Numerical Algorithms **37:** 311–323, 2004.

An efficient approach to the numerical verification for solutions of elliptic differential equations

Mitsuhiro T. Nakao [a] and Yoshitaka Watanabe [b]

[a] *Faculty of Mathematics, Kyushu University, Fukuoka 812-8581, Japan*
E-mail: mtnakao@math.kyushu-u.ac.jp
[b] *Computing and Communications Center, Kyushu University, Fukuoka 812-8581, Japan*
E-mail: watanabe@cc.kyushu-u.ac.jp

The authors and their colleagues have developed numerical verification methods for solutions of second-order elliptic boundary value problems based on the infinite-dimensional fixed-point theorem using the Newton-like operator with appropriate approximation and constructive a priori error estimates for Poisson's equations. Many verification results show that the authors' methods are sufficiently useful when the equation has no first-order derivative. However, in the case that the equation includes the term of a first-order derivative, there is a possibility that the verification algorithm does not work even though we adopt a sufficiently accurate approximation subspace. The purpose of this paper is to propose an alternative method to overcome this difficulty. Numerical examples which confirm the effectiveness of the new method are presented.

Keywords: elliptic equations, numerical verification, fixed-point theorem

AMS subject classification: 35J65, 65G20, 47H10

1. Introduction

Consider the following nonlinear elliptic boundary value problem:

$$\begin{aligned} -\Delta u &= f(x, u, \nabla u), & x &\in \Omega, \\ u &= 0, & x &\in \partial\Omega, \end{aligned} \tag{1}$$

where Ω is a bounded convex domain in $\mathbb{R}^n$ ($n = 1, 2, 3$) with piecewise smooth boundary $\partial\Omega$ and the map f is assumed to satisfy appropriate conditions described later.

The aim of *numerical verification* is to verify by computer the existence of the solution u of (1) around a given approximate solution u_h with guaranteed error bounds. The verification principle was originated by one of the authors [4] and several improvements have been made up to now. First, problem (1) is rewritten in infinite-dimensional fixed-point form, and then the equation is decomposed into the finite-dimensional part and the infinite-dimensional error part. When both the finite and the infinite-dimensional parts are simultaneously contraction maps under suitable assumptions, an infinite-dimensional fixed-point theorem implies the existence of the solution in a certain function set (see [5,6], etc.). For the finite-dimensional part, self-validating Newton-like iterations are performed by the computer. On the other hand, for the infinite-dimensional part, successive iterations using the a priori error estimates of the projection for the solution of Poisson's equations are applied. These two processes are connected with each other at every iteration step.

Many verification results up to the present show that this procedure is sufficiently useful if applied to equations having no first-order derivative term [6]. This is based on a principle of contractiveness, namely, local contraction of the Newton operator for the finite-dimensional part and the naturally contractive property for the error part from the degree of approximation when the dimension of the finite-dimensional approximate subspace increases. However, in the case that f includes a first-order derivative, it was proved that there exist some cases for which the effect of error estimations no longer work in the Newton-like iterations, and as a result, the verification could fail even if very fine approximate subspaces are used.

In this paper, in order to overcome this difficulity, we propose an improvement of the verification method by using norm estimates instead of solving the interval linear system in the Newton-like iterations for the finite-dimensional part. Particularly note that, in these norm estimations, it is not necessary to bound the inverse operator of linearlized problems which is needed in another verification approach proposed by Plum [8].

The contents of this paper are as follows. In section 2, we define some function spaces and notations, and then the fixed-point formulation and the constructive a priori error estimates of the H_0^1-projection for Poisson's equations are described. Next, the basic verification principle using a Newton-like iteration is considered. The current and improved computable verification conditions are given in section 3. Numerical examples which show advantages of the new method are presented in section 4.

2. Fixed-point formulation

For an integer m, let $H^m(\Omega)$ denote the L^2-Sobolev space of order m on Ω. We define

$$H_0^1(\Omega) := \{u \in H^1(\Omega) \mid u = 0,\ x \in \partial\Omega\}$$

with the inner product $(\nabla u, \nabla v)_{L^2}$ and the norm $\|u\|_{H_0^1(\Omega)} := \|\nabla u\|_{L^2(\Omega)}$, where $(u, v)_{L^2}$ implies the L^2-inner product on Ω. The nonlinear operator $f : H_0^1(\Omega) \to L^2(\Omega)$ is supposed to be continuous, Fréchet differentiable on $H_0^1(\Omega)$ and also map a bounded set in $H_0^1(\Omega)$ to a bounded set in $L^2(\Omega)$. We rewrite (1) in the following weak form: find $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla v)_{L^2} = \big(f(\cdot, u, \nabla u), v\big)_{L^2}, \quad \forall v \in H_0^1(\Omega). \tag{2}$$

It is well known [3] that for any $\xi \in L^2(\Omega)$ Poisson's equations

$$\begin{aligned} -\Delta\psi &= \xi, \quad x \in \Omega, \\ \psi &= 0, \quad x \in \partial\Omega, \end{aligned} \tag{3}$$

have a unique solution $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$ and the estimate

$$|\psi|_{H^2(\Omega)} \leqslant \|\xi\|_{L^2(\Omega)} \tag{4}$$

holds, where $|\cdot|_{H^2(\Omega)}$ means the semi-norm on $H^2(\Omega)$ defined by

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

For $\xi \in L^2(\Omega)$ let $A\xi$ be the solution of (3), then the operator $A : L^2(\Omega) \to H_0^1(\Omega)$ is compact because of the compactness of the imbedding $H^2(\Omega) \hookrightarrow H^1(\Omega)$ [1]. From assumptions of f, the nonlinear operator F defined by

$$F := A \circ f$$

is also a compact map on $H_0^1(\Omega)$. Then the weak form (2) can be rewritten equivalently as the fixed-point form

$$u = F(\cdot, u, \nabla u) \tag{5}$$

in $H_0^1(\Omega)$. In the following, for simplicity, we denote $f(u) := f(\cdot, u, \nabla u)$ and $Fu := F(\cdot, u, \nabla u)$.

Next we introduce a Newton-like method for verification [5]. Let S_h be an approximate finite-dimensional subspace of $H_0^1(\Omega)$ dependent on the parameter h. For example, S_h is taken to be a finite element subspace with mesh size h. Also let $P_h : H_0^1(\Omega) \to S_h$ denote the H_0^1-projection defined by, for each element $\phi \in H_0^1(\Omega)$,

$$\big(\nabla(\phi - P_h\phi), \nabla v\big)_{L^2} = 0, \quad \forall v \in S_h. \tag{6}$$

Now we suppose the following approximation property of P_h:

$$\|v - P_h v\|_{H_0^1(\Omega)} \leqslant Ch|v|_{H^2(\Omega)}, \quad \forall v \in H_0^1(\Omega) \cap H^2(\Omega), \tag{7}$$

where $C > 0$ is a positive constant numerically determined. This assumption holds for many finite element subspaces of $H_0^1(\Omega)$ [2,6,7] or function spaces of Fourier series with finite truncation [13]. Since S_h is the closed subspace of $H_0^1(\Omega)$, each element of $H_0^1(\Omega)$ can be uniquely represented as the direct sum of the elements of S_h and $S_h^\perp$. Here $S_h^\perp$ stands for the orthogonal complement subspace of S_h in $H_0^1(\Omega)$. Therefore, the fixed-point equation $u = Fu$ in $H_0^1(\Omega)$ can also be uniquely decomposed as the finite-dimensional (projection) part and the infinite-dimensional (error) part of the form

$$\begin{aligned} P_h u &= P_h F u, \\ (I - P_h)u &= (I - P_h)Fu. \end{aligned} \tag{8}$$

In order to obtain a solution satisfying (8), we fix an approximate solution $u_h \in S_h$ of (2) and define the nonlinear operator $N_h : H_0^1(\Omega) \to S_h$ by

$$N_h u := P_h u - \big[I - P_h F'(u_h)\big]_h^{-1} P_h(u - Fu),$$

where $F'(u_h)$ means the Fréchet derivative of F at u_h and $[I - P_hF'(u_h)]_h^{-1}$ denotes the inverse on S_h of the restriction operator $P_h(I - P_hF'(u_h))|_{S_h}$. Note that existence of $[I - P_hF'(u_h)]_h^{-1}$ is equivalent to the nonsingularity of a matrix G defined later by (14) in section 3, which is able to be numerically confirmed in the actual verification process. Because of the equivalency of $P_hu = P_hFu$ and $P_hu = N_hu$, defining the operator T on $H_0^1(\Omega)$ by

$$Tu := N_hu + (I - P_h)Fu, \tag{9}$$

two fixed-point problems $u = Tu$ and $u = Fu$ are also equivalent. Therefore Schauder's fixed-point theorem asserts that if, for a nonempty, bounded, convex and closed set $U \subset H_0^1(\Omega)$,

$$TU = \{Tu \mid u \in U\} \subset U$$

holds, then there exists a fixed-point of T in U. We call such a set U, expected to be $TU \subset U$, as a *candidate set*. If a candidate set U is chosen such as

$$U := u_h + U_h + U_*, \quad U_h \subset S_h,\ U_* \subset S_h^{\perp}, \tag{10}$$

then the verification condition $TU \subset U$ can be written in the form

$$\begin{aligned} N_hU - u_h &\subset U_h, \\ (I - P_h)FU &\subset U_*. \end{aligned} \tag{11}$$

Note that, from the definition of the operator T, when the approximate solution $u_h \in S_h$ is sufficiently good, the finite-dimensional part of T will possibly be contractive. On the other hand, the magnitude of the infinite-dimensional part of T, i.e., $(I - P_h)Fu$, is expected to be small when the parameter h of S_h is sufficiently small because of the approximation property (7) with (4) of P_h.

3. Verification algorithms

This section is devoted to the construction of the current verification algorithm and a new computable verification algorithm which generate candidate sets expected to satisfy condition (11).

3.1. *Infinite-dimensional part*

The following general criterion of verification for the infinite-dimensional part is led by the constructive a priori error estimates for Poisson's equations and the Aubin–Nitsche trick [4].

Theorem 1. Let the infinite-dimensional part of the candidate set U_* in (10) be a ball with radius $\alpha > 0$ such as

$$U_* := \left\{u_* \in S_h^{\perp} \mid \|u_*\|_{H_0^1(\Omega)} \leqslant \alpha,\ \|u_*\|_{L^2(\Omega)} \leqslant Ch\alpha\right\}. \tag{12}$$

For a bounded closed subset $U_h \subset S_h$, if the candidate set $U := u_h + U_h + U_*$ satisfies

$$Ch \sup_{u \in U} \| f(u) \|_{L^2(\Omega)} \leqslant \alpha,$$

then the second part of condition (11) is satisfied.

3.2. *Finite-dimensional part using interval coefficients*

We now briefly describe the existing method to compute the finite-dimensional part, i.e., the enclosing method for the former part of the left-hand side of (11) for a candidate set U. Let $K := \dim S_h$ and let $\{\phi_i\}_{1 \leqslant i \leqslant K}$ be a basis of S_h. Also let $\mathbb{I}\mathcal{T}$ denote the interval extension for $\mathcal{T} \in \{\mathbb{R}, \mathbb{R}^K\}$. The set U_h is taken to be a set of linear combinations of base functions in S_h with interval coefficients such as

$$U_h := \sum_{i=1}^{K} [\underline{A}_i, \overline{A}_i] \phi_i. \tag{13}$$

Defining the $K \times K$ matrix $G = (G_{ij})$ by

$$G_{ij} = (\nabla \phi_j, \nabla \phi_i)_{L^2} - \big(f'(u_h)\phi_j, \phi_i\big)_{L^2}, \tag{14}$$

we obtain the following sufficient condition of the first part of (11) [6].

Theorem 2. Assume that the candidate set $U \subset H_0^1(\Omega)$ is defined by (12) and (13) with (10), and denote an arbitrary element $u \in U$ by

$$u = u_h + \hat{u}_h + u_*, \quad \hat{u}_h \in U_h, \ u_* \in U_*.$$

Let $\boldsymbol{d} = (d_i) \in \mathbb{IR}^K$ denote the interval enclosure of the set whose ith component consists of

$$\big\{\big(f(u) - f'(u_h)\hat{u}_h, \phi_i\big)_{L^2} - (\nabla u_h, \nabla \phi_i)_{L^2} \in \mathbb{R} \mid u \in U\big\}, \quad 1 \leqslant i \leqslant K. \tag{15}$$

If, for a K-dimensional interval vector $\boldsymbol{v} = (v_i) \in \mathbb{IR}^K$ enclosing the solution $\boldsymbol{x} \subset \mathbb{R}^K$ for the linear equation

$$G\boldsymbol{x} = \boldsymbol{d}, \tag{16}$$

the conditions

$$v_i \subset A_i, \quad 1 \leqslant i \leqslant K, \tag{17}$$

hold, then the former half of condition (11) is satisfied.

The above interval vector $\boldsymbol{v} \in \mathbb{IR}^K$ is obtained with guaranteed accuracy by various kinds of methods, e.g., [9]. Based on theorems 1 and 2, we usually adopt the following verification algorithm AL-1 using the acceleration method with $\varepsilon > 0$ which is called "ε-inflation".

Verification algorithm AL-1.

- $k = 0$
  Set initial values $A_i^{(0)} \in \mathbb{IR}$ $(1 \leqslant i \leqslant K)$ and $\alpha^{(0)} > 0$.
- $k \geqslant 1$

  1. For a fixed small constant $\varepsilon > 0$ set
  $$\hat{A}_i^{(k)} := (1+\varepsilon)A_i^{(k-1)} \quad (1 \leqslant i \leqslant K), \quad \hat{\alpha}^{(k)} := (1+\varepsilon)\alpha^{(k-1)}.$$
  2. The kth candidate set $U^{(k)}$ is defined by
  $$\begin{aligned} U_h^{(k)} &:= \sum_{i=1}^{K} \hat{A}_i \phi_i \subset S_h, \\ U_*^{(k)} &:= \left\{ v_* \in S_h^{\perp} \mid \|v_*\|_{H_0^1(\Omega)} \leqslant \hat{\alpha}^{(k)},\ \|v_*\|_{L^2(\Omega)} \leqslant Ch\hat{\alpha}^{(k)} \right\}, \\ U^{(k)} &:= u_h + U_h^{(k)} + U_*^{(k)}. \end{aligned}$$
  3. Compute values of the kth iteration by
  $$\begin{aligned} A_i^{(k)} &:= v_i \quad \text{of (16) in theorem 2,} \\ \alpha^{(k)} &:= Ch \sup_{u \in U^{(k)}} \left\| f(u) \right\|_{L^2(\Omega)}. \end{aligned}$$
  4. If $A_i^{(k)} \subset \hat{A}_i^{(k)}$ $(1 \leqslant i \leqslant K)$ and $\alpha^{(k)} \leqslant \hat{\alpha}^{(k)}$ then stop, and there exists a desired solution in $U^{(k)} \subset H_0^1(\Omega)$.
  5. Set $k := k + 1$ and return to step 1. If k reaches a maximum iteration number or some norm of $A_i^{(k)}$ and $\alpha^{(k)}$ exceed certain criteria then stop, and the verification fails.

3.3. Some problems in AL-1

A lot of verification results prove that AL-1 is actually effective if $f(\cdot)$ in (1) has *no* first-order derivative [6]. However, in the case that there is a first-order term in f, the verification algorithm does not always work [5,12]. We now give a simple example. When $f(u) \equiv u'$, each element of the set in (15) becomes

$$-\left(u_h + u_h{}', \phi_i{}'\right)_{L^2} + \left(u_*{}', \phi_i\right)_{L^2}. \tag{18}$$

Here, the first term $-(u_h + u_h{}', \phi_i{}')_{L^2}$ of (18) could be estimated by a narrow interval using usual interval arithmetic. On the other hand, the second term $(u_*{}', \phi_i)_{L^2}$ would be estimated as $|(u_*{}', \phi_i)_{L^2}| \leqslant \alpha\|\phi_i\|_{L^2(\Omega)}$, which generally causes an overestimation and $\boldsymbol{d} = (d_i) \in \mathbb{IR}^K$ is defined as the interval enclosing:

$$d_i := -\left(u_h + u_h{}', \phi_i{}'\right)_{L^2} + [-1, 1]\alpha\|\phi_i\|_{L^2(\Omega)}.$$

Table 1
Computed norms for the linear combination with interval coefficients.

δ	Linear	Fourier	Cubic Hermite
		$\lVert v_h \rVert^2$	
10^{-10}	9.87×10^{-21}	9.91×10^{-19}	9.89×10^{-21}
10^{-5}	9.87×10^{-11}	9.91×10^{-9}	9.89×10^{-11}
10^{-4}	9.87×10^{-9}	9.91×10^{-7}	9.89×10^{-9}
10^{-3}	9.87×10^{-7}	9.91×10^{-5}	9.89×10^{-7}
10^{-2}	9.87×10^{-5}	9.91×10^{-3}	9.89×10^{-5}
10^{-1}	9.87×10^{-3}	9.91×10^{-1}	9.89×10^{-3}
		$\lVert v_h{}' \rVert^2$	
10^{-10}	3.95×10^{-16}	3.29×10^{-15}	4.74×10^{-16}
10^{-5}	3.95×10^{-6}	3.29×10^{-5}	4.74×10^{-6}
10^{-4}	3.95×10^{-4}	3.29×10^{-3}	4.74×10^{-4}
10^{-3}	3.95×10^{-2}	0.329	4.74×10^{-2}
10^{-2}	3.95	32.9	4.74
10^{-1}	395	3283	474

If we take $\Omega = (0, 1)$ and S_h, the set of piecewise linear functions with uniform mesh size h, then $\lVert \phi_i \rVert_{L^2(\Omega)}$ is $(2h/3)^{1/2}$, and note that d_i is $\mathrm{O}(h^{1/2})$. Taking into account that the dimension of the matrix G is proportional to $1/h$, one can deduce that $A_i^{(k)} \subset \hat{A}_i^{(k)}$ might not occur even if h tends to be very small.[1]

Moreover, there might be a possibility that the norm of $\lVert f(u) \rVert_{L^2(\Omega)}$ could be explosive because of the properties of interval arithmetic. For example, table 1 shows the computed norms for $\lVert v_h \rVert^2$ and $\lVert v_h{}' \rVert^2$ of the function v_h of the form $v_h = \sum_{i=1}^{K} [-\delta, \delta]\, \phi_i$ for small $\delta > 0$, for three kinds of base functions $\{\phi_i\}$ on $\Omega = (0, 1)$. Namely, base functions are chosen as: piecewise linear functions (100 uniform partitions) "linear", $\sin(\pi i x)$ $(1 \leqslant i \leqslant 100)$ "Fourier", piecewise cubic Hermite functions (100 uniform partitions) "cubic Hermite". Note that, as indicated in table 1, the norm of the derivatives could be very much larger than we might expect, even if the norm of the function itself and the width of the coefficient intervals are relatively small.

3.4. Improvement of computation for finite-dimensional part

In order to overcome difficulties described in the previous subsection, we propose a method to compute more effectively the finite-dimensional part of the candidate set by using norm estimations. Let U_h be the finite-dimensional term of the candidate set in (10) defined as a ball with radius $\gamma > 0$ of the form

$$U_h := \{\hat{u}_h \in S_h \mid \lVert \hat{u}_h \rVert_{H_0^1(\Omega)} \leqslant \gamma\}. \tag{19}$$

[1] The situation is the same even if one uses estimates such as $|((u_*{}', \phi_i)_{L^2})| \leqslant \lVert u_* \rVert_{L^2(\Omega)} \lVert \phi_i{}' \rVert_{L^2(\Omega)}$.

We define the $K \times K$ positive definite matrix D by $D_{ij} = (\nabla\phi_j, \nabla\phi_i)_{L^2}$, the $K \times K$ lower triangle matrix L by the Cholesky decomposition: $D = LL^T$, and the matrix norm $\rho > 0$ by

$$\rho = \sup_{\|\boldsymbol{x}\|_E=1} \left\| L^{\mathrm{T}} G^{-1} L\boldsymbol{x} \right\|_E, \tag{20}$$

where G is defined in (14) and $\|\cdot\|_E$ is the Euclidean norm of $\mathbb{R}^K$. Then we obtain the following verification condition different from (17) in theorem 2 for the finite-dimensional part.

Theorem 3. Define the candidate set $U = u_h + U_h + U_*$ by (19) and (12), and denote any element $u \in U$ by

$$u = u_h + \hat{u}_h + u_*, \quad \hat{u}_h \in S_h,\ u_* \in S_h^{\perp}.$$

If it holds that

$$\rho \sup_{u\in U} \left\| P_h F u - P_h F'(u_h)\hat{u}_h - u_h \right\|_{H_0^1(\Omega)} \leqslant \gamma, \tag{21}$$

then the former half of condition (11) is satisfied.

Proof. From (11), it is sufficient to show that

$$\|N_h u - u_h\|_{H_0^1(\Omega)} \leqslant \rho \left\| P_h F u - P_h F'(u_h)\hat{u}_h - u_h \right\|_{H_0^1(\Omega)} \tag{22}$$

for each $u = u_h + \hat{u}_h + u_* \in U$.

It is easily shown that for a K-dimensional vector $\boldsymbol{v} := (v_i)$ and $v_h = \sum_{i=1}^K v_i\phi_i \in S_h$, $\|v_h\|_{H_0^1(\Omega)} = \|L^{\mathrm{T}}\boldsymbol{v}\|_E$ holds. Since

$$N_h u - u_h = \left[I - P_h F'(u_h)\right]_h^{-1} P_h\left(F u - F'(u_h)\hat{u}_h - u_h\right), \tag{23}$$

by setting

$$v_h := N_h u - u_h = \sum_{i=1}^{K} v_i\phi_i, \quad \boldsymbol{v} := (v_i) \in \mathbb{R}^K,$$

$$w_h := P_h\left(F u - F'(u_h)\hat{u}_h - u_h\right) = \sum_{i=1}^{K} w_i\phi_i, \quad \boldsymbol{w} := (w_i) \in \mathbb{R}^K,$$

equation (23) is written as

$$P_h\left(I - F'(u_h)\right)v_h = w_h. \tag{24}$$

From the definition of P_h, equation (24) is equivalent to

$$\sum_{i=1}^{K}\left\{(\nabla\phi_i, \nabla\phi_j)_{L^2} - \left(f'(u_h)\phi_i, \phi_j\right)_{L^2}\right\}v_i = \sum_{i=1}^{K}(\nabla\phi_i, \nabla\phi_j)_{L^2} w_i, \quad 1 \leqslant j \leqslant K. \tag{25}$$

Then equation (25) is represented by the matrix and vector form

$$\boldsymbol{v} = G^{-1}A\boldsymbol{w}.$$

Consequently,

$$\begin{aligned}\|N_h u - u_h\|_{H_0^1(\Omega)} &= \|L^{\mathrm{T}}\boldsymbol{v}\|_E = \|L^{\mathrm{T}}G^{-1}A\boldsymbol{w}\|_E \leqslant \|L^{\mathrm{T}}G^{-1}L\|\,\|L^{\mathrm{T}}\boldsymbol{w}\|_E \\ &= \rho\|P_hFu - P_hF'(u_h)\hat{u}_h - u_h\|_{H_0^1(\Omega)},\end{aligned}$$

which proves the theorem. □

In (22), $P_hFu - P_hF'(u_h)\hat{u}_h - u_h$ corresponds to the higher-order residual for the Taylor expansion of P_hFu at u_h. Therefore, inequality (21) would be expected to hold if the radii of U_h and U_* are sufficiently small. In the actual calculation on computer, we would compute the value of $\|P_hFu - P_hF'(u_h)\hat{u}_h - u_h\|_{H_0^1(\Omega)}$ in an over-estimated sense using α, γ and a priori constant C.

The estimate of ρ in (20) is generally reduced to a computation of the largest singular value of a matrix. There are some computational algorithms with result verification to estimate rigorous bounds for the largest (or smallest) singular value (e.g., see [10]). Moreover, the interval Cholesky decomposition algorithm for $D = LL^{\mathrm{T}}$ is usually feasible because of the positive definiteness of the matrix D.

Based on theorem 3, we formulate the following new verification algorithm AL-2.

Verification algorithm AL-2.

- $k = 0$
  Set initial values $\gamma^{(0)} > 0$ and $\alpha^{(0)} > 0$.
- $k \geqslant 1$
  1. For a fixed small constant $\varepsilon > 0$ set
  $$\hat{\gamma}^{(k)} := (1+\varepsilon)\gamma^{(k-1)}, \qquad \hat{\alpha}^{(k)} := (1+\varepsilon)\alpha^{(k-1)}.$$
  2. The kth candidate set $U^{(k)}$ is defined by
  $$\begin{aligned}U_h^{(k)} &:= \{\hat{v}_h \in S_h \mid \|\hat{v}_h\|_{H_0^1(\Omega)} \leqslant \hat{\gamma}^{(k)}\},\\ U_*^{(k)} &:= \{v_* \in S_h^{\perp} \mid \|v_*\|_{H_0^1(\Omega)} \leqslant \hat{\alpha}^{(k)},\ \|v_*\|_{L^2(\Omega)} \leqslant Ch\hat{\alpha}^{(k)}\},\\ U^{(k)} &:= u_h + U_h^{(k)} + U_*^{(k)}.\end{aligned}$$
  3. Compute values of the kth iteration by
  $$\begin{aligned}\gamma^{(k)} &:= \sup_{u\in U^{(k)}} \rho\|P_hFu - P_hF'(u_h)\hat{u}_h - u_h\|_{H_0^1(\Omega)},\\ \alpha^{(k)} &:= Ch \sup_{u\in U^{(k)}} \|f(u)\|_{L^2(\Omega)}.\end{aligned}$$
  4. If $\gamma^{(k)} \leqslant \hat{\gamma}^{(k)}$ and $\alpha^{(k)} \leqslant \hat{\alpha}^{(k)}$ hold then stop, and there exists a desired solution in $U^{(k)} \subset H_0^1(\Omega)$.

5. Set $k := k + 1$ and return to step 1. If k reaches a maximum iteration number or $\gamma^{(k)}$ and $\alpha^{(k)}$ exceed certain criteria then stop, and the verification fails.

The essential difference between algorithms AL-1 and AL-2 is that the finite-dimensional part U_h of the candidate set is taken as a ball U_h in $H_0^1(\Omega)$ in the latter. Using the norm estimates, as shown in numerical examples in the next section, we can expect to avoid the drawbacks of AL-1. Namely, we can overcome the difficulty caused by local overestimates in the calculation of the vector $\boldsymbol{d}$ and the risk of explosive enlargement of the candidate set caused by interval computations.

4. Numerical examples

We now give some numerical examples which confirm the effectiveness of algorithm AL-2. The interval arithmetic in each verification step was implemented using Sun Forte Fortran Desktop Edition 6 update 1 [11][2] on FUJITSU GP7000F model 900 (CPU: SPARC64-GP 400 MHz, OS: Solaris 7).

4.1. Example 1

Consider the following two point boundary value problem:

$$\begin{aligned} -u'' + au' + u &= g, \\ u(0) = u(1) &= 0, \end{aligned} \tag{26}$$

where $a \geqslant 0$ is a parameter, and g is chosen such that $u = \sin(\pi x)$ is the exact solution of (26). The interval $\Omega = (0, 1)$ is divided into N equal parts and S_h is taken as the set of piecewise linear functions on (0, 1). Then $\dim S_h = N - 1, h = 1/N$ and a priori constant C in (7) can be taken as $1/\pi$ [5]. Table 2 shows the verification results. The solution is enclosed in a candidate set $u_h + U_h + U_*$, where U_h and U_* are represented as

Table 2
Verification results for piecewise linear basis.

	AL-1				AL-2			
	$\max_i \lvert A_i\rvert$		$\Vert U_*\Vert_{H_0^1(\Omega)}$		$\Vert U_h\Vert_{H_0^1(\Omega)}$		$\Vert U_*\Vert_{H_0^1(\Omega)}$	
a	$N = 100$	$N = 200$	$N = 100$	$N = 200$	$N = 100$	$N = 200$	$N = 100$	$N = 200$
0	0.00009	0.00004	0.02309	0.01173	0.00003	0.00001	0.02222	0.01111
0.1	0.00868	0.01536	0.07346	0.09360	0.00083	0.00041	0.02223	0.01111
0.2	0.03727	0.07194	0.16059	0.22073	0.00163	0.00081	0.02224	0.01112
0.5	0.36826	0.72471	0.64288	0.89934	0.00402	0.00200	0.02227	0.01112
1	3.50529	11.2050	3.10837	7.04551	0.00796	0.00397	0.02232	0.01114
5	×	×	×	×	0.03825	0.01870	0.02325	0.01136
10	×	×	×	×	0.08381	0.03867	0.02584	0.01195
20	×	×	×	×	0.27512	0.09404	0.04287	0.01465

[2] At present, the name has been changed to Sun ONE Studio 7, Compiler Collection Fortran 95.

the magnitude of the H_0^1 norm in S_h and $S_h^{\perp}$, respectively. In the table, $\max_i |A_i|$ means the maximum value of each coefficient $A_i^{(k)}$ obtained by AL-1, the last digit in the mantissa for each of the norm values is rounded-up, and × indicates that the corresponding verification failed.

One can see that algorithm AL-1 fails when the terms of the first-order derivative tend to be large, while AL-2 works well.

4.2. *Example 2*

Consider the following two-dimensional problem:

$$\begin{aligned} -\Delta u &= k \cdot \nabla u + u + g \quad && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{27}$$

where $\Omega = (0,1) \times (0,1)$, $k = (k_1, k_2)^{\mathrm{T}} \in \mathbb{R}^2$ is a parameter, and g is chosen such that $u = \sin(\pi x)\sin(\pi y)$ is the exact solution of (27). The approximate subspace S_h is taken to be a double finite Fourier series of the form:

$$S_h = \left\{ \sum_{m,n=1}^{N} a_{mn} \sin(\pi m x) \sin(\pi n y);\ a_{mn} \in \mathbb{R} \right\}.$$

Then $\dim S_h = N^2, h = 1/N$ and we can choose a priori constant C as $N/\sqrt{((N+1)^2+1)}\pi$ [12]. Table 3 shows the comparison of the two methods, and the results seem to be similar to those of example 1.

4.3. *Example 3*

Consider the one-dimensional Burgers equation

$$\begin{aligned} -\nu u'' + uu' + g &= 0, \\ u(0) = u(1) &= 0, \end{aligned} \tag{28}$$

Table 3
Verification results for the two-dimension case $k = k_1 = k_2$.

	AL-1				AL-2			
	$\max_i \lvert A_i \rvert$		$\lVert U_* \rVert_{H_0^1(\Omega)}$		$\lVert U_h \rVert_{H_0^1(\Omega)}$		$\lVert U_* \rVert_{H_0^1(\Omega)}$	
k	$N=10$	$N=30$	$N=10$	$N=30$	$N=10$	$N=30$	$N=10$	$N=30$
0.0	0.00089	0.00012	0.28468	0.10131	0.00198	0.00026	0.28468	0.10131
0.5	0.03883	0.02083	0.29325	0.10263	0.05153	0.01755	0.29208	0.10222
1.0	0.09043	0.03809	0.31275	0.10660	0.10350	0.03496	0.30204	0.10341
1.5	0.15381	0.06118	0.36187	0.11753	0.15899	0.05251	0.31492	0.10489
2.0	0.30024	0.10475	0.51491	0.15025	0.22030	0.07068	0.33127	0.10667
2.5	2.85612	0.36382	3.83422	0.40620	0.29205	0.08969	0.35260	0.10878
3.0	×	×	×	×	0.37641	0.10966	0.38025	0.11125
4.0	×	×	×	×	0.60983	0.15309	0.46481	0.11740
5.0	×	×	×	×	1.02990	0.20575	0.63113	0.12604

Table 4
Verification results for the trigonometric basis.

$\mu 1/\nu$	$N = 100$	$N = 200$	$N = 500$	$N = 100$	$N = 200$	$N = 500$
	AL-1					
	$\max_i \lvert A_i \rvert$			$\Vert U_* \Vert_{H_0^1(\Omega)}$		
0.1	0.0041	0.0041	0.0041	0.0221	0.0111	0.0045
0.5	0.0210	0.0207	0.0206	0.0222	0.0112	0.0045
1.0	0.0459	0.0451	0.0445	0.0226	0.0114	0.0046
2.0	×	0.2884	0.2619	×	0.1458	0.0057
2.5	×	×	×	×	×	×
3.0	×	×	×	×	×	×
4.0	×	×	×	×	×	×
5.0	×	×	×	×	×	×
	AL-2					
	$\Vert U_h \Vert_{H_0^1(\Omega)}$			$\Vert U_* \Vert_{H_0^1(\Omega)}$		
0.1	0.0008	0.0004	0.0002	0.0221	0.0111	0.0045
0.5	0.0038	0.0019	0.0008	0.0221	0.0111	0.0045
1.0	0.0078	0.0039	0.0016	0.0222	0.0111	0.0045
2.0	0.0166	0.0080	0.0032	0.0224	0.0112	0.0045
2.5	0.0215	0.0102	0.0040	0.0226	0.0112	0.0045
3.0	0.0267	0.0125	0.0048	0.0228	0.0113	0.0045
4.0	0.0389	0.0174	0.0066	0.0234	0.0114	0.0045
5.0	0.0549	0.0228	0.0084	0.0243	0.0116	0.0046

where $\nu \geqslant 0$ is the kinetic constant and g is chosen such that $u = \sin(\pi x)$ is the exact solution of (28). Table 4 shows the verification results for the approximate subspace S_h using the set of finite Fourier series of the form:

$$S_h = \left\{ \sum_{n=1}^{N} a_n \sin(\pi n x);\ a_n \in \mathbb{R} \right\}.$$

Then $\dim S_h = N, h = 1/N$ and a priori constant C can be taken as $N/(\pi(N+1))$.

Though the examples are rather artificial, the results are sufficient to show that algorithm AL-2 is superior to AL-1 when concerning equations including a first-order derivative. On the other hand, we achieved a successful verification using AL-1 for a more complicated forth order elliptic problem [13]. In that case, we actually utilized special techniques which enabled us to avoid the influence of the derivatives, but which are only applicable to the particular Fourier basis. We will apply this new algorithm to more realistic nonlinear elliptic equations, for example, the Orr–Sommerfeld equation or the Navier–Stokes equations in forthcoming papers.

References

[1] R.A. Adams, *Sobolev Spaces* (Academic Press, New York, 1975).

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

[3] P. Grisvard, *Elliptic Problems in Nonsmooth Domains* (Pitman, London, 1985).

[4] M.T. Nakao, A numerical approach to the proof of existence of solutions for elliptic problems, Japan J. Appl. Math. 5 (1988) 313–332.

[5] M.T. Nakao, A numerical verification method for the existence of weak solutions for nonlinear boundary value problems, J. Math. Anal. Appl. 164 (1992), 489–507.

[6] M.T. Nakao, Numerical verification methods for solutions of ordinary and partial differential equations, Numer. Funct. Anal. Optim. 22 (2001) 321–356.

[7] M.T. Nakao, N. Yamamoto and S. Kimura, On best constant in the optimal error estimates for the H_0^1-projection into piecewise polynomial spaces, J. Approx. Theory 93 (1998) 491–500.

[8] M. Plum, Computer-assisted enclosure methods for elliptic differential equations, Linear Algebra Appl. 324 (2001) 147–187.

[9] S.M. Rump, On the solution of interval linear systems, Computing 47 (1992) 337–353.

[10] S.M. Rump, Verification methods for dense and sparse systems of equations, in: *Topics in Validated Computations*, ed. J. Herzberger (Elsevier Science/North-Holland, Amsterdam, 1994) pp. 63–135.

[11] Sun Microsystems, Fortran 95 interval arithmetic programming reference, `http://docs.sun.com/source/816-2462/index.html`.

[12] K. Toyonaga, M.T. Nakao and Y. Watanabe, Verified numerical computations for multiple and nearly multiple eigenvalues of elliptic operators, J. Comput. Appl. Math. 147 (2002) 175–190.

[13] Y. Watanabe, N. Yamamoto, M.T. Nakao and T. Nishida, A numerical verification of nontrivial solutions for the heat convection problem, J. Math. Fluid Mech. 6 (2004) 1–20.