# Numerical Verifications of Solutions for Nonlinear Elliptic Equations

Yoshitaka Watanabe and Mitsuhiro T. Nakao

Department of Mathematics, Kyushu University 33,

Fukuoka 812, Japan

**Abstract**

A numerical technique which enables us to verify the existence of weak solutions for nonlinear elliptic boundary value problems is proposed. It is based on the infinite dimensional fixed point theorems using the Newton-like operator and the error estimates for finite element approximations. We also present an effective verification procedure which automatically generates the set including the exact solution in a computer. Some numerical examples are illustrated.

## 1 Introduction

In recent years, there have been several approaches to the numerical proof of existence of solutions using computers for various differential equations in mathematical science. Concerning ordinary differential equations, in addition to the analytical methods ( [6], [14], etc.), there are several methods based on interval arithmetic ( [5], [8], etc.). Also in [12], a verification method was formulated for the solutions of nonlinear two-point boundary value problems. It is a kind of Newton-like method using an enclosing technique combined with the explicit error estimates for finite element approximations.

For partial differential equations, in [9] and [10], the computer assisted verification procedures were presented for the weak solutions of Dirichlet problems of the second order, and in [11], the method was extended to nonlinear elliptic equations. Also, in [13], another verification technique was proposed for the solution for the elliptic boundary value problems using a $C^1$-class approximate solution with high accuracy and the exact eigenvalue enclosure for a linearized operator.

In this paper, we extend the verification method for the solutions of ordinary differential equations presented by [12] to nonlinear elliptic problems. In [12], the general verification algorithm was described, but it was not so efficient for obtaining the verification condition. We describe a new procedure which enables us to get a better convergence property than in [12], and we also present some numerical examples to confirm its great efficiency. Moreover, following a concrete example, we consider a general procedure which achieves greater ability in verification.

In the following section, we formulate, in a way similar to [12], a numerical verification method for parametrized nonlinear elliptic problem. In section 3, we introduce two concepts, rounding and rounding error, to deal with the verification in a computer. And in section 4, we construct a computing algorithm for verification, and show that this algorithm gives good convergence result according to the property of the interval arithmetic. Finally, some numerical examples are presented in section 5.

# 2 Formulation of the problem

In the below, for an operator $A$ and a function set $W$, including the set represented by intervals e.g. (9) in section 3, we naturally define $AW = \{Aw \mid w \in W\}$.

Consider the parametrized nonlinear elliptic problem:

$$\begin{cases} -\Delta u &= \lambda f(u) & \text{in } \Omega, \\ u &= 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\Omega$ is a bounded and convex domain in $\mathbf{R}^n$ $(1 \le n \le 3)$ with piecewise smooth boundary $\partial\Omega$, and $\lambda \ge 0$ is a real parameter.

We suppose that the nonlinear real-valued function $f(\cdot)$ satisfies the following assumptions.

**A 1** *$f$ is the continuous map from $H_0^1(\Omega)$ to $L^2(\Omega)$.*

**A 2** *For each bounded subset $U$ in $H_0^1(\Omega)$, $f(U)$ is also bounded in $L^2(\Omega)$.*

Here, let $H^m(\Omega)$ denote $m$-th order $L^2$-Sobolev space on $\Omega$, and set

$$H_0^1(\Omega) = \{\, v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega \,\}.$$

The inner product on $H_0^1(\Omega)$ is defined as $(\nabla u\,,\, \nabla v)$, where $(\,\cdot\,,\,\cdot\,)$ is the $L^2(\Omega)$-inner product. To verify the existence of a solution of (1) in a computer, we use the fixed point formulation. First, we rewrite (1) in weak form:

$$\text{find } u \in H_0^1(\Omega) \cap H^2(\Omega) \quad \text{s.t.} \quad (\,\nabla u\,,\, \nabla v\,) = \lambda(\,f(u)\,,\, v\,), \qquad \forall v \in H_0^1(\Omega). \tag{2}$$

It is well known that for any $\psi \in L^2(\Omega)$, the boundary value problem:

$$\begin{cases} -\Delta \phi &= \psi & \text{in } \Omega, \\ \phi &= 0 & \text{on } \partial\Omega, \end{cases} \tag{3}$$

has a unique solution $\phi \in H_0^1(\Omega) \cap H^2(\Omega)$ and that the following estimate holds:

$$|\,\phi\,|_{H^2(\Omega)} \le C_1 \|\psi\|_{L^2(\Omega)}, \tag{4}$$

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

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

Now, for $\psi \in L^2(\Omega)$, let $G\psi$ be the solution of (3). Then the operator $G : L^2(\Omega) \to H_0^1(\Omega)$ is compact because of the compactness of the imbedding $H^2(\Omega) \hookrightarrow H^1(\Omega)$. Therefore, from assumptions A1 and A2, when we define the nonlinear operator $F$ by

$$F \equiv G \lambda f,$$

$F$ is a compact operator on $H_0^1(\Omega)$, and thus we can rewrite the weak form (2) as a fixed point form:

$$u = Fu \quad \text{in } H_0^1(\Omega).$$

Next, we introduce the Newton-like method proposed in [12]. Let $S_h$ be an appropriate finite element subspace of $H_0^1(\Omega)$ dependent on a parameter $h$ $(0 < h < 1)$ and let $u_h \in S_h$ be an approximate solution to (2). Further, let $P_h$ be an orthogonal projection from $H_0^1(\Omega)$ into $S_h$ in $H_0^1(\Omega)$-sense determined by

$$( \nabla(u - P_h u) , \nabla v ) = 0, \qquad \forall v \in S_h. \tag{5}$$

We assume that there exists a Fréchet derivative $F'(u_h)$ for $F$ at $u_h$, and suppose that

**A 3** *The restriction of the operator $P_h(I - F'(u_h)) : H_0^1(\Omega) \to S_h$ to $S_h$ has the inverse operator $[I - F'(u_h)]_h^{-1} : S_h \to S_h$, where $I$ denotes the identity map on $H_0^1(\Omega)$.*

For a small parameter $\varepsilon\,(0 < \varepsilon < 1)$, we define a nonlinear operator $T : H_0^1(\Omega) \to H_0^1(\Omega)$ by

$$Tu \equiv \{ \ I - ( \ [ \ I - F'(u_h) \ ]_h^{-1} P_h + \varepsilon I \ )( \ I - F \ ) \}u. \tag{6}$$

By a simple calculation for (6), we can show that the operator $T$ is a condensing map. Then, under the assumptions A1 $-$ A3, if we find a non-empty, bounded, convex and closed set $U \subset H_0^1(\Omega)$ such that $TU \subset U$, then there exists a fixed point $u$ of $T$ in $U$ by Sadovskii's fixed point theorem [16]. Moreover, if the operator $[I - F'(u_h)]_h^{-1} P_h + \varepsilon I$ is invertible, $u$ is also a fixed point of $F$, i.e. a solution of (2), see [12] for details.

# 3 Rounding and verification condition

Since the operator $T$ defined by (6) is a mapping on infinite dimensional space $H_0^1(\Omega)$, it is impossible to calculate directly $TU$ for a given subset $U \subset H_0^1(\Omega)$. In order to deal with such an operator in a computer, we introduce, analogous to [12], two concepts, rounding and rounding error.

**rounding**

For $u \in H_0^1(\Omega)$, the approximation of $Tu$ by an element of finite element subspace is called rounding. Let $P_h$ be the $H_0^1$-projection defined by (5). Then, for $Tu \in H_0^1(\Omega)$, the rounding $\tilde{T}u = P_h T u$ is defined by

$$\tilde{T}u \quad \equiv \quad \{\tilde{I} - ([ \ I - F'(u_h) \ ]_h^{-1} + \varepsilon\tilde{I})(\tilde{I} - \tilde{F})\}u,$$

where $\tilde{I} = P_h I$ and $\tilde{F} = P_h F$. Next, for a set $U$, we define the rounding $R(TU)$ as

$$R(TU) \equiv \{v \in S_h \mid v = \tilde{T}u \ , \ u \in U\}.$$

**rounding error**

We call the error bounds between an element $u$ of $H_0^1(\Omega)$ and its rounding element $P_h u$ of $S_h$ rounding error. We assume, as the approximation property of $P_h$, that

**A 4** $\quad \|u - P_h u\|_{H_0^1(\Omega)} \leq C_2 h |u|_{H^2(\Omega)}, \quad \forall u \in H_0^1(\Omega) \cap H^2(\Omega),$

where $C_2$ is a positive constant independent of $u$ and $h$ which can be numerically determined. In fact, it is known that the assumption A4 is valid for many finite element subspaces (e.g. [3] ). For a set $U$, the rounding error $RE(TU)$ is defined by

$$RE(TU) \equiv \{\phi \in S_h^{\perp} \mid \|\phi\|_{H_0^1(\Omega)} \leq \alpha \ \text{ and } \ \|\phi\|_{L^2(\Omega)} \leq Ch\alpha\},$$

$$\text{where} \quad \alpha \equiv \sup_{u \in U} \|Tu - \tilde{T}u\|_{H_0^1(\Omega)} \quad \text{and}$$

$$C \equiv C_1 C_2. \tag{7}$$

Here, $C_1$ is the same constant in (4). Then the following lemma holds (proof is similar to that in [12] ).

**Lemma 1** *Let $U \subset H_0^1(\Omega)$ be a non-empty, bounded, convex, and closed subset such that for some $\varepsilon$ $(0 < \varepsilon < 1)$,*

$$R(TU) \ \oplus \ RE(TU) \ \overset{\circ}{\subset} \ U, \tag{8}$$

*then there exists a solution of $u = Fu$ in $U$.*
*Here, $\oplus$ denotes the direct sum in the sense of $H_0^1(\Omega)$ and $M_1 \overset{\circ}{\subset} M_2$ implies $\overline{M_1} \subset \overset{\circ}{M_2}$ for any sets $M_1$, $M_2$.*

Next, we propose a computer algorithm to construct the set $U$ which satisfies the verification condition (8). In order to realize it, we use the iteration method described below.

Let $\{\phi_j\}_{j=1}^M$ be a basis of finite element subspace $S_h$, and $\Theta_h$ be the set of all linear combinations of $\{\phi_j\}_{j=1}^M$ with interval coefficients. Note that each element of $\Theta_h$ is defined as a subset of $S_h$, namely, $\omega \in \Theta_h$ means that

$$\omega = \sum_{j=1}^{M} A_j \phi_j \equiv \{ \ \sum_{j=1}^{M} a_j \phi_j \mid a_j \in A_j \ \}, \tag{9}$$

where $A_j$ are real intervals.
Now, we denote the set of all nonnegative real numbers by $\mathbf{R}^+$ and for any $\alpha \in \mathbf{R}^+$, set

$$[\alpha] \equiv \{ \ \phi \in S_h^{\perp} \mid \|\phi\|_{H_0^1(\Omega)} \leq \alpha \ \text{ and } \ \|\phi\|_{L^2(\Omega)} \leq Ch\alpha \ \}.$$

This set corresponds to the rounding error defined above. Let $u_h \in S_h$ be a given approximation of the solution to (2), and let $\sigma$ be a positive constant ($\sigma \ll 1$). We also choose $\alpha_0 \in \mathbf{R}^+$ and $\delta u_h^0 \in \Theta_h$ as appropriate initial value and element, respectively, and set as $\alpha_0 = 0$ and $\delta u_h^0 = \{0\}$. We now define the following iteration for $n \geq 1$ :

For $\delta u_h^{n-1}$ and $\alpha_{n-1}$, we set

$$\left\{ \begin{array}{lcl} \delta \tilde{u}_h^{n-1} & \equiv & \delta u_h^{n-1} + \sum_{j=1}^{M} [\,-1, 1\,]\, \sigma\, \phi_j \ , \\ \tilde{\alpha}_{n-1} & \equiv & \alpha_{n-1} + \sigma \ , \end{array} \right. \tag{10}$$

which is called $\sigma$-*inflation*. And, for $U^{n-1} \equiv u_h + \delta \tilde{u}_h^{n-1} + [\tilde{\alpha}_{n-1}]$, define $\delta u_h^n$ and $\alpha_n$ by

$$\left\{ \begin{array}{lcl} \delta u_h^n & \equiv & \tilde{T} U^{n-1} - u_h \ , \\ \alpha_n & \equiv & Ch\lambda \sup_{u \in U^{n-1}} \| f(u) \|_{L^2(\Omega)} \ . \end{array} \right. \tag{11}$$

Here, $C$ is the constant defined in (7). Then using Lemma 1, the following theorem holds ( the proof is a direct consequence of the arguments in [12] ).

**Theorem 1** *If for some $n \geq 1$, two relationships*

$$\left\{ \begin{array}{lcl} \delta u_h^n & \overset{\circ}{\subset} & \delta \tilde{u}_h^{n-1}, \\ \alpha_n & < & \tilde{\alpha}_{n-1} \end{array} \right. \tag{12}$$

*hold, then there exist a solution of* (2) *in* $u_h + \delta u_h^n + [\alpha_n]$.
*Here, the first term of* (12) *means the strict inclusion in the sense of each coefficient interval in* $\delta \tilde{u}_h^{n-1}$ *and* $\delta u_h^n$.

# 4 Computation procedure for verification

In the previous section, we presented a general idea of the verification procedure as a straightforward extension of the method described in [12], but the detailed computing algorithm is not yet clear. A concrete algorithm to realize the above procedure is described in [12]. But it appears that this computing procedure is not always useful from the view point of interval arithmetic. In this section, we give an efficient computing algorithm to implement the iterative sequence defined in (11).

First, we rewrite (11) as

$$\begin{array}{rcl} \delta u_h^n & = & -([I - F'(u_h)]_h^{-1} + \varepsilon \tilde{I})(\tilde{I} - \tilde{F}) u_h + ((1-\varepsilon)\tilde{I} - [I - F'(u_h)]_h^{-1}) \delta \tilde{u}_h^{n-1} \\ & & + ([I - F'(u_h)]_h^{-1} + \varepsilon \tilde{I})(\tilde{F} U^{n-1} - \tilde{F} u_h). \end{array} \tag{13}$$

Note that the first term of (13):

$$-([I - F'(u_h)]_h^{-1} + \varepsilon \tilde{I})(\tilde{I} - \tilde{F}) u_h$$

is independent of iteration and can be determined as a fixed single element in $S_h$ according to $u_h$. Therefore, using the real vector $\{a_j\}_{j=1}^M$ , we denote this term by

$$\sum_{j=1}^{M} a_j\, \phi_j \ . \tag{14}$$

On the other hand, the second and the third terms of (13):

$$((1-\varepsilon)\tilde{I}-[I-F'(u_h)]_h^{-1})\delta\tilde{u}_h^{n-1}+([I-F'(u_h)]_h^{-1}+\varepsilon\tilde{I})(\tilde{F}U^{n-1}-\tilde{F}u_h)$$

can be enclosed as an element of the power set $2^{S_h}$ in the sense of (9). We write this enclosure as

$$\sum_{j=1}^{M} B_j^n\,\phi_j\ . \tag{15}$$

Here, $\{\ B_j^n\ \}_{j=1}^M$ is an interval vector. Thus, $\delta u_h^n$ is determined as $\sum_{j=1}^{M}(\ a_j+B_j^n\ )\phi_j$ in a computer.

Next, we consider the explicit form to obtain $\{a_j\}_{j=1}^M$ and $\{B_j^n\}_{j=1}^M$ . First, we shall determine $a_j$. Operating $[I-F'(u_h)]_h \equiv P_h[I-F'(u_h)]$ on both sides of

$$\sum_{j=1}^{M} a_j\,\phi_j\ =\ -([I-F'(u_h)]_h^{-1}+\varepsilon\tilde{I})(\tilde{I}-\tilde{F})u_h,$$

yields

$$\sum_{j=1}^{M} a_j\,[I-F'(u_h)]_h\phi_j=-(\tilde{I}-\tilde{F})u_h-\varepsilon[I-F'(u_h)]_h\,(\tilde{I}-\tilde{F})u_h$$

which implies that for each $\phi_i$ ( $i=1,2,...,M$ ),

$$\sum_{j=1}^{M} a_j(\ \nabla[I-F'(u_h)]_h\phi_j,\nabla\phi_i)=-(\ \nabla(\tilde{I}+\varepsilon[I-F'(u_h)]_h)(u_h-\tilde{F}u_h),\nabla\phi_i)\,.$$

Since the initial value $u_h$ is an element in $S_h$, and $\tilde{F}u_h$ is also determined as an element of $S_h$ by $f$, we can write them by the following forms :

$$u_h\equiv\sum_{j=1}^{M} b_j\phi_j \qquad (b_j\in\mathbf{R}\quad 1\le j\le M), \tag{16}$$

$$\tilde{F}u_h\equiv\sum_{j=1}^{M} c_j\phi_j \qquad (c_j\in\mathbf{R}\quad 1\le j\le M). \tag{17}$$

Thus, from (16) and (17), we obtain

$$\sum_{j=1}^{M} a_j(\ \nabla[I-F'(u_h)]_h\phi_j,\nabla\phi_i)=\sum_{j=1}^{M}(c_j-b_j)(\ \nabla(\tilde{I}+\varepsilon[I-F'(u_h)]_h)\phi_j,\nabla\phi_i). \tag{18}$$

Next, in order to make the computing procedure clearer, we rewrite (18) by the product of matrix and vector. We denote for each $1\le i,j\le M$,

$$g_{ji}\equiv(\ \nabla[I-F'(u_h)]_h\phi_j,\nabla\phi_i\ ), \tag{19}$$

$$d_{ji}\equiv(\ \nabla\phi_j,\nabla\phi_i\ ), \tag{20}$$

and the corresponding $M\times M$ matrices to the components $g_{ji}$ and $d_{ji}$ by $G$ and $D$, respectively. From the assumption A3, $G$ is invertible, and the matrix $D$ also has an inverse matrix by property of finite element subspace $S_h\subset H_0^1(\Omega)$ ( e.g.[2] ). Then, the following result holds.

**Proposition 1** *The coefficient $a_j$ of (14) is determined by*

$$(a_j) = (\, G^{-1}D + \varepsilon E)(c_i - b_i),$$

*where $E$ is the unit matrix and $(\, a_i\,), (\, b_i - c_i\,)$ denote vectors which correspond to the quantities which appeared in (18).*

In order to determine the interval vectors $B_j^n$, we need some additional consideration. Let $\psi$ be a subset of $S_h$ defined by

$$\psi \equiv ((1-\varepsilon)\tilde{I} - [I - F'(u_h)]_h^{-1})\delta\tilde{u}_h^{n-1} + ([I - F'(u_h)]_h^{-1} + \varepsilon\tilde{I})(\tilde{F}U^{n-1} - \tilde{F}u_h),$$

and we shall enclose $\psi$ by $\sum_{j=1}^{M} B_j^n \phi_j \in \Theta_h$. Operating $[I - F'(u_h)]_h$ on both sides of the above, we obtain

$$[\, I - F'(u_h)\,]_h\psi = (\, (1-\varepsilon)[\, I - F'(u_h)\,]_h - \tilde{I}\,)\delta\tilde{u}_h^{n-1} + (\tilde{I} + \varepsilon[\, I - F'(u_h)\,]_h\,)(\, \tilde{F}U^{n-1} - \tilde{F}u_h\,). \quad (21)$$

We denote the enclosure of $\tilde{F}U^{n-1} - \tilde{F}u_h$ in $\Theta_h$ using interval coefficients $K_j^{n-1}$ as

$$\sum_{j=1}^{M} K_j^{n-1}\phi_j\ , \quad (22)$$

and also set $\delta\tilde{u}_h^{n-1} \equiv \sum_{j=1}^{M} A_j^{n-1}\phi_j$. Then from (21), it suffices to determine $\psi$ satisfying, for each $\phi_i\ \ (i = 1, 2, ..., M)$,

$$\begin{aligned}(\, \nabla[I - F'(u_h)]_h\psi\,,\, \nabla\phi_i\,) \ \subset\ & \sum_{j=1}^{M} A_j^{n-1}(\, \nabla\{(1-\varepsilon)[I - F'(u_h)]_h - \tilde{I}\,\}\phi_j\,,\, \nabla\phi_i\,) \\ & + \sum_{j=1}^{M} K_j^{n-1}(\, \nabla\{\tilde{I} + \varepsilon[I - F'(u_h)]_h\,\}\phi_j\,,\, \nabla\phi_i\,).\end{aligned}$$

Thus, we can determine $B_j^n$ by the following products of vector and matrix using (19) , (20)

$$(B_j^n) = (\, (1-\varepsilon)E - G^{-1}D\,)(A_i^{n-1}) + (G^{-1}D + \varepsilon E)(K_i^{n-1}). \quad (23)$$

Each term of (23) is represented as the product of scalar valued matrix and interval vector. However, in the case of the form (23), from the nature of interval arithmetic, it is not easy to obtain the converging sequence which satisfies the verification condition, because both terms in (23) contain interval vectors with the same order of magnitude. Therefore, we modify it as follows:

According to the properties of $\tilde{F}$,

$$(\, \nabla(\, \tilde{F}U^{n-1} - \tilde{F}u_h\,), \nabla\phi_i\,) = \lambda(\, f(U^{n-1}) - f(u_h), \phi_i\,).$$

Since $U^{n-1} = u_h + \delta\tilde{u}_h^{n-1} + [\tilde{\alpha}_{n-1}]$, $f(U^{n-1}) - f(u_h)$ can be represented as $f'(u_h)\delta\tilde{u}_h^{n-1} + k(\, u_h, \delta\tilde{u}_h^{n-1}, [\tilde{\alpha}_{n-1}]\,)$ with a suitable function representing the higher order term in $\delta\tilde{u}_h^{n-1}$. Thus

$$(\, \nabla(\, \tilde{F}U^{n-1} - \tilde{F}u_h\,), \nabla\phi_i\,) = \sum_{j=1}^{M} A_j^{n-1}\,\lambda(\, f'(u_h)\,\phi_j, \phi_i\,) + \lambda(\, k(\, u_h, \delta\,\tilde{u}_h^{n-1}, [\tilde{\alpha}_{n-1}]\,), \phi_i\,).$$

On the other hand, noting that the definition of Fréchet derivative implies

$$\begin{aligned} g_{ji} &= (\,\nabla[I - F'(u_h)]_h\phi_j, \nabla\phi_i\,) \\ &= d_{ji} - \lambda(\,f'(u_h)\,\phi_j, \phi_i\,), \end{aligned}$$

we can determine interval coefficients $K_j^{n-1}$ as

$$(K_i^{n-1}) = (E - D^{-1}G)(A_j^{n-1}) + D^{-1}(\hat{K}_j^{n-1}), \tag{24}$$

where we used an interval vector $\{\hat{K}_j^{n-1}\}_{j=1}^M$ such that $\hat{K}_i^{n-1}$ enclose $\lambda(\,k(\,u_h, \delta\tilde{u}_h^{n-1}, [\tilde{\alpha}_{n-1}]\,), \phi_i\,)$ for each $i$. Then, we substitute (24) into (23). Noting that two intervals with the same symbol denote originally the same value, we can use the distributive law which is against the usual interval arithmetic. Thus we obtain the following result.

**Proposition 2** *Interval coefficients $B_i^n$ in (15) are determined by*

$$(B_i^n) = -\varepsilon D^{-1}G(A_j^{n-1}) + (G^{-1} + \varepsilon D^{-1})(\hat{K}_j^{n-1}). \tag{25}$$

**Remark 1** *Since $\varepsilon$ in the right-hand side of (25) is a small parameter, the values of $(B_i^n)$ are nearly equal to $G^{-1}(\hat{K}_j^{n-1})$. Hence, it is expected that the computation by (25) yields the improvement in convergence property as compared with (23), because of the characteristic of interval arithmetic. Indeed, this will be confirmed by the numerical examples in the following section.*

# 5 Numerical examples

## 5.1 One dimensional case

Consider the following two point boundary value problem which is related to the equation in mathematical biology.

$$\begin{cases} -u'' = \lambda u(u-a)(1-u) \quad \text{in} \quad (\,0,1\,), \\ u(0) = 0 \;\;,\;\; u(1) = 0. \end{cases} \tag{26}$$

We divide the interval $\Omega = (\,0,1\,)$ into N equal parts and set

$$x_i = \frac{i}{N} \quad (i = 0, 1, ...., N),$$

$$\Omega_i = (\,x_{i-1}, x_i\,) \quad (i = 1, 2, ...., N),$$

$$h = \frac{1}{N}\,.$$

Also, let $P_1(\Omega_i)$ denote the set of linear polynomials on $\Omega_i$ and define the finite element subspace $S_h$ by

$$S_h \equiv \{\, v \in C(\Omega) \mid\; v|_{\Omega_i} \in P_1(\Omega_i)\,,\; 1 \le i \le N\,,\; v(0) = v(1) = 0\,\}. \tag{27}$$

Then, $\dim S_h = N - 1$, and we now choose the basis of $S_h$ as the following hat functions :

$$\phi_j(x_k) = \begin{cases} 1 \quad (k = j), \\ \\ 0 \quad (k \neq j), \end{cases} \qquad \text{for} \;\; 1 \le j, k \le N\,.$$

We can take the constant $C$ which previously appeared as $\frac{1}{\pi}$ ( cf. [12] ).

In the present case, the $H_0^1-$rounding $R\phi$ coincides with the interpolation of $\phi$ at each node $x_i$ ( cf. [9] ). Hence, the value of $\alpha_n$ at $x_i$ may be taken as zero for each iteration step. Therefore, when we complete the verification with the set $u_h + \delta u + [\alpha]$, the values of exact solution $u$ at each node $x_i$ are included in the interval $u_h(x_i) + \delta u(x_i)$ and we do not need to take into account of the error $[\alpha]$. We choose $\sigma = 10^{-6}$, $\varepsilon = 10^{-4}$, $\delta u_h^0 = 0$ and $\alpha_0 = 0$.

In order to obtain the approximate solution $u_h$, we use the following method. First, starting from the approximate solution $v_h$ which is a finite sum of eigenfunctions of the Laplace operator ( [4] ), we iterate the scheme (10), (11) with setting identically $\alpha_n = 0$ and $\sigma = 0$. Therefore, the coefficients of $\delta u_h^n = \sum_{j=1}^{M} A_j^n \phi_j$ are determined as a point vector, and if $\max_{1 \le j \le M} |A_j^n - A_j^{n-1}| < 10^{-8}$ holds, then we stop the iteration and adopt $v_h + \delta u_h^n$ as the initial value $u_h$.

It is known that the equation (26) has two non-trivial solutions for arbitrary $\lambda > \lambda^*$ with a certain positive $\lambda^*$ whose exact value is known as $\frac{4\pi^2}{(1-a)^2}$ provided that $0 < a < \frac{1}{2}$ ( cf. [15] ). Fig.1 shows the outline of the bifurcation diagram derived by the spectral method [4]. The axes of abscissa and ordinate show the value of $\lambda$ and the $L^\infty$-norm of the solutions, respectively.

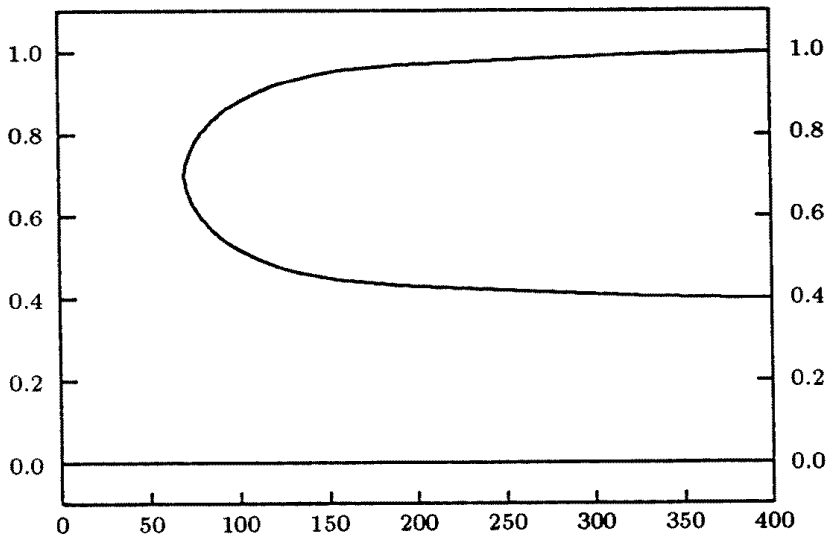


Fig.1. Bifurcation diagram for the solution of (27) , $a$=0.25.

We could verify the solution to (26) with several combinations od data of $\lambda$ and $a$.

**Example 1**

$\lambda = 400$, $a = 0.25$, number of partitions $N = 100$, iteration numbers: 15, $\alpha_{15} = 0.1042$. Fig.2 shows the assured intervals for an exact solution on the upper branch at each mesh point. Namely, it is verified that there exists a solution whose range at mesh points are included between two curves.

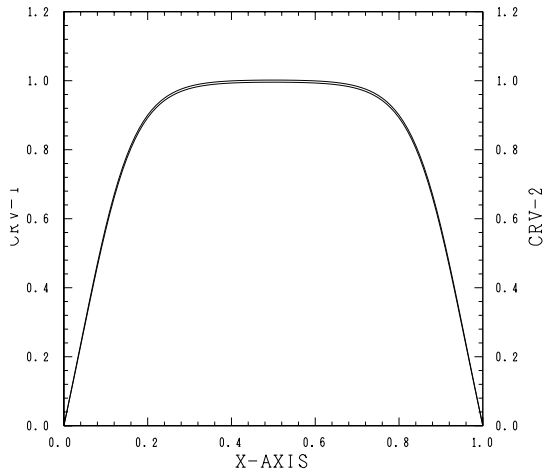


Fig.2. Range of the upper branch solution for $a$=0.25, $\lambda$=400.

**Example 2**

$\lambda = 400$, $a = 0.25$, number of partitions $N = 300$, iteration numbers: 6, $\alpha_6 = 0.0077$. Fig.3 shows the shape of the lower branch solution with the same data as in Example 1.

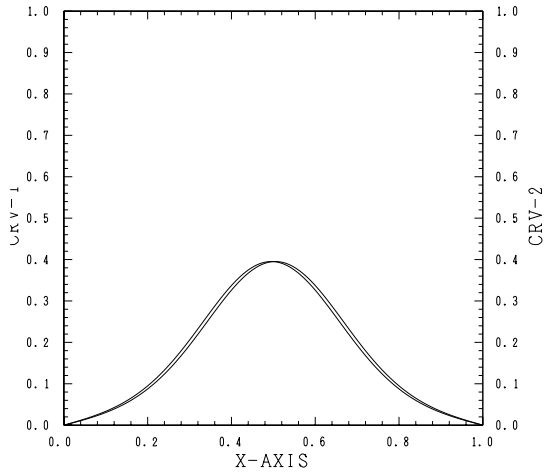


Fig.3. Range of the lower branch solution for $a$=0.25, $\lambda$=400.

Other examples can be verified for this problem, e.g. $\lambda = 80$, $a$=0.25 with nearly turning point, and $\lambda = 150$, $a$=0.01 etc.

**Remark 2** *We could not get these solutions by the original algorithm described in* [12] *owing to the divergence of the iterative sequence* (10), (11). *This fact shows that our procedure defined by Proposition 1 and 2 provides a significant improvement to that of* [12] .

## 5.2 Two dimensional case

First, we consider the same type equation as the one dimensional problem :

$$\begin{cases} -\Delta u &= \lambda u(u-a)(1-u) & \text{in} \quad \Omega, \\ u &= 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{28}$$

Let $\Omega$ be a rectangular domain in $\mathbf{R}^2$ such that $\Omega = (\,0,1\,) \times (\,0,1\,)$. Also let $\delta_x : 0 = x_0 < x_1 < \cdots x_n = 1$ be a uniform partition, and let $\delta_y$ be the same partition as $\delta_x$ for $y$ direction. We define the partition of $\Omega$ by $\delta \equiv \delta_x \otimes \delta_y$. Further, we define the finite element subspace $S_h$ by $S_h \equiv \mathcal{M}_0^1(x) \otimes \mathcal{M}_0^1(y)$ , where $\mathcal{M}_0^1(x)$ , $\mathcal{M}_0^1(y)$ are sets of piecewise linear polynomials on $(\,0,1\,)$ defined by (27) in the variables $x$ and $y$, respectively. We can also take the constant $C$ as $\dfrac{1}{\pi}$ ( cf.[9] ).

We set the parameters $\sigma$, $\varepsilon$ and initial data $\delta u_h^0$, $\alpha_0$ the same as in the one dimensional case, and the initial value is determined by an analogous method. It is known that for $\lambda > 0$, the bifurcation diagram of the solutions are similar to that of the one dimensional case ( cf.[7] ), but the exact value $\lambda^*$ corresponding to the turning point is unknown.

**Example 3**

$\lambda = 150$, $a = 0.01$, number of partitions $N = 80$, iteration numbers: 14, $\alpha_{14} = 0.0622$. Fig.4 shows the outline of the shape for the upper branch solution of (28) along the line $y$=0.5 . That is, there exists a solution between those two curves with additional $H_0^1$-error $\alpha_{14}$ .

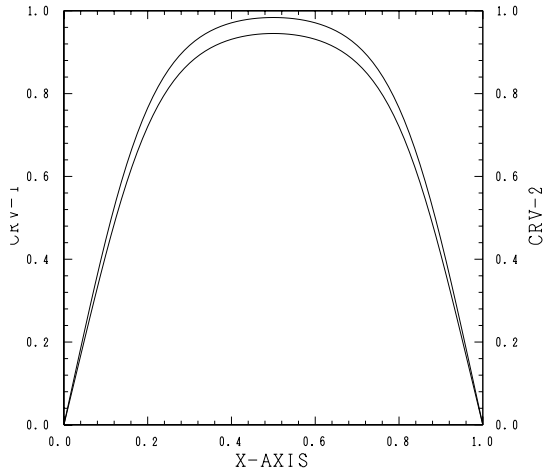


Fig.4. Range of the upper branch solution for $a$=0.01, $\lambda$=150.

**Example 4**

$\lambda = 150$, $a = 0.01$, number of partitions $N = 80$, iteration numbers: 12, $\alpha_{12} = 0.01066$.

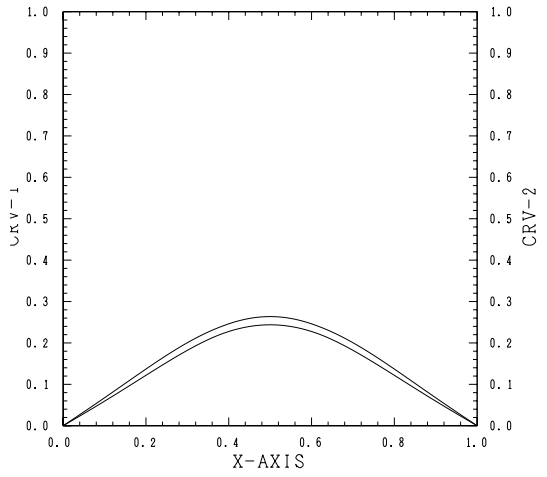


Fig.5. Range of the lower branch solution for $a$=0.01, $\lambda$=150.

Next, we consider Emden's equation:

$$\begin{cases} -\Delta u &= u^2 & \text{in} \quad \Omega, \\ u &= 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{29}$$

Let $\Omega$ be $\Omega = (0,1) \times (0,1)$. In this case, the approximate solution $u_h \in S_h$ of (29) yields that $||u_h||_{L^\infty(\Omega)} \approx 30$, then because of the large starting error $\alpha_1$ (for example, $\alpha_1 \approx 4$ with $N = 80$), it is difficult to obtain a convergent iterative sequence (10),(11), if $h$ is not so small. For very small $h$ ( e.g. $h \leq 10^{-3}$), we may expect the convergence, but such a mesh size could not be used because of the limitation of our computer facility. To overcome this difficulty, we take the approximate solution $u_h$ as an element in $H^2(\Omega)$, and consider the following problem:

$$\begin{cases} -\Delta v &= v^2 + 2u_h v + \Delta u_h + {u_h}^2 & \text{in} \quad \Omega, \\ v &= 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{30}$$

It is easily seen that $u = v + u_h$ is the solution of (29), and therefore, if we verify the weak solution of (30), it means the verification of (29). Note that if $u_h$ is chosen as a sufficiently good approximation of the exact solution to (30), then the norm of the right-hand side becomes very small, and thus the error $\alpha$ as well.

In the actual program, for the first approximate solution $\hat{u}_h$ of (29), we took a piecewise linear function in $S_h$ and set $u_h$ as a pseudo-Hermite interpolation of $\hat{u}_h$ ( cf.[1] ). Thus we could verify the solution of (29) as below.

**Example 5**

Number of partitions $N = 80$, iteration numbers: 12, $\alpha_{12} = 0.06661$. The residual error for initial approximation: $\|\Delta u_h + {u_h}^2\|_{L^2(\Omega)} \approx 14.7$. Fig.6 shows the outline of the shape for the solution of (29) along the line $y$=0.5.

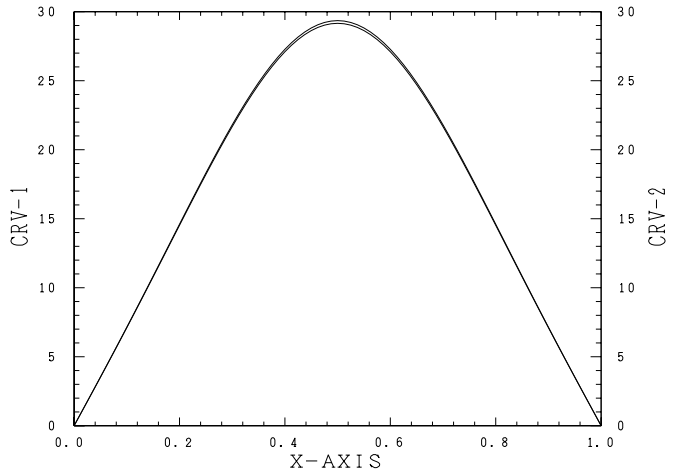


Fig.6. Range of the nontrivial solution of Emden's Equation.

**Remark 3** *In these examples, we used the piecewise linear elements for each iteration. This means that our method can also be applied to the more general domain $\Omega$, e.g. polygonal domain with triangulation and that the verification program is simpler than that of using smooth function such as in* [13].

**Remark 4** *In the above calculations, we used computer arithmetic with double precision instead of strict interval computations (e.g. ACRITH , PASCAL-SC etc.). But from our experiences, the order of magnitude for the effect of round-off is under $10^{-10}$. Therefore, it is almost negligible compared with the truncation error which amounts to $10^{-3} \sim 10^{-2}$. Of course, we have to use those verification software systems if it is necessary to verify the problems whose solutions are mathematically unknown.*

# References

[1] H.Akima : Bivariate Interpolation and Smooth Surface Fitting Based on Local Procedures, *Comm. of the ACM*, **17** (1974), 26—31.

[2] O.Axelsson and V.A.Barker : Finite Element Solution of Boundary Value Problems, Theory and Computation, *Academic Press* (1984).

[3] P.G.Ciarlet : The Finite Element Method for Elliptic Problems, *North-Holland, Amsterdam* (1978).

[4] J.C.Eilbeck : The Pseudo-Spectral Method and Following in Reaction-Diffusion Bifurcation Studies, *SIAM J.Sci.Stat.Comput.*, **17** (1986), 599—610.

[5] E.W.Kaucher and W.L.Miranker : Self-Validating Numerics for Function Space Problems, *Academic Press, New York* (1984).

[6] G.Kedem : A Posteriori Error Bounds for Two-point Boundary Value Problems, *SIAM J.Numer.Anal.*, **18** (1981), 431—448.

[7] P.L.Lions : On the Existence of Positive Solutions of Semilinear Elliptic Equations, *SIAM Review,* **24** (1982), 441—467.

[8] R.J.Lohner : Enclosing the Solutions of Ordinary Initial and Boundary Value Problems, *Computerarithmetic ( eds. E.Kaucher et al.), B.G.Teubner, Stuttgart* (1987), 255—286.

[9] 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.

[10] M.T.Nakao : A Numerical Approach to the Proof of Existence of Solutions for Elliptic Problems Ⅱ, *Japan Journal of Applied Mathematics,* **7** (1990), 477—488.

[11] M.T.Nakao : A Computational Verification Method of Existence of Solutions for Nonlinear Elliptic Equations, *Lecture Notes in Num. Appl. Anal.,* **10** (1988), 101—120. *In proc. Recent Topics in Nonlinear PDE* 4, *Kyoto, 1988,North-Holland / Kinokuniya,* (1989).

[12] M.T.Nakao : A Numerical Verification Method for the Existence of Weak Solutions for Nonlinear BVP, *to appear in Journal of Mathematical Analysis and Applications,* **163** (1992).

[13] M.Plum : Numerical Existence Proofs and Explicit Bounds for Solutions of Nonlinear Elliptic Boundary Value Problems, *preprint.*

[14] J.Schröder : A Method for Producing Verified Results for Two-Point Boundary Value Problems, *Computing Suppl.,* **6** (1988), 9—22.

[15] J.Smoller : Shock Waves and Reaction-Diffusion Equations, *Springer, NewYork,* (1983).

[16] E.Zeidler : Nonlinear Functional Analysis and its Applications 1, *Springer, NewYork* (1986).