Numerical Verification Methods for Solutions of Ordinary and Partial Differential Equations

Mitsuhiro T. Nakao

Graduate School of Mathematics, Kyushu University 33
Fukuoka 812-8581, Japan
e-mail: mtnakao@math.kyushu-u.ac.jp

Abstract.
In this article, we describe on a state of the art of validated numerical computations for solutions of differential equations. A brief overview of the main techniques in self-validating numerics for initial and boundary value problems in ordinary and partial differential equations including eigenvalue problems will be presented. A fairly detailed introductions are given for the author’s own method related to second-order elliptic boundary value problems. Many references which seem to be useful for readers are supplied at the end of the article.

Key words. Numerical verification, nonlinear differential equation, computer assisted proof in analysis

1 Introduction

If we denote an equation for unknown \( u \) by

\[ F(u) = 0 \] (1.1)

then the problem of finding the solution \( u \) generally implies an \( n \) dimensional simultaneous system of equations provided \( u \) is in some finite dimensional (\( n \)-dimension) space. Since very large scale, e.g. several thousand, linear system of equations can be easily solved on computers nowadays, it is not too surprising that, even if (1.1) is nonlinear, we can verify the existence and uniqueness of the solution as well as its domain by some numerical computations on the digital computer. However, in the case where (1.1) is a differential equation, it becomes a simultaneous equation which has infinitely many unknowns because the dimension of the potential function space containing \( u \) is infinite. Therefore, we might naturally feel that it is impossible to study the existence or uniqueness of the solutions by finite procedures based upon the computer arithmetic. Actually, when the author first ran across an assertion about the possibility of such arguments by computer, in Kaucher-Miranker [13] about fifteen years ago, he could not believe it and seriously attempted to find their theoretical mistakes. When it became clear that the grounds of their arguments could not be denied, though the application area seemed rather narrow, the author was greatly impressed and was also convinced that such a study must be one of the most important research areas in computational mathematics, particularly in numerical analysis. Subsequently, the author has learned that there had been similar researches centered around ordinary differential equations (ODEs). However, with his research background it was natural for this writer to be particularly interested in numerical verification methods for partial differential equations (PDEs).
In this exposition, we will survey the state of the art, fairly emphasizing the author's own works, about the verification methods for the existence, uniqueness and enclosure of solutions for differential equations based on numerical computations.

Many differential equations appearing in the mathematical sciences such as physics or technology are numerically (approximately) solved by the use of contemporary supercomputers, and those computed results are supplied for simulations of phenomena, independently of the guarantee of existence and/or uniqueness of solutions. Naturally, there are not a few theoretical studies done by mathematicians for such equations. However, because of the nonlinearity or variety of the problem and so on, applications of unified theory are quite difficult and it is hard to say that those results sufficiently satisfy demands of the people who are working on the numerical analysis in various areas. Also even if the existence and uniqueness are already known, in general, we cannot confirm the order of magnitude of the difference between the computed solutions and the exact solutions. Actually in such a case, we are obliged to be satisfied with the ambiguous validity which may be interpreted as the comparison with experimental data. On the other hand, in pure mathematics, the problem proving the existence of solutions for particular differential equations can often arise. The principal purpose of numerical verification methods here is the mathematically and numerically rigorous grasp of solutions of differential equations appearing in various mathematical sciences including pure mathematics. Therefore, we note that the term 'numerical' does not mean 'approximate'. Although it is not too long time since this kind of research was started, a great number of developments will be expected in the 21st century as a new approach in numerical analysis which exceeds the existing numerical methods in the sense of assurance of numerical qualities for infinite-dimensional problems.

In the followings, in Section 2, we describe the case studies of numerical verification methods for ODEs, stressing on Lohner's method for initial value problems. And, in Section 3, the methods for partial differential equations for elliptic and evolutional problems, around the author's method, will be surveyed. Next, in section 4, we will treat the eigenvalue problems of elliptic operators. Finally, we will give some concluding remarks in Section 5.

2 Ordinary differential equations (ODEs)

A germ of numerical approaches to the verification of solutions for ODEs already appeared in Cesari [7] in the first half of 1960; and since then not a few of case studies have been done. Generally speaking, the functional equation is equivalent to the simultaneous equations with infinite dimension. Therefore, it should be the essential point to consider the errors caused by the truncation or discretization of the original problems. The enclosure methods for function space problems, in the author's opinion, will be classified into the following three groups according to their verification techniques.

(1) Analytic method:

This is a method such that, by using the estimations of constants, norms of functions including approximate solutions, and operators appearing in the equation, one prove
that they satisfy a certain condition, e.g., the convergence condition in the Newton-Kantorovich theorem in the below. Particularly, the norm estimation of an inverse linearized operator for the original problem is most important and essential task in this method. The interval analysis is used only in the simple arithmetic calculations.

(2) Interval method:

An interval can be considered as the set of functions whose ranges are contained that interval. For example, \([a, b]\) can be considered as the set:

\[\{ f \in C(\alpha, \beta) \mid f(x) \in [a, b], \forall x \in (\alpha, \beta) \} .\]

In this method, using this infinite dimensional property of the interval, the truncation errors are essentially grasped by the interval arithmetic. In order to apply this method, usually, we need a transformation of the original differential equation to an equivalent integral equation, but there are no norm estimations of the inverse linearized operator.

(3) Mixed method:

This is an intermediate method between the above two. The interval plays an essential role in this method, but the analytic arguments in a certain function space as well. Although some kind of inverse linearized operator is used, in general, but it is not directly evaluated as an infinite dimensional operator.

We now introduce some of the typical studies of these categories.

**Analytic methods** The following Newton-Kantorovich theorem is used for the existence and local uniqueness of the solution of functional equations (e.g., [87] for proof).

**Theorem 2.1** Let \( X, Y \) be Banach spaces, and let \( D \subset X \) be a convex set. For a map \( F : D \subset X \to Y \), \( F', F'' \) denote the Fréchet derivatives. Assume that, for some \( u_0 \in D \) and constants \( B, \eta, \kappa \) and \( r \),

(i) \([F'(u_0)]^{-1} \) exists, and \( ||[F'(u_0)]^{-1}|| \leq B \) and \( ||[F'(u_0)]^{-1}F(u_0)|| \leq \eta .\)

(ii) \( ||F''(u)|| \leq \kappa , \) \( ||u-u_0|| \leq r .\)

(iii) \( h = \eta \cdot B \cdot \kappa < 1/2 .\)

(iv) \( r \geq r_0 = \eta (1 - \sqrt{1-2h})/h .\)

Then, the equation \( F(u) = 0 \) has one and only one solution \( u^* \) in the ball: \( ||u-u_0|| \leq r_0 .\)

Kedem[15], McCarthy[18] applied this theorem to the verification of solutions for nonlinear two point boundary value problems. On the other hand, in Japan, Urabe [78], [79] developed a useful version of the Kantorovich-like theorem for the verification of solutions for multipoint and periodic boundary value problems. Plum's method [59] is also classified into this category. In his method, the norm of the inverse linearized operator is bounded by using eigenvalue enclosing technique with homotopy method (see also Section 3 and 4). Oishi [55] derived a version of the Theorem 2.1 by using some finite dimensional projection and its a priori error estimates, and proved, as an example, the existence of periodic solutions for the Duffing equation. Also, in [57], his method was extended for the finite element projection in one dimension.
Sinai's work [74] on the verification of the existence of a closed orbit is based upon the fixed point problem of a Poincaré map and the numerical proof of a Kantorovich type convergence condition. He verified a closed orbit of the following Lorenz model:

\[
\begin{align*}
    x' &= a_1 x + b_1 y z + b_1 x z \\
    y' &= a_2 y - b_1 y z - b_1 x z \\
    z' &= -a_3 z + (x + y)(b_2 x + b_3 y)
\end{align*}
\]  

(2.1)

Also, Nishida [52] solved, applying the theory of pseudo trajectory similar in [74], some bifurcation problems appeared in nonlinear fluid dynamics by numerical enclosing technique of the critical eigenvalues of linearized problem.

**Interval methods** In Moore's initial work [24], a basic idea of this method was already presented. Such a primitive technique was, however, not so effective for the practical problems. Nowadays, by the view point of the verification principle, the most famous method will be Lorner's technique [17] which we describe an outline in the below. In what follows, let \( IR \) and \( IR^n \) denote the set of real intervals and \( n \)-dimensional interval vectors, respectively.

Consider the following initial value problem of autonomous system:

\[ u' = f(u) \]  

(2.2)

Here, \( u = u(t) \in R^n \), \( f \) is a \( C^p \) map on \( R^n \) for a fixed positive integer \( p \), and the initial value is given as

\[ u(0) \in [u_0] = u_0 + [z_0]. \]  

(2.3)

Here, \([u_0],[z_0]\) denote the interval vectors, \( u_0 \) point vector. Note that, in case of non-autonomous, according to add a new dependent variable \( u_{n+1} = t \) and an equation \( u'_{n+1} = 1 \), it can be reduced to the type (2.2). Our purpose is, for fixed \( T > 0 \), to get an interval enclosure \([u(t)]\) for a solution \( u \) of (2.2), (2.3) such that

\[ u(t) \in [u(t)], \quad t \in [0,T]. \]

First, we consider an explicit one step method for (2.2), (2.3) with step size \( h \) based on a function \( \Phi = \Phi(u) \). Setting \( t_j \equiv jh \) and \( u_j \equiv u(t_j) \), where \( u(t) \) stands for an exact solution. When we denote the local discretization error between \( t_j \) and \( t_{j+1} \) by \( z_{j+1}/h \), it holds that

\[ u_{j+1} = u_j + h\Phi(u_j) + z_{j+1}, \quad j \geq 0, \]  

(2.4)

where \( u_0 \in [u_0] \). Suppose that \( \Phi \) is chosen so that \( z_{j+1} \) can be estimated by \( u \) and its derivative. For example, if the scheme (2.4) is order \( p \), then we have, using the Taylor expansion of \( u_{j+1} = u(t_j + h) \) at \( t_j \),

\[ z_{j+1} = \frac{h^p}{p!} u^{(p)}(\hat{t}_{j+1}), \quad \hat{t}_{j+1} \in (t_j, t_{j+1}). \]  

(2.5)

When an assured interval \([u_j]\) at \( t_j \) is obtained, we can enclose \( u_{j+1} \) by using (2.5). But such a simple method would yield the monotone and rapid increasing of the width
of the assured interval. The basic idea of Lohner's method consists in the representation of \( u_{j+1} \) as a function of the \( n \)-dimensional vector variable \( z_0, z_1, \ldots, z_{j+1} \). That is, the exact value at \( t_{j+1} \) is determined by all the local discretization errors up to \( j \) instead of \( u_j \) and \( z_{j+1} \) only.

If \( \Phi \) is continuously differentiable with respect to \( j+2 \) variables \( z_0, z_1, \ldots, z_{j+1} \), then \( u_{j+1} = u_{j+1}(z_0, z_1, \ldots, z_{j+1}) \) as well. Therefore, by using the mean value theorem around \((s_0, s_1, \ldots, s_{j+1})\), we have the following representation

\[
  u_{j+1} = \tilde{u}_{j+1} + \sum_{k=0}^{j+1} \frac{\partial u_{j+1}(\tilde{z})}{\partial z_k} (z_k - s_k),
\]

where \( s_k \) is usually chosen as the midpoint of \( z_k \). And, \( \tilde{z} \) is an unknown vector which is determined by \( z = (z_0, z_1, \ldots, z_{j+1}) \) and \( s = (s_0, s_1, \ldots, s_{j+1}) \). Since \( \tilde{u}_{j+1} = u_{j+1}(s_0, s_1, \ldots, s_{j+1}) \), we have

\[
  \tilde{u}_{j+1} = \tilde{u}_j + h \Phi(\tilde{u}_j) + s_{j+1}, \quad j \geq 0,
\]

where, \( \tilde{u}_0 = u_0 \) and \( s_0 = 0 \).

Next, we briefly mention about an interval enclosing algorithm of (2.6). Assume that we have already got the enclosure until \( t_j \).

1. First step: Calculation of a rough enclosure \([u_{j+1}^0]\) for \( u(t) \) on \([t_j, t_{j+1}]\).
   This can be done by computing a constant interval which encloses the solution of the equivalent integral equation on \([t_j, t_{j+1}]\) of the form:

\[
  u(t) = u_j + \int_{t_j}^{t} f(u(s)) ds, \quad u_j \in [u_j], t \in [t_j, t_{j+1}],
\]

where \([u_j]\) implies the interval enclosure of the solution at \( t = t_j \).

For an interval \( X \) such that \([u_j] \subset X\), we define \( Y \) as

\[
  Y \equiv [u_j] + [0, h] \cdot f(X).
\]

If \( Y \subset X \) then, by Schauder's fixed point theorem, it holds that

\[
  u(t) \in Y, \quad \forall t \in [t_j, t_{j+1}].
\]

Therefore, we set \([u_{j+1}^0] \equiv Y\). If \( Y \subset X \) does not hold, then, for an appropriately small \( \varepsilon > 0 \), setting

\[
  X := (1 + \varepsilon)Y - \varepsilon Y
\]

(\( \varepsilon \)-inflation), we retry the computation (2.9) for this \( X \) and check the same relation. If we could not get the desired inclusion within the definite times, then we adopt a smaller step size.

2. Second step: Calculate the local discretization error \([z_{j+1}]\) by the use of \([u_{j+1}^0]\) in the first step.

3. Third step: Compute a new (narrow) enclosure \([u_{j+1}]\) by using \([z_{j+1}]\).

4. Forth step: Replace the rough enclosure \([u_{j+1}^0]\) obtained in the first step by a new interval \([u_{j+1}]\) in the third step, and repeat the second-forth step until we get the interval as small as possible.
5. Fifth step: Determine $s_{j+1} \in [z_{j+1}]$ and calculate $\overline{u}_{j+1}$.

In [17], this method was applied to enclose a solution of the following Kepler-Ellipse problem, by using Taylor’s method of order 18 and step size 0.1,

\[
\begin{align*}
    u'_1 &= u_3, \\
    u'_2 &= u_4, \\
    u'_3 &= -u_1(u_1^2 + u_2^2)^{-1.5}, \\
    u'_4 &= -u_2(u_1^2 + u_2^2)^{-1.5},
\end{align*}
\]

(2.11)

with initial condition: $u_1(0) = 1.2$, $u_2(0) = 0$, $u_3(0) = 0$, $u_4(0) = \sqrt{\frac{2}{3}}$.

He got the enclosure of solution within $10^{-5}$ accuracy for each $t \in [0, 70]$. This method can also be applied to the boundary value problems combining with the shooting technique.

There are another works on the interval methods for initial value problems, e.g., [2], [13], [75] etc. Especially, Rihm[68] described a good survey of the techniques including the fundamental ideas.

**Mixed methods** The author’s method (e.g.,[33]), which originally proposed for PDEs, belongs to this category. The method uses both the constructive error analysis for the finite element method and the interval coefficient functions of a finite dimensional space, which will be described in detail in the next section. [51] and [73], which uses some properties of monotone operator, are regarded as another examples for the mixed method.

Oishi [54] proposed a kind of mixed method combining an approximate fundamental matrix with the interval representation of functions. He recently applied the method to the numerical verification of the existence of solutions for a connecting orbit in the continuous dynamics[56]. Mrozek et al.[25] presented an interesting computer assisted approach on the proof of a chaos property of the Lorenz model based on the Conley index theory.

3 Partial differential equations (PDEs)

There has been not so many works on the numerical verification for PDEs. As far as the author is concerned, it was hard to find any methods except for Plum and the author’s own work up to recently. As mentioned before, the former is an analytic method and the latter a mixed method. There is no interval method for PDEs, for it is difficult to transform a PDE to an equivalent integral equation. In the below, we present the outline of both methods for second-order elliptic problems, particularly around the author's method. Moreover, since, quite recently, some case studies have appeared for computer assisted proof in the actual nonlinear analysis, we will also briefly refer to them.
3.1 elliptic problems

We consider the following nonlinear elliptic boundary value problem on a bounded convex domain $\Omega$ in $\mathbb{R}^n, 1 \leq n \leq 3$:

\[
\begin{cases}
-\Delta u = f(x, u, \nabla u) & x \in \Omega, \\
\quad u = 0 & x \in \partial\Omega,
\end{cases}
\]

(3.1)

where the map $f$ is assumed to satisfy appropriate conditions on the smoothness. For an integer $m$, let $H^m(\Omega) \equiv H^m$ denote $L^2$-Sobolev space of order $m$ on $\Omega$. And set $H^1_0 \equiv \{ \phi \in H^1 | \text{tr}(\phi) = 0 \text{ on } \partial\Omega \}$ with the inner product $\langle \phi, \psi \rangle \equiv (\nabla \phi, \nabla \psi)$, where $(\cdot, \cdot)$ means the inner product on $L^2(\Omega)$.

In the below, we denote $|| \cdot ||_{L^2(\Omega)} \equiv || \cdot ||_{L^2}$ by $|| \cdot ||$. And define

\[
|\phi|^2_{H^2} \equiv \sum^n_{i,j=1} ||\frac{\partial^2 \phi}{\partial x_i \partial x_j}||^2_{L^2}.
\]

3.1.1 The author’s method

The verification principle of this method is first originated in 1988 by [29] and, in the meantime, several improvements have been done up to now.

In what follows, the map $f$ in (3.1) is assumed to be continuous from the Sobolev space $H^1_0(\Omega)$ into $L^2(\Omega)$ such that having a bounded image in $L^2(\Omega)$ on a bounded set in $H^1_0(\Omega)$. For example, when $n = 2$, $f(u) \equiv f(x, u, \nabla u) := g_1 \cdot \nabla u + g_2 u^p$ satisfies above assumption, where $g_1 = (g_1^1, g_1^2)$ and $g_2$ are in $L^\infty(\Omega)$, and $p$ an arbitrary nonnegative integer. And, for $n = 3$, the same assumption holds for any $p$ such that $1 \leq p \leq 3$ by the Sobolev imbedding theorem(e.g., [1]).

Now let $S_h$ be a finite dimensional subspace of $H^1_0$ dependent on $h$ ($0 < h < 1$). Usually, $S_h$ is taken to be a finite element subspace with mesh size $h$. And let $P_h : H^1_0 \longrightarrow S_h$ denote the $H^1$-projection defined by

\[
(\nabla u - \nabla (P_h u), \nabla \hat{\phi}) = 0, \quad \hat{\phi} \in S_h.
\]

(3.2)

We now suppose the following approximation property of $P_h$.

For any $\phi \in H^2 \cap H^1_0$,

\[
||\phi - P_h \phi||_{H^1_0} \leq C_0 h ||\phi||_{H^2},
\]

(3.3)

where $C_0$ is a positive constant numerically determined and independent of $h$. This assumption holds for many finite element subspace of $H^1_0$ defined by piecewise polynomials with quasi-uniform mesh(e.g., [8],[53]). For example, it can be taken as $C_0 = \frac{1}{\pi}$ and $\frac{1}{2\pi}$ for bilinear and biquadratic element, respectively, in case of the rectangular mesh [42]. For the triangular and uniform mesh of the domain in $\mathbb{R}^2$, we can take, e.g., $C_0 = 0.81$ for linear element [49], and for the more fine constant, see the arguments in the end of this sub-subsection.
Now, it is well known [11] that for arbitrary $\psi \in L^2(\Omega)$ there exists a unique solution $\phi \in H^2 \cap H^1_0$ of the following Poisson's equation:

$$\begin{cases}
-\Delta \phi &= \psi, \quad x \in \Omega, \\
\phi &= 0, \quad x \in \partial \Omega.
\end{cases} \quad (3.4)$$

When we denote the solution of (3.4) by $\phi \equiv K\psi$, the map $K : L^2 \to H^1_0$ is compact as well as the following estimate holds:

$$|\phi|_{H^2} \leq ||\psi||.$$ \quad (3.5)

Defining the nonlinear map $F(u) := Kf(u)$, $F$ is a compact map on $H^1_0$ and we get the following fixed point equation of the operator $F$ equivalent to (3.1):

$$u = F(u). \quad (3.6)$$

Therefore, if we find a nonempty, bounded, convex and closed subset $U$ in $H^1_0$ satisfying

$$F(U) = \{F(u)|u \in U\} \subset U,$$ \quad (3.7)

then by the Schauder fixed point theorem, there exists an element $u \in F(U)$ such that $u = F(u)$. Usually, we choose such a set $U$, which is referred a candidate set of solutions, of the form $U = U_h \oplus U_\perp$, where $U_h \subset S_h$ and $U_\perp \subset S_h^\perp$. Here, $S_h^\perp$ stands for the orthogonal complement subspace of $S_h$ in $H^1_0$. Then, the verification condition can be written as

$$\begin{cases}
P_hF(U) &\subset U_h \\
(I-P_h)F(U) &\subset U_\perp.
\end{cases} \quad (3.8)$$

Sometimes we call the quantities $R(F(U)) := P_hF(U)$ and $RE(F(U)) := (I-P_h)F(U)$ as the rounding into $S_h$ and the rounding error of $F(U)$, respectively. Then (3.8) implies that

$$R(F(U)) \oplus RE(F(U)) \subset U,$$ \quad (3.9)

which is the basic principle of our verification method. The set $U_h$ is taken to be a set of linear combinations of base functions in $S_h$ with interval coefficients, while $U_\perp$ a ball in $S_h^\perp$ with radius $\alpha \geq 0$.

Namely,

$$U_h = \{\phi_h \in S_h \mid \phi_h = \sum_{j=1}^M[\underline{A}_j, \overline{A}_j]\phi_j\} \quad (3.10)$$

and

$$U_\perp = \{\phi \in S_h^\perp \mid ||\phi||_{H^1_0} \leq \alpha\}, \quad (3.11)$$
respectively, where \( \{ \phi_j \}_{j=1}^M \) is a basis of \( S_h \). Here, \( \sum_{j=1}^{M} [A_j, \overline{A}_j] \phi_j \) is interpreted as the set of functions in which each element is a linear combination of \( \{ \phi_j \}_{j=1}^M \) whose coefficient of \( \phi_j \) belongs to the corresponding interval \([A_j, \overline{A}_j]\) for each \( 1 \leq j \leq M \). We denote the set of such interval functions by \( S_{h,I} \), that is,

\[
S_{h,I} := \{ \hat{\phi} \in S_h | \hat{\phi} = \sum_{j=1}^{M} A_j \phi_j, A_j \in IR \}.
\]

Then it can be considered as \( S_h \subset S_{h,I} \). Note that each element \( \phi_h \in P_h F(U) \) satisfies the following finite element solution for some \( \psi \in U \)

\[
(\nabla \phi_h, \nabla \hat{\phi}) = (f(\psi), \hat{\phi}), \quad \forall \hat{\phi} \in S_h.
\]  

(3.12)

Therefore, it can be easily seen that \( P_h F(U) \) is directly computed or enclosed from \( U_h \) and \( U_\perp \) by solving a linear system of equations with interval right-hand side using some interval arithmetic approaches, e.g., [3], [50]. On the other hand, \( (I - P_h) F(U) \) is unknown but can be evaluated by the following constructive a priori error estimates for the finite element solution of Poisson’s equation:

\[
\| (I - P_h) F(U) \|_{H_0^1} \leq C_0 h \sup_{u \in U} \| f(u) \|.
\]  

(3.13)

which is obtained by (3.3) and (3.5).

Thus, the former condition in (3.8) is validated as the inclusion relations of corresponding coefficient intervals, and the latter part can be confirmed by comparing two nonnegative real numbers which correspond to the radii of balls. In the actual computation, we use an iterative method for both part of \( P_h F(U) \) and \( (I - P_h) F(U) \) as below.

(1) Verification by the simple iteration method

As stated above, we usually find a candidate set of the form

\[
U = U_h \oplus U_\perp.
\]  

(3.14)

In the below, we fix an approximate solution \( \hat{u}_h \) of (3.1) such that \( \hat{u}_h = \sum_{i=1}^{n} u_i \phi_i \in S_h \).

We consider the set of functions \( U_h \in S_{h,I} \) of the form

\[
U_h = \sum_{i=1}^{n} (u_i + A_i) \phi_i,
\]  

(3.15)

where \( A_i \) are intervals, in general, centered at 0. And the set \( U_\perp \) is same as (3.11).

Then we take an interval vector \( (B_i) \) satisfying

\[
P_h F(U) \subset \sum_{i=1}^{n} B_i \phi_i,
\]  

(3.16)

where \( (B_i) \) is usually determined by a solution of the linear system of equations with interval right-hand side. Namely, for the \( M \times M \) matrix \( G := ( (\nabla \phi_i, \nabla \phi_j) ) \) and the \( M \)
dimensional interval vector $b := ( (f(U), \phi_i) )$, the interval vector $(B_i)$ can be computed as a solution of the following equation

$$G \cdot (B_i) = b.$$  

(3.17)

Here, $(f(U), \phi_i) \in IR$ stands for the interval enclosure of the set $\{(f(u), \phi_i) | u \in U\}$. And we set

$$\beta := C_0 h \sup_{u \in U} \|f(u)\|_{L^2}.$$  

(3.18)

On the actual and detailed computational procedures for the determining the interval vector $b$ and the estimation of the right-hand side in (3.18), refer [29], [30], [34] etc. Now the computable verification condition is described as

**Theorem 3.1** For the sets defined by (3.14), (3.15) and (3.11), if the following conditions hold

$$B_i \subset u_i + A_i, \quad i = 1, \cdots, n,$$

$$\beta \leq \alpha,$$  

(3.19)

then there exists a solution $u$ of $u = F(u)$ in $U$.

Based on Theorem 3.1, we obtain the following verification algorithm by using the simple iteration method with $\delta$-inflation (cf. [69]). In what follows, we define $[\alpha] \equiv \{ \phi \in S_h^1 \parallel |\phi|_{H^1} \leq \alpha \}$ for a nonnegative real number $\alpha$.

**Verification algorithm S-1**

1. Setting $A_i^{(0)} := [0, 0]$ $(i = 1, \cdots, n)$ and $\alpha^{(0)} := 0$, initial candidate set is defined by $U^{(0)} := \hat{u}_h$.

2. For the candidate set $U^{(k)}$ determined by $(A_i^{(k)})$ and $(k)$, compute $(B_i^{(k)})$ and $\beta^{(k)}$ from (3.17) and (3.18), respectively.

   If (3.19) holds, then there exist a desired solution in the set

   $$U^{(k)} = \hat{u}_h + \sum_{i=1}^{M} A_i^{(k)} \phi_i + [\alpha^{(k)}].$$

   Otherwise, go to the next step.

3. Using some fixed small constant $\delta > 0$, after setting

   $$A_i^{(k+1)} := [-1, 1] \delta + A_i^{(k)}, \quad i = 1, \cdots, n,$$

   $$\alpha^{(k+1)} := (1 + \delta) \beta^{(k)};$$

   return to the previous step.

**Remark 3.1** The above algorithm using Theorem 3.1 could work only for limited case. However, we can say that not only the implementation of the procedures is quite simple and easy but also the essential point of our verification principle, i.e., the direct solving a finite dimensional problem with additional error estimates, is clearly shown in this algorithm.
Numerical example 1([30]).

The algorithm S-1 has not so wide applications, because it needs that the map $F$ is retractive in some neighborhood of the fixed point to be verified. But, we actually succeeded the verification for several realistic problems, e.g., as below [30]:

$$
\begin{align*}
-\Delta u &= f(x)u^2 + g(x) \quad x \in \Omega \equiv (0,1) \times (0,1), \\
u &= 0 \quad x \in \partial \Omega,
\end{align*}
$$

(3.20)

where $f(x)$ and $g(x)$ are arbitrary $L^\infty$-functions on $\Omega$ whose ranges are in $[-2,2]$ and $[0,7]$, respectively. Here, as the finite element subspace $S_h$, we used the bilinear elements on the uniform rectangular mesh with mesh size $h = 1/15$.

(2) Verification by Newton-like method

In order to apply our verification method for more general problems, we introduce a kind of Newton-like method.

First, note that (3.6) can also be rewritten as the following decomposed form in $S_h$ and $S_h^\perp$:

$$
\begin{align*}
P_h u &= P_h F(u) \\
(I-P_h) u &= (I-P_h) F(u)
\end{align*}
$$

(3.21)

In order to consider the Newton type operator for (3.21), define the nonlinear operator $N$ on $H_0^1(\Omega)$ by

$$
N(u) := u - [P_h - P_h A'(\hat{u}_h)]^{-1}_h (P_h u - P_h F(u)),
$$

where $A'(\hat{u}_h) \equiv (-\Delta)^{-1}f'(\hat{u}_h)$ and $'$ means the Fréchet derivative of $f$ at $\hat{u}_h$. Here, $[P_h - P_h A'(\hat{u}_h)]^{-1}_h$ denotes the inverse on $S_h$ of the restriction operator $(P_h - P_h A'(\hat{u}_h))|S_h$.

The existence of such a finite dimensional inverse operator can be validated by the usual invertibility of a matrix corresponding to the restriction operator (e.g., [70]). Also note that we can replace $P_h A'(\hat{u}_h)$ by some approximate operator.

We now define

$$
T(u) := P_h N(u) + (I-P_h) F(u).
$$

Then $T$ is considered as the Newton-like operator for the former part of (3.21) but the simple iterative operator for the latter part. Moreover, $T$ is compact on $H_0^1(\Omega)$ by compactness of $F$. Furthermore, it can readily be seen that

Proposition 3.1 The fixed point equation

$$
u = T(u)
$$

(3.22)

is equivalent to (3.6).

Indeed, if (3.21) holds, then we have, by using the former part,

$$
0 = [P_h - P_h A'(\hat{u}_h)]^{-1}_h (P_h u - P_h F(u)),
$$

(3.23)
which yields

\[ P_h u = P_h N(u). \]  \hfill (3.24)

Therefore, \( u = Tu \) follows by adding (3.24) to each side of the latter part of (3.21). Conversely, if \( u = Tu \), then we immediately obtain (3.24), and thus (3.23) follows. The conclusion would be now straightforward. ☐

If we find a nonempty, bounded, convex and closed subset \( U \) in \( H_0^1(\Omega) \) satisfying \( T(U) = \{ T(u) \mid u \in U \} \subset U \), then by the Schauder fixed point theorem there exists an element \( u \in T(U) \) such that \( u = T(u) \).

Conversely, if \( u = Tu \), then we immediately obtain (3.24), and thus (3.23) follows.

The computational procedure for \( P_h N(U) \) (rounding) consists of the solving linear system of equations with interval right-hand side which is similar to that in the case of simple iteration method. But concerned matrix is a Newton type one which is exactly the same matrix as in the usual simplified Newton method for the discretized problem of (3.1) determined by the following nonlinear system of equations:

\[ (\nabla \tilde{u}_h, \nabla v_h) = (f(\tilde{u}_h), v_h), \quad v_h \in S_h. \]  \hfill (3.27)

We consider the more detailed procedure as below.

Observe that, for arbitrary \( u = u_h \oplus u_\perp \in U = U_h \oplus U_\perp \),

\[
P_h N(u) = P_h u - [I - P_h A'(\tilde{u}_h)]_{h}^{-1}(P_h u - P_h F(u))
\]  

\[
= [I - P_h A'(\tilde{u}_h)]_{h}^{-1}(P_h F(u) - P_h A'(\tilde{u}_h)u_h)
\]  

\[
= [I - P_h A'(\tilde{u}_h)]_{h}^{-1}P_h K(f(u) - f'(\tilde{u}_h)u_h),
\]

where we used the fact that \( P_h u = u_h \), and in the last right-hand side, we supposed that \( A'(\tilde{u}_h) = f'(\tilde{u}_h) \) for simplicity. It is not necessary but, usually, we take \( A'(\tilde{u}_h) \approx f'(\tilde{u}_h) \). Therefore, as in the previous paragraph, we choose the interval vector \( (B_N)_i \) satisfying

\[
P_h N(U) \subset \sum_{i=1}^{M} (B_N)_i \phi_i. \]  \hfill (3.28)

Actually, if we define the \( M \times M \) matrix \( G_N := (\nabla \phi_i, \nabla \phi_j) - (f'(\tilde{u}_h)u_h, \phi_j) \) and the \( M \) dimensional interval vector \( b_N := (f(U) - f'(\tilde{u}_h)u_h, \phi_i) \), then \( (B_N)_i \) is determined by solving the linear equation

\[
G_N \cdot (B_N)_i = b_N.
\]  \hfill (3.29)
Here, \( (f(U) - f'(\hat{u}_h)u_h, \phi_i) \in IR \) means the interval enclosure of the set \( \{(f(U) - f'(\hat{u}_h)u_h, \phi_i) \in R \mid u = u_h \oplus u_\perp \in U \} \) as before.

On the other hand, the error bound (rounding error) \( \beta_N \) is determined exactly same as (3.18), i.e.,

\[
\beta_N = C_0h \sup_{u \in U} \|f(u)\|_{L^2}.
\] (3.30)

Then, we get the following computable verification condition of the same type as in Theorem 3.1.

**Theorem 3.2** For the sets (3.14), (3.15) and (3.11), let \( (B_N)_i \) be a solution of (3.29) and \( \beta_N \) a real number defined by (3.30). If the following conditions hold

\[
(B_N)_i \subset u_i + A_i, \quad i = 1, \cdots, n,
\]

\[
\beta_N \leq \alpha,
\] (3.31)

then there exists a solution \( u \) of \( u = F(u) \) in the set \( V := \hat{u}_h + \sum_{i=1}^{M}(BN)_i\emptyset_i + [\beta_N] \).

By using the above theorem, one can readily obtain a verification procedure based on the Newton-like iteration, which is similarly described to that of the simple iteration algorithm S-1.

**Numerical example 2([80]).**

We considered the following two dimensional Allen-Cahn equation which plays an important role in the mathematical biology:

\[
\begin{aligned}
-\Delta u &= \lambda u(u-a)(1-u) \quad \text{in } \Omega, \\
\phantom{-\Delta u} &= 0 \quad \text{on } \partial \Omega.
\end{aligned}
\] (3.32)

Here, \( \Omega = (0,1) \times (0,1) \) and the constant \( a \) is taken to be \( 0 < a < 1/2 \) by the reason in actual problems. And \( \lambda \) is a positive parameter. For fixed \( a \), it is known that the equation (3.32) has two non-trivial solutions for each \( \lambda \geq \lambda^* \) with a certain positive \( \lambda^* \). But the exact value for the critical \( \lambda^* \) corresponding to the turning point is unknown by any theoretical approaches.

We verified several upper and lower bifurcated solutions on the approximate bifurcation diagram by using the same finite element subspace \( S_h \) as in the previous paragraph. For example, We enclosed an upper branch solution with the following data:

- conditions: \( a = 0.01, \lambda = 150, \) mesh size: \( h = 1/80. \)
- results: \( \|\hat{u}_h\|_{L^\infty} \approx 0.96, \) Maximum width of \( (B_N)_i \leq 0.025, \ H^1_0 \) error bound: \( \alpha \leq 0.0107. \)

**Accuracy improvement by the residual method** In this paragraph, we present some improvement of accuracy and efficiency of verification by some residual technique based on an a posteriori error estimation for the higher order finite element([84]).
Let \( \hat{u}_h \) be an approximate solution of (3.1) satisfying (3.27). We take an element \( \hat{u} \in H^2(\Omega) \cap H^1_0(\Omega) \) as the solution to the following Poisson equation:

\[
\begin{align*}
-\Delta \hat{u} & = f(\hat{u}_h) \quad \text{in} \quad \Omega, \\
\hat{u} & = 0 \quad \text{on} \quad \partial \Omega.
\end{align*}
\]  

(3.33)

We intend to find the exact solution \( u \) around \( \hat{u} \). Then, notice that \( v_0 \equiv \hat{u} - \hat{u}_h \in S^{\perp}_h \), which also implies that \( P_h \hat{u} \) coincides with \( \hat{u}_h \). Therefore, while the explicit form of \( v_0 \) is unknown, the norm can be estimated as follows, by using (3.3) and (3.5):

\[
\|v_0\|_{H^1_0} \leq C_0 h \|f(\hat{u}_h)\|_{L^2}.
\]  

(3.34)

Thus, in this case, we consider the candidate set \( U \) of the form

\[
U = \hat{u}_h + v_0 + \sum_{i=1}^{M} W_i \phi_i + [\alpha],
\]

where \( W_i \in IR \).

Setting \( u = \hat{u} + w \), (3.1) is rewritten as the following residual form finding \( w \).

\[
\begin{align*}
-\Delta w & = f(\hat{u}_h + v_0 + w) - f(\hat{u}_h) \quad \text{in} \quad \Omega, \\
w & = 0 \quad \text{on} \quad \partial \Omega.
\end{align*}
\]  

(3.35)

Now, let \( S^*_h \subset H^1 \) be a finite element subspace whose basis consists of the basis of \( S_h \) and the base functions having nonzero values on the boundary \( \partial \Omega \). We define the two dimensional vector valued function \( \nabla \hat{u}_h \in S^*_h \times S^*_h \) by the \( L^2 \)-projection of \( \nabla \hat{u}_h \in L^2(\Omega) \times L^2(\Omega) \) into \( S^*_h \times S^*_h \) and set \( \Delta \hat{u}_h \equiv \nabla \cdot \nabla \hat{u}_h \).

Then we can easily obtain the following identity:

\[
(\nabla \hat{u}_h, \nabla \phi) + (\hat{\Delta} \hat{u}_h, \phi) = 0, \quad \forall \phi \in H^1_0(\Omega).
\]  

(3.36)

By using the above equality, we have

\[
(\nabla v_0, \nabla \phi) = (\nabla \hat{u}_h - \nabla \hat{u}_h, \nabla \phi) + (\hat{\Delta} \hat{u}_h + f(\hat{u}_h), \phi), \quad \forall \phi \in H^1_0(\Omega).
\]

Choosing \( \phi = v_0 \) and using the Aubin-Nitsche inequility \( \|v_0\|_{L^2} \leq C_0 h \|v_0\|_{H^1_0} \), we get the following estimate:

\[
\|v_0\|_{H^1_0} \leq \|\nabla \hat{u}_h - \nabla \hat{u}_h\|_{L^2} + C_0 h \|\Delta \hat{u}_h + f(\hat{u}_h)\|_{L^2}.
\]  

(3.37)

Notice that, if we use the higher order element, the right-hand side of (3.37) will converge to 0 with higher than \( O(h) \). Thus we can expect that the a posteriori error estimates for the right-hand side can be actually smaller than the a priori estimates (3.34), i.e., \( O(h) \). Therefore, it will be possible that, if we use a higher order finite element to obtain the approximation \( \hat{u}_h \), then it will enable us to verify with high accuracy even for the relatively rough mesh. Moreover, this technique can also be extended for the nonconvex and nonsmooth domain such as \( L \)-shape domain in which the solution has low regularity and we could no longer apply the a priori estimates ([83]).
Remark 3.2 We consider the meaning of the initial error $v_0$. Let denote the dual space of $H^1_0$ by $H^{-1}$ and let $\langle \cdot, \cdot \rangle$ be the duality pairing on $H^{-1} \times H^1_0$. Taking account of $\Delta \hat{u}_h \in H^{-1}$, observe that

\[
\|\Delta \hat{u}_h + f(\hat{u}_h)\|_{H^{-1}} = \sup_{0 \neq \phi \in H^1_0} \frac{\langle \Delta \hat{u}_h + f(\hat{u}_h), \phi \rangle}{\|\phi\|_{H^1_0}} = \|v_0\|_{H^1_0}.
\]

Therefore, we can say that $v_0$ stands for an element in $H^1_0(\Omega)$ which is determined by the Riesz representation theorem from the residual functional $\Delta \hat{u}_h + f(\hat{u}_h) \in H^{-1}$.

Numerical example 3([84]).

We verified a solution of the following Emden's equation on the unit square in $\mathbb{R}^2$.

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

(3.38)

We used the biquadratic finite element subspace $S_h$ on the uniform rectangular mesh with mesh size $h$. Therefore, we can use the constant: $C_0 = \frac{1}{2\pi}$. We show the verification results in Table 1. for several mesh sizes $h = 1/12, \cdots, 1/20$. Note that, in case of the linear element with some other residual technique, we needed at least $h = 1/80$ for the verification due to the large right hand side, i.e., $\|\hat{u}_h^2\|_{L^\infty} \approx 900 ([80])$, which confirms us the effectiveness of the present a posteriori technique. On furher comparison with linear a posteriori method, see [84].

Table 1. Verification results for Emden's equation

| $h$ | $max |W_i|$ | $\alpha$ | $\|v_0\|_{H^1_0(\Omega)}$ |
|-----|------------|--------|-----------------|
| $\frac{1}{12}$ | 0.96132 | 0.48296 | 0.36265 |
| $\frac{1}{14}$ | 0.43282 | 0.18168 | 0.26593 |
| $\frac{1}{16}$ | 0.28121 | 0.10248 | 0.20278 |
| $\frac{1}{18}$ | 0.20335 | 0.06575 | 0.15943 |
| $\frac{1}{20}$ | 0.15434 | 0.04489 | 0.12847 |

Some other extentions: We briefly mention about the improvements or extentions of the present method.

(i) $L^\infty$ error bounds This can be done by using the following constructive a priori $L^\infty$ error estimates for the $H^1_0$-projection of the Poisson equation (3.4) :

\[
\|\phi - P_h \phi\|_{L^\infty(\Omega)} \leq C_0^{(\infty)} h \|\psi\|.
\]

(3.39)

For example, it can be taken as $C_0^{(\infty)} = 1.054([35])$ and $0.831([41])$ for the uniform bilinear and biquadratic element with rectangular mesh, respectively. For the triangular case, we can take, e.g., $C_0^{(\infty)} = 1.818$ for the linear element with uniform mesh.
([35]). The residual method using a posteriori $L^\infty$ error estimates was also presented in [41]. As a numerical example, in [41], the solution $u$ for Emden's equation (3.38) was verified with the following error bound in the neighborhood of the approximate solution $\hat{u}_h$:

$$
\|u - \hat{u}_h\|_{L^\infty} \leq 0.814 \quad (\frac{\|u - \hat{u}_h\|_{L^\infty}}{\|\hat{u}_h\|_{L^\infty}} \leq 0.0275),
$$

where the same finite element subspace as in Numerical example 3. was adopted with $h = 1/20$.

(ii) **Verification with local uniqueness** Based on the verification method described above, one can formulate a method to prove the existence and local uniqueness by using the Banach fixed point theorem. We now, by using the Newton-like operator $T$, write the residual equation (3.35) as:

$$
w = T(w).
$$

Then we consider the following set of functions

$$
T(0) + T'(W)W := \{ v \in H^1_0(I) \mid v = T(0) + T'(\tilde{w})w, \quad \tilde{w}, w \in W \},
$$

where $T'$ stands for the Fréchet derivative of $T$. By appropriate bounding of this set, one can prove that $T$ is a contraction map on the candidate set $W = W_h \oplus W_\perp$ satisfying $T(W) \subset W$. Therefore, by Banach's fixed point theorem, we obtain the inclusion result that there exists a unique solution $w \in W$ with $w = T(w)$. For details and examples, see [85].

(iii) **Parameter dependent equations** For the parameter dependent elliptic problems such as (3.32), it is important to enclose the solution curve itself or to verify the existence of some singular points, e.g., turning point or bifurcation point. It is also possible to apply our verification method to these problems by using some additional techniques such as the solution curve enclosing for regular solutions [43] and the bordering technique for turning points [76], [21]. The verification for simple bifurcation points would also be possible by applying the similar method to that in [77], [66].

(iv) **Navier-Stokes equations** In [45], an a posteriori and a constructive a priori error estimates for finite element solutions of the Stokes equations were presented. By using these results a prototype verification has been derived in [82] for the solutions of Navier-Stokes problems with small Reynolds number and small solutions. Recently, the present method was applied to verify the bifurcated periodic solutions of the Rayleigh-Bénard problem for the heat convection in a two dimensional domain [47]. By these validated computational results, several properties which were not yet proved by the theoretical approaches were numerically verified.

(v) **Variational inequalities** Our method can also be applied to the enclosing solutions of the variational inequalities with nonlinear right-hand side of the form [71]:

$$
\begin{cases}
\text{Find } u \in K \text{ such that } \\
(\nabla u, \nabla (v - u)) \geq (f(u), v - u), \forall v \in K,
\end{cases}
$$

(3.40)
which appears in arguments on the obstacle problems([10]). Here, \( K = \{ v \in H^1_0(\Omega) \mid v \geq 0 \} \). Furthermore, in [46] and [72], somewhat different kind of variational inequalities are treated.

(vi) **Estimation of the optimal constant** When we apply our verification method to the boundary value problems in non-rectangular polygonal domains, we need the constant \( C_0 \) in (3.3) as small as possible. Therefore, it is important to estimate the optimal value of the constant \( C_0 \). The problem of finding such a constant is reduced to the following an eigenvalue-like problem of the Laplacian:

\[
\begin{cases}
-\Delta u & = \lambda u + \lambda \psi + \Delta \psi \quad \text{in} \quad \Omega, \\
\frac{\partial u}{\partial n} & = 0 \quad \text{on} \partial \Omega, \\
\int_{\Gamma_3} u\, dS & = 0,
\end{cases}
\tag{3.41}
\]

where \( \Omega \) is the standard triangle in \( R^2 \) whose vertices are points \((0,0),(0,1)\) and \((1,0)\), and \( \Gamma_3 \) is the edge from \((0,0)\) to \((0,1)\) on \( \partial \Omega \) on the \( y \)-axis. When we denote the minimal value \( \lambda \) in the set of the all solutions of (3.41) by \( \lambda_0 \), the optimal constant \( C_0 \) looking for satisfies \( C_0 \leq \frac{1}{\sqrt{\lambda_0}} \). By applying our numerical verification method for the solution of (3.41), we obtained the estimate \( C_0 \leq 0.494 \) [48] which is considered as a really significant improvement of the existing best constant \( C_0 \leq 0.81 \). On the other hand, since it is known that \( 0.467 \leq C_0 \), we can say this would be ‘nearly’ optimal estimate of \( C_0 \).

### 3.1.2 Plum’s method

Plum presented a verification principle, which belongs to the category of analytic method, for the solutions of nonlinear elliptic problems in [61], [62] incorporating with his results on the enclosing eigenvalues for elliptic operators.

Let assume that the nonlinear function \( f \) in (3.1) is sufficiently smooth in with respect to each variable.

Also assume that an approximate solution \( \omega \) of (3.1) with \( H^2(\Omega) \) smoothness is obtained. This means that we use the \( C^1 \)-element when the finite element approximation is adopted for (3.1). And the residual is bounded by a sufficiently small \( \delta \) as follows:

\[
|| - \Delta \omega + f(\omega) ||_{L^2} \leq \delta. \tag{3.42}
\]

Also, suppose that there exists a constant \( K \) satisfying

\[
||u||_{L^\infty} \leq K||Lu||_{L^2}, \quad \forall u \in H^2, \tag{3.43}
\]

where \( Lu = -\Delta u + \frac{\partial f}{\partial u}(\omega)u \). The above constant \( K \) is determined by the arguments described in later.

Next, assume that the following non-decreasing function \( g : [0, \infty) \rightarrow [0, \infty) \) can be chosen as

\[
|f(\omega(x) + y) - f(\omega(x)) - J(x)y| \leq g(|y|) \quad (x \in \Omega, y \in R), \tag{3.44}
\]
where $g(t) = o(t)$ as $t \to 0$ and $J(x) \equiv \frac{\partial f}{\partial u}(\omega(x))$.

Moreover, let us suppose that there exist positive constants $K_i$, $i = 0, 1, 2$ such that, for any $u \in H^2$,

$$
||u||_{L^2} \leq K_0 ||Lu||_{L^2}, \quad ||\nabla u||_{L^2} \leq K_1 ||Lu||_{L^2}, \quad |u|_{H^2} \leq K_2 ||Lu||_{L^2}.
$$

Here, $K_i$ can be determined by the later consideration. Then the following theorem provides the verification condition of the exact solution $u$ to (3.1) in a neighborhood of $\omega$.

**Theorem 3.3** For $\alpha \geq 0$, if the inequality

$$
\delta \leq \frac{\alpha}{K} - \sqrt{\Omega} \cdot g(\alpha),
$$

(3.46)

holds, then there exists a solution $u$ of (3.1) satisfying $||u - \omega||_{L^\infty} \leq \alpha$. Here, $|\Omega|$ means the measure of $\Omega$.

The condition (3.46) implies that the map on $L^\infty(\Omega)$ defined by the residual simplified Newton operator for the equation (3.1) at $\omega$ is retractive on the set $D \equiv \{u \in L^\infty(\Omega) \mid ||u||_{L^\infty} \leq \alpha\}$. Therefore, the verification is based on the Schauder fixed point theorem in $L^\infty(\Omega)$. In order to verify the solution by this method, we need a fine approximation $\omega \in H^2 \cap H^1_0$ as well as the estimates of the constants appeared in the above. Particularly, the constant $K$ in (3.43), which provides the inverse norm for the linearized operator of the original equation, plays the most essential role. This constant is estimated as follows (see [62] for details):

First, we need an estimate of the following positive constant $\sigma$.

$$
\sigma \leq \min\{||\lambda|| \mid \lambda \text{ is the eigenvalue of operator } L \text{ on } H^2 \cap H^1_0\}.
$$

In order to get a lower bound of the eigenvalues for $L$, he uses the homotopy method starting with some simple operator, e.g., $L = \Delta$, whose lower bound of eigenvalue is already known ([58]).

Next, using this $\sigma$, the constants $K_i$ in (3.45) are determined as follows. We only describe here for the case that $f$ is independent of $\nabla u$.

First, it can be taken as $K_0 = 1/\sigma$. And defining the constants $c$ and $\bar{c}$ such that

$$
c \leq J(x) \leq \bar{c}, \quad \forall x \in \bar{\Omega},
$$

(3.47)

$K_1$ is given by

$$
K_1 = \begin{cases} 
K_0(1 - cK_0)^{1/2}, & \text{if } cK_0 \leq 1/2, \\
\frac{1}{2\sqrt{\bar{c}}}, & \text{otherwise}.
\end{cases}
$$

(3.48)

Also, $K_2$ is decided as:

$$
K_2 = 1 + K_0 \cdot \max\{-\bar{c}, \frac{1}{2}(\bar{c} - c)\}.
$$
Thus, if the constants $C_j$, $j = 0, 1, 2$ in the Sobolev inequality:

$$\|u\|_{L^\infty} \leq C_0\|u\|_{L^2} + C_1\|\nabla u\|_{L^2} + C_2|u|_{H^2}$$  \hspace{1cm} (3.49)

can be numerically estimated, then the desired constant $K$ in (3.43) is obtained by $K \equiv C_0K_0 + C_1K_1 + C_2K_2$. On the other hand, $C_j$ in (3.49) is computed as follows [62]:

$$C_j = \frac{\gamma_j}{|\Omega|} \cdot \max_{x_0 \in \Omega} \int_{\Omega} |x - x_0|^{2\nu} dx^{1/2}, \quad j = 0, 1, 2,$$  \hspace{1cm} (3.50)

where, $(\gamma_0, \gamma_1, \gamma_2) = (1, 1.1548, 0.22361)$ for $n = 2$, and $(\gamma_0, \gamma_1, \gamma_2) = (1.0708, 1.6549, 0.41413)$ for $n = 3$.

As the numerical examples, in [61], he verified the solutions of $-\Delta u = \lambda e^u$, for several parameter $\lambda$, on the unit square in $R^2$ with homogeneous Dirichlet condition. The approximate solution $\omega$ is constituted as the piecewise biquintic $C^1$-spline on the $8 \times 8$ finite element mesh.

In the meantime, he extended the method to the equation having turning points and bifurcation points in [65] and [66], respectively, as well as the problem on the nonconvex nonsmooth domain in [64].

### 3.1.3 Heywood’s method

In [12], Heywood et al. presented a numerical verification method for the spatially periodic solutions of the steady Navier-Stokes equations by using the spectral techniques. They considered the following problem on the fundamental domain $\Omega = (0,1) \times (0,1)$ of periodicity:

$$\begin{cases}
-\nu \Delta u + (u \cdot \nabla)u + \nabla p - f = 0, & x \in \Omega, \\
\nabla \cdot u = 0 & x \in \Omega, \\
u(x + K) = u(x), & \forall x \in R^2, K \in Z^2, \\
\int_{\Omega} u(x) dx = 0,
\end{cases}$$  \hspace{1cm} (3.51)

where $f$ is a prescribed force satisfying the same periodic condition as $u$, and $\nu$ is the kinematic viscosity.

Under some appropriate setting of the spatially periodic function space, i.e., $V \approx H^2(\Omega) \cap \{\text{periodic and divergence free}\}$, the original equation (3.51) can be written as:

find $\exists u \in V$ such that

$$F(u) \equiv -\nu \Delta u + (u \cdot \nabla)u - f = 0.$$  \hspace{1cm} (3.52)

Now let $M$ be a constant satisfying

$$\|(u \cdot \nabla)v\|_{L^2} \leq M\|u\|_V\|v\|_V, \quad \forall u, v \in V.$$

For some approximate solution $\hat{u}$ of (3.52), denoting the Fréchet derivative of $F$ at $\hat{u}$ by $L \equiv F'(\hat{u}) : V \rightarrow L^2$, the following theorem is obtained.
Theorem 3.4 If $L$ is invertible with a bound $\|L^{-1}\|$ and
\[ 4M\|L^{-1}\|^2\|F(\tilde{u})\|_{L^2} \leq 1 \]
holds, then (3.52) has a locally unique solution $u$ satisfying
\[ \|u - \tilde{u}\|_{V} \leq 2\|L^{-1}\|\|F(\hat{u})\|_{L^2}. \]

This theorem can be proved by the fact that the simplified Newton operator at $\tilde{u}$ defines a contraction mapping on a neighborhood of it under the above condition.

The main task of the verification procedure by the application of Theorem 3.4 is concerned with estimating the norm $\|L^{-1}\|$. This is done by a reduction in two steps. The first step is to approximate $L$ with a simpler, but still infinite dimensional, linear operator $L_{N}$ having the same finite part as $L$, where $N$ is a positive integer, and converging to $L$ as $N \to \infty$. Next, a finite dimensional operator $\tilde{L}_{N}$ is introduced as the spectral Galerkin approximation of $L$, which is the restriction of $L_{N}$ to some finite dimensional space, and it is shown that $L_{N}$ is invertible and $\|L_{N}^{-1}\|$ is bounded through a reduction to the property with respect to $\tilde{L}_{N}$. Thus, the general perturbation theorem yields the invertibility of $L$ and the bound $\|L^{-1}\|$ by using the error estimate $\|L - L_{N}\|$.

They applied the method to the case that $\nu = 0.001$ and $f(x) \equiv f(x_1, x_2)$ is a simple sine function in $x_2$ and got some verified results by constructing $\tilde{L}_{120}$ numerically which is the spectral Galerkin approximation of $L$ restricted to a finite dimensional space with dimension 2,820. Then, the quantities in Theorem 4 were estimated as follows:

\[ M = 15.42493, \quad \|L^{-1}\| < 390.16, \quad \|F(\tilde{u})\|_{L^2} = 3.421 \times 10^{-8}. \]

and thus they obtained
\[ M\|L^{-1}\|^2\|F(\hat{u})\|_{L^2} \leq 0.32132, \]
which implies by the theorem that there exists a locally unique solution $u$ of (3.52) with the error bound
\[ \|u - \tilde{u}\|_{V} \leq 2.7 \times 10^{-5}. \]

In these numerical computations, they neglected the round-off error of floating point arithmetic with double precision as well as used commercially available software concerned with linear algebra without verification.

3.2 Evolution equations

The study for the numerical verification method for evolution problems has been still made less progress than for the elliptic case.

We consider the following parabolic problems

\[ \begin{aligned}
\frac{\partial u}{\partial t} - \Delta u &= f(x,t,u), \quad (x,t) \in \Omega \times J, \\
u(x,t) &= 0, \quad (x,t) \in \partial \Omega \times J, \\
u(x,0) &= 0, \quad x \in \Omega,
\end{aligned} \quad (3.54) \]
where $\Omega$ is a convex domain in $R^1$ or $R^2$ and $J = (0, T)$ ($T > 0$). If the time $T$ is fixed, then the equation (3.54) can be treated as a kind of stationary problem. Thus, for example, the the author’s method in the previous subsection can also be applied, in principle, to the verification of solutions of this problem. In such applications, the simple linear problem which corresponds to the Poisson equation in the elliptic case is as follows:

$$
\begin{align*}
\frac{\partial \phi}{\partial t} - \Delta \phi &= g, \quad (x, t) \in \Omega \times J, \\
\phi(x, t) &= 0, \quad (x, t) \in \partial \Omega \times J, \\
\phi(x, 0) &= 0, \quad x \in \Omega,
\end{align*}
$$

(3.55)

where $g$ is the prescribed function. Therefore, if one obtain a fixed point formulation in the appropriate function space and the constructive a priori error estimates for the finite element solution of (3.55), then the arguments in the elliptic case can be applied to the verification for the problem (3.54). In [32] and [38], the verification examples were presented based on the simple iteration method for one and two space dimensional cases, respectively.

On the other hand, noting that it is readily possible to estimate the norm of the inverse of a linearized operator for (3.54), Plum’s method can also be applied to the same problem without any complicated work concerning the eigenvalue enclosing for the operator. In line with this direction, Minamoto [19], [22] formulated and presented some verification examples based on the residual Newton type operator. Also for the case of the following hyperbolic equations, in [20], [23], by using the similar arguments to parabolic case, several verification results were obtained.

$$
\begin{align*}
\frac{\partial^2 u}{\partial t^2} - \Delta u &= f(x, t, u), \quad (x, t) \in \Omega \times J, \\
u(x, t) &= 0, \quad (x, t) \in \partial \Omega \times J, \\
u(x, 0) &= 0, \quad x \in \Omega, \\
\frac{\partial \nu}{\partial t}(x, 0) &= 0, \quad x \in \Omega.
\end{align*}
$$

(3.56)

Since, in these verification procedures by analytic methods, the residual Newton method are utilized, the accuracy of the approximate solution essentially affects the success of verification. Usually, due to the difficulty to improve the accuracy for time direction, it is not so easy to compute an approximation with sufficiently small residue for the practical problems.

Recently, in [14], Kawanago proposed a method to verify a time periodic solution for some bifurcation problem on the following semilinear dissipative wave equation:

$$
\begin{align*}
u_{tt} - \nu_{xx} + \nu_t + \nu^3 &= \lambda \sin x \cos t, \quad (x, t) \in (0, \pi) \times (0, \infty), \\
u(0, t) &= \nu(\pi, t) = 0, \quad t \in (0, \infty),
\end{align*}
$$

(3.57)

where $\lambda$ is a positive parameter. He numerically proved that the above equation has a symmetry-breakung pitchfork bifurcation point $(\lambda_0, u_0)$ around an approximate solution pair $(\tilde{\lambda}_0, \tilde{u}_0)$. He used a kind of analytic method of Newton type by using spectral basis for both of space and time.
4 Eigenvalue problems

Consider the following self-adjoint elliptic eigenvalue problem:

\[
\begin{cases}
Au \equiv -\Delta u + qu = \lambda u, & x \in \Omega, \\
u = 0, & x \in \partial \Omega,
\end{cases}
\]

(4.1)

where \(\Omega\) is a bounded convex domain in \(R^2\) and \(q \in L^\infty(\Omega)\). We normalize the problem (4.1) as:

\[
\exists (u, \lambda) \in H_0^1(\Omega) \times R
\]

find \(\not\exists (u, \lambda) \in H_0^1(\Omega) \times R\) satisfying

\[
\begin{cases}
-\Delta u + (q - \lambda)u = 0, & x \in \Omega, \\
\int_{\Omega} u^2 = 1.
\end{cases}
\]

(4.2)

As in the usual elliptic problems, it is readily seen that the problem (4.2) can be rewritten as a fixed point equation of a compact map on \(H_0^1(\Omega) \times R\), and that, in order to enclose the eigenpair \((u, \lambda)\) around an approximate pair \((\tilde{u}_h, \lambda_h) \in S_h \times R\), we can also apply the author's verification method described in the previous section (see [44], [26], [27] for details). Namely, the exact eigenvalue \(\lambda\) and the corresponding eigenfunction \(u\) are enclosed in an interval \(\Lambda \in IR\) and in a candidate set \(U \subset H_0^1(\Omega)\) of the form \(U = \tilde{u}_h + v_0 + U_h + \{[\alpha]\}\), respectively. Here, as in the previous section, \(U_h \in S_{h,I}\), \(v_0 \in S_h^1\) and \([\alpha] \subset S_h^1\). Furthermore, in the present case, we can assert the uniqueness of the eigenvalue in \(\Lambda\) by the almost same verification condition as in Theorem 3.1 or 3.2([27]). Note that this method gives the enclosure of eigenvalues as well as verifies the corresponding eigenfunctions in contrast to other methods, described later, which present only eigenvalue enclosing.

We now briefly remark on the method to enclose the eigenvalues in order of magnitude. In such a case, an eigenvalue excluding procedure plays an essentially important role. This can be done as follows.

We consider an sufficiently narrow interval \(\Lambda\), and set, for a \(\lambda \in \Lambda\),

\[
L(\lambda) \equiv -\Delta u + (q - \lambda)u.
\]

Then, since \(L(\lambda)\) is a linear elliptic operator, the following equation has a trivial solution \(u = 0\):

\[
\begin{cases}
L(\lambda)u = 0, & x \in \Omega, \\
u = 0, & x \in \partial \Omega.
\end{cases}
\]

(4.3)

Therefore, if we validate the uniqueness of the solution in (4.3), then it implies that \(\lambda\) is not an eigenvalue of (4.1). The uniqueness property can be proved, taking into account that the operator \(L(\lambda)\) is linear, by the method analogous to that in the previous section. Thus, by using some interval computing techniques, it is easily to verify that there is no eigenvalue of (4.3) in \(\Lambda\). Thus the eigenvalue excluding process advances from the one to the next, backward or forward to the adjacent intervals.

We now note that, by some eigenvalue shift, we can easily present the lower bound of the spectrum of (4.1). Therefore, by appropriately combining this excluding procedure.
with the enclosing technique described before, we make the eigenvalue ordering as far as each eigenvalue is geometrically simple([28]).

A numerical example:

Particularly, the eigenvalue with smallest absolute value(ESAV), which is important to verify the solution of the corresponding nonlinear elliptic equations, was verified for the following problem([44]):

\[
\begin{aligned}
-\Delta u + \nu(3\tilde{v}_h^2 - 2(a + 1)\tilde{v}_h + a)u &= \lambda u, & x \in \Omega, \\
uu &\quad u = 0, & x \in \partial\Omega.
\end{aligned}
\] (4.4)

This is the eigenvalue problem for the linearized operator of the Allen-Cahn equation (3.32) at $\tilde{v}_h$. Here $\tilde{v}_h$ is an approximate lower branch solution of the original problem in the finite element subspace $S_h$ of biquadratic polynomials as in the numerical example 2 in the previous section with $\nu = 150$, $a = 0.01$ and mesh size $h = 1/20$. The ESAV of (4.4) were enclosed using the present method in the interval:

$\Lambda = \left[-16.67017, -16.55062\right]$

and the corresponding eigenfunction was enclosed in the set: $U = \tilde{u}_h + v_0 + U_h + [\alpha]$, where $||\tilde{u}_h||_{L^\infty} \approx 2.308$, $||v_0||_{L^2} \leq 0.0001372$, [maximum width of the coefficient intervals in $U_h$] $\leq 0.00321$ and $\alpha \leq 0.00105$.

Remark 4.1 It is seen that the above method can also be applied to the non-selfadjoint eigenvalue problems of the form:

\[
\begin{aligned}
-\Delta u + p \cdot \nabla u + qu &= \lambda u, & x \in \Omega, \\
uu &\quad u = 0, & x \in \partial\Omega.
\end{aligned}
\] (4.5)

Now, we will mention about other methods. The following proposition is well known as the Weinstein bounds for the eigenvalues of the form $Au = \lambda u$:

Proposition 4.1 Let $(\tilde{u}, \tilde{\lambda})$ be an approximate eigenpair for $Au = \lambda u$ such that $\tilde{u} \in D(A)$, where $D(A)$ is the domain of $A$, and let $\epsilon := \frac{||A\tilde{u} - \tilde{\lambda}\tilde{u}||}{||\tilde{u}||}$. Then there exists at least one exact eigenvalue in the interval $[\tilde{\lambda} - \epsilon, \tilde{\lambda} + \epsilon]$.

This proposition was extended to more practical version known as Kato's bound([67]). On the other hand, Lehman-Goerisch method [9], [4] is well-known, and is based on the Rayleigh-Ritz method which gives only upper bounds to the eigenvalues. We now describe the outline of this method. Let denote the N-th eigenvalue by $\lambda_N$ ordered from the smallest one, and let $\Lambda_N$ be the largest eigenvalue of the discretized problem for (4.1) by using the approximate eigenfunctions $\{\tilde{u}_i\}_{i=1}^N$ as a set of trial functions. We assume that some $\rho \in R$ is known satisfying

\[\Lambda_N < \rho < \lambda_{N+1}.\] (4.6)
Moreover, define the following matrices:

\[
(A_{1})_{ij} := (A\bar{u}_{i}, A\tilde{u}_{j}),
\]
\[
(A_{2})_{ij} := ((A - \rho)\bar{u}_{i}, \tilde{u}_{j}),
\]
\[
(A_{3})_{ij} := ((A - \rho)\bar{u}_{i}, (A - \rho)\tilde{u}_{j}),
\]

where \(i, j = 1, \ldots, \dot{N}\) and \((\cdot, \cdot)\) stands for the \(L^{2}\) inner product on \(\Omega\). Then, by the assumption (4.6), the generalized eigenvalue problem: \(A_{2}x = \mu A_{3}x\) has negative eigenvalues \(\{\mu_{i}\}_{i=1}^{N}\). Then we have

\[\lambda_{N+1-i} \geq \rho + \frac{1}{\mu_{i}}\]

for \(i = 1, \ldots, N\).

By using this theorem one can determine the lower bounds of eigenvalues provided that the spectral parameter is a priori known by some other consideration.

An example:

In [6], Behnke applied some extended method of the above to the following eigenvalue problem of forth order related to the vibrations of a clamped plate:

\[
\frac{\partial^{4}u}{\partial x^{4}} + P\frac{\partial^{4}u}{\partial x^{2}\partial y^{2}} + Q\frac{\partial^{4}u}{\partial y^{4}} = \lambda u \quad x \in \Omega,
\]

\[u = 0 \quad \text{and} \quad \frac{\partial u}{\partial n} = 0 \quad x \in \partial \Omega,
\]

where \(P, Q\) are positive constants and \(\Omega = (-\frac{a}{2}, \frac{a}{2}) \times (-\frac{b}{2}, \frac{b}{2})\). He considered the eigenvalues as functions of \(s = a/b\) and numerically proved a curve veering phenomena (cf. [5]).

Now, the Lehman-Goerisch method a priori needs a spectral parameter \(\rho\) in (4.6). It is, in general, not necessarily easy to decide such a parameter. Plum [58], [60], [67] introduced a homotopy method which overcomes this difficulty by connecting the given problem (4.1) with a simple problem whose spectrum is already known. For example, the problem (4.1) is connected to the simple eigenvalue problem for the Laplacian by the following homotopy, for \(t \in [0, 1]\),

\[
\begin{cases}
A_{t}u = -\Delta u + tu = \lambda u, & x \in \Omega, \\
u = 0, & x \in \partial \Omega.
\end{cases}
\]

When we denote the \(n\)-th eigenvalue of (4.8) by \(\lambda_{n}^{t}\) the homotopy is set such that \(\lambda_{n}^{t}\) is monotonically non-decreasing with respect to \(t\). Therefore, basically both spectra have invariant structure. Starting at some \(t = t_{1} \in (0, 1]\), by repeated applications of the Lehmann-Goerisch method or other bounding methods with additional procedures, ordered eigenvalues \(\lambda_{n}^{t_{1}}\) are enclosed, then increase \(t\), e.g., \(t = t_{2} \in (t_{1}, 1]\), little by little up to the final stage: \(t = 1\).
Examples:
In [60], he enclosed several eigenvalues of the following problems on the rectangular domain $\Omega = (0, \pi) \times (0, \pi)$ with Dirichlet boundary condition:

1. $-\Delta u + 10 \cos^2 \left[ \frac{1}{6} (2x + y) \right] u = \lambda u.$

2. $-\Delta u = \frac{\lambda u}{1 + \frac{1}{\pi^2} (x_1^2 + 2x_2^2)}.$

5 Conclusions

We have surveyed numerical verification methods for differential equations, especially around PDEs and the author’s works. But the period of this research is shorter than the history of the numerical methods for differential equations by computer and we can say it is still in the stage of case studies. Indeed, recently, this kind of studies have been referred little by little for practical applications in PDEs but there are many open problems to be resolve. Therefore, we can make no prediction that these approaches will grow into really useful methods for various kinds of equations in mathematical analysis. Also, since the program description of the verification algorithm is very complicated in general, there is another problem like software technology associated with assurance for the correctness of the verification program itself. Actually, some of the mathematician would not give credit the computer assisted proof in analysis as correct as they believe the theoretical proof, which might cause a kind of seriously emotional problem in the methodology of mathematical sciences. And there is another difficulty from the huge scale of numerical computations which often exceed the capacity of the concurrent computing facilities.

However, in the 21st century, the computing environment would make more and more rapid progress, which should be beyond conception in the present state. The author believes that the above difficulties should be overcome by this evolution and that the computer would greatly contribute to the theory of analysis in mathematics and create a new research area which should be called computer aided analysis. On the other hand, numerical methods with guaranteed accuracy for differential equations would highly improve the reliability in the numerical simulation of the complicated phenomena in science and technology.

参考文献


[14] Kawanago, T., Generalization of bifurcation theorems aiming for the study of nonlinear forced vibrations described by PDEs, preprint.


[27] Nagatou, K., A numerical method to verify the elliptic eigenvalue problems including a uniqueness property, Computing 63 (1999), 109-130.


[34] —, Solving nonlinear elliptic problems with result verification using an $H^{-1}$ residual iteration, Computing Suppl. 9 (1993), 161-173.


[57] —, Method of numerical inclusion for solutions of fixed point type equations, preprint.


[77] Tsuchiya, T., Numerical verification of simple bifurcation points, Kokyuroku RIMS Kyoto University 831(1993), 129-140.


