sup_{\phi \in u_h} \|\phi\|_{H_0^1}. \tag{4.22}$$

Using (4.20) and (4.21) we determine a map $\tilde{\Phi} : 2^{S_h} \times R^+ \longrightarrow 2^{S_h} \times R^+$ as

$$\tilde{\Phi}(u_h,\alpha) = (T_1(u_h,\alpha),T_2(u_h,\alpha)). \tag{4.23}$$

Then $\tilde{\Phi}$ is essentially the same as the map Φ defined by (4.16), although Φ is slightly overestimated in comparison with $\tilde{\Phi}$. We now consider about the convergence property for the sequence $\{(u_h^{(i)},\alpha_i)\}$ which is given by the following iteration

$$(u_h^{(i)},\alpha_i) = \tilde{\Phi}(u_h^{(i-1)},\alpha_{i-1}), \tag{4.24}$$

when $(u_h^{(0)},\alpha_0)$ is appropriately chosen.

Theorem 3. On the assumptions A1 - A4, when $(I-A)^{-1}$ exists and ε is taken as appropriately small, for sufficiently small h, the iterative sequence $\{(u_h^{(i)},\alpha_i)\}$ defined by (4.24) converges to a unique limit (u_h,α), with an arbitrary initial value $(u_h^{(0)},\alpha_0)$, which is also a unique fixed point of $\tilde{\Phi}$.

Proof. As there are many arguments analogous to that in [2],

Theorem 2, we will omit details of the similar discussions. We show that $\{(u_h^{(i)},\alpha_i)\}$ is the Cauchy sequence in $2^{S_h} \times R^+$. In the below, C_i', $i = 1, 2, \cdots$ denote positive constants which are independent of h.

First, observe that from (4.20)

$$\sup_{\phi\in u_h^{(i)}} d(\phi,u_h^{(i-1)}) \le \sup_{(i-1)(i-2)} \inf (\|\varepsilon(I_h - A_h)(\hat{u}_h^{(i-1)} - \hat{u}_h^{(i-2)})\|_{H_0^1} + \|([I-A]_h^{-1} + \varepsilon)A_h(\hat{\alpha}_{i-1} - \hat{\alpha}_{i-2})\|_{H_0^1}), \tag{4.25}$$

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

$$(i-1) \equiv \begin{cases} \hat{u}_h^{(i-1)} \in u_h^{(i-1)} \\ \hat{\alpha}_{i-1} \in [\alpha_{i-1}] \end{cases} \quad \text{and} \quad (i-2) \equiv \begin{cases} \hat{u}_h^{(i-2)} \in u_h^{(i-2)} \\ \hat{\alpha}_{i-2} \in [\alpha_{i-2}] \end{cases},$$

respectively.

Also we can deduce that, by the result of error estimates, for sufficiently small h,

$$\|[I-A]_h^{-1}P_h - (I-A)^{-1}\| \le C_1'\|(I-A)^{-1}\|,$$

where $\|\cdots\|$ means the operator norm associated with H_0^1-norm. Therefore, using the triangle inequality, we have

$$\|[I-A]_h^{-1}P_h\| \le C_2'\|(I-A)^{-1}\|. \tag{4.26}$$

Furthermore, we get by arguments similar to that in [2]

$$\sup_{(i-1)(i-2)} \inf \|A_h(\hat{\alpha}_{i-1} - \hat{\alpha}_{i-2})\|_{H_0^1} \le C_3'h|\alpha_{i-1} - \alpha_{i-2}|\sqrt{M}. \tag{4.27}$$

Thus, from (4.25) - (4.27), it holds that

$$D(u_h^{(i)},u_h^{(i-1)}) \le C_4'\varepsilon D(u_h^{(i-1)},u_h^{(i-2)}) + C_5'h\sqrt{M}|\alpha_{i-1} - \alpha_{i-2}|, \tag{4.28}$$

where we have used the fact that $\|I_h - A_h\|$ is bounded by C_4' independently of h.

On the other hand, from (4.21)

$$|\alpha_i - \alpha_{i-1}| \le C_6'\varepsilon h D(u_h^{(i-1)}, u_h^{(i-2)}) + (1 - \varepsilon + \tilde{C}_2\varepsilon h)|\alpha_{i-1} - \alpha_{i-2}|. \tag{4.29}$$

We now rewrite (4.28) and (4.29) as the following matrix form.

$$\begin{bmatrix} D(u_h^{(i)}, u_h^{(i-1)}) \\ |\alpha_i - \alpha_{i-1}| \end{bmatrix} \le \begin{bmatrix} C_4'\varepsilon & C_5'h\sqrt{M} \\ C_6'\varepsilon h & 1-\varepsilon+\tilde{C}_2\varepsilon h \end{bmatrix} \begin{bmatrix} D(u_h^{(i-1)}, u_h^{(i-2)}) \\ |\alpha_{i-1} - \alpha_{i-2}| \end{bmatrix} . \tag{4.30}$$

Let P denote the square matrix of the right hand side of (4.30). Then it is not difficult to show that, when we choose ε as $1 - \varepsilon > \varepsilon C_4'$, for sufficiently small h, the spectral radius of P is less than 1. This fact implies that $\{(u_h^{(i)}, \alpha_i)\}$ becomes the Cauchy sequence. Uniqueness of the limit independent of the initial value also follows by the same arguments as in [2].

Furthermore, it will be also expected that we can discuss, similarly in [2], about the attainability of the verification condition (4.17) on the same assumptions as in Theorem 3.

§5. Numerical examples

We now illustrate some examples which confirm that the verification method described here is available independently of the spectral radius of the operator A.

(i) One dimensional case

We considered the folowing simple two point boundary value problem with constant coefficients :

$$(5.1)\qquad \begin{cases} -u'' - Ku = (\pi - K/\pi)\sin\pi x, & x \in I = (0,1), \\ \\ u(0) = u(1) = 0, \end{cases}$$

where K is a positive parameter. Notice that (5.1) has a solution $u(x) = \frac{1}{\pi}\sin\pi x$ independently of K.

We now take the finite element subspace S_h of $H_0^1(I)$ as the same as in [2]. Let $\delta_x : 0 = x_0 < x_1 < \cdots < x_L = 1$ be a uniform partition of the interval $I = (0,1)$. Set $I_i = (x_{i-1}, x_i)$ and $h = 1/L$. Also let $P_1(I_i)$ denote the set of linear polynomials on I_i and define S_h by

$$(5.2)\qquad S_h \equiv \mathcal{M}_0^1(x) = \{v \in C(I) \ ; \ v|_{I_i} \in P_1(I_i), \ 1 \le i \le L, \ v(0) = v(1) = 0\}.$$

Then, $M = \dim S_h = L - 1$ and we can take as $\tilde{C} = 1$ in previous sections. We define the basis $\{\phi_j\}$, $j = 1, \cdots, M$ of S_h by the set of hat functions as in [2]. Further we choose $\varepsilon = 10^{-1}$, $\delta_1 = 10^{-3}$ and $\delta_2 = 10^{-1}$ in (2.5), (4.13) and (4.14), respectively. Table 1 shows the iteration numbers, for various meshes, required to attain the condition (4.13). Also note that the corresponding spectral radius $r(A)$ to (5.1) becomes $r(A) = K/\pi^2$. Hence, we have $r(A) > 1$ for each case in Table 1.

Table 1. Iteration numbers for verification

L	K = 15	K = 30	K = 45
10	×	×	×
20	70	×	×
40	32	76	×
80	21	35	233

Here, in Table 1, × means that iteration was divergent.

Remark 2. Since in the present examples each spectral radius is not less than 1, we cannot use the previous scheme proposed in [2]. In case of $r(A) < 1$, however, the former will be more efficient than the present scheme. For example, when $K = \pi$ in (5.1), i.e. $r(A) = 1/\pi < 1$, we needed 26 times of iterations to verify by the present method, while only 7 times were required for the scheme in [2] under the same conditions.

(ii) Two dimensional case

Consider the following problem with interval coefficient.

$$(5.3)\qquad \begin{cases} -\Delta u + [c_1, c_2]u = [f_1, f_2] & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\Omega = (0,1) \times (0,1) \subset R^2$ and $[c_1,c_2]$, $[f_1,f_2]$ are intervals which mean the sets of L^∞-functions whose ranges are included in $[c_1,c_2]$ and $[f_1,f_2]$, respectively. According to the similar arguments in [2], we can easily extend the techniques in the preceding sections to the case of interval coefficients such as (5.3). Further we take the finite element subspace S_h of $H_0^1(\Omega)$ as the tensor product of one dimensional case as in [2], that is, $\delta_x = \delta_y$ and $S_h \equiv \mathcal{M}_0^1(x) \times \mathcal{M}_0^1(y)$. Then dim $S_h = (L-1)^2$. Also, as described in [2], we can also choose $\tilde{C} = 1$.

The numerical results for the concrete problem which is verified are as follows :

Problem :

$$(5.4) \quad \begin{cases} -\Delta u = [-21,5]u + [0,1] & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Conditions : $\varepsilon = 10^{-1}$, $\delta_1 = 10^{-2}$, $\delta_2 = 10^{-1}$

Mesh size : $h = 0.025$ $(L = 40)$, dim $S_h = 1521$

Initial values : $u_h^{(0)} = 0$, $\alpha_0 = 0$

Results : Iterations : $N = 11$

L^2 error bound : $\alpha = 0.3123$

Coefficient intervals : $\min_{1 \le j \le M} \underline{A}_j = -0.7061.$

$\max_{1 \le j \le M} \bar{A}_j = +0.7951.$

Here, $\underline{A}_j$ and $\bar{A}_j$ are the infimum and supremum, respectively, of the coefficient intervals for ϕ_j in u_h which appears in Theorem 2. Note that (5.4) includes the problem with $r(A) \ge 1$, for it contains the equation of the form $-\Delta u = Ku + [0,1]$, where K is a

constant such that $-21 \leq K \leq -2\pi^2$, which has the associated spectral radius $r(A) = |K|/2\pi^2 \geq 1$. In case of $h > 0.025$, the iteration (4.16) diverged and we could not verify the problem. This fact suggests that the assumption for the smallness of the mesh size must be crucial similarly in the one dimensional case.

Remark 3. Owing to the limitation of our computer facility, in two dimensional problem, we had to use the iterative method for the solution of linear equations arised in the verification process. We used the SOR method with stopping parameter 10^{-7}. Hence, it may be not sure that, by the truncation errors for iteration, the verification condition (4.17) is strictly satisfied. As we also calculated all numerical computation by the usual double precision computer arithmetic, there may be some round off errors in each step. The author believes, however, the above example is really significant as an numerical experiment to show that the present technique is applicable for the case $r(A) \geq 1$.

Faculty of Science
Kyushu University 33
Hakozaki Fukuoka 812
JAPAN

References

[1] E. W. Kaucher & W. L. Miranker, Self-validating numerics for function space problems, Academic Press, New York, 1984.

[2] M. T. Nakao, A numerical approach to the proof of existence of solutions for elliptic problems, Japan Journal of Applied Mathematics, 5 (1988), 313-332.

[3] M. T. Nakao, A computational verification method of existence of solution for nonlinear elliptic problems, to appear.

[4] J. T. Oden & J. N. Reddy, An introduction to the mathematical theory of finite elements, John Wiley & Sons,New York, 1976.

[5] S. M. Rump, Solving algebraic problems with high accuracy, Proceedings of the IBM symposium, Academic Press,New York, 1983.

[6] E. Zeidler, Nonlinear functional analysis and its applications, Springer-Verlag, New York, 1986.