Studies on verified computations for solutions to elliptic boundary value problems

楕円型境界値問題の解の精度保証付き数値計算に関する研究

February 2012

Waseda University
Graduate School of Fundamental Science and Engineering
Major in Pure and Applied Mathematics
Research on Numerical Analysis

Akitoshi Takayasu

Acknowledgments

The author express his sincere thanks to his supervisor Prof. Shin'ichi Oishi for his precise advice, joint study and careful referee as the chief examiner. Sincere thanks are also due to Prof. Masahisa Tabata and Prof. Masahide Kashiwagi for their review as the second reader and valuable comments in detail. He also expresses his sincere gratitude to Dr. Takayuki Kubo at University of Tsukuba, who is one of joint researchers with respect to this study, for spending much time in discussion; Dr. Xuefeng Liu at Waseda University, who is also one of joint researchers corresponding to the mixed finite element method and error constants for FEM; Prof. Siegfried M. Rump at Hamburg Institute of Technology for his valuable comments and discussions for verified computations of minimal spectrum evaluations; Dr. Takeshi Ogita at Tokyo Woman's Christian University, who advised him the efficient method of calculating the constant τ and frequent advises for making a code of computation. He also wants to acknowledge with many thanks to the support of many indispensable people: Dr. Katsuhisa Ozaki, who also advised powerful yet simple code with respect to MATLAB; Dr. Naoya Yamanaka has spent so much time in our discussion; all members of Oishi Laboratory in Waseda university and fellow pupils in Laboratory. Finally, he owes thanks to his parents Michio and Noriko Takayasu for their unwavering and unfailing support. This research is supported by Research Fellow of the Japan Society for the Promotion of Science (JSPS).

Takadanobaba, Tokyo in January 2012.

Akitoshi Takayasu

Contents

Chapter 1

Introduction

The main concern of this thesis is to give a verified computation method for nonlinear two-point boundary value problems and semilinear elliptic boundary value problems on arbitrary polygonal domains. The verified computation gives the existence of solution to a differential equation with certain boundary condition. The guaranteed upper bound of absolute error between the solution and its approximation is also verified. As an introduction, let us kick off this thesis by background below.

1.1. Background

Studies on numerical computations are well brought under review. In the field of scientific computations and numerical algorithms in engineering, they help us to predict unknown phenomena or analyze an object itself. Most of them play important role in the industrial world and the academic society. Properties of most phenomena are modeled as differential equations so that it is worthwhile to compute these differential equations. Lately, the field of numerical analysis has been developed explosively along with the significant advance of computers. It is innovative approach that differential equations in the infinite dimension are solved on the computer by finite procedure, while convergence properties are proved by various error theories. Computations on the computer, however, do not guarantee the correctness of their results explicitly. It has been hard to say that the computational result is mathematically correct. The correctness of computation is usually checked by experiments or simulations of considered phenomena. Years after the progress of numerical analysis, the accuracy of computational results are improving. In the sense of simulation procedure, the accuracy is enough to *verify* the result but it is said that *It does not matter whether the computer simulation is correct or not.* Such statement represents the situation of reliability in the present numerical analysis. The rack of reliability is tragic circumstances in the field of scientific computing. Under such a situation, the correctness of numerical computation is becoming desirable. Leading an answer to the procedure of representing a computation error, the interval arithmetic [**34**] becomes an important tool for verified computations, which can guarantee the correctness of computation results. That is the dawn of verified numerical computations. In the 60s, the interval arithmetic has been applied to several numerical algorithms, *e.g.* Gaussian elimination etc. Meanwhile, this implementation consumes resources of computers. At that time, the performance of computation is much worse than that of these days. This disadvantage influences discussions critically. Eventually, the interval arithmetic didn't come under the spotlight of scientific computations in those days.

Several years later, studies on verified computations reach a turning point in the struggle for these difficulties. Oishi and Rump [**24**] give a fast verification algorithm for getting the accuracy of numerical solutions to linear simultaneous equations, where the direct rounding to operate the interval arithmetic is used. Pointing out the usefulness of functional-analysis-like approach, the authors cut the computational cost drastically. Here, important point is that authors use approximate solution to obtain the verified error bound.

In this thesis, we consider verified computation procedures to two-point boundary value problems and semilinear elliptic equations. Studies on verified computations for differential equations are the main topic of this thesis. The purpose of this thesis is that we will provide an algorithm to prove the existence of solution for the considered differential equation. If we have a *good* approximate solution of considered differential equation in a certain function space, we will try to validate the existence of solution with verified error bounds. Computer-assisted proofs are known as another name of verified computations for differential equations. All computational errors need to be bounded using the error analysis of discretization and interval arithmetic. Taking into account every errors of numerical computations, a mathematically guaranteed error estimation is obtained. In this thesis, we use FEM error analysis for bounding discretization errors. Our proposal method is based on Newton-Kantorovich theorem, which is related to fixed-point theorem. Now let us introduce the short history of computer-assisted proofs.

Computer-assisted proofs for the existence of solutions to two-point boundary value problems and elliptic Dirichlet problems have been started by pioneering works of Kantorovich [**10**] and Urabe [**37**]. The works of McCarthy, Tapia [**17**] and Kedem [**11**] have followed. In 1988, M. T. Nakao [**18**] has presented a method of computer-assisted proof for the existence of solutions to the elliptic problems including two-point boundary value problems. This method has been shown to be quite useful to generate a tight numerical inclusion of solutions [**22**]. Nakao's method can be considered as an interval extension of the finite element method. In 1991, Plum [**25**] has presented another method of proving the existence and uniqueness of solutions for elliptic boundary value problems. In his method, the norm of the inverse of linearized operator is bounded by an eigenvalue enclosing technique based on the homotopy method. In these two decades, both Nakao's method and Plum's method have been demonstrated to be quite useful for the computer-assisted proof of existence solutions to various boundary value problems [**18, 19, 21, 22, 38, 25, 26**]. For more discussions, see Section 1.4.

1.2. Sketch of FEM for the Poisson equation

In this section, we have a briefly review for the finite element method for the Poisson equation. This follows [**9**] for explanation. We will consider the following boundary value problems for the Poisson equation:

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

where Ω is a bounded open domain on the plane ($\mathbb{R}^2 = \{(x_1, x_2) : x_i \in \mathbb{R}\}$) with its boudary $\partial\Omega$, f is a given function. As usual, it follws

$$\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}.$$

The Poisson equation (1) is the famous mathematical model in physics and mechanics. For example, u represents a temperature, an electro-magnetic potential or the displacement of an elastic membrane fixed at the boundary under a transversal load of intensity f.

Let us recall Green's formula which will be fundamental importance. Before that we will start from the divergence theorem in two dimensions:

$$\int_\Omega \text{div p } dx = \int_{\partial\Omega} \text{p} \cdot n ds,$$

where $\text{p} = (p_1, p_2)$ is a vector function defined on Ω,

$$\text{div p} := \frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2},$$

and $n = (n_1, n_2)$ is the outward unit normal to $\partial\Omega$. Here, dx is the element of area and ds is the element of arc length along $\partial\Omega$. If $\text{p} = (vw, 0)$ and $\text{p} = (0, vw)$ are applied to the divergence theorem, we have

$$\int_\Omega \frac{\partial v}{dx_i} w dx + \int_\Omega v \frac{\partial w}{dx_i} dx = \int_{\partial\Omega} vwn_i ds, \ \ i = 1, 2. \tag{2}$$

Denoting the gradient of v by ∇v, *i.e.*

$$\nabla v := \left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2} \right),$$

we obtain the following Green's formula from (2):

$$\begin{aligned}\int_\Omega \nabla v \cdot \nabla w dx &:= \int_\Omega \left(\frac{\partial v}{\partial x_1}\frac{\partial w}{\partial x_1} + \frac{\partial v}{\partial x_2}\frac{\partial w}{\partial x_2} \right) dx \\ &= \int_{\partial\Omega} \left(v\frac{\partial w}{\partial x_1} n_1 + v\frac{\partial w}{\partial x_2} n_2 \right) ds - \int_\Omega v \left(\frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} \right) dx \\ &= \int_{\partial\Omega} v\frac{\partial w}{\partial n} ds - \int_\Omega v \Delta w dx,\end{aligned}$$

i.e.

$$\int_\Omega \nabla v \cdot \nabla w dx = \int_{\partial\Omega} v\frac{\partial w}{\partial n} ds - \int_\Omega v \Delta w dx, \tag{3}$$

where

$$\frac{\partial w}{\partial n} = \frac{\partial w}{\partial x_1} n_1 + \frac{\partial w}{\partial x_2} n_2$$

is the normal derivative.

We will give a variational formulation of problem (1), which becomes a basic formulation of our verification method in Section 2.3. Firstly, we will show that if u satisfies (1), then u is the solution of the following variational problem: Find $u \in V$ such that

$$A(u, v) = (f, v), \ \forall v \in V, \tag{4}$$

where

$$\begin{aligned}A(u, v) &:= \int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega \left(\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} \right) dx, \\ (f, v) &:= \int_\Omega f v dx,\end{aligned}$$

$V = H_0^1(\Omega)$ is the function space defined in Section 2.1. To see that (4) follows from (1) we multiply equation (1) with an arbitrary test function $v \in V$. Then we integrate over Ω. According to Green's formula (3), we obtain

$$(f, v) = -\int_\Omega \Delta u v dx = -\int_{\partial\Omega} \frac{\partial u}{\partial n} v ds + \int_\Omega \nabla u \cdot \nabla v dx = A(u, v),$$

where the boundary integral vanishes since $v = 0$ on $\partial\Omega$. On the other hand, if $u \in V$ satisfies (4) and u is sufficiently regular, then we see that u also satisfies (1), c.f. [**3, 9, 33**], etc.

Remark 1.1 (Claes [**9**]). *The formulation (4) is said to be a weak formulation of (1). The solution of (4) is said to be a weak solution of (1). If u is a weak solution of (4) then it is not immediately clear that u is also a classical solution of (1). Since this requires u to be sufficiently regular so that Δu is defined in a classical sence. The advantage of the weak formulation (4) is that it is easy to prove the existence of a solution to (4), whereas it is relatively difficult to prove the existence of a classical solution of (1). To prove the existence of a classical solution of (1), one usually starts with the weak solution of (1). Then one shows that this solution is sufficiently regular to be also a classical solution. On non-convex domain, it is known that the weak solution of (4) is not sufficient regular on the non-convex corner, see* [**8**].

Let us construct a finite dimensional subspace $V_h \subset V$. For simplicity, we will assume that $\partial\Omega$ is a polygonal shape. Let us make a triangulation of Ω, by subdividing Ω into a set $T^h = K_1, ..., K_m$ of non-overlapping triangles K_i,

$$\Omega = \bigcup_{K_h \in T^h} K_h = K_1 \cup K_2 \cup ... \cup K_m,$$

such that no vertex of one triangle lies on the edge of another triangle. We introduce the mesh parameter

$$h = \max_{K_h \in T^h} \text{diam}(K_h), \ \text{diam}(K_h) = \text{diameter of } K_h = \text{longest side of } K_h.$$

Let us define V_h as

$$V_h = \{v : v \text{ is continuous on } \Omega, \ v|_{K_h} \text{ is linear functions for } K_h \in T^h, \ v = 0 \text{ on } \partial\Omega\},$$

where $v|_{K_h}$ denotes the restriction of v to K_h. The space V_h consists of all continuous functions that are linear on each triangle K_h and vanish on $\partial\Omega$. Furthermore, $V_h \subset V$

is obtained. As parameters to describe a function $v_h \in V_h$, we choose freedoms as $v_h(x_i, y_i)$ at node points $N_i = (x_i, y_i),\ (i = 1, ..., n)$ of T^h. They exclude nodes on the boundary since $v = 0$ on $\partial\Omega$. Base functions $\phi_i \in V_h,\ (i = 1, ..., n)$ are defined by

$$\phi_j(N_i) = \delta_{ij} := \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases},\ i, j = 1, ..., n.$$

We see that the support of ϕ_i consists of triangles with common node N_j. A function $v_h \in V_h$ is represented by

$$v_h(x, y) = \sum_{j=1}^{n} v_j \phi_j(x, y),\ v_j = v(N_j),\ \text{for } (x, y) \in \Omega \cup \partial\Omega.$$

Let us formulate the following finite element method for (1) starting from the variational formulation (4):

$$\text{Find } u_h \in V_h \text{ such that } A(u_h, v_h) = (f, v_h),\ \forall v_h \in V_h. \tag{5}$$

It is easy to see that (5) is equivalent to the linear system

$$\mathrm{Au} = \mathrm{f},$$

where $\mathrm{A} = [a_{ij}]_{i,j=1}^{n}$ is $n \times n$ matrix (stiffness matrix) whose i-j elements are given by $a_{ij} = A(\phi_i, \phi_j)$ and $\mathrm{u} = [u_i]_{i=1}^{n}$, $\mathrm{f} = [f_i]_{i=1}^{n}$ are n-vectors with elements $u_i = u_h(N_i)$, $f_i = (f, \phi_i)$. For the Poisson equation, A is symmetric and positive definite. Then this linear system is solvable, which admits a unique solution u. Moreover, A is sparse matrix, *i.e.* $a_{ij} = 0$ unless N_i and N_j are nodes of the same triangle K_h.

Note that $u_h \in V_h$ is the best approximation of the exact solution u in the sense that

$$\|\nabla(u - u_h)\| \leq \|\nabla(u - v_h)\|,\ \forall v_h \in V_h,$$

where $\|\nabla v\| = A(v, v)^{1/2} = (\int_\Omega |\nabla v|^2 dx)^{1/2}$.

1.3. Verified computations for differential equations

In this thesis, the following two type of differential equation problems are considered. Let $\mathbb{R}$ and $\mathbb{N}$ be sets of reals and natural numbers, respectively. Let $(0,1)$ be an open interval on $\mathbb{R}$. One topic of this thesis is that we are concerned with two-point boundary value problems of second order ordinary differential equations:

$$\begin{cases} -(au')' = f(u', u, x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{6}$$

where $u = u(x)$ is a function to be determined, $u'(x) = du(x)/dx$ and $a(x)$ is a smooth function on $(0,1)$ with $a(x) \geq a_0 > 0$ for some $a_0 \in \mathbb{R}$. In this thesis, we assume that $a'(x) \in L^\infty(0,1)$. Here, $f : H_0^1(0,1) \to L^2(0,1)$ is assumed to be Fréchet differentiable. For example, the following function

$$f(u', u, x) = bu' + c_1 u + c_2 u^2 + c_3 u^3 + ... + c_N u^N + g$$

with $N \in \mathbb{N}$, $b(x), c_i(x) \in L^\infty(0,1)$, $(i = 1, ..., N)$ and $g(x) \in L^2(0,1)$ satisfies this condition.

Further, let Ω be a bounded polygonal domain on $\mathbb{R}^2$ with arbitrary shape. The other is that we are also concerned with the Dirichlet boundary value problem of the semilinear elliptic equation of the form:

$$\begin{cases} -\Delta u = f(\nabla u, u, x), & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{7}$$

where $f : H_0^1(\Omega) \to L^2(\Omega)$ is assumed to be Fréchet differentiable. For example,

$$f(\nabla u, u, x) = b \cdot \nabla u + c_1 u + c_2 u^2 + c_3 u^3 + ... + c_N u^N + g$$

with $b(x) \in (L^\infty(\Omega))^2$, $c_i \in L^\infty(\Omega)$, $(i = 1, ..., N)$ and $g(x) \in L^2(\Omega)$ satisfies this condition.

Verified computations for differential equations prove the existence of solution with guaranteed error bounds. It starts with their approximate solutions. In this thesis, we propose a verified computation approach. Our verified computation approach will be adopted to explore the existence and local uniqueness of the solution for (6) and (7). Namely, if an approximate solution is given by certain numerical method, we will try to validate the existence of exact solution in the neighborhood of the approximation. Here, we assume that u is exact solution of (6) or (7) and $\hat{u}$ is its approximation. Our aim of this thesis is to prove the existence of solution in the neighborhood

$$\|u - \hat{u}\|_{H_0^1} \leq \rho.$$

We will explain how to evaluate ρ with verification. In order to treat verified computations, we have to consider computational errors for differential equations. First one is discretized error, which is occurred by approximating $H_0^1(\Omega)$ functional space. We define an orthogonal projection $\mathcal{P}_h$, mapping $H_0^1(\Omega)$ to its functional subspace $V_h \subset H_0^1(\Omega)$. For given $f \in L^2(\Omega)$, a certain error estimation

$$\|u - \mathcal{P}_h u\|_{H_0^1} \leq C_M \|f\|_{L^2} \tag{8}$$

treats the discrete error of computational errors. Second one is the rounding error to use floating point arithmetic. The simplest way to treat rounding error is using interval arithmetic, *e.g.*

$$1/3 \in [0.333, 0.334],\ \sqrt{2} \in [1.4142, 1.4143],\ \pi \in [3.1415, 3.1416].$$

These are a correct expression of irrational numbers on the computer. In this thesis, we use INTLAB (INterval LABoratory) [**29**] for implementing the interval arithmetic in a computational code. This is a software on MATLAB to use interval arithmetic in our computation. Users get explicit computational result using floating point arithmetic. To treat all computational errors, our computer-assisted approach can be realizable.

1.4. Previous works

This section is devoted to describe previous methods by Nakao [**18**] and Plum [**26**]. These methods can be applied to the following operator equation. Using linear/nonlinear operator $\mathcal{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$ and $\mathcal{N} : H_0^1(\Omega) \to H^{-1}(\Omega)$, we can define a nonlinear operator equation

$$\mathcal{F}(u) = \mathcal{A}u - \mathcal{N}(u) = 0, \tag{9}$$

where $\mathcal{F} : H_0^1(\Omega) \to H^{-1}(\Omega)$. Problems (6) and (7) are transformed into this operator equation equivalently, see Subsection 2.3. Plum's method considers (9) directly. In Nakao's method, (9) is transformed into the invariance form:

$$\mathcal{A}^{-1}\mathcal{F}(u) = u - \mathcal{A}^{-1}\mathcal{N}(u) = 0. \tag{10}$$

Then, using some orthogonal projection $\mathcal{P}_h : H_0^1(\Omega) \to V_h \subset H_0^1(\Omega)$, Nakao's method transforms the following equation

$$u = \mathcal{A}^{-1}\mathcal{N}(u),$$

which is equivalent to (10), into

$$\begin{aligned} \mathcal{P}_h u &= \mathcal{P}_h \mathcal{A}^{-1}\mathcal{N}(u), \\ (\mathcal{I} - \mathcal{P}_h)u &= (\mathcal{I} - \mathcal{P}_h)\mathcal{A}^{-1}\mathcal{N}(u). \end{aligned}$$

For a certain approximate solution $\hat{u} \in V_h$ of (10), Nakao's method further defines $N_h : H_0^1(\Omega) \to V_h$

$$N_h(u) := \mathcal{P}_h u - \left[(\mathcal{I} - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h}\right]^{-1} \mathcal{P}_h(u - \mathcal{A}^{-1}\mathcal{N}(u)).$$

Using this, in Nakao's method

$$T_N(u) = N_h(u) + (\mathcal{I} - \mathcal{P}_h)\mathcal{A}^{-1}\mathcal{N}(u)$$

is considered. Then Nakao's method searches a non-empty bounded convex closed set $U \subset H_0^1(\Omega)$ satisfying $T_N(U) \subset U$. If we can find such a U, then Schauder's fixed-point theorem states that the set U includes at least one solution of (10). This is a simple outline of Nakao's method.

On the other hand, in Plum's method, constants δ and K_P are calculated explicitly such that

$$\|\mathcal{F}(\hat{u})\|_{H^{-1}} \leq \delta \tag{11}$$

and

$$\|u\|_{H_0^1} \leq K_P \|\mathcal{F}'[\hat{u}]u\|_{H^{-1}}. \tag{12}$$

Furthermore, assuming that there exists a non-decreasing function $g : [0, \infty) \to [0, \infty)$ such that

$$\|\mathcal{F}'[\hat{u} + w] - \mathcal{F}'[\hat{u}]\|_{H_0^1, H^{-1}} \leq g(\|w\|_{H_0^1}),\ \forall w \in H_0^1(\Omega). \tag{13}$$

In Plum's method, the existence of a solution for (9) is proved using the following theorem, which is similar to Newton-Kantorovich theorem:

THEOREM 1.2. *Let δ, K_P and g satisfy conditions (11)-(13). Suppose that some $\alpha_P > 0$ exists such that*

$$\delta \leq \frac{\alpha_P}{K_P} - G(\alpha_P),$$

where $G(t) := \int_0^t g(s)ds$, and

$$K_P g(\alpha_P) < 1.$$

Then, there exists a solution $u \in H_0^1(\Omega)$ of the equation $\mathcal{F}(u) = 0$ satisfying

$$\|u - \hat{u}\|_{H_0^1} \leq \alpha_P. \tag{14}$$

The solution is moreover unique under the side condition (14).

This theorem is proved by Banach's fixed-point theorem in [**26**]. Otherwise, our method directly uses Newton-Kantorovich theorem. Accordingly, for three methods by Nakao, Plum and us, they don't have a mathematically difference in the sense that every method uses fixed-point theorem to prove the existence and uniqueness of solutions. The main difference, which distinguish three methods, is their own technique to obtain the verification conditions. In particular, there is a different technique to obtain the norm of inverse operators corresponding to linearized operators $\mathcal{F}'[\hat{u}]$, see Section 4.1. Our method is characterized by Theorem 4.1 or Theorem 4.3. For other methods, an inverse operator norm bound also occurs in other methods by Nakao and Plum. In Plum's case, used in (12), it is verified by enclosing eigenvalues with a homotopy technique. In Nakao's case (e.g. [**19**]), the norm of the inverse is also bounded by a splitting into a bound for a finite dimensional matrix part (which is precisely our constant τ in Section 4.1) and a projection error bound. In our method, we need a residual evaluation and Lipschitz constant of linearized operator, explicitly. The following remarks (Remark 1.3 and 1.4) mention some relations with respect to each explicit evaluation.

REMARK 1.3. *The computation of a residual bound (Section 4.2) with respect to an operator equation also occurs in other methods. In Plum's method, where it is (11), the general smoothing technique is used with $H(\mathrm{div}, \Omega)$ elements. In Nakao's method* [**38**]*, the smoothing method with $(H^1(\Omega))^2$ elements is proposed. For more arguments, see Section 4.2.*

REMARK 1.4. *Lipschitz constant also has its analogue in Plum's method in the form of inequality (13). In the literature* [**26**]*, more general expression has been treated instead of the "classical" Lipschitz condition in Newton-Kantorovich theorem. In Nakao's method, such a condition also occurs more implicitly in the contractive condition.*

1.5. CHALLENGES

Our proposal method is specialized for the finite element method (FEM). We assume an approximate solution $\hat{u} \in V_h$ is a certain finite element solution, using

P_1 (piecewise linear) or P_2 (piecewise quadratic) finite elements. The merit of the finite element method is that the classical error analysis of an orthogonal projection can be obtained easily. For example, a priori constant of the finite element method, which satisfies (8), is well studied especially for P_1 elements. Detailed discussions are explained in Section 3.2. Especially, a method of getting concrete a priori constant is given in the case of non-convex domain. It plays essentially important role in verification methods.

In the following, we will try to

- introduce our verification framework for Problems (6) and (7).
- obtain a certain variational formulation for the verification framework.
- explain how to evaluate some constants explicitly on arbitrary polygonal domains.
- prove the invertibility of a linearized operator with respect to original problems.
- improve a residual evaluation for an operator equation.
- illustrate features of our verified computation method numerically.

In Chapter 2, firstly we prepare some notations and basic theorems. Then an abstract of our verification framework is introduced. It is shown that variational notations are applied to our framework naturally. In Chapter 3, methods of explicit evaluations are indicated. For elliptic equations, this evaluation has a key role in verification theories. Three constants are needed for the verification framework. Some methods to obtain these constants are put together in Chapter 4. Finally, in Chapter 5, some examples are demonstrated. For certain problems, concrete evaluations corresponding to some constants are described. Numerical results are also presented. It implies features of our verified computation method.

Chapter 2

Preliminaries

Verified computations for differential equations are based on the following preliminaries. Let us introduce notations, frameworks and formulations in this chapter.

2.1. Notations and basic theorems

Here, we would like to introduce several notations. These are used throughout this thesis. Let $\mathbb{R}$ and $\mathbb{N}$ be sets of real and natural numbers, respectively. Let Ω be bounded domain in $\mathbb{R}^d$ with $d = 1, 2$. Let $L^p(\Omega)$, $p \in [1, \infty)$ denote the function space of p-th power Lebesgue integrable functions on Ω. It follows that for $u \in L^p(\Omega)$,

$$\int_\Omega |u(x)|^p dx < \infty.$$

Especially, in case of $p = 2$, we denote L^2-inner product and L^2-norm as

$$(u, v) := \int_\Omega u(x)v(x)dx, \ \ \|u\|_{L^2} := \sqrt{(u, u)},$$

respectively. For vector functions $\mathrm{u}, \mathrm{v} \in \left(L^2(\Omega)\right)^2$, L^2-inner product also denotes

$$(\mathrm{u}, \mathrm{v}) := \sum_{i=1}^{2} (u_i, v_i), \ \text{ for } \mathrm{u} = (u_1, u_2)^T, \ \mathrm{v} = (v_1, v_2)^T.$$

Let $L^\infty(\Omega)$ denote the space of functions that are essentially bounded on Ω with the norm

$$\|u\|_{L^\infty} := \operatorname*{ess\,sup}_{x \in \Omega} |u(x)|.$$

The n-dimensional vector for $\mathrm{u} = (u_1, ..., u_n)^T$ in $\mathbb{R}^n$, let $|\mathrm{u}|_{l^2}$ be the Euclidean norm

$$|\mathrm{u}|_{l^2} = \sqrt{u_1^2 + u_2^2 + ... + u_n^2}$$

and for a matrix $A \in \mathbb{R}^{n \times n}$, the norm $\|A\|_2$ denotes the spectral norm of matrix A.

Let $H^m(\Omega)$ denote the L^2-Sobolev space of order $m \in \mathbb{N}$ with the inner product

$$\langle u, v \rangle_m := \sum_{|k|=0}^{m} (D^{(k)}u, D^{(k)}v).$$

Here, if $d = 2$, $D^{(k)}$ denotes the partial differentiation with respect to the multi-index $k = (k_1, k_2)$ with $|k| = k_1 + k_2$:

$$D^{(k)}u := \frac{\partial^{|k|}u}{\partial x_1^{k_1} \partial x_2^{k_2}}.$$

The H^m-norm and semi-norm are defined for $u \in H^m(\Omega)$ by

$$\|u\|_{H^m} := \left(\sum_{|k| \leq m} (D^{(k)}u, D^{(k)}u) \right)^{1/2}, \ |u|_{H^m} := \left(\sum_{|k|=m} (D^{(k)}u, D^{(k)}u) \right)^{1/2}.$$

Let us further define $H_0^1(\Omega)$ as

$$H_0^1(\Omega) := \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$$

with the inner product

$$(\nabla u, \nabla v) := \int_\Omega \nabla u \cdot \nabla v dx$$

and the norm

$$\|u\|_{H_0^1} := |u|_{H^1} = \|\nabla u\|_{L^2} = \left(\int_\Omega |\nabla u|^2 dx \right)^{1/2}.$$

Here, $u = 0$ on $\partial\Omega$ is in the trace sense.

If $d = 1$, let $H^m(0, 1)$ inner product be

$$\langle u, v \rangle_m = (u, v) + (u^{(1)}, v^{(1)}) + \cdots + (u^{(m)}, v^{(m)})$$

and the norm

$$\|u\|_{H^m} = \sqrt{\langle u, u \rangle_m} = \sqrt{\|u\|_{L^2}^2 + \|u^{(1)}\|_{L^2}^2 + \cdots + \|u^{(m)}\|_{L^2}^2}.$$

Here, $u^{(m)}$ is the m-th derivative of u with respect to x. Let further

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

with the inner product $(u^{(1)}, v^{(1)})$ and the norm $\|u\|_{H_0^1} = \|u^{(1)}\|_{L^2}$.

Generally, for $p \in [1, \infty]$, $W^{m,p}(\Omega)$ denotes the L^p-Sobolev space of order $m \in \mathbb{N}$.

Let $H^{-1}(\Omega)$ be the topological dual space of $H_0^1(\Omega)$, *i.e.*, the space of linear continuous functionals on $H_0^1(\Omega)$. Let $T \in H^{-1}(\Omega)$ and $u \in H_0^1(\Omega)$. We denote $Tu \in \mathbb{R}$ as $\langle T, u \rangle$. The norm of $T \in H^{-1}(\Omega) : H_0^1(\Omega) \to \mathbb{R}$ is defined as

$$\|T\|_{H^{-1}} := \sup_{0 \neq u \in H_0^1(\Omega)} \frac{|\langle T, u \rangle|}{\|u\|_{H_0^1}}.$$

Further, let X and Y be Banach spaces. The set of a bounded linear operator is denoted by $\mathcal{L}(X, Y)$. For $L \in \mathcal{L}(X, Y)$, operator norm is denoted by

$$\|L\|_{X,Y} := \sup_{0 \neq u \in X} \frac{\|Lu\|_Y}{\|u\|_X}.$$

Here, $\|\cdot\|_X$ is the norm of X and $\|\cdot\|_Y$ is the norm of Y respectively.

Let us introduce Sobolev's embedding theorem. For Banach spaces X and Y, the embedding $X \hookrightarrow Y$ means that a natural embedding operator $u \in X \mapsto u \in Y$ is continuous, *i.e.* $\|u\|_Y \leq C\|u\|_X$ holds for a constant C. Let us show the following embedding theorem which is related to the compactness of the embedding operator.

THEOREM 2.1 (Rellich-Kondrashov [**1**]). *Let $\Omega \subset \mathbb{R}^d$ be a bounded polygonal domain. The compactness of the embedding operator is given as follows*

(i) For $1 \leq p < d$, let $p^* = dp/(d-p)$. The embedding $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ is compact with $\forall q \in [1, p^*)$.

(ii) For $p = d$, the embedding $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ is compact with $\forall q \in [1, \infty)$.

(iii) For $p > d$, the embedding $W^{1,p}(\Omega) \hookrightarrow C(\Omega)$ is compact.

In this thesis, we mainly consider the case of $d = 1, 2$ and $p = 2$. Using this theorem, the following corollary is obtained.

COROLLARY 2.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded polygonal domain.*

(i) For $d = 1$, the embedding $H^1(\Omega) \hookrightarrow C(\Omega)$ is compact.

(ii) For $d = 2$, the embedding $H^1(\Omega) \hookrightarrow L^p(\Omega)$ is compact with $\forall p \in [1, \infty)$. Then, it follows for $v \in H^1(\Omega)$ and $p \in [1, \infty)$,

$$\|v\|_{L^p} \leq C_{e,p}\|v\|_{H_0^1}. \tag{15}$$

Values of constants $C_{e,p}$ depend on the shape of the domain Ω. For their concrete values, see Section 3.3 below. Now we choose spaces $X := L^2(\Omega)$, $V := H_0^1(\Omega)$ and $V^* := H^{-1}(\Omega)(\in \mathcal{L}(H_0^1(\Omega), \mathbb{R}))$ for simplicity.

2.2. Framework of verified computations

This section is devoted to explain a computer-assisted approach for the following abstract problem:

$$\text{Find } u \in V \text{ satisfying } \mathcal{F}(u) = 0, \tag{16}$$

where $\mathcal{F} : V \to V^*$ denotes some Fréchet differentiable mapping. Let $\hat{u} \in V_h \subset V$ be an approximate solution to (16), and the Fréchet derivative of $\mathcal{F}$ at $\hat{u}$ denotes $\mathcal{F}'[\hat{u}] : V \to V^*$, *i.e.*

$$\|\mathcal{F}(\hat{u} + \nu) - \mathcal{F}(\hat{u}) - \mathcal{F}'[\hat{u}]\nu\|_{V^*} = o(\|\nu\|_V).$$

where $o(\|\nu\|_V)$ means faster convergence than $\|\nu\|_V \to 0$,

$$\lim_{\|\nu\|_V \to 0} \frac{o(\|\nu\|_V)}{\|\nu\|_V} = 0.$$

In order to verify the existence and local uniqueness of the exact solution in the neighborhood of $\hat{u}$, we consider to apply the Newton-Kantorovich theorem [**7, 10**] to (16).

Theorem 2.3. *Assuming that the Fréchet derivative $\mathcal{F}'[\hat{u}]$ is nonsingular and satisfies*

$$\|\mathcal{F}'[\hat{u}]^{-1}\mathcal{F}(\hat{u})\|_V \leq \alpha,$$

for a certain positive α*. Then, let* $\overline{B}(\hat{u}, 2\alpha) := \{v \in V : \|v - \hat{u}\|_V \leq 2\alpha\}$ *be a closed ball centered at* $\hat{u}$ *with radius* 2α*. Let also* $D \supset \overline{B}(\hat{u}, 2\alpha)$ *be an open ball on* V*. We assume that for a certain positive* ω*, the following holds:*

$$\|\mathcal{F}'[\hat{u}]^{-1}(\mathcal{F}'[v] - \mathcal{F}'[w])\|_{V,V} \leq \omega \|v - w\|_V, \ \forall v, w \in D.$$

If

$$\alpha\omega \leq \frac{1}{2}$$

holds, then there is a solution $u \in V$ *of (16) satisfying*

$$\|u - \hat{u}\|_V \leq \rho := \frac{1 - \sqrt{1 - 2\alpha\omega}}{\omega}. \tag{17}$$

Furthermore, the solution u *is unique in* $\overline{B}(\hat{u}, 2\alpha)$*.*

REMARK 2.4. *To apply Newon-Kantorovich theorem, we will calculate the following constants explicitly.*

$$\|\mathcal{F}'[\hat{u}]^{-1}\|_{V^*,V} \leq C_1, \tag{18}$$

$$\|\mathcal{F}(\hat{u})\|_{V^*} \leq C_{2,h}, \tag{19}$$

$$\|\mathcal{F}'[v] - \mathcal{F}'[w]\|_{V,V^*} \leq C_3 \|v - w\|_V, \ \forall v, w \in D \subset V. \tag{20}$$

Therefore, if $C_1^2 C_{2,h} C_3 \leq 1/2$ *is confirmed by verified computations, then the existence and local uniqueness of the solution are proved numerically based on Newton-Kantorovich theorem. Our main task in this thesis is the calculation of these constants explicitly.*

REMARK 2.5. *The uniqueness of the solution is also available in the ball* $\overline{B}(\hat{u}, 2\alpha)$ [**7**]*. So that there is a nonexistence area of solution:*

$$\overline{B}(\hat{u}, 2\alpha) \setminus \overline{B}(\hat{u}, \rho) = \{v \in V : \rho < \|v - \hat{u}\|_V \leq 2\alpha\}.$$

Remark 2.6. *In our approach, it does not require the invertibility of $\mathcal{F}'[\hat{u}]$ explicitly. In fact, we will numerically check the invertibility of $\mathcal{F}'[\hat{u}]$ through the condition (18) even if the invertibility is difficult to be proved by the "analytic" way.*

2.3. Variational formulation

In this part, we provide variational formulations for (ordinary/partial) differential equations. We would like to deduce the equation (16) from following problems. Firstly, let $\Omega = (0,1)$ be an open interval in $\mathbb{R}$. We consider two-point boundary value problems of second order ordinary differential equations:

$$\begin{cases} -(au')' = f(u', u, x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{21}$$

where $u = u(x)$ is a function to be determined, $u'(x) = du(x)/dx$ and $a(x)$ is a smooth function on Ω with $a(x) \geq a_0 > 0$ for some $a_0 \in \mathbb{R}$. We assume $a(x) \in W^{1,\infty}(\Omega)$. Here, $f : V \to X$ is assumed to be the Fréchet differentiable. For example, the following function

$$f(u', u, x) = bu' + c_1 u + c_2 u^2 + c_3 u^3 + ... + c_N u^N + g$$

with $N \in \mathbb{N}$, $b(x), c_i(x) \in L^\infty(\Omega)$, $(i = 1, ..., N)$ and $g(x) \in X$ satisfies this condition.

Secondly, let Ω be a bounded polygonal domain on $\mathbb{R}^2$ with arbitrary shape. Let us also be concerned with the Dirichlet boundary value problem of a semilinear elliptic equation of the form:

$$\begin{cases} -\Delta u = f(\nabla u, u, x), & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{22}$$

where $f : V \to X$ is assumed to be the Fréchet differentiable. For example, the following function

$$f(\nabla u, u, x) = b \cdot \nabla u + c_1 u + c_2 u^2 + c_3 u^3 + ... + c_N u^N + g$$

with $b(x) \in (L^\infty(\Omega))^2$, $c_i(x) \in L^\infty(\Omega)$, $(i = 1, ..., N)$ and $g(x) \in X$ satisfies this condition. Our verified computation approach can help to prove the existence and local uniqueness for a weak solution of (21) and (22). Namely, if a *good* approximate solution is given in a certain function subspace of V, we will try to validate the existence of solution in the neighborhood of its approximation.

Here, we would consider problems (21) and (22) with same variational forms. We will only discuss the 2D case in details, while the 1D case can easily be done analogously. Further we rewrite $f(\nabla u, u, x)$ and $f(u', u, x)$ as $f(u)$ for simple notation.

In the classical analysis of variational theory, the solution to the Dirichlet boundary problem (22) satisfies the variational problem: Find $u \in V$ such that

$$(\nabla u, \nabla v) = (f(u), v), \text{ for all } v \in V. \tag{23}$$

For $u, v \in V$, let us define a continuous bilinear form $A(u, v)$ as

$$A(u, v) := (\nabla u, \nabla v).$$

For fixed $u \in V$, $A(u, \cdot) \in V^*$ is a linear functional. It enables us to define an operator $\mathcal{A} : V \to V^*$ by

$$\langle \mathcal{A}u, v \rangle := A(u, v), \text{ for all } v \in V.$$

We know $A(u, v)$ is an inner product of V. Then, for given $T \in V^*$, Riesz's representation theorem states the existence of a unique solution $u \in V$ such that

$$A(u, v) = \langle T, v \rangle, \text{ for all } v \in V,$$

especially, $\|u\|_V = \|\mathcal{A}u\|_{V^*}$ holds. This declares the invertibility of $\mathcal{A}$. We denote the inverse of $\mathcal{A}$ as $\mathcal{A}^{-1} : V^* \to V$. Thus, the operator $\mathcal{A}$ becomes isometric isomorphism.

Remark 2.7. *In case of (21), the continuous bilinear form $A(u, v)$ is defined by*

$$A(u, v) := (au', v'), \ u, v \in V.$$

Thus, we can define an operator $\mathcal{A} : V \to V^*$ *by*

$$\langle \mathcal{A}u, v\rangle := A(u, v).$$

It is noted that the bilinear form A *is coercive, i.e.,*

$$A(u, u) \geq a_0 \|u\|_V^2. \tag{24}$$

Then, for given $T \in V^*$*, Lax-Milgram's theorem states the existence of a unique solution for the following equation:*

$$A(u, v) = \langle T, v\rangle, \ \forall v \in V. \tag{25}$$

If we denote the operator which maps T *to the solution* u *of (25) by* $\mathcal{A}^{-1} : V^* \to V$*, this theorem declares that* $\mathcal{A}^{-1}$ *is the inverse of* $\mathcal{A} : V \to V^*$*. Further, we note that the bilinear form* $A(u, v)$ *is an inner product of* V*. There exist positive constants* C_a *and* c_a *satisfying*

$$c_a \|u\|_V \leq \|u\|_a \leq C_a \|u\|_V \text{ for } u \in V \tag{26}$$

where $\|u\|_a = \sqrt{A(u, u)}$*. In fact, we can choose* $c_a = \sqrt{a_0}$ *and* $C_a = \sqrt{\|a\|_\infty}$*.*

For fixed $u \in V$, $(f(u), \cdot)$ becomes a linear functional. Then, we can define a nonlinear operator $\mathcal{N} : V \to V^*$ by

$$\langle \mathcal{N}(u), v\rangle = (f(u), v), \text{ for all } v \in V.$$

Using these operators, the variational problem (23) can be transformed into

$$\mathcal{A}u = \mathcal{N}(u). \tag{27}$$

Furthermore, we define the operator $\mathcal{F} : V \to V^*$ by $\mathcal{F}(u) := \mathcal{A}u - \mathcal{N}(u)$. Then, (27) can be written as

$$\mathcal{F}(u) = 0. \tag{28}$$

This is nothing but the abstract problem (16).

In order to apply Newton-Kantorovich theorem, the Fréchet derivative of $\mathcal{F}$ is needed. The Fréchet differentiability of $\mathcal{F}$ is derived by that of f. We now show that $\mathcal{F} : V \to V^*$ is the Fréchet differentiable. For fixed $u, \hat{u} \in V$, $(f'(\hat{u})u, \cdot)$ is a linear functional on V. Here, $f'(\hat{u}) : V \to X$ is the Fréchet derivative of $f : V \to X$ at $\hat{u}$. Hence, we can define an operator $\mathcal{N}'[\hat{u}] : V \to V^*$ by

$$\langle \mathcal{N}'[\hat{u}]u, v\rangle := (f'(\hat{u})u, v), \ \forall v \in V. \tag{29}$$

For a given $\hat{u} \in V$, the Fréchet derivative $\mathcal{F}'[\hat{u}] : V \to V^*$ of $\mathcal{F} : V \to V^*$ at $\hat{u}$ is given as

$$\mathcal{F}'[\hat{u}] = \mathcal{A} - \mathcal{N}'[\hat{u}].$$

In fact, we have

$$\begin{aligned} \|\mathcal{F}(\hat{u}+v) - \mathcal{F}(\hat{u}) - (\mathcal{A} - \mathcal{N}'[\hat{u}])v\|_{V^*} &= \sup_{0\neq w\in V} \frac{|\langle \mathcal{N}(\hat{u}+v) - \mathcal{N}(\hat{u}) - \mathcal{N}'[\hat{u}]v, w\rangle|}{\|w\|_V} \\ &= \sup_{0\neq w\in V} \frac{|(f(\hat{u}+v) - f(\hat{u}) - f'(\hat{u})v, w)|}{\|w\|_V} \\ &\leq C_{e,2}\|\mu(\hat{u}, v)\|_X \end{aligned}$$

where $\hat{u}, v \in V$ and

$$\mu(\hat{u}, v) = f(\hat{u}+v) - f(\hat{u}) - f'(\hat{u})v.$$

From the Fréchet differentiability of $f : V \to X$, we have

$$\frac{\|\mu(\hat{u}, v)\|_X}{\|v\|_V} \to 0, \ (\|v\|_V \to 0).$$

This shows the Fréchet differentiability of $\mathcal{F} : V \to V^*$ at $\hat{u} \in V$.

Now, we define the natural embedding operator $i_{(X\to V^*)} : X \to V^*$. For fixed $w \in X$, $(w, \cdot) \in V^*$ is also linear functional. Then, we can define

$$\langle i_{(X\to V^*)}w, v\rangle := (w, v) \text{ for all } v \in V.$$

Since the embedding operator $i_{(V \to X)} : V \to X$ is compact from Corollary 2.2, its adjoint operator $i_{(X \to V^*)} : X \to V^*$ becomes compact by Schauder's theorem [**4**]. The operator $i_{(X \to V^*)} : X \to V^*$ is compact and $f'(\hat{u}) : V \to X$ is continuous so that the composite operator

$$\mathcal{N}'[\hat{u}] = i_{(X \to V^*)} \circ f'(\hat{u}) : V \to V^* \tag{30}$$

is compact.

Remark 2.8. *Actually, the nonlinear operator $\mathcal{N} : V \to V^*$ is presented using this embedding operator such that*

$$\mathcal{N}(u) = i_{(X \to V^*)} \circ f(u) \in V^*, \ \textit{for } f(u) \in X.$$

Chapter 3

Explicit evaluations

Based on fixed-point theorem, getting explicit error bounds yields the existence of solutions, see Figure 3.1. There are two constants with respect to our framework. One is error constant of an approximation. The other is the embedding constant. Here, we use the finite element approximation for getting approximate solutions. An error analysis of the finite element method is well studied so far. Recently, an explicit value of error constants is given by [**12, 15**]. Moreover, Sobolev's embedding constant $C_{e,p}$ in (15) plays important role in our framework. In this chapter, we will explain how to get explicit values of these constants.

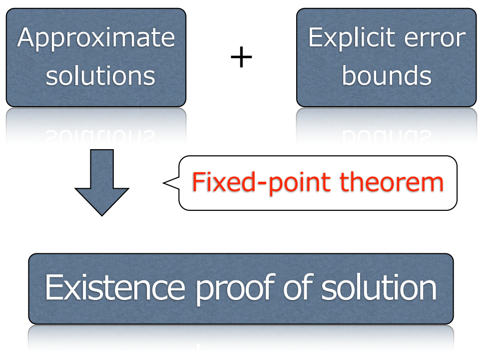

Figure 3.1: Short sketch of frame work

3.1. Error constants of finite elements (ODE)

Firstly, we consider ODE case in (21). Let us define the finite element approximation with respect to the mesh size h. Let $x_i := ih$, $0 \leq i \leq n+1$, $n \in \mathbb{N}$ with $h := 1/(n+1)$ be an equidistant partition of interval $[0,1]$. Let V_h denote a finite dimensional subspace of V spanned by linearly independent V-conforming finite element base functions. For the piecewise linear base functions ϕ_i^l, we define V_h^l as

$$V_h^l = \text{span}\{\phi_1^l, \phi_2^l, ..., \phi_n^l\} \subset V.$$

On the other hand, for piecewise quadratic base functions ϕ_i^q, we define V_h^q as

$$V_h^q = \mathrm{span}\{\phi_1^q, \phi_2^q, ..., \phi_{2n+1}^q\} \subset V$$

with midpoint of each intervals. If we use the piecewise linear or quadratic base functions, $V_h = V_h^l$ or $V_h = V_h^q$, respectively. In the following, by ϕ_i we designate ϕ_i^l or ϕ_i^q according to the base function being linear or quadratic, respectively.

An orthogonal projection $\mathcal{P}_h : V \to V_h$ is defined by

$$(a(x)(u' - (\mathcal{P}_h u)'), v_h') = 0, \quad \forall v_h \in V_h. \tag{31}$$

For $u \in V \cap H^2(\Omega)$ and its approximation $\mathcal{P}_h u \in V_h$, the error estimate is given as

$$\|u - \mathcal{P}_h u\|_V \leq C_M \|f(u)\|_X. \tag{32}$$

When $a(x) = 1$, one can take $C_M = h/\pi$ for piecewise linear elements and $C_M = h/2\pi$ for piecewise quadratic elements [**3**]. Here, we discuss how to evaluate the constant C_M in case of $a(x) \neq 1$. We assume that $\Pi_h : V \to V_h$ is the orthogonal projection defined by

$$(u' - (\Pi_h u)', v_h') = 0, \ \forall v_h \in V_h.$$

If u is smooth enough, we can assume that the constant C_h satisfying

$$\|u - \Pi_h u\|_V \leq C_h \|u''\|_X,$$

is given, *e.g.* in case of piecewise linear finite elements on grid points, one can take $C_h = h/\pi$ as mentioned above. From (24), facts that $\mathcal{P}_h u, \Pi_h u \in V_h$ and continuity of the bilinear form $A(\cdot,\cdot)$, it follows

$$\begin{aligned}
c_a^2\|u-\mathcal{P}_h u\|_V^2 &\leq A(u-\mathcal{P}_h u, u-\mathcal{P}_h u)\\
&= A(u-\mathcal{P}_h u, u) - A(u-\mathcal{P}_h u, \mathcal{P}_h u)\\
&= A(u-\mathcal{P}_h u, u)\\
&= A(u-\mathcal{P}_h u, u) - A(u-\mathcal{P}_h u, \Pi_h u)\\
&= A(u-\mathcal{P}_h u, u-\Pi_h u)\\
&\leq C_a^2\|u-\mathcal{P}_h u\|_V\|u-\Pi_h u\|_V\\
&\leq C_a^2\|u-\mathcal{P}_h u\|_V C_h\|u''\|_X.
\end{aligned}$$

Then we have Céa's lemma:

$$\|u-\mathcal{P}_h u\|_V \leq \left(\frac{C_a}{c_a}\right)^2 C_h\|u''\|_X. \tag{33}$$

Putting $-(au')' =: g_d$, it follows

$$\begin{aligned}
\|u''\|_X &= \left\|\frac{a'u'+g_d}{a}\right\|_X\\
&\leq \frac{1}{a_0}\left(\|a'u'\|_X + \|g_d\|_X\right)\\
&\leq \frac{1}{a_0}\left(\|a'\|_\infty\|u'\|_X + \|g_d\|_X\right).
\end{aligned}$$

On the other hand, from (24) it follows that

$$c_a^2\|u'\|_X^2 \leq A(u,u) = (g_d,u) \leq \|g_d\|_X\|u\|_X \leq C_{e,2}\|g_d\|_X\|u'\|_X.$$

Therefore, the inequality between $\|u''\|_X$ and $\|g_d\|_X$ is given as

$$\|u''\|_X \leq \frac{1}{a_0}\left(\frac{C_{e,2}}{c_a^2}\|a'\|_\infty + 1\right)\|g_d\|_X. \tag{34}$$

Putting

$$C' = \frac{1}{a_0}\left(\frac{C_{e,2}}{c_a^2}\|a'\|_\infty + 1\right),$$

from (33) and (34), we have the desired estimate of the constant C_M for the case of $a(x) \neq 1$ as

$$C_M = \left(\frac{C_a}{c_a}\right)^2 C_h C'.$$

Remark 3.1. *Actually, the same argument can be performed for the PDE notation on convex domain. In such a case, the bilinear form is defined by*

$$A(u, v) := (a(x)\nabla u, \nabla v),\ \forall v \in V$$

with coercivity $a(x) \geq a_0 > 0$. The orthogonal projection $\mathcal{P}_h : V \to V_h$ is defined by

$$A(u - \mathcal{P}_h u, v_h) = 0,\ \forall v_h \in V_h.$$

The error constant corresponding to (32) is given by

$$C_M = \left(\frac{C_a}{c_a}\right)^2 C_h C',\ C' = \frac{1}{a_0}\left(\frac{C_{e,2}}{c_a^2}\||\nabla a|_E\|_{L^\infty} + 1\right),$$

where C_h satisfies $\|u - \pi_{h,1}u\|_V \leq C_h\|\Delta u\|_X$, e.g. in case of the linear and equilateral triangle mesh, one can take $C_h = 0.493h$ as mentioned below.

3.2. Error constants of finite elements (PDE)

In case of PDE, evaluation of the error constant becomes more complicated. It deeply depends on the shape of domain. Let us define some notations corresponding to mesh triangulations. Let T^h be the mesh triangulation of Ω. A triangle element of T^h denotes $K_h \in T^h$. Since $V = H_0^1(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$, we define a finite element approximation of V, depending on the mesh size h

$$V_h := \text{span}\{\phi_1, \phi_2, ..., \phi_n\} \subset V. \tag{35}$$

V_h is spanned by V-conforming finite elements. n is the number of node points in $\Omega \setminus \partial\Omega$. Let us consider the following Poisson's equation for given $f \in X$,

$$\text{Find } u \text{ satisfies } -\Delta u = f,\ u = 0 \text{ on } \partial\Omega.$$

The weak formulation is defined by

$$\text{Find } u \in V \text{ satisfies } (\nabla u, \nabla v) = (f, v),\ \text{ for } v \in V. \tag{36}$$

From the finite element theory, we have a priori error estimation of Poisson's equation. Let us define the orthogonal projection which maps V to its approximation V_h as

$$(\nabla(u - \mathcal{P}_h u), \nabla v_h) = 0,\ \forall v_h \in V_h. \tag{37}$$

The classical error estimation theory gives a priori estimation of Poisson equation, for projection $\mathcal{P}_h : V \to V_h$,

$$\|u - \mathcal{P}_h u\|_V \le C_M \|f\|_X. \tag{38}$$

In case of Drichlet boudary condition with convex domains, we know the solution of (36) belongs to $H^2(\Omega)$. This is called H^2-regularity. However, such regularity is not obtained over non-convex domains. When we treat convex domains, the classical error estimate works well. It is not obtained on non-convex domains. The lack of $H^2(\Omega)$ regularity causes several problems. In order to treat arbitrary polygonal domain, we introduce some techniques to treat non-convex domain.

The concrete value of C_M is calculated by verified numerical computations. We will explain how to compute the constant explicitly. Let us begin with the classical error analysis for FEM orthogonal projections. The interpolation constants on triangular elements are evaluated by Kikuchi, Liu [**12**] and Kobayashi [**14**]. On non-convex domain, a posteriori error estimate with lowest order Raviart-Thomas mixed finite elements works alternatively.

3.2.1. A priori error estimate with H^2-regularity. Here, we will introduce two constants $C_{h,i}$ $(i = 0, 1)$ which play an important role throughout this chapter. These are related to function interpolations π_i $(i = 0, 1)$ over triangle element $K_h \in T^h$. For $u \in L^2(K_h)$, $\pi_0 u$ is constant function defined by

$$\pi_0 u := \left(\int_{K_h} u dx \right) / \left(\int_{K_h} 1 dx \right).$$

Let $\pi_1 u$ be a linear function defined for $u \in H^2(K_h)$

$$(\pi_1 u)(x) := u(x) \text{ on the vertex of } K_h.$$

Let global interpolations $\pi_{h,0}$ and $\pi_{h,1}$ be an extension of π_0 and π_1, which is $(\pi_{h,i} u)|_{K_h} = \pi_i(u|_{K_h}),\ (i = 0, 1)$. Here, we define $C_{h,0}$ and $C_{h,1}$ over triangulation T^h

$$C_{h,i} := \max_{K_h \in T^h} C_i(K_h),\ i = 0, 1 \tag{39}$$

where

$$C_0(K_h) := \sup_{0 \neq v \in H^1(K_h)} \frac{\|\pi_0 u - u\|_X}{\|u\|_V},\ C_1(K_h) := \sup_{0 \neq v \in H^2(K_h)} \frac{|\pi_1 u - u|_{H^1}}{|u|_{H^2}}.$$

These constants $C_i(K_h)$ $(i = 0, 1)$ are corresponding to eigenvalue of differential operators. Kikuchi and Liu [**12**] give the upper bound of constants in the following lemma.

LEMMA 3.2 (Kikuchi and Liu, 2007). *For $\alpha \in (0, 1)$ and $\theta \in (0, \pi)$,*

$$C_0(K_h) \leq \frac{h}{\pi} \sqrt{\frac{\nu_+(\alpha, \theta)}{2}},\ C_1(K_h) \leq 0.493h \frac{\nu_+(\alpha, \theta)}{\sqrt{2\nu_-(\alpha, \theta)}}$$

with

$$\begin{aligned} \nu_-(\alpha, \theta) &= 1 + \alpha^2 - \sqrt{1 + 2\alpha^2 \cos 2\theta + \alpha^4}, \\ \nu_+(\alpha, \theta) &= 1 + \alpha^2 + \sqrt{1 + 2\alpha^2 \cos 2\theta + \alpha^4}. \end{aligned}$$

Here, $h = |OA|$, $\alpha = |OB|/|OA|$ *and* $\theta = \angle AOB$ *on Figure 3.2.*

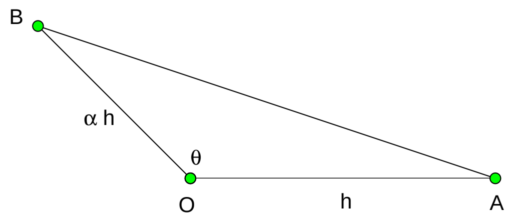

Figure 3.2: Triangle element K_h for Lemma 3.2.

Particular,

$$C_0 \le \frac{h}{\pi}, C_1 \le 0.493h$$

hold on the unit isosceles right triangle. Using this lemma, the concrete value of constants $C_{h,i}$ $(i = 0, 1)$ is evaluated explicitly with verified computations. Aside from this, another upper bounds for $C_i(K_h)$ $(i = 0, 1)$ are introduced by Kobayashi [**14**].

LEMMA 3.3 (Kobayashi). *For arbitrary triangle element,*

$$C_0(K_h) < \sqrt{\frac{a^2+b^2+c^2}{28} - \frac{S^4}{a^2b^2c^2}}$$

and

$$C_1(K_h) < \sqrt{\frac{a^2b^2c^2}{16S^2} - \frac{a^2+b^2+c^2}{30} - \frac{S^2}{5}\left(\frac{1}{a^2}+\frac{1}{b^2}+\frac{1}{c^2}\right)}$$

hold where $a = |BC|$, $b = |AC|$, $c = |AB|$ *and* S *is area of* K_h *on Figure 3.3.*

The classical a priori error estimate is given by the following theorem.

THEOREM 3.4. *Let* Ω *be* **convex** *polygonal domain. For a given* $f \in X$, *let* u *be the solution of the variational problem in (36). The error estimate between* u *and its*

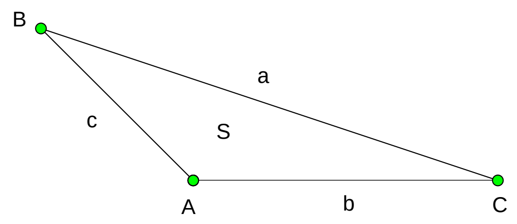

Figure 3.3: Triangle element K_h for Lemma 3.3.

approximation $\mathcal{P}_h u \in V_h$ *is given as*

$$\|u - \mathcal{P}_h u\|_V \le C_{h,1}\|f\|_X,\ \ \|u - \mathcal{P}_h u\|_X \le C_{h,1}\|u - \mathcal{P}_h u\|_V \le (C_{h,1})^2\|f\|_X.$$

PROOF. Under the given assumptions, the solution u belongs to $H^2(\Omega)$. By using the interpolation error estimate for $\pi_{h,1}$, the minimization principle gives

$$\|u - \mathcal{P}_h u\|_V \le |u - \pi_{h,1}u|_{H^1} \le C_{h,1}|u|_{H^2} \le C_{h,1}\|f\|_X,$$

where the constant $C_{h,1}$ is the one defined in (39). Here we use the fact [**8**] that, for $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $f = -\Delta u$, we have $|u|_{H^2} \le \|\Delta u\|_X = \|f\|_X$. Furthermore, by adopting Aubin-Nitsche's trick, we can deduce

$$\|u - \mathcal{P}_h u\|_X \le C_{h,1}\|u - \mathcal{P}_h u\|_V \le (C_{h,1})^2\|f\|_X.$$

□

Thus, one can take $C_M = C_{h,1}$ in (38) when we choose V_h as piecewise linear finite subspace.

3.2.2. A posteriori error estimate without H^2-regularity. For the solution with singularity ($u \notin H^2(\Omega)$), the classical error estimate is not obtained. Avoiding this difficulty, we will show the novel way to get the error estimate as below. It requires only the first derivative of the solution ($u \in H^1(\Omega)$). Thus, the following approach treats lack of H^2-regularity. Here, we follow the briefly sketch by X. Liu and S. Oishi. One can see the full discussion in [**16**].

Let us define some functional spaces corresponding to Raviart-Thomas mixed finite elements, see Appendix A for detail. Raviart-Thomas mixed finite elements are given as the subspace of $H(\text{div}, \Omega)$:

$$W_h := \left\{p_h \in H(\text{div}, \Omega) : p_h = (a_k + c_k x, b_k + c_k y)^T \text{ in } K_h\right\},$$

where a_k, b_k, c_k are constants on element K_h and

$$H(\text{div}, \Omega) := \left\{\psi \in \left(L^2(\Omega)\right)^2 : \text{div } \psi \in L^2(\Omega)\right\}.$$

W_h is spanned by $H(\text{div}, \Omega)$-conforming Raviart-Thomas mixed finite elements ψ_i

$$W_h = \text{span}\{\psi_1, \psi_2, ..., \psi_l\}, \tag{40}$$

where l denotes the number of edges on T^h. The set of piecewise constant functions on T^h is defined as

$$M_h := \left\{v \in L^2(T^h) : v \text{ is constant on each element of } T^h\right\}.$$

M_h is spanned by piecewise constant functions q_i

$$M_h = \text{span}\{q_1, q_2, ..., q_m\}, \tag{41}$$

where the number of elements on T^h declares m. The classical analysis shows $\text{div}(W_h) = M_h$ [**28**]. Corresponding to $f_h \in M_h$, a subspace of W_h is denoted by

$$W_{f_h} := \{p_h \in W_h : \text{div } p_h + f_h = 0, \text{ on each } K_h\}.$$

Further, we define another orthogonal projection $\mathcal{P}_{h,0} : L^2(\Omega) \to M_h$ satisfying

$$(u - \mathcal{P}_{h,0}u, \mu_h) = 0, \ \forall \mu_h \in M_h.$$

The property of orthogonality says

$$\|u\|_X^2 = \|\mathcal{P}_{h,0}u\|_X^2 + \|\mathcal{P}_{h,0}u - u\|_X^2, \ \forall u \in L^2(\Omega). \tag{42}$$

From the definition (39), the error estimate between $u \in H^1(\Omega)$ and its approximation $\mathcal{P}_{h,0}u \in M_h$ is given as

$$\|u - \mathcal{P}_{h,0}u\|_X \leq C_{h,0}\|u\|_V.$$

In order to evaluate the error estimate for FEM solutions without the second derivatives, we need a computable quantity κ such that

$$\kappa := \max_{0 \neq f_h \in M_h} \min_{v_h \in V_h} \min_{p_h \in W_{f_h}} \frac{\|p_h - \nabla v_h\|_X}{\|f_h\|_X}.$$

LEMMA 3.5. *For a given $f_h \in M_h$, let $\bar{u} \in H^1(\Omega)$ and $u_h \in V_h$ be the solution of variational problems,*

$$(\nabla \bar{u}, \nabla v) = (f_h, v), \ \forall v \in V \ \textit{and} \ (\nabla u_h, \nabla v_h) = (f_h, v_h), \ \forall v_h \in V_h,$$

respectively. Then we have a computable error estimate

$$\|\bar{u} - u_h\|_V \leq \kappa \|f_h\|_X. \tag{43}$$

PROOF. From Prager-Synge's theorem [**27**], for $\bar{u}$, any $v_h \in V_h$ and $p_h \in W_{f_h}$, it follows

$$\|\nabla \bar{u} - \nabla v_h\|_X^2 + \|\nabla \bar{u} - p_h\|_X^2 = \|p_h - \nabla v_h\|_X^2,$$

which is called *hypercircle equation*. This can be checked by noticing the vanishment of cross terms. Then, the following inequality holds

$$\|\nabla \bar{u} - \nabla v_h\|_X \leq \|p_h - \nabla v_h\|_X, \ \forall v_h \in V_h, \ \forall p_h \in W_{f_h}.$$

From the minimization principle, we obtain the error estimate between $\bar{u}$ and u_h,

$$\|\nabla \bar{u} - \nabla u_h\|_X \leq \|\nabla \bar{u} - \nabla v_h\|_X \leq \min_{v_h \in V_h} \min_{p_h \in W_{f_h}} \|p_h - \nabla v_h\|_X .$$

Further the definition of κ yields

$$\|\nabla \bar{u} - \nabla u_h\|_X \leq \kappa \|f_h\|_X .$$

□

THEOREM 3.6 (Liu and Oishi 2010). *For $f \in L^2(\Omega)$, let $u \in V$ and $\mathcal{P}_h u \in V_h$ be solutions of*

$$(\nabla u, \nabla v) = (f, v),\ \forall v \in V \ \textit{and}\ (\nabla(\mathcal{P}_h u), \nabla v_h) = (f, v_h),\ \forall v_h \in V_h,$$

respectively. One can set $C_M := \sqrt{(C_{h,0})^2 + \kappa^2}$ in (38). Then, the following a posteriori estimation is obtained

$$\|u - \mathcal{P}_h u\|_V \leq C_M \|f\|_X,\ \|u - \mathcal{P}_h u\|_X \leq C_M \|u - \mathcal{P}_h u\|_V \leq (C_M)^2 \|f\|_X .$$

PROOF. We follow analogous framework with Kikuchi and Saito [**13**] to finish the proof. Let $\bar{u}$ and u_h be the ones defined in Lemma 3.5 with $f_h = \mathcal{P}_{h,0} f \in M_h$. The minimization principle leads $\|u - \mathcal{P}_h u\|_V \leq \|u - u_h\|_V$. Decomposing $u - u_h$ by $(u - \bar{u}) + (\bar{u} - u_h)$, we have

$$\|u - \mathcal{P}_h u\|_V \leq \|u - u_h\|_V \leq \|u - \bar{u}\|_V + \|\bar{u} - u_h\|_V . \tag{44}$$

From definitions of u and $\bar{u}$, it follows for $\forall v \in V$,

$$(\nabla(u - \bar{u}), \nabla v) = (f - \mathcal{P}_{h,0} f, v) = ((I - \mathcal{P}_{h,0}) f, (I - \mathcal{P}_{h,0}) v) .$$

Take v be $u - \bar{u}$ and apply the error estimate for projection $\mathcal{P}_{h,0}$, we have

$$\begin{aligned}\|u-\bar{u}\|_V^2 &\leq \|(I-\mathcal{P}_{h,0})f\|_X \, \|(I-\mathcal{P}_{h,0})(u-\bar{u})\|_X \\ &\leq \|(I-\mathcal{P}_{h,0})f\|_X \cdot C_{h,0}\|u-\bar{u}\|_V.\end{aligned}$$

Hence, we have

$$\|u-\bar{u}\|_V \leq C_{h,0} \, \|(I-\mathcal{P}_{h,0})f\|_X \,. \tag{45}$$

From (42), (43) and (45) the error estimate (44) is bounded by

$$\begin{aligned}\|u-\mathcal{P}_h u\|_V &\leq \|u-\bar{u}\|_V + \|\bar{u}-u_h\|_V \\ &\leq \kappa\|\mathcal{P}_{h,0}f\|_X + C_{h,0}\,\|(I-\mathcal{P}_{h,0})f\|_X \\ &\leq \sqrt{(C_{h,0})^2+\kappa^2}\,\|f\|_X.\end{aligned}$$

Furthermore, by adopting Aubin-Nitsche's trick, the estimate for $\|u-\mathcal{P}_h u\|_X$ can be obtained. Define $e := (u-\mathcal{P}_h u) \in L^2(\Omega)$ and ζ satisfying

$$(\nabla\zeta, \nabla v) = (e, v), \ \forall v \in V.$$

Thus, we have

$$(e,e) = (\nabla\zeta, \nabla e) = (\nabla(\zeta-\mathcal{P}_h\zeta), \nabla e) \leq \|\nabla(\zeta-\mathcal{P}_h\zeta)\|_X \cdot \|\nabla e\|_X \leq C_M\|e\|_X\|\nabla e\|_X,$$

which leads to

$$\|u-\mathcal{P}_h u\|_X \leq C_M|u-\mathcal{P}_h u|_{H^1} \leq (C_M)^2\|g\|_X.$$

□

Computation of κ. In this part, we explain how to evaluate the quantity κ. The discussion will be divided into two steps. First we derive the explicit form of $u_h \in V_h$ and $p_h \in W_{f_h}$ which optimize $\|p_h - \nabla u_h\|_X$. Then, we find $f_h \in M_h$ that maximizes the value of $\|p_h-\nabla u_h\|_X/\|f_h\|_X$.

For given $f_h \in M_h$, we consider the optimization problem,

$$\inf_{u_h \in V_h} \inf_{p_h \in W_{f_h}} \|p_h - \nabla u_h\|_X. \tag{46}$$

The classical theory on Raviart-Thomas finite element method [**28, 2**] implies that the minimizer of (46) is given by solutions of following two problems

a) Find $p_h \in W_h$ and $\lambda_h \in M_h$ such that

$$\begin{cases} (p_h, q_h) + (\lambda_h, \text{div } q_h) = 0, & \forall q_h \in W_h, \\ (\text{div } p_h, \mu_h) + (f_h, \mu_h) = 0, & \forall \mu_h \in M_h. \end{cases}$$

b) Find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v_h) = (f_h, v_h), \ \forall v_h \in V_h.$$

Let the base functions of FEM spaces be ones in (35), (40) and (41). Define some matrices $P_{l\times l}$, $G_{n\times l}$, $S_{n\times n}$, $B_{n\times m}$, $M_{m\times m}$ and $N_{m\times l}$ for inner products of base functions,

$$\begin{aligned} P_{l\times l} &= (\psi_i, \psi_j), \quad G_{n\times l} = (\nabla \phi_i, \psi_j), \\ S_{n\times n} &= (\nabla \phi_i, \nabla \phi_j), \quad B_{n\times m} = (\phi_i, q_j), \\ M_{m\times m} &= (q_i, q_j), \quad N_{m\times l} = (q_i, \text{div } \psi_j). \end{aligned}$$

Additionally, suppose that $x \in \mathbb{R}^l$, $y \in \mathbb{R}^n$, $z \in \mathbb{R}^m$ and $g_v \in \mathbb{R}^m$ are vectors and let $p_h \in W_h$, $u_h \in V_h$, $\lambda_h \in M_h$ and $f_h \in M_h$ be the elements such that

$$\begin{aligned} x &= (x_1, ..., x_l)^T \in \mathbb{R}^l, \quad p_h = (\psi_1, ..., \psi_l) \cdot x \in W_h, \\ y &= (u_1, ..., u_n)^T \in \mathbb{R}^n, \quad u_h = (\phi_1, ..., \phi_n) \cdot y \in V_h, \\ z &= (z_1, ..., z_m)^T \in \mathbb{R}^m, \quad \lambda_h = (q_1, ..., q_l) \cdot z \in M_h, \\ f_v &= (g_1, ..., g_m)^T \in \mathbb{R}^m, \quad f_h = (q_1, ..., q_l) \cdot f_v \in M_h. \end{aligned}$$

By using matrix notations, problems **a)** and **b)** can be characterized by

$$\textbf{a)} \begin{cases} Px + N^T z = 0 \\ Nx + Mf_v = 0 \end{cases}, \ \textbf{b)} \ Sy = Bf_v.$$

There are various methods on solving this system. Adopting block matrix arithmetic, coefficient vectors of minimizer $p_h \in W_h$ and $u_h \in V_h$ are given by

$$x := -P^{-1}N^T(NP^{-1}N^T)^{-1}Mf_v = Hf_v \text{ and } y := S^{-1}Bf_v = Kf_v,$$

if $NP^{-1}N^T$ be nonsinguler. Then, the following is obtained

$$\begin{aligned} \|\nabla u_h - p_h\|_X^2 &= (\nabla u_h, \nabla u_h) + (p_h, p_h) - 2(\nabla u_h, p_h) \\ &= y^T S y + x^T P x - 2y^T G x \\ &= f_v^T (K^T S K + H^T P H - 2K^T G H) f_v \\ &= f_v^T Q f_v. \end{aligned}$$

Here, we put $Q := K^T SK + H^T PH - 2K^T GH \in \mathbb{R}^{m\times m}$. Note that $x^T Gy = y^T Gx = f_v^T K^T GH f_v = f_v^T H^T GK f_v$, we see Q is symmetric easily. Finally, κ^2 is given as

$$\begin{aligned} \kappa^2 &= \max_{0\neq f_h \in M_h} \min_{u_h \in V_h} \min_{p_h \in W_{f_h}} \frac{\|\nabla u_h - p_h\|_X^2}{\|f_h\|_X^2} \\ &= \max_{0\neq f_h \in M_h} \frac{f_v^T Q f_v}{f_v^T M f_v}. \end{aligned}$$

This is nothing but Rayleigh quotient form of general matrix eigenvalue problem

$$Qf_v = \lambda M f_v. \tag{47}$$

Thus, κ^2 is given by the maximum eigenvalue of (47).

Remark 3.7. *The assertion in Subsection 3.2.2 is obtained for higher order finite element, such as piecewise quadratic or more higher oder spline base functions. Using the same discussion, we have the error constant in (38). Over convex domains, the error constant $C_{h,1}$ can be used alternatively. However, this estimation becomes overestimation.*

3.3. Embedding constant

Another task of explicit evaluation is Sobolev's embedding constant $H^1(\Omega) \hookrightarrow L^p(\Omega)$ on arbitrary polygonal domain. Sobolev's embedding constant is related to the minimal spectrum of Laplacian $(-\Delta)$, which is verified by Liu, Oishi [**16**]. For $p \in [2, \infty)$, the upper bound of the constant $C_{e,p}$ satisfying

$$\|u\|_{L^p} \leq C_{e,p} \|u\|_V$$

are given by the following lemma. This is introduced by M. Plum [**26**]. He pointed out "*This is not always optimal but easy to compute*".

Lemma 3.8. *Let $\sigma \in [0, \infty)$ denote the point of the minimal spectrum of $-\Delta$ on V. Let $p \in [2, \infty)$ and ν denote the largest integer less than $p/2$. We have*

$$C_{e,p} := \left(\frac{1}{2}\right)^{\frac{1}{2}+\frac{2\nu-3}{p}} \left[\frac{p}{2}\left(\frac{p}{2}-1\right)\cdots\left(\frac{p}{2}-\nu+2\right)\right]^{\frac{2}{p}} \sigma^{-\frac{1}{p}},$$

where the bracket term is put equal to 1 if $\nu = 1$.

Here, we need the verified lower bound of the minimal spectrum of Laplacian $(-\Delta)$ on treated domain. The following theorem gives desired lower bounds, which is derived by X. Liu and S. Oishi [**16**].

Theorem 3.9 (Liu and Oishi 2010). *Let λ_k be spectrums of $-\Delta$. $\tilde{\lambda}_k$ is assumed to be its discretized approximation with verified computations. C_M declares the error constant satisfies (38). Suppose*

$$1 - (C_M)^2 \lambda_k > 0,$$

then each spectrum of $-\Delta$ is bounded by

$$\frac{\tilde{\lambda}_k}{1 + (C_M)^2 \tilde{\lambda}_k} \leq \lambda_k \leq \tilde{\lambda}_k.$$

Using this result, we can take

$$\sigma \geq \frac{\tilde{\lambda}_1}{1 + (C_M)^2 \tilde{\lambda}_1}, \tag{48}$$

where $\tilde{\lambda}_1$ is the first approximate eigenvalue in finite element discretized systems of the eigenvalue problem

$$-\Delta u = \lambda u$$

with Dirichlet boundary condition: $u = 0$ on $\partial\Omega$. For the verified method for eigenvalue problem of Laplacian $(-\Delta)$, one can see full discussions in [**16**]. It treats verified eigenvalue evaluation for elliptic operators on arbitrary polygonal domain. It also depends on the error constant C_M, which appears in (38).

Chapter 4

Verification theories

Our computer-assisted approach needs explicit values of (18)-(20) in Section 2.2. In this chapter, we treat how to calculate each constant with verification.

4.1. Invertibility of linearized operator

Let $\hat{u} \in V_h$ be approximate solution of (23). Here, we evaluate the verified upper bound of C_1 in (18), which is the norm estimation corresponding to the inverse of the Freécht derivative operator $\mathcal{F}'[\hat{u}] = \mathcal{A} - \mathcal{N}'[\hat{u}]$. Let V_h be a finite element approximation of V and $V_c := V \setminus V_h$ be its orthogonal complement. For the estimation of $\|(\mathcal{A} - \mathcal{N}'[\hat{u}])^{-1}\|_{V^*,V}$, we will present two theorems.

4.1.1. Initial theorem with ODE notations.

Theorem 4.1 (Oishi 1995 [**23**]). *Let $\hat{u} \in V_h$ and $\mathcal{N}'[\hat{u}] : V \to V^*$ be a linear compact operator. Let V_h be a finite dimensional subspace of V. Let $\mathcal{P}_h : V \to V_h$ be the orthogonal projection defined in (31). Assuming $\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] : V \to V$ is bounded and satisfies*

$$\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]\|_{V,V} \leq K,$$

the difference between $\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]$ and $\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]$ is bounded and enjoys

$$\|(\mathcal{A}^{-1} - \mathcal{P}_h\mathcal{A}^{-1})\mathcal{N}'[\hat{u}]\|_{V,V} \leq L_h,$$

and the finite dimensional operator $\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ is invertible with

$$\|(\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h})^{-1}\|_{V,V} \leq \tau.$$

Here, $\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ is the restriction of the operator $\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}]) : V \to V_h$ on V_h. If $(1 + \tau K)L_h < 1$, then the operator $\mathcal{A} - \mathcal{N}'[\hat{u}]$ is also invertible and

$$\|(\mathcal{A} - \mathcal{N}'[\hat{u}])^{-1}\|_{V^*,V} \leq \frac{1}{c_a^2}\frac{1 + \tau K}{1 - (1 + \tau K)L_h} =: C_1.$$

□

PROOF. Since

$$u = (\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u + (\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u + \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u,$$

we have

$$\begin{aligned} &\|u\|_V \\ \leq\ & \|(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V + \|(\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])\|_{V,V}\|u\|_V + \|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u\|_V \\ \leq\ & \|(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V + L_h\|u\|_V + \|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u\|_V. \end{aligned} \tag{49}$$

From

$$\begin{aligned} &\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u \\ =\ & \mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])(\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u + \mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u \\ =\ & \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}](\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u + \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}](\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u \end{aligned}$$

and the invertibility of $\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ with

$$\|(\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h})^{-1}\|_{V,V} \leq \tau,$$

we have

$$\begin{aligned} \|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u\|_V \ \leq\ & \tau\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]\|_{V,V}\|(\mathcal{A}^{-1} - \mathcal{P}_h\mathcal{A}^{-1})\mathcal{N}'[\hat{u}]\|_{V,V}\|u\|_V \\ & +\tau\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]\|_{V,V}\|(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V \\ \leq\ & \tau K L_h\|u\|_V + \tau K\|(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V. \end{aligned} \tag{50}$$

Substituting (50) into (49), we have

$$\|u\|_V \leq (1+\tau K)\|(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V + (1+\tau K)L_h\|u\|_V.$$

Thus, if $(1+\tau K)L_h < 1$, then we obtain

$$\|u\|_V \le \frac{1+\tau K}{1-(1+\tau K)L_h}\|(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u\|_V. \tag{51}$$

From (51), if $(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u = 0$, $u = 0$ follows. This implies that the operator $(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]) : V \to V$ is injective. Since the operator $(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]) : V \to V$ is Fredholm type with the index 0, it is also surjective. Thus, $\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}] : V \to V$ is invertible and enjoys

$$\|(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])^{-1}\|_{V,V} \le \frac{1+\tau K}{1-(1+\tau K)L_h}.$$

From the coercivity,

$$c_a^2\|u\|_V^2 \le \|u\|_a^2 = A(u,u) = \langle \mathcal{A}u, u\rangle \le \|\mathcal{A}u\|_{V^*}\|u\|_V,$$

it follows

$$\|\mathcal{A}^{-1}\|_{V^*,V} \le \frac{1}{c_a^2}.$$

If we note

$$(\mathcal{A}-\mathcal{N}'[\hat{u}])^{-1} = (\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])^{-1}\mathcal{A}^{-1},$$

we have

$$\|(\mathcal{A}-\mathcal{N}'[\hat{u}])^{-1}\|_{V^*,V} \le \frac{1}{c_a^2}\frac{1+\tau K}{1-(1+\tau K)L_h}.$$

This completes the proof. □

Actually, this theorem is related to the perturbation lemma of linear operators. Then, another proof is given by a well known lemma as below.

Lemma 4.2. *Let V and W be normed spaces with at least one of them being complete. Assume $\mathcal{L} \in \mathcal{L}(V,W)$ has a bounded inverse $\mathcal{L}^{-1} : W \to V$ and $\mathcal{M} \in \mathcal{L}(V,W)$ satisfies*

$$\|\mathcal{L}-\mathcal{M}\|_{V,W} < \frac{1}{\|\mathcal{L}^{-1}\|_{W,V}}.$$

Then, $\mathcal{M} : V \to W$ is a bijection, $\mathcal{M}^{-1} \in \mathcal{L}(W, V)$ and

$$\|\mathcal{M}^{-1}\|_{W,V} \leq \frac{\|\mathcal{L}^{-1}\|_{W,V}}{1 - \|\mathcal{L}^{-1}\|_{W,V}\|\mathcal{L} - \mathcal{M}\|_{V,W}}. \tag{52}$$

Now we consider to apply this lemma to the proof of Theorem 4.1. Let $V = W$, $\mathcal{L} = I - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]$ and $\mathcal{M} = I - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}]$. Then, it follows

$$\|\mathcal{L} - \mathcal{M}\|_{V,W} = \|(\mathcal{A}^{-1} - \mathcal{P}_h\mathcal{A}^{-1})\mathcal{N}'[\hat{u}]\|_{V,V} \leq L_h$$

and

$$\|\mathcal{L}^{-1}\|_{W,V} = \|(I - \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])^{-1}\|_{V,V} \leq 1 + \tau K.$$

Thus, if $(1 + \tau K)L_h < 1$, from (52), it turns out that $\mathcal{M} = I - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}] : V \to V$ is invertible and enjoys

$$\|(I - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])^{-1}\|_{V,V} \leq \frac{1 + \tau K}{1 - (1 + \tau K)L_h}.$$

This argument becomes another proof of Theorem 4.1.

Next, we review briefly how to compute K, L_h and τ in Theorem 4.1. Detailed arguments can be seen in [**36**]. In the first place, we show how to calculate K. Equations (26), (31) and the definition of the operator $\mathcal{A}$ yield for $v \in V$

$$\begin{aligned}
\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v\|_V^2 &\leq \frac{1}{c_a^2}(a(x)(\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v)', (\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v)') \\
&= \frac{1}{c_a^2}(a(x)(\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v)', (\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v)') \\
&= \frac{1}{c_a^2}(f'(\hat{u})v, \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v) \\
&\leq \frac{C_{e,2}}{c_a^2}\|f'(\hat{u})v\|_X\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v\|_V.
\end{aligned}$$

Then we have

$$\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]\|_{V,V} = \sup_{v \in V\setminus\{0\}} \frac{\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]v\|_V}{\|v\|_V} \leq \frac{C_{e,2}}{c_a^2}\|f'(\hat{u})\|_{V,X}.$$

Thus, we can take

$$K := \frac{C_{e,2}}{c_a^2}\|f'(\hat{u})\|_{V,X}.$$

Secondly, we calculate the constant L_h. It follows from (32),

$$\begin{aligned}\|(\mathcal{A}^{-1} - \mathcal{P}_n\mathcal{A}^{-1})\mathcal{N}'[\hat{u}]v\|_V &\leq C_M\|f'(\hat{u})v\|_X \\ &\leq C_M\|f'(\hat{u})\|_{V,X}\|v\|_V.\end{aligned}$$

Thus, one can put

$$L_h := C_M\|f'(\hat{u})\|_{V,X}.$$

Let S and B be $n \times n$ matrices whose (i,j)-elements are given by

$$(a(x)\phi_j', \phi_i') \text{ and } (a(x)\phi_j', \phi_i') - (f'(\hat{u})\phi_j, \phi_i).$$

Let a lower triangular matrix L be the Cholesky decomposition of S, $S = LL^T$. We note that Nakao, Hashimoto and Watanabe [**19**] have shown that τ is given as

$$\tau := \frac{C_a}{c_a}\|L^T B^{-1} L\|_2.$$

We give the method of calculating τ with verified computations in Subsection 4.1.4.

4.1.2. Improved theorem with PDE notations. Theorem 4.1 is enough evaluation for ODE case. However, it sometimes becomes quite overestimation in case of PDE. The following theorem is presumed as an improvement of former theorem. It uses the structure of orthogonal properties. This theorem is a modification of the main theorem in Nakao, Hashimoto and Watanabe [**19**] in 2005. We here give another proof of this theorem.

THEOREM 4.3. *Let $\mathcal{N}'[\hat{u}] : V \to V^*$ be the linear compact operator defined in (29) and V_h be the finite dimensional subspace of V spanned by finite element base functions. Let $\mathcal{P}_h : V \to V_h$ be the orthogonal projection defined in (37). For three* *constants K_1, K_2 and K', we assume*

$$\|f'(\hat{u})u\|_X \le K_1\|u\|_V,\ \forall u \in V,$$

$$\|f'(\hat{u})u_c\|_X \le K_2\|u_c\|_V,\ \forall u_c \in V_c$$

and

$$\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c\|_V \le K'\|u_c\|_V,\ \forall u_c \in V_c.$$

Assuming that the finite dimensional operator $\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ is invertible with

$$\left\|\left(\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h}\right)^{-1}\right\|_{V,V} \le \tau.$$

Here, $\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ is the restriction of $\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]) : V \to V_h$ to V_h. Moreover, the same as (38), the error estimate of $\mathcal{P}_h$ is obtained for given $f \in X$:

$$\|u - \mathcal{P}_h u\|_V \le C_M\|f\|_X.$$

If $C_M(K_1\tau K' + K_2) < 1$, then $(\mathcal{A}-\mathcal{N}'[\hat{u}]) : V \to V^*$ is invertible and enjoys

$$\|(\mathcal{A}-\mathcal{N}'[\hat{u}])^{-1}\|_{V^*,V} \le \sqrt{r^2+s^2}.$$

Here,

$$r := \frac{\sqrt{(C_M K_1\tau)^2+1}}{1-C_M(K_1\tau K'+K_2)} \text{ and } s := \tau(K'r+1).$$

Proof. We fix $u \in V$. Putting $\varphi \in V^*$ as

$$(\mathcal{A}-\mathcal{N}'[\hat{u}])u = \varphi, \tag{53}$$

and

$$u_h := \mathcal{P}_h u,\ u_c := (\mathcal{I}-\mathcal{P}_h)u,$$

$$\varphi_h := \mathcal{P}_h\mathcal{A}^{-1}\varphi,\ \varphi_c := (\mathcal{I}-\mathcal{P}_h)\mathcal{A}^{-1}\varphi.$$

Obviously, the following is obtained

$$u = u_h + u_c,\ \mathcal{A}^{-1}\varphi = \varphi_h + \varphi_c.$$

Further, the property of the orthogonality says

$$\|u_h\|_V^2 + \|u_c\|_V^2 = \|u\|_V^2,\ \|\varphi_h\|_V^2 + \|\varphi_c\|_V^2 = \|\mathcal{A}^{-1}\varphi\|_V^2 = \|\varphi\|_{V^*}^2.$$

From (53), we have

$$\begin{aligned} & \mathcal{P}_h\mathcal{A}^{-1}(\mathcal{A} - \mathcal{N}'[\hat{u}])(u_h + u_c) = \varphi_h \\ \Longleftrightarrow\quad & \mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u_h = \mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c + \varphi_h. \end{aligned}$$

From the assumption, it holds

$$\left\|\left(\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h}\right)^{-1}\right\|_{V,V} \leq \tau.$$

So that the following inequality holds

$$\begin{aligned} \|u_h\|_V &\leq \tau\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c + \varphi_h\|_V \\ &\leq \tau\left(K'\|u_c\|_V + \|\varphi_h\|_V\right). \end{aligned} \tag{54}$$

On the other hand, from (53), it follows

$$\begin{aligned} & (\mathcal{I} - \mathcal{P}_h)\mathcal{A}^{-1}(\mathcal{A} - \mathcal{N}'[\hat{u}])(u_h + u_c) = \varphi_c \\ \Longleftrightarrow\quad & u_c = (\mathcal{I} - \mathcal{P}_h)\mathcal{A}^{-1}\mathcal{N}'[\hat{u}](u_h + u_c) + \varphi_c. \end{aligned}$$

For given $f \in X$, we note that the solution of variational problem $(\nabla u, \nabla v) = (f, v),\ \forall v \in V$, can be denoted as $u = \mathcal{A}^{-1}i_{(X\to V^*)} \circ f$. The error estimate (38) is rewritten by

$$\|u - \mathcal{P}_h u\|_V = \|(\mathcal{I} - \mathcal{P}_h)\mathcal{A}^{-1}i_{(X\to V^*)} \circ f\|_V \leq C_M\|f\|_X.$$

The representation of $\mathcal{N}'[\hat{u}]$ in (30) follows

$$\begin{aligned}\|(\mathcal{I}-\mathcal{P}_h)\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u\|_V &= \|(\mathcal{I}-\mathcal{P}_h)\mathcal{A}^{-1}i_{(X\to V^*)}\circ f'(\hat{u})u\|_V \\ &\le C_M\|f'(\hat{u})u\|_X.\end{aligned}$$

Thus, it turns out with (54)

$$\begin{aligned}\|u_c\|_V &= \left\|(\mathcal{I}-\mathcal{P}_h)\mathcal{A}^{-1}\mathcal{N}'[\hat{u}](u_h+u_c)+\varphi_c\right\|_V \\ &\le C_M\|f'(\hat{u})(u_h+u_c)\|_X+\|\varphi_c\|_V \\ &\le C_M\left(\|f'(\hat{u})u_h\|_X+\|f'(\hat{u})u_c\|_X\right)+\|\varphi_c\|_V \\ &\le C_M\left(K_1\|u_h\|_V+K_2\|u_c\|_V\right)+\|\varphi_c\|_V \\ &\le C_M\left(K_1\tau\left(K'\|u_c\|_V+\|\varphi_h\|_V\right)+K_2\|u_c\|_V\right)+\|\varphi_c\|_V \\ &\le C_M(K_1\tau K'+K_2)\|u_c\|+\sqrt{(C_MK_1\tau)^2+1}\|\mathcal{A}^{-1}\varphi\|_V.\end{aligned}$$

If the assumption

$$C_M(K_1\tau K'+K_2)<1 \tag{55}$$

holds, then we have

$$\|u_c\|_V\le\frac{\sqrt{(C_MK_1\tau)^2+1}}{1-C_M(K_1\tau K'+K_2)}\|\mathcal{A}^{-1}\varphi\|_V=:r\|\varphi\|_{V^*}. \tag{56}$$

Under the condition (55), substituting (56) into (54), it follows

$$\|u_h\|_V\le\tau\left(K'r\|\varphi\|_{V^*}+\|\varphi_h\|_V\right)\le\tau(K'r+1)\|\varphi\|_{V^*}=:s\|\varphi\|_{V^*}.$$

Summing up above arguments, we have

$$\|u\|_V\le\sqrt{r^2+s^2}\|(\mathcal{A}-\mathcal{N}'[\hat{u}])u\|_{V^*} \tag{57}$$

provided that $C_M(K_1\tau K'+K_2)<1$. From (57), if $(\mathcal{A}-\mathcal{N}'[\hat{u}])u=0$ in V^*, it follows $u=0$. This implies the operator $(\mathcal{A}-\mathcal{N}'[\hat{u}]):V\to V^*$ is injective. Since the operator

$(\mathcal{A}-\mathcal{N}'[\hat{u}]) : V \to V^*$ is Fredholm type with the index 0, it is also surjective. Thus, $(\mathcal{A}-\mathcal{N}'[\hat{u}]) : V \to V^*$ is invertible and enjoys

$$\|(\mathcal{A}-\mathcal{N}'[\hat{u}])^{-1}\|_{V^*,V} \le \sqrt{r^2+s^2}.$$

This completes the proof. □

Therefore one can put $C_1 := \sqrt{r^2+s^2}$ in (18).

4.1.3. Several constants. Three constants K_1, K_2 and K' are able to compute explicitly. For K_1 and K_2, it is obvious by the definition

$$K_1 = \|f'(\hat{u})\|_{V,L^2} \text{ and } K_2 = \|f'(\hat{u})\|_{V_c,L^2},$$

respectively. It depends on the concrete notation of Frécht derivative $f'(\hat{u})$. Further, the norm of $\mathcal{P}_h\mathcal{A}^1\mathcal{N}'[\hat{u}] : V_c \to V_h$ is satisfying for $u_c \in V_c$

$$\begin{aligned}
\|\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c\|_V &= \sup_{0\neq v_h\in V_h} \frac{A\left(\mathcal{P}_h\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c, v_h\right)}{\|v_h\|_V} \\
&= \sup_{0\neq v_h\in V_h} \frac{A\left(\mathcal{A}^{-1}\mathcal{N}'[\hat{u}]u_c, v_h\right)}{\|v_h\|_V} \\
&= \sup_{0\neq v_h\in V_h} \frac{\langle \mathcal{N}'[\hat{u}]u_c, v_h\rangle}{\|v_h\|_V} \\
&= \sup_{0\neq v_h\in V_h} \frac{(f'(\hat{u})u_c, v_h)}{\|v_h\|_V} \\
&\le C_{e,2}\|f'(\hat{u})u_c\|_X \\
&\le C_{e,2}K_2\|u_c\|_X.
\end{aligned}$$

Thus, one can put $K' := C_{e,2}K_2$. In Section 5.2, practical notations with respect to K_1 and K_2 are introduced as Example.

4.1.4. Method of calculating τ. The invertibility of $\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h} : V_h \to V_h$ can be checked by verified computations. In the following, the upper bound of τ will be intruduced. Putting $\mathcal{B}_h := \mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h}$ and for $u_h \in V_h$, it follows

$$\begin{aligned}
\|\mathcal{P}_h(\mathcal{I}-\mathcal{A}^{-1}\mathcal{N}'[\hat{u}])|_{V_h}u_h\|_V &= \|\mathcal{B}_h u_h\|_V \\
&= \sup_{0\neq v_h\in V_h}\frac{A(\mathcal{B}_h u_h, v_h)}{\|v_h\|_V} \\
&\geq \inf_{0\neq u_h\in V_h}\sup_{0\neq v_h\in V_h}\frac{A(\mathcal{B}_h u_h, v_h)}{\|u_h\|_V\|v_h\|_V}\|u_h\|_V.
\end{aligned}$$

From the inf-sup condition, if a nonnegative value

$$\eta := \inf_{0\neq u_h\in V_h}\sup_{0\neq v_h\in V_h}\frac{A(\mathcal{B}_h u_h, v_h)}{\|u_h\|_V\|v_h\|_V} > 0$$

is obtained, then $\mathcal{B}_h$ is invertivle and enjoys

$$(\|\mathcal{B}_h^{-1}\|_{V,V})^{-1} = \eta.$$

Then, one can put $\tau := \eta^{-1}$. The verified evaluation of η is introduced as follows. Let $X, Y \in \mathbb{R}^n$ be be a real vector and $u_h, v_h \in V_h$ be the elements satisfying

$$\begin{aligned}
X = (u_1, ..., u_n)^T \in \mathbb{R}^n, &\qquad u_h = (\phi_1, ..., \phi_n)\cdot X \in V_h \\
Y = (v_1, ..., v_n)^T \in \mathbb{R}^n, &\qquad v_h = (\phi_1, ..., \phi_n)\cdot Y \in V_h,
\end{aligned}$$

respectively. Let $B_{n\times n}$ and $S_{n\times n}$ be real matrices whose i-j elements are given by

$$\begin{aligned}
B_{ij} &= (\nabla\phi_j, \nabla\phi_i) - (f'(\hat{u})\phi_j, \phi_i), \\
S_{ij} &= (\nabla\phi_j, \nabla\phi_i),
\end{aligned}$$

for $1 \leq i, j \leq n$. Therefore, we have

$$\begin{aligned}
\eta &= \inf_{0\neq u_h\in V_h} \sup_{0\neq v_h\in V_h} \frac{A(\mathcal{B}_h u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \\
&= \inf_{0\neq u_h\in V_h} \sup_{0\neq v_h\in V_h} \frac{A(\mathcal{P}_h(\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \\
&= \inf_{0\neq u_h\in V_h} \sup_{0\neq v_h\in V_h} \frac{A((\mathcal{I} - \mathcal{A}^{-1}\mathcal{N}'[\hat{u}])u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \\
&= \inf_{0\neq u_h\in V_h} \sup_{0\neq v_h\in V_h} \frac{(\nabla u_h, \nabla v_h) - (f'(\hat{u})u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \\
&= \inf_{0\neq X\in\mathbb{R}^n} \sup_{0\neq Y\in\mathbb{R}^n} \frac{X^T B Y}{|X^T S X|^{1/2} |Y^T S Y|^{1/2}}.
\end{aligned}$$

Since S is symmetric positive definite, there exists a lower triangular matrix L forming the Cholesky decomposition, $S = LL^T$. Here we denote $\tilde{Y} = L^T Y$, then

$$\begin{aligned}
\eta &= \inf_{0\neq X\in\mathbb{R}^n} \sup_{0\neq Y\in\mathbb{R}^n} \frac{X^T B (L^{-t} L^T) Y}{|X^T S X|^{1/2} |Y^T (LL^T) Y|^{1/2}} \\
&= \inf_{0\neq X\in\mathbb{R}^n} \sup_{0\neq \tilde{Y}\in\mathbb{R}^n} \frac{(X^T B L^{-t})\tilde{Y}}{|X^T S X|^{1/2} |\tilde{Y}^T \tilde{Y}|^{1/2}} \\
&= \inf_{0\neq X\in\mathbb{R}^n} \frac{X^T B S^{-1} B^T X}{|X^T S X|^{1/2} |X^T B S^{-1} B^T X|^{1/2}}, \ (\text{putting } \tilde{Y} := L^{-1} B^T X) \\
&= \inf_{0\neq X\in\mathbb{R}^n} \frac{|X^T B S^{-1} B^T X|^{1/2}}{|X^T S X|^{1/2}}.
\end{aligned}$$

This is nothing but Rayleigh quotient form of general matrix eigenvalue problem. Thus, η^2 is the smallest eigenvalue of

$$BS^{-1}B^T x = \lambda S x, \ \lambda \in \mathbb{R}, \ x \in \mathbb{R}^n.$$

We now discuss how to get a rigorous upper bound of τ by verified numerical computation. For a matrix $A, \ B \in \mathbb{R}^{n\times n}$, we define

$$\lambda_{\min}(A) := \min\{|\lambda| : \lambda \in \mathrm{Spec}(A)\}, \ \lambda_{\max}(A) := \max\{|\lambda| : \lambda \in \mathrm{Spec}(A)\},$$

where Spec(A) is the set of eigenvalues of A. Further, let $\sigma_{\min}(A)$ be the minimum of the singular values of A. It is known that always

$$\sigma_{\min}(A) \leq \lambda_{\min}(A), \ \sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B).$$

Since $\tau = \eta^{-1}$, the lower bound of η gives the upper bound of τ. As an efficient method of evaluation to the lower bound of η by verified numerical computation, we use the following lemma, which exploits effectively the sparsity of B and S. This is based on the method in [**31**]

Lemma 4.4. *Let $\gamma > 0$ be an estimate of lower bound of $\sigma_{\min}(S^{-1}B)$. Check*

$$BB^T - \gamma^2 S^2 \succeq 0, \tag{58}$$

where $A \succ 0$ ($\succeq 0$) means that $A \in \mathbb{R}^{n\times n}$ is symmetric positive (semi-)definite. [1] *If the condition (58) is satisfied, then*

$$\sigma_{\min}(S^{-1}B) \geq \gamma > 0. \tag{59}$$

Proof. Note that from

$$\nu := \sigma_{\min}(S^{-1}B)^2 = \lambda_{\min}(S^{-1}B(S^{-1}B)^T) = \lambda_{\min}(S^{-2}BB^T),$$

it follows that there exists $y \in \mathbb{R}^n\backslash\{0\}$ satisfying

$$BB^T y = \nu S^2 y. \tag{60}$$

Suppose $\gamma^2 > \nu$, we have

$$BB^T - \nu S^2 = BB^T - \gamma^2 S^2 + (\gamma^2 - \nu)S^2 \succ 0,$$

because S is symmetric positive definite. ν is no eigenvalue of $S^{-2}BB^T$. This contradicts to (60). □

[1] The condition (58) is suggested by S.M. Rump [**32**].

It should be noted here that (58) can be checked using Rump's method (`isspd`) [**30**] by performing the sparse Cholesky decomposition with the floating point arithmetic once. The sparse Cholesky decomposition algorithm is stable and efficient.

Now, let us consider the case $B \in \mathbb{R}^{n\times n}$ being symmetric. In this case, from (59), we have

$$\eta = \lambda_{\min}(S^{-1}BS^{-1}B^T)^{1/2} = \lambda_{\min}(S^{-1}B) \geq \sigma_{\min}(S^{-1}B) \geq \gamma.$$

The upper bound of τ is evaluated as $\tau = \eta^{-1} \leq \gamma^{-1}$.

Next, let us consider the case $B \in \mathbb{R}^{n\times n}$ being general. In this case, we have

$$\eta = \lambda_{\min}(S^{-1}BS^{-1}B^T)^{1/2} \geq \sigma_{\min}(S^{-1}BS^{-1}B^T)^{1/2} \geq \left(\sigma_{\min}(S^{-1}B)\sigma_{\min}(S^{-1}B^T)\right)^{1/2}. \tag{61}$$

If further we check

$$B^TB - \gamma'^2S^2 \succeq 0$$

as above, then

$$\sigma_{\min}(S^{-1}B^T) \geq \gamma'$$

holds. So that it follows form (61),

$$\tau = \eta^{-1} \leq \frac{1}{\sqrt{\gamma\gamma'}}.$$

Remark 4.5. *η is represented by*

$$\eta = (\|L^TB^{-1}L\|_2)^{-1}.$$

It is proved as follows. Since $LL^T = S$ denotes the Cholesky decomposition of S, the symmetry of S implies

$$\begin{aligned}
\eta^2 &= \lambda_{\min}(S^{-1}BS^{-1}B^T) \\
&= \lambda_{\max}(B^{-t}SB^{-1}S)^{-1} \\
&= \lambda_{\max}(B^{-t}LL^TB^{-1}LL^T)^{-1} \\
&= \lambda_{\max}(L^TB^{-t}L \cdot L^TB^{-1}L)^{-1} \\
&= \lambda_{\max}((L^TB^{-1}L)^T \cdot L^TB^{-1}L)^{-1} \\
&= \|L^TB^{-1}L\|_2^{-2}.
\end{aligned}$$

Therefore $\tau = \|L^TB^{-1}L\|_2$.

4.2. Residual bounds

In this section, we would like to consider how to calculate a residual evaluation (19) such that

$$\begin{aligned}
\|\mathcal{F}(\hat{u})\|_{V^*} &= \sup_{0\neq v\in V} \frac{|\langle \mathcal{A}\hat{u} - \mathcal{N}(\hat{u}), v\rangle|}{\|v\|_V} \\
&= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}, \nabla v) - (f(\hat{u}), v)|}{\|v\|_V}
\end{aligned}$$

in several ways. If an approximate solution satisfies $\hat{u} \in H^2(\Omega) \cap V_h$, it follows

$$\begin{aligned}
\|\mathcal{F}(\hat{u})\|_{V^*} &= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}, \nabla v) - (f(\hat{u}), v)|}{\|v\|_V} \\
&= \sup_{0\neq v\in V} \frac{|(-\Delta\hat{u}, v) - (f(\hat{u}), v)|}{\|v\|_V} \\
&\leq C_{e,2}\|\Delta\hat{u} + f(\hat{u})\|_X. \qquad (62)
\end{aligned}$$

Here, $C_{e,p}$ means Sobolev's embedding constant, which satisfies $\|u\|_{L^p} \leq C_{e,p}|u|_{H^1}$, $(2 \leq p < \infty)$ for $u \in V$. We point out that the evaluation (62) does not work when V_h is taken as C^0 finite element functions, such as P_1 (piecewise linear) or

P_2 (piecewise quadratic) elements. This is because $\Delta\hat{u}$ does not belong to $L^2(\Omega)$ anymore. Furthermore, in case of V_h being C^1 element, we don't have *good* bound for (62) in non convex domain. It is famous fact that the weak solution has a singularity on non convex corner [**8**]. The lack of $H^2(\Omega)$-regularity causes (62) unbounded [**20**].

To weaken the condition on $\hat{u}$, we will introduce several methods that do not need the H^2-regularity of approximate solution. The first method to be introduced is fast to compute but gives little rough bound. The second one has accurate estimation with smoothing technique. The third one is based on Raviart-Thomas mixed finite elements [**2, 5, 28**], which can provide better bound for residue if higher order elements are used.

4.2.1. Simple bounds. Let V_h be a finite element subspace of V, such that $V_h := \text{span}\{\phi_1, ..., \phi_n\}$. Let $u_h := \mathcal{P}_h u \in V_h$ be an orthogonal projection of $u \in V$, defined as

$$(\nabla(u - u_h), \nabla v_h) = 0, \ \forall v_h \in V_h.$$

In this part, we will show a simple upper bound of residue. In the following, we denote v_h by the projection of v, *i.e.* $\mathcal{P}_h v$. From the classical error analysis, such as Aubin-Nitsche's trick, we have

$$\|v - v_h\|_X \leq C_M \|v - v_h\|_V, \tag{63}$$

$$\|v - v_h\|_V \leq \|v\|_V \text{ and } \|v_h\|_V \leq \|v\|_V. \tag{64}$$

Here C_M is a priori error constant for projection $\mathcal{P}_h$. The full discussion of this constant on arbitrary domain is shown in [**16**]. For $v_h \in V_h$, the residual bound of

(19) is given using inequalities (63) and (64)

$$
\begin{aligned}
\|\mathcal{F}(\hat{u})\|_{V^*} &= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u},\nabla v)-(f(\hat{u}),v)|}{\|v\|_V} \\
&= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u},\nabla(v-v_h))-(f(\hat{u}),v-v_h)+(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v\|_V} \\
&\leq \sup_{0\neq v\in V} \frac{|(f(\hat{u}),v-v_h)|}{\|v\|_V} + \sup_{0\neq v_h\in V_h} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v_h\|_V} \\
&\leq C_M\|f(\hat{u})\|_X + C_r \qquad (65)
\end{aligned}
$$

where the second term C_r is defined by the following procedure

$$
\begin{aligned}
&\sup_{0\neq v\in V} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v\|_V} \\
= &\sup_{\substack{0\neq v\in V\\ 0=v_h\in V_h}} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v\|_V} + \sup_{\substack{0\neq v\in V\\ 0\neq v_h\in V_h}} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v_h\|_V}\cdot\frac{\|v_h\|_V}{\|v\|_V} \\
\leq &\sup_{0\neq v_h\in V_h} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v_h\|_V} =: C_r.
\end{aligned}
$$

Let ε_i be $\varepsilon_i := (\nabla\hat{u},\nabla\phi_i)-(f(\hat{u}),\phi_i),\ (i=1,...,n)$. Since $\forall v_h\in V_h$, we can express v_h as $v_h := \sum_{i=1}^n c_i\phi_i$. Let us put $c := (c_1,...,c_n)^T$ and $\varepsilon := (\varepsilon_1,...,\varepsilon_n)^T$. Let further D be $n\times n$ matrix whose (i,j)-elements are given by $(\nabla\phi_i,\nabla\phi_j)$. Then, C_r follows

$$
C_r = \sup_{0\neq v_h\in V_h} \frac{|(\nabla\hat{u},\nabla v_h)-(f(\hat{u}),v_h)|}{\|v_h\|_V} = \sup_{c\in\mathbb{R}^n} \frac{|\sum_{i=1}^n c_i\varepsilon_i|}{\sqrt{c^TDc}} \leq \sup_{c\in\mathbb{R}^n} \frac{|c|_{l^2}|\varepsilon|_{l^2}}{\sqrt{c^TDc}} \leq \|D^{-1}\|_2|\varepsilon|_{l^2}. \qquad (66)
$$

From (65) and (66), we obtain

$$
\|\mathcal{F}(\hat{u})\|_{V^*} \leq C_M\|f(\hat{u})\|_X + \|D^{-1}\|_2|\varepsilon|_{l^2}. \qquad (67)
$$

4.2.2. Accurate bounds with a smoothing technique. As mentioned above, the simple bound (67) seems a rough bound. Overestimation often causes false in verification. Next, another method for evaluating the residual is introduced. This is based on the smoothing technique proposed by N. Yamamoto et. al. [**38**]. Here, *smoothing* means to an approximate vector $\nabla\hat{u}$ by a smooth function. According to

[**38**], if P_1 (piecewise linear) elements are used for approximate solutions, the residual evaluation becomes almost the same as the rough bound in (67). On the other hand, using higher order element, this smoothing technique works very well [**35**]. Let $X_h \subset H^1(\Omega)$ be a finite element subspace that does not vanish on boundary of Ω. Let $p_h \in (X_h)^2$ be the vector function defined by

$$(p_h - \nabla\hat{u}, v^*) = 0, \ \forall v^* \in (X_h)^2. \tag{68}$$

Namely it is the L^2-projection of $\nabla\hat{u} \in (X)^2$ to $p_h \in (X_h)^2$. p_h makes the quantity $\|p_h - \nabla\hat{u}\|_X$ small. Further the following Green's formula holds for p_h [**38**]:

$$(p_h, \nabla v) + (\text{div } p_h, v) = 0, \ \forall v \in V. \tag{69}$$

Therefore, using p_h and inequalities (63), (64), (66) and (69), we have

$$\begin{aligned}
\|\mathcal{F}(\hat{u})\|_{V^*} &= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}, \nabla v) - (f(\hat{u}), v)|}{\|v\|_V} \\
&= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}, \nabla(v - v_h)) - (f(\hat{u}), v - v_h) + (\nabla\hat{u}, \nabla v_h) - (f(\hat{u}), v_h)|}{\|v\|_V} \\
&\leq \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}, \nabla(v - v_h)) - (f(\hat{u}), v - v_h)|}{\|v\|_V} + C_r \\
&\leq \sup_{0\neq v\in V} \frac{\big|\,(\nabla\hat{u} - p_h, \nabla(v - v_h)) + (p_h, \nabla(v - v_h)) - (f(\hat{u}), v - v_h)\,\big|}{\|v\|_V} + C_r \\
&\leq \sup_{0\neq v\in V} \frac{\|\nabla\hat{u} - p_h\|_X\|v - v_h\|_V + \|\text{div } p_h + f(\hat{u})\|_X\|v - v_h\|_X}{\|v\|_V} + C_r \\
&\leq \|\nabla\hat{u} - p_h\|_X + C_M\|\text{div } p_h + f(\hat{u})\|_X + \|D^{-1}\|_2|\varepsilon|_{l^2}.
\end{aligned} \tag{70}$$

One can use the bound (70) instead of (67). The smoothing element p_h is obtained by solving an additional linear equation (68), which takes extra computational costs. Meanwhile, for a certain *good* approximate solution, *e.g.* using P_2 (piecewise quadratic) elements, residual bound (70) becomes drastically small [**35**].

Remark 4.6. *One can consider another evaluation with $H(\mathrm{div},\Omega)$-smoothing elements* [**26**]. *A smoothing function $q \in H(\mathrm{div},\Omega)$ satisfying $q \approx \nabla\hat{u}$ and $\mathrm{div}\ q + f(\hat{u}) \approx 0$ yields*

$$\|\mathcal{F}(\hat{u})\|_{V^*} \leq \|\nabla\hat{u} - q\|_X + C_{e,2}\|\mathrm{div}\ q + f(\hat{u})\|_X.$$

One feature of this estimation is that it seeks the smoothing function in $q \in H(\mathrm{div},\Omega) \supset (H^1(\Omega))^2$, which can provide better approximation of $\nabla\hat{u}$, compared with the one in eq.(68).

4.2.3. Raviart-Thomas mixed finite element on triangle element. Inspired by Remark 4.6, we are concerned with a smoothing technique using mixed finite elements as below. Here, we would like to introduce Raviart-Thomas mixed finite element [**2, 5, 28**]. We follow discussions in [**2, 5**]. Let $H(\mathrm{div},\Omega)$ denote the space of vector functions such that

$$H(\mathrm{div},\Omega) := \left\{\psi \in (L^2(\Omega))^2 : \mathrm{div}\ \psi \in L^2(\Omega)\right\}.$$

Let K_h be a triangle element in triangulation of Ω. We define

$$P_k(K_h)\ :\ \text{the space of polynomials of degree less than } k \text{ on } K_h,$$

$$R_k(\partial K_h) := \{\varphi \in L^2(\partial K_h) : \varphi|_{e_i} \in P_k(e_i)\},\ \text{for any edge } e_i \text{ of } \partial K_h.$$

Functions of $R_k(\partial K_h)$ are polynomials of degree $\leq k$ on each side e_i of K_h $(i = 1, 2, 3)$. For $k \geq 0$, we define

$$RT_k(K_h) \ := \ \left\{ q \in (L^2(K_h))^2 \ : \ q = \begin{pmatrix} a_k \\ b_k \end{pmatrix} + c_k \cdot \begin{pmatrix} x \\ y \end{pmatrix},\ a_k, b_k, c_k \in P_k(K_h) \right\}.$$

The dimension of $RT_k(K_h)$ is $(k+1)(k+3)$. We now introduce basic result about $RT_k(K_h)$ spaces.

Proposition 4.7. *Let e_i be subtense of vertex i $(= 1, 2, 3)$ and $\vec{n}_{|e_i} = (n_1^{(i)}, n_2^{(i)})^T$ be an outward unit normal vector on boundary e_i. For $q \in RT_k(K_h)$, it follows*

$$\begin{cases} \text{div } q \in P_k(K_h), \\ q \cdot \vec{n}_{|e_i} \in R_k(\partial K_h). \end{cases}$$

Moreover, the divergence operator is surjective from $RT_k(K_h)$ onto $P_k(K_h)$, i.e.

$$\text{div}(RT_k(K_h)) = P_k(K_h).$$

Proof. $q \in RT_k(K_h)$ is written by $q = q_k + p_k(x, y)^T$ with $q_k \in (P_k(K_h))^2$ and $p_k \in P_k(K_h)$. It is clear that div q becomes a polynomial of degree k. On the other hand, let $\vec{n}|_{e_i} = (n_1^{(i)}, n_2^{(i)})^T$ be the normal side. We have

$$q \cdot \vec{n}|_{e_i} = q_k \cdot \vec{n}|_{e_i} + p_k(xn_1^{(i)} + yn_2^{(i)}).$$

Since $xn_1^{(i)} + yn_2^{(i)}$ becomes constant on each edge, $q \cdot \vec{n}|_{e_i}$ is a polynomial of degree k. □

Proposition 4.8. *For $k \geq 0$ and any $q \in RT_k(K_h)$, the following relations imply $q = 0$.*

$$\int_{\partial K_h} q \cdot \vec{n} \ \varphi_k ds = 0, \ \forall \varphi_k \in R_k(\partial K_h),$$

$$\int_{K_h} q \cdot q_{k-1} dx = 0, \ \forall q_{k-1} \in (P_{k-1}(K_h))^2.$$

Proof. Using Green's formula,

$$\int_{K_h} (\text{div } q) p_k dx = -\int_{K_h} q \cdot \nabla p_k dx + \int_{\partial K_h} q \cdot \vec{n} p_k ds = 0, \ \forall p_k \in P_k(K_h).$$

Since div $q \in P_k(K_h)$, we can choose $p_k =$ div q. Then the statement is obtained. □

The Raviart-Thomas finite element space RT_k is given by

$$RT_k \;:=\; \Big\{\, p_h \in (L^2(\Omega))^2 \;:\; p_h|_{K_h} = \begin{pmatrix} a_k \\ b_k \end{pmatrix} + c_k \cdot \begin{pmatrix} x \\ y \end{pmatrix},\; a_k, b_k, c_k \in P_k(K_h),$$
$$p_h \cdot \vec{n} \text{ is continuous on the inter-element boundaries.} \Big\} \tag{71}$$

It is a finite dimensional subspace of $H(\mathrm{div}, \Omega)$. Further let us define

$$M_h := \{ v \in L^2(\Omega) \;:\; v|_{K_h} \in P_k(K_h) \}. \tag{72}$$

It follows $\mathrm{div}(RT_k) = M_h$ (cf. Chapter IV.1 of [**5**]).

4.2.4. Proposal bounds with RT_k element. For the residual bound estimation, the smoothing technique in Subsection 4.2.2 works well to give accurate bounds. Some general smoothing techniques have been proposed in [**21, 26, 38**], etc, where smoothing functions $p_h \in (H^1(\Omega))^2$ or $H(\mathrm{div}, \Omega)$ are often used. One feature of proposal method is that we can use the basic property of Raviart-Thomas element, $\mathrm{div}(RT_k) = M_h$, for getting effective residual estimation. For given $f_h \in M_h$, this property enbables us to define a subspace of RT_k as

$$W_{f_h} = \{\; p_h \in RT_k : \mathrm{div}\; p_h + f_h = 0 \text{ for } f_h \in M_h \;\}.$$

Furthermore, we define $v_h \in M_h$ by an orthogonal projection of $v \in L^2(\Omega)$ such that

$$(v - v_h, w_h) = 0, \; \forall w_h \in M_h.$$

Assuming an error estimate

$$\|v - v_h\|_X \le C_{M_h} \|v\|_V \text{ for } v_h \in M_h$$

is obtained. Also we define $f_h(\hat{u}) \in M_h$ by the projection of $f(\hat{u}) \in L^2(\Omega)$. Finally, inequalities (63) and (64) give the following evaluation of the residual bound using

$p_h \in W_{f_h(\hat{u})}$,

$$
\begin{aligned}
\|\mathcal{F}(\hat{u})\|_{V^*} &= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u},\nabla v)-(f(\hat{u}),v)|}{\|v\|_V} \\
&= \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}-p_h,\nabla v)+(p_h,\nabla v)-(f(\hat{u}),v)|}{\|v\|_V} \\
&\leq \sup_{0\neq v\in V} \frac{|(\nabla\hat{u}-p_h,\nabla v)|}{\|v\|_V} + \sup_{0\neq v\in V} \frac{|(\mathrm{div}\ p_h+f(\hat{u}),v)|}{\|v\|_V} \\
&\leq \|\nabla\hat{u}-p_h\|_X + \sup_{0\neq v\in V} \frac{|(\mathrm{div}\ p_h+f_h(\hat{u})+f(\hat{u})-f_h(\hat{u}),v)|}{\|v\|_V} \\
&\leq \|\nabla\hat{u}-p_h\|_X + \sup_{0\neq v\in V} \frac{|(f(\hat{u})-f_h(\hat{u}),v-v_h)|}{\|v\|_V} \\
&\leq \|\nabla\hat{u}-\tilde{p}_h\|_X + C_{M_h}\ \|f(\hat{u})-f_h(\hat{u})\|_X \qquad (73)
\end{aligned}
$$

where $\tilde{p}_h$ is an interval function $p_h \in \tilde{p}_h$ obtained by verified computations.

REMARK 4.9. *Proposed estimation (73) holds for $k \geq 0$. If the approximate solution $\hat{u}$ is obtained from V_h, which has member function to be piecewise $(k+1)$-th polynomial. An effective choice of functional space W_{f_h} is to choose W_{f_h} is subspace of RT_k and M_h spanned by P_k elements. The rate of convergence can be expect to be $\|\nabla\hat{u}-p_h\|_X = o(h^{k+1})$ and $\|f(\hat{u})-f_h(\hat{u})\|_X = o(h^{k+1})$.*

4.2.5. How to determine p_h. In this part, we would like to explain the procedure of determining the smoothing element $p_h \in W_{f_h(\hat{u})}$ in Subsection 4.2.4. Using a verified computation of linear system, we will have an interval function $\tilde{p}_h$. This includes the smoothing element $p_h \in \tilde{p}_h$ with verification. The mixed method for the Poisson equation is applied to our procedure. First of all, we write the original problem (22) as the system

$$
\begin{cases} \nabla u = p, \\ -\mathrm{div}\ p = f(u). \end{cases}
$$

This system leads directly to the following saddle point problem: Find $(p, u) \in H(\text{div}, \Omega) \times X$ such that

$$\begin{cases} (p, q) + (u, \text{div } q) &= 0, \qquad\qquad \forall q \in H(\text{div}, \Omega), \\ (\text{div } p, v) &= -(f(u), v), \quad \forall v \in X. \end{cases} \tag{74}$$

Since the inf-sup condition of the general saddle point framework is obtained [**2**], this saddle point problem (74) is stable. Let M_h be defined in (72). As mentioned above, we determine $f_h(\hat{u}) \in M_h$ such that

$$(f(\hat{u}) - f_h(\hat{u}), v_h) = 0, \ \forall v_h \in M_h.$$

In order to obtain $p_h \in W_{f_h(\hat{u})}$ for given $f_h(\hat{u})$, we consider an approximation of the problem (74), we seek $(p_h, u_h) \in RT_k \times M_h$ defined in (71) and (72) satisfying

$$\begin{cases} (p_h, q_h) + (u_h, \text{div } q_h) &= 0, \qquad\qquad \forall q_h \in RT_k, \\ (\text{div } p_h, v_h) &= -(f(\hat{u}), v_h), \quad \forall v_h \in M_h. \end{cases} \tag{75}$$

Let spaces RT_k and M_h be

$$RT_k = \text{span}\{\psi_1, ..., \psi_l\}, \ M_h = \text{span}\{q_1, ..., q_m\},$$

respectively. Define matrices $P_{l\times l}$ and $N_{m\times l}$ as inner products of base functions

$$P_{l\times l} = [(\psi_i, \psi_j)]_{i,j=1,\ldots,l}, \ N_{m\times l} = [(q_i, \text{div } \psi_j)]_{i=1,\ldots,m, \ j=1,\ldots l}.$$

Additionally, suppose that $x \in \mathbb{R}^l$, $z \in \mathbb{R}^m$ and $f_v \in \mathbb{R}^m$ are vectors. Using these notations, let $p_h \in RT_k$, $u_h \in M_h$ be elements described as

$$\begin{aligned} x = (x_1, ..., x_l)^T \in \mathbb{R}^l, &\quad p_h = (\psi_1, ..., \psi_l) \cdot x \in RT_k, \\ z = (z_1, ..., z_m)^T \in \mathbb{R}^m, &\quad u_h = (q_1, ..., q_l) \cdot z \in M_h, \\ f_v = [(f(\hat{u}), q_i)]_{i=1,\ldots,m}. & \end{aligned}$$

By using matrix notations, problem (75) is finally characterized by

$$\begin{cases} Px + N^T z &= 0, \\ Nx &= -f_v. \end{cases}$$

In order to obtain $p_h \in W_{f_h(\hat{u})}$, we need to obtain the vector $x \in \mathbb{R}^l$ with verified computations. Here, we will use a basic algorithm to solve the linear system:

$$\text{Find } x_z \in \mathbb{R}^{l+m} \text{ s.t. } \tilde{A}x_z = \tilde{f},\ \tilde{A} = \begin{pmatrix} P & N^T \\ N & 0 \end{pmatrix},\ x_z = \begin{pmatrix} x \\ z \end{pmatrix},\ \tilde{f} = \begin{pmatrix} 0 \\ -f_v \end{pmatrix}.$$

The following theorem yields verified result of this linear system:

THEOREM 4.10 (Yamamoto [**39**]). *Let R be some approximate inverse of the matrix $\tilde{A}$ together with an approximate solution $\hat{x}_z$. For $G = R\tilde{A} - I$, let $\kappa \in \mathbb{R}^{l+m}$ obtain*

$$\kappa_i \geq \sum_{j=1}^{l+m} |\tilde{A}_{i,j}|,\ \textit{for } i = 1, ..., l+m.$$

If $\|\kappa\|_\infty \leq 1$ is satisfied, then $\tilde{A}^{-1}$ exists and

$$\delta = |x_z - \hat{x}_z| \leq |R(A\hat{x}_z - \tilde{f})| + \frac{\|R(A\hat{x}_z - \tilde{f})\|_\infty}{1 - \|R\tilde{A} - I\|_\infty}\kappa$$

holds for $\delta \in \mathbb{R}^{l+m}$.

The verified result is an interval vector $\tilde{x}_z := <\hat{x}_z, \delta>$, which is centered at $\hat{x}_z$ with radius δ. It includes the solution x_z. Using the verified result $\tilde{x}_z$, we can express the interval function $\tilde{p}_h$ in (73).

$$\tilde{p}_h = (\psi_1, ..., \psi_l)^T \cdot \tilde{x}_z(1 : l),$$

where $\tilde{x}_z(1 : l)$ denotes first l-elements of the interval vector $\tilde{x}_z$.

4.3. Lipschitz constant

Finally, we estimate the Lipschitz constant of $\mathcal{F}'[u] : V \to V^*$. Here, we assume that $f' : V \to \mathcal{L}(V, X)$ is Lipschitz continuous on the open ball $D \supset \overline{B}(\hat{u}, 2\alpha)$. Namely, there exists a positive constant C_L satisfying

$$|((f'(v) - f'(w))u, \psi)| \le C_L \|v - w\|_V \|u\|_V \|\psi\|_V \tag{76}$$

for $v, w \in D$ and $u, \psi \in V$. Usually, the optimal estimation depends on the definition of f. We will discuss the estimation of C_L in Subsection 5.2 for a model case. For $v, w \in D$, we have

$$\begin{aligned}
\|\mathcal{F}'[v] - \mathcal{F}'[w]\|_{V,V^*} &= \sup_{0 \ne u \in V} \sup_{0 \ne \psi \in V} \frac{|\langle (\mathcal{N}'[v] - \mathcal{N}'[w])u, \psi \rangle|}{\|u\|_V \|\psi\|_V} \\
&= \sup_{0 \ne u \in V} \sup_{0 \ne \psi \in V} \frac{|((f'(v) - f'(w))u, \psi)|}{\|u\|_V \|\psi\|_V} \\
&\le C_L \|v - w\|_V.
\end{aligned}$$

Therefore, one can put $C_3 := C_L$.

Chapter 5

Results

Summarizing this thesis, we would like to show computational results. In the following, we first present numerical examples for two-point boundary value problems. After that we present elliptic boundary problems for a model problem on arbitrary polygonal domains.

5.1. Computational results (ODE)

In this section, we shall present four numerical examples corresponding to two-point boundary problems. In these examples, we use piecewise quadratic (P_2) finite elements to get approximate solutions. Following results demonstrate the usefulness of the method using (70). Namely, a remarkable improvement in accuracy is attained. All computations are carried out on Mac OS X, 2.26GHz Quad-Core Intel Xeon (Nehalem) with 32 GB RAM by using MATLAB 2010a with a toolbox for verified computations, INTLAB [**29**].

5.1.1. Emden equation on a interval. Let $\Omega = (0, 1)$. We consider the following quadratic nonlinear two-point boundary value problem.

$$\begin{cases} -u'' = u^2, & 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{77}$$

Obviously, the Fréchet derivative of $f(u) = u^2$ is given by $f'(u) = 2u$. An approximate solution $\hat{u}$ is computed by the finite element method with one-dimensional piecewise quadratic elements. The calculated approximate solution $\hat{u}$ is bounded on Ω so that $\hat{u} \in L^\infty(\Omega)$ in this case. Therefore, for $\hat{u}, v, w \in V$ it follows

$$\|f'(\hat{u})\|_{V,X} \leq 2C_{e,2}\|\hat{u}\|_\infty,$$

and

$$\|f'(v) - f'(w)\|_{V,X} \leq 2C_{e,4}^2\|v - w\|_V.$$

The proposed method can be applied to approximate solution $\hat{u}$. Results with piecewise quadratic elements are shown in Table 5.1. For the mesh size $h = 1/32$, our verification method yields

$$C_1 \leq 6.288,\ C_{2,h} \leq 0.016,\ C_3 \leq 0.368.$$

Then we have $C_1^2 C_{2,h} C_3 \leq 0.225$. Consequently, it follows that there exists an unique solution in the ball $\overline{B}(\hat{u}, \rho)$ with the radius

$$\|u - \hat{u}\|_V \leq \rho = 0.112.$$

Since $V \hookrightarrow C^0(\Omega)$ in Theorem 2.1, we can obtain a verified error bound in maximum norm by Poincaré's inequality. Note that for $u \in V$,

$$|u(x)| = \left| \int_0^x u' dt \right| \leq \int_0^x |u'| dt \text{ and } \left| \int_x^1 u' dt \right| \leq \int_x^1 |u'| dt,$$

imply

$$2|u(x)| \leq \int_0^1 |u'| dt \leq \|u'\|_X = \|u\|_V.$$

Therefore, $\|u\|_\infty \leq \frac{1}{2}\|u\|_V$ holds. The verification procedure proves the existence and uniqueness of the exact solution of (77) between two curves in Figure 5.1.

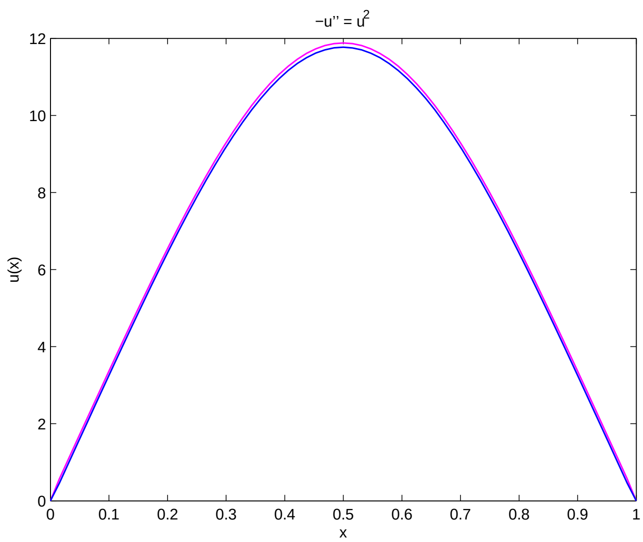

Figure 5.1: Verified Bounds for (77), mesh size 1/32.

Results with piecewise linear elements are also shown in Table 5.2. Comparing Table 5.1 with Table 5.2, we can see the improvement of smoothing technique which is discussed in Subsection 4.2.2. The value of $C_{2,h}$ in Table 5.2 is calculated by (67) with piecewise linear elements. On the other hand, for the calculation of $C_{2,h}$ in Table 5.1, we used (70) with piecewise quadratic elements. For the same mesh size $h = 1/(n+1)$, node points on Ω are $n \times n$ in piecewise linear elements and $(2n + 1) \times (2n + 1)$ in piecewise quadratic elements respectively. Furthermore, the sparse structure of the matrices becomes tridiagonal matrix for piecewise linear elements. Meanwhile, the sparse structure of piecewise quadratic elements becomes penta-diagonal matrix. Comparing with the method in [**36**], the smoothing technique in (70) works very well.

Table 5.1: Results using (70) with P_2 element

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ
5	4.974×10^{-3}	6.288	1.544×10^{-2}	3.676×10^{-1}	2.244×10^{-1}	1.115×10^{-1}
6	2.487×10^{-3}	5.626	4.373×10^{-3}	3.676×10^{-1}	5.087×10^{-2}	2.527×10^{-2}
7	1.244×10^{-3}	5.345	1.365×10^{-3}	3.676×10^{-1}	1.433×10^{-2}	7.345×10^{-3}
8	6.217×10^{-4}	5.215	5.561×10^{-4}	3.676×10^{-1}	5.558×10^{-3}	2.908×10^{-3}
9	3.109×10^{-4}	5.171	5.191×10^{-4}	3.676×10^{-1}	5.098×10^{-3}	2.691×10^{-3}

Table 5.2: Results using (67) with P_1 element

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ
5	9.948×10^{-3}	9.137	9.571×10^{-1}	3.676×10^{-1}	29.37	Failed
6	4.974×10^{-3}	6.808	4.776×10^{-1}	3.676×10^{-1}	8.134	Failed
7	2.487×10^{-3}	6.039	2.387×10^{-1}	3.676×10^{-1}	3.201	Failed
8	1.244×10^{-3}	5.717	1.194×10^{-1}	3.676×10^{-1}	1.434	Failed
9	6.217×10^{-4}	5.568	5.966×10^{-2}	3.676×10^{-1}	6.798×10^{-1}	Failed
10	3.11×10^{-4}	5.497	2.983×10^{-2}	3.676×10^{-1}	3.313×10^{-1}	2.075×10^{-1}

We also present a comparison of verified numerical error estimations obtained by using (67) and using (70). Using (67), the guaranteed error bound ρ has been only obtained when the mesh size is $1/2^{10}$. On the other hand, (70) yields ρ from $1/2^5$. As a result, the verified error can be obtained from rough mesh partition.

5.1.2. Solution branches. Next, we consider the following quadratic nonlinear two-point boundary value problem with one-parameter family σ for $\sigma \geq 0$:

$$\begin{cases} -u'' = u^2 \ + \ \sigma, & 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{78}$$

For different values of $\sigma \geq 0$, we have verified the existence of exact solutions. Figure 5.2 shows the result of our computer-assisted proof. In this figure, asterisks are plotted when an exact solution is found. This figure implies that there are two solution branches, upper branch and lower branch, for the case of $\sigma < 23$. For the verification of the existence of solutions on the upper branch, we need fine meshes with around $h = 1/2^{10}$ if we use residual evaluation in (67). Especially, at $\sigma = 0$, there is a solution around $\max(|u|) = 11.9$. We need a fine mesh ($h = 1/2^{10}$) to get a verified result as mentioned in previous section. At $\sigma = 0$, the solution on the lower branch becomes a trivial solution $u = 0$. The refined method using (70) needs only a mesh size $h = 1/2^5$ for verifying all solutions exhibited in Figure 5.2.

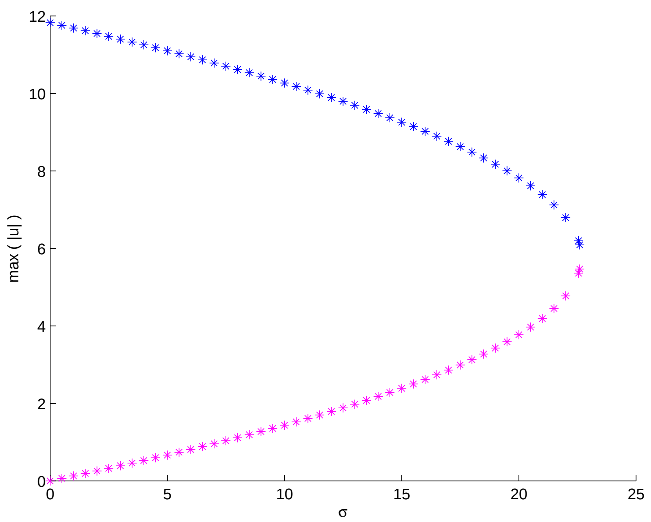

Figure 5.2: Verified exact solutions to the problem (78)

5.1.3. Three solutions. Let us further consider a nonlinear two-point boundary value problem of the following form:

$$\begin{cases} -u'' = u^3 + 3.0, & 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{79}$$

By numerical computations, it is easy to see that there exist at least three approximate solutions $\hat{u}_{-1}, \hat{u}_0$ and $\hat{u}_1$ for the problem (79). Here, one of them, $\hat{u}_0$, is the solution close to zero. Thus, the existence of exact solution around $\hat{u}_0$ is easy to validate. The existence of exact solutions around rest two approximate solutions is relatively difficult to verify, because $\max(|u_i|), (i = -1, 1)$ are relatively large. Figure 5.3 shows function shapes of approximate solutions $\hat{u}_{-1}$ and $\hat{u}_1$ together with $\hat{u}_0$.

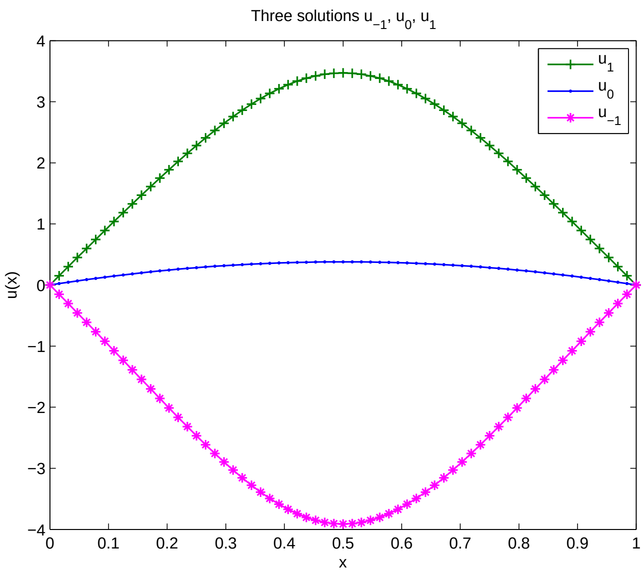

Figure 5.3: Three approximate solutions of (79)

Using our method, it is proved that there exist three local unique exact solutions u_{-1}, u_0 and u_1 in the neighborhood of approximate solutions $\hat{u}_{-1}, \hat{u}_0$ and $\hat{u}_1$, respectively. Table 5.3 exhibits verification results expressing the upper bound of distances between three exact solutions and their approximate solutions, $\|u_{-1} - \hat{u}_{-1}\|_V \le \rho_{-1}$, $\|u_0 - \hat{u}_0\|_V \le \rho_0$ and $\|u_1 - \hat{u}_1\|_V \le \rho_1$.

Table 5.3: Verified results for (79)

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_0
5	4.974×10^{-3}	1.047	6.798×10^{-4}	8.376×10^{-1}	6.231×10^{-4}	7.114×10^{-4}
6	2.487×10^{-3}	1.046	2.404×10^{-4}	8.367×10^{-1}	2.201×10^{-4}	2.514×10^{-4}
7	1.244×10^{-3}	1.046	8.516×10^{-5}	8.364×10^{-1}	7.787×10^{-5}	8.905×10^{-5}
8	6.217×10^{-4}	1.046	3.189×10^{-5}	8.363×10^{-1}	2.914×10^{-5}	3.334×10^{-5}
9	3.109×10^{-4}	1.046	2.069×10^{-5}	8.363×10^{-1}	1.891×10^{-5}	2.163×10^{-5}
10	1.555×10^{-4}	1.046	3.082×10^{-5}	8.363×10^{-1}	2.817×10^{-5}	3.222×10^{-5}

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_1
5	4.974×10^{-3}	13.68	4.973×10^{-3}	7.277	6.764	Failed
6	2.487×10^{-3}	9.827	1.248×10^{-3}	7.171	8.641×10^{-1}	Failed
7	1.244×10^{-3}	8.615	3.342×10^{-4}	7.152	1.774×10^{-1}	3.194×10^{-3}
8	6.217×10^{-4}	8.115	1.352×10^{-4}	7.149	6.359×10^{-2}	1.134×10^{-3}
9	3.109×10^{-4}	7.886	1.466×10^{-4}	7.149	6.514×10^{-2}	1.196×10^{-3}
10	1.555×10^{-4}	7.776	2.649×10^{-4}	7.151	1.145×10^{-1}	2.193×10^{-3}

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_{-1}
5	4.974×10^{-3}	60.41	7.213×10^{-3}	8.791	231.4	Failed
6	2.487×10^{-3}	18.91	1.777×10^{-3}	8.022	5.091	Failed
7	1.244×10^{-3}	14.07	4.644×10^{-4}	7.971	7.322×10^{-1}	Failed
8	6.217×10^{-4}	12.48	1.723×10^{-4}	7.962	2.134×10^{-1}	2.446×10^{-3}
9	3.109×10^{-4}	11.81	1.651×10^{-4}	7.961	1.831×10^{-1}	2.168×10^{-3}
10	1.555×10^{-4}	11.51	2.928×10^{-4}	7.962	3.081×10^{-1}	4.156×10^{-3}

5.1.4. Another multiple solutions. Finally, we treat the case of $a(x) \neq 1$. Let us be concerned with the following problem of the form

$$\begin{cases} -(au')' = f(u), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{80}$$

where $a(x) = 1 + x^2 \geq 1 > 0$, $f(u) = u^2 + g$ and $g \in X$. For each constants, the following are obtained.

$$a_0 = 1,\ c_a = 1,\ C_a = \sqrt{2},\ \|a'\|_\infty = 2.$$

We set $g = -(\sin^2 2\pi x + (2\pi(1+x^2)\cos 2\pi x)')$ so that the exact solution becomes $u = \sin 2\pi x$. Then two approximate solutions are given. u_0 resembles the exact solution $u = \sin(2\pi x)$. In addition, another solution u_1 is found. The maximum value of u_1 is much higher than that of u_0. So that the solution u_1 is comparatively difficult to verify the existence. Our computer-assisted proof method proves that there are multiple solutions of (80). In Figure 5.4, it shows the verified inclusion of two exact solutions.

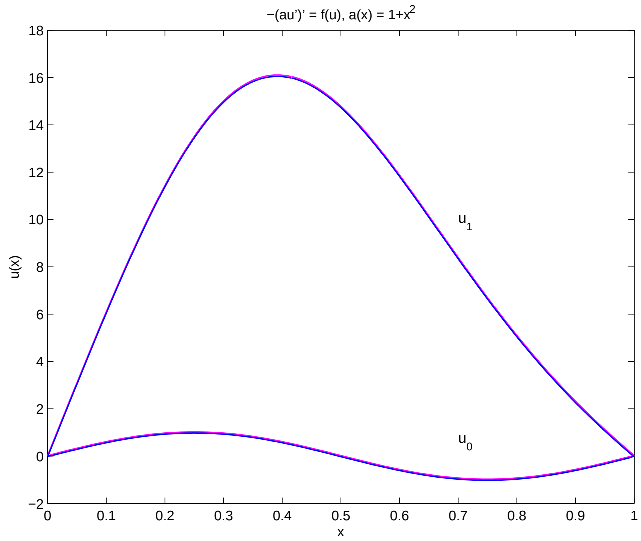

Figure 5.4: Verified inclusions for solutions to (80)

Furthermore, Table 5.4 shows verified numerical error bounds. Upper bounds of $\|u_1 - \hat{u}_1\|_V \leq \rho_1$ and $\|u_0 - \hat{u}_0\|_V \leq \rho_0$ are presented in Table 5.4.

Table 5.4: Verified results for (80)

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ_0
5	1.628×10^{-2}	1.331	1.627×10^{-1}	3.676×10^{-1}	1.053×10^{-1}	2.282×10^{-1}
6	8.141×10^{-3}	1.321	3.571×10^{-2}	3.676×10^{-1}	2.288×10^{-2}	4.769×10^{-2}
7	4.071×10^{-3}	1.316	6.203×10^{-3}	3.676×10^{-1}	3.947×10^{-3}	8.177×10^{-3}
8	2.035×10^{-3}	1.314	1.733×10^{-3}	3.676×10^{-1}	1.099×10^{-3}	2.277×10^{-3}
9	1.018×10^{-3}	1.314	1.009×10^{-3}	3.676×10^{-1}	6.401×10^{-4}	1.326×10^{-3}
10	5.088×10^{-4}	1.307	6.235×10^{-4}	3.676×10^{-1}	3.914×10^{-4}	8.151×10^{-4}

$1/2^\gamma$	C_M	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ_1
6	8.141×10^{-3}	37.02	8.029×10^{-2}	3.676×10^{-1}	40.43	Failed
7	4.071×10^{-3}	14.56	1.982×10^{-2}	3.676×10^{-1}	1.544	Failed
8	2.035×10^{-3}	11.18	5.241×10^{-3}	3.676×10^{-1}	2.405×10^{-1}	6.806×10^{-2}
9	1.018×10^{-3}	10.11	3.194×10^{-3}	3.676×10^{-1}	1.198×10^{-1}	3.446×10^{-2}
10	5.088×10^{-4}	9.041	5.225×10^{-3}	3.676×10^{-1}	1.571×10^{-1}	5.167×10^{-2}

5.2. Formulation for verification example (PDE)

In this section, we are concerned with practical formulation of a certain example. Let us consider following Drichlet boundary value problems such that

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

with

$$f(u) = b \cdot \nabla u + c_1 u + c_2 u^2 + c_3 u^3 + g.$$

Here, $b(x) \in (L^\infty(\Omega))^2$, $c_i \in L^\infty(\Omega)$, $(i = 1, 2, 3)$ and $g \in X$.

In order to show the applicability of our verification theory to this problem, we must check that f is Fréchet differentiable at $\hat{u} \in V_h$ as a map $f : V \to X$. This can be shown as follows. The candidate of $f'(\hat{u}) : V \to X$ is obviously

$$f'(\hat{u}) = b \cdot \nabla + c_1 + 2c_2\hat{u} + 3c_3\hat{u}^2.$$

Recall $V = H_0^1(\Omega)$ and Sobolev's embedding theorem states $V \subset L^p(\Omega)$ for $p \geq 1$ with

$$\|v\|_{L^p} \leq C_{e,p}\|v\|_V, \ \forall v \in V.$$

Similarly, it is easily seen that for $u, v, w \in V$ from Hölder's inequality,

$$\|uvw\|_{L^2} \leq \|u\|_{L^6}\|v\|_{L^6}\|w\|_{L^6} \leq C_{e,6}^3\|u\|_V\|v\|_V\|w\|_V.$$

Then, we have for $\nu \in V$

$$\begin{aligned}\|f(\hat{u}+\nu) - f(\hat{u}) - f'(\hat{u})\nu\|_X &= \|(c_2 + 3c_3\hat{u})\nu^2 + c_3\nu^3\|_X \\ &\leq C_{e,6}^3\left(\|c_2\|_{L^\infty} + \|c_3\|_{L^\infty}(3\|\hat{u}\|_V + \|\nu\|_V)\right)\|\nu\|_V^2.\end{aligned}$$

This shows the Fréchet differentiability of $f : V \to X$ at $\hat{u} \in V$.

For the inverse operator norm estimation, we have following constants. We can assume that the approximate solution $\hat{u} \in V_h$ is essentially bounded by computing result. So that $\hat{u} \in L^\infty(\Omega) \cap V$ is obtained.

$$\begin{aligned}\|f'(\hat{u})\|_{V,X} &= \sup_{0\neq v\in V}\frac{\|f'(\hat{u})v\|_X}{\|v\|_V} \\ &= \sup_{0\neq v\in V}\frac{\|b\cdot\nabla v + c_1 v + 2c_2\hat{u}v + 3c_3\hat{u}^2 v\|_X}{\|v\|_V} \\ &\leq \||b(x)|_E\|_{L^\infty} + C_{e,2}\left(\|c_1\|_{L^\infty} + 2\|c_2\|_{L^\infty}\|\hat{u}\|_{L^\infty} + 3\|c_3\|_{L^\infty}\|\hat{u}\|_{L^\infty}^2\right) \\ &=: K_1,\end{aligned}$$

where $b = (b_1, b_2)^T$ and $|b(x)|_E = (b_1(x)^2 + b_2(x)^2)^{\frac{1}{2}}$. Furthermore, we have

$$\begin{aligned}\|f'(\hat{u})\|_{V_c,X} &= \sup_{0\neq v_c\in V_c}\frac{\|f'(\hat{u})v_c\|_X}{\|v_c\|_V} \\ &= \sup_{0\neq v_c\in V_c}\frac{\|b\cdot\nabla v_c + c_1 v_c + 2c_2\hat{u}v_c + 3c_3\hat{u}^2 v_c\|_X}{\|v_c\|_V} \\ &\leq \||b(x)|_E\|_{L^\infty} + C_M\left(\|c_1\|_{L^\infty} + 2\|c_2\|_{L^\infty}\|\hat{u}\|_{L^\infty} + 3\|c_3\|_{L^\infty}\|\hat{u}\|_{L^\infty}^2\right) =: K_2.\end{aligned}$$

Since $v_c = v - \mathcal{P}_h v$ in Theorem 4.3, C_M is the quantity defined in (38), which is discussed in Subsection 3.2.1 and 3.2.2. Thus, one can get the explicit value of K_1 and K_2 by verified computations.

Let us describe Lipschitz continuity of $\mathcal{F}'[\hat{u}] : V \to V^*$ by checking inequality (76). $\overline{B}(\hat{u}, 2\alpha)$ is assumed to be an open ball centered at $\hat{u} \in V_h$ with radius 2α. For $v, w \in D \supset \overline{B}(\hat{u}, 2\alpha)$ and $u, \psi \in V$, we have

$$\begin{aligned} |((f'(v) - f'(w))u, \psi)| &= |(2c_2(v-w)u, \psi) + (3c_3(v+w)(v-w)u, \psi)| \\ &\leq \left(2C_{e,3}^3 \|c_2\|_{L^\infty} + 3C_{e,4}^4 \|c_3\|_{L^\infty} \|v+w\|_V\right) \|v-w\|_V \|u\|_V \|\psi\|_V. \end{aligned}$$

Since $v, w \in D$, it follows that

$$\|v+w\|_V < \ 2\|\hat{u}\|_V + 4 \ \mathrm{succ}(C_1 C_{2,h}),$$

where $\mathrm{succ}(C_1 C_{2,h})$ denotes the successor of $C_1 C_{2,h}$ in floating point, *e.g* we can take $\mathrm{succ}(C_1 C_{2,h}) = \frac{C_1 C_{2,h}}{1-2^{-53}}$ for double precision. Thus, it follows

$$|((f'(v) - f'(w))u, \psi)| < C_L \|v-w\|_V \|u\|_V \|\psi\|_V, \text{ for } v, w \in D$$

with

$$C_L := 2 \ C_{e,3}^3 \|c_2\|_{L^\infty} + 6 \ C_{e,4}^4 \|c_3\|_{L^\infty} \left(\|\hat{u}\|_V + 2 \ \mathrm{succ}(C_1 C_{2,h})\right).$$

5.3. Square domain

Now, we will present numerical results on square domain. All computations are carried out on Mac OS X, 2.26GHz Quad-Core Intel Xeon (Nehalem) with 32 GB RAM by using MATLAB 2011a with a toolbox for verified computations, INTLAB [**29**].

5.3.1. Example 1. We consider the following semilinear Dirichlet boundary value problem on square domain $\Omega = (0,1) \times (0,1)$:

$$\begin{cases} -\Delta u = u^2 + 10, & \text{in } \Omega, \\ u = 0, & \partial\Omega. \end{cases} \tag{81}$$

An approximate solution $\hat{u}$ is calculated by piecewise linear finite elements on uniform mesh triangulation (size $\frac{1}{16}$). The verification procedure in Chapter 4 is applied for

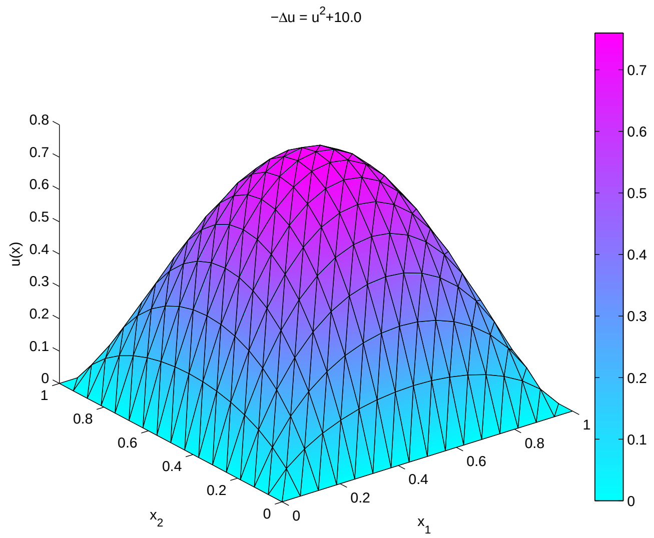

Figure 5.5: Approximation $\hat{u}$ in (81), mesh size: $\frac{1}{16}$.

(81). Our computer-assisted proof method gives the following bounds:

$$C_1 \leq 1.095, \; C_{2,h} \leq 0.318, \; C_3 \leq 0.717,$$

where we use (67) evaluation to bound $C_{2,h}$. Thus, we have

$$C_1^2 C_{2,h} C_3 \leq 0.028.$$

It turns out that there exists a solution in the closed ball $\overline{B}(\hat{u}, \rho)$ with

$$\|u - \hat{u}\|_V \leq \rho = 0.353.$$

By increasing the number of grid points, guaranteed error bounds are improved by the ratio $O(h)$. The guaranteed error bound is presented in Table 5.5.

Table 5.5: Verification results for (81).

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ
4	3.082×10^{-2}	2.251×10^{-1}	1.095	3.181×10^{-1}	7.165×10^{-2}	2.729×10^{-2}	3.529×10^{-1}
5	1.541×10^{-2}	2.251×10^{-1}	1.089	1.591×10^{-1}	7.165×10^{-2}	1.351×10^{-2}	1.743×10^{-1}
6	7.704×10^{-3}	2.251×10^{-1}	1.086	7.953×10^{-2}	7.165×10^{-2}	6.713×10^{-3}	8.662×10^{-2}
7	3.852×10^{-3}	2.251×10^{-1}	1.084	3.977×10^{-2}	7.165×10^{-2}	3.348×10^{-3}	4.318×10^{-2}
8	1.926×10^{-3}	2.251×10^{-1}	1.084	1.989×10^{-2}	7.165×10^{-2}	1.672×10^{-3}	2.156×10^{-2}

Next, we show that there exists another solution of (81). Figure 5.6 shows the shape of such a solution. The maximum value of this solution becomes 30 times

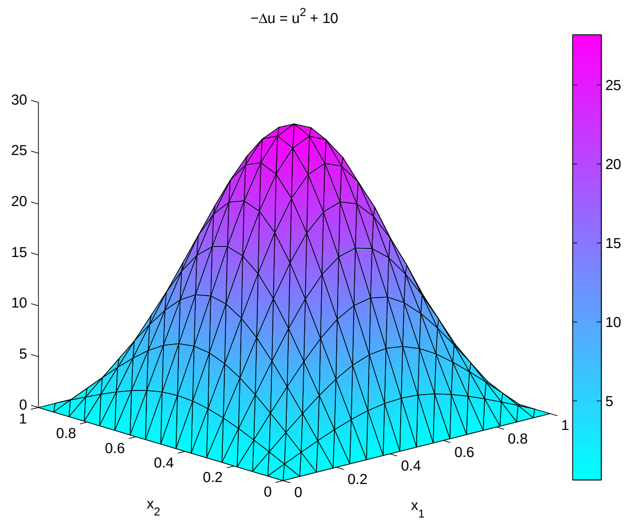

Figure 5.6: Another approximate solution of (81), mesh size: $\frac{1}{16}$.

larger than that of the approximate solution shown in Figure 5.5. Since the residual evaluation (67) includes the term $C_M\|f(\hat{u})\|_{L^2} = C_M\|\hat{u}^2 + 10\|_{L^2}$, $C_{2,h}$ becomes quite large. Thus, in order to satisfy the condition of Newton-Kantorovich theorem: $C_1^2 C_{2,h} C_3 \leq 1/2$, the mesh size h should be taken sufficiently small such that $C_M\|f(\hat{u})\|_{L^2} \ll 1$ holds. This causes a problem of increasing computational costs. In order to overcome such difficulty, we use the smoothing technique with piecewise quadratic finite elements introduced in Subsection 4.2.2. For a *good* approximate solution, $C_{2,h}$ becomes quite small. Then, the condition $C_1^2 C_{2,h} C_3 \leq 1/2$ is easier to be fulfilled. Table 5.6 presents verified results for (81) shown in Figure 5.6 by using smoothing techniques. Here, we use piecewise quadratic finite elements on an uniform triangular mesh.

Table 5.6: Verification results for another solution of (81).

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ
5	8.126×10^{-3}	2.251×10^{-1}	3.257	1.492×10^{-1}	7.165×10^{-2}	1.134×10^{-1}	5.171×10^{-1}
6	4.063×10^{-3}	2.251×10^{-1}	3.008	3.759×10^{-2}	7.165×10^{-2}	2.435×10^{-2}	1.145×10^{-1}
7	2.032×10^{-3}	2.251×10^{-1}	2.904	1.009×10^{-2}	7.165×10^{-2}	6.093×10^{-3}	2.938×10^{-2}
8	1.016×10^{-3}	2.251×10^{-1}	2.887	6.874×10^{-3}	7.165×10^{-2}	4.104×10^{-3}	1.989×10^{-2}

5.3.2. Example 2. Let us consider another semilinear Dirichlet boundary value problem on $\Omega = (0,1)\times(0,1)$:

$$\begin{cases} -\Delta u = u^3 + 5, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{82}$$

An approximate solution $\hat{u}$ is calculated by using piecewise quadratic elements on the uniform mesh. There are three approximations $\hat{u}_{-1}$, $\hat{u}_0$ and $\hat{u}_1$. Shapes of them are shown in Figure 5.7. For the approximation $\hat{u}_0$ with mesh size 1/32, our computer-assisted proof method yields the following bounds:

$$C_1^2 C_{2,h} C_3 \leq 3.742 \times 10^{-3}.$$

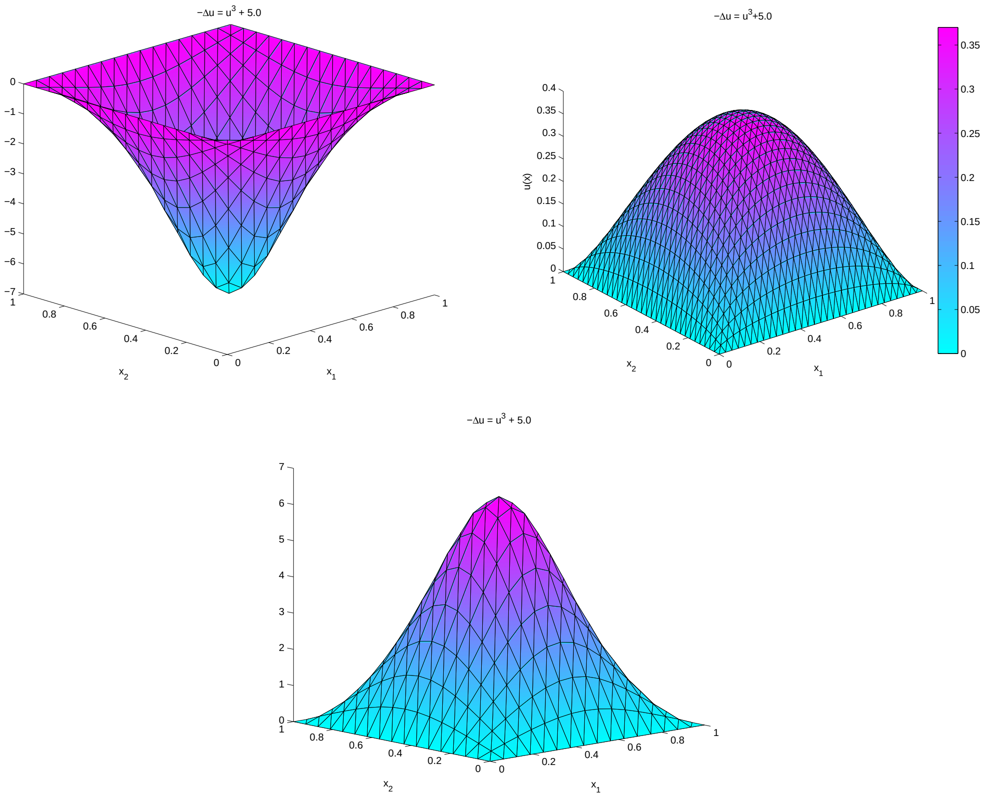

Figure 5.7: Approximate solutions $\hat{u}_{-1}$, $\hat{u}_0$ and $\hat{u}_1$

It is obtained that there exists an exact solution in the closed ball $\overline{B}(\hat{u}_0, \rho_0)$ with

$$\|u_0 - \hat{u}_0\|_V \leq \rho_0 = 4.551 \times 10^{-3}.$$

Guaranteed error bounds are improved by decreasing the mesh size h, presented in Table 5.7. Upper bounds of guaranteed error are $\|u_{-1} - \hat{u}_{-1}\|_V \leq \rho_{-1}$, $\|u_0 - \hat{u}_0\|_V \leq \rho_0$ and $\|u_1 - \hat{u}_1\|_V \leq \rho_1$.

Table 5.7: Verification results for (82).

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_{-1}
5	8.126×10^{-3}	2.251×10^{-1}	11.91	5.678×10^{-2}	8.538	68.71	Failed
6	4.063×10^{-3}	2.251×10^{-1}	7.051	1.443×10^{-2}	7.839	5.621	Failed
7	2.032×10^{-3}	2.251×10^{-1}	6.083	3.715×10^{-3}	7.743	1.065	Failed
8	1.016×10^{-3}	2.251×10^{-1}	5.796	1.705×10^{-3}	7.728	4.423×10^{-1}	1.475×10^{-2}

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_0
5	8.126×10^{-3}	2.251×10^{-1}	1.426	3.186×10^{-3}	5.777×10^{-1}	3.742×10^{-3}	4.551×10^{-3}
6	4.063×10^{-3}	2.251×10^{-1}	1.426	8.562×10^{-4}	5.736×10^{-1}	9.978×10^{-4}	1.221×10^{-3}
7	2.032×10^{-3}	2.251×10^{-1}	1.426	2.344×10^{-4}	5.725×10^{-1}	2.725×10^{-4}	3.341×10^{-4}
8	1.016×10^{-3}	2.251×10^{-1}	1.426	1.191×10^{-4}	5.723×10^{-1}	1.383×10^{-4}	1.696×10^{-4}

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2C_{2,h}C_3$	ρ_1
5	8.126×10^{-3}	2.251×10^{-1}	6.127	4.278×10^{-2}	7.537	12.11	Failed
6	4.063×10^{-3}	2.251×10^{-1}	4.697	1.085×10^{-2}	7.281	1.742	Failed
7	2.032×10^{-3}	2.251×10^{-1}	4.274	2.831×10^{-3}	7.233	3.739×10^{-1}	1.611×10^{-2}
8	1.016×10^{-3}	2.251×10^{-1}	4.201	1.518×10^{-3}	7.226	1.934×10^{-1}	7.146×10^{-3}

5.3.3. Example 3. For $\Omega = (0,1)\times(0,1)$, let us consider another example of the form:

$$\begin{cases} -\nabla\cdot(a\nabla u) + b\cdot\nabla u = f_1(u), & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{83}$$

where $a(x) = 1 + (1-x_1)^2$, $b(x) = (-1,-1)^T$ and $f_1(u) = u^2 + 10$. In this case, we have

$$a_0 = 1,\ c_a = 1,\ C_a = \sqrt{2},\ \||\nabla a|_E\|_{L^\infty} = 2,\ \||b(x)|_E\|_{L^\infty} = \sqrt{2}.$$

Figure 5.8 shows an approximate solution $\hat{u}$ using linear finite elements on the uniform mesh (size 1/64). Our verification method yields

$$C_1 \le 1.728,\ C_{2,h} \le 0.267,\ C_3 \le 0.144.$$

Thus, we have

$$C_1^2C_{2,h}C_3 \le 0.114.$$

It is demonstrated that there exists an exact solution in the closed ball $\overline{B}(\hat{u},\rho)$ with

$$\|u-\hat{u}\|_V \le \rho = 0.491.$$

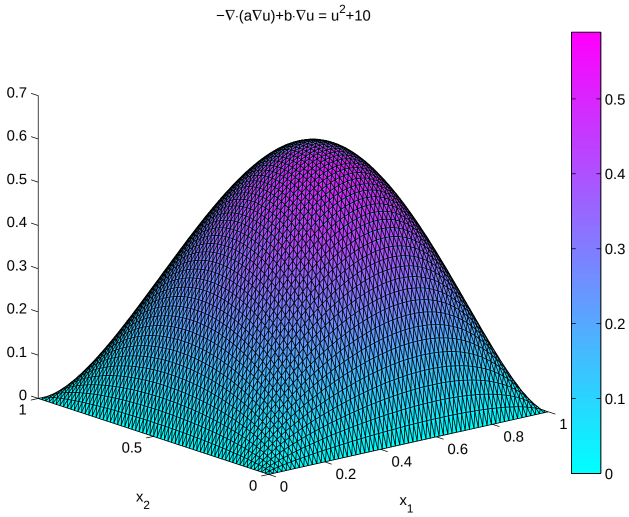

Figure 5.8: Approximate solution $\hat{u}$ of (83).

Convergence ratio is $O(h)$, which can be seen form Table 5.8.

Table 5.8: Verification results for (83).

$1/2^{\gamma}$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ
4	8.937×10^{-2}	2.251×10^{-1}	2.121	1.06470	1.433×10^{-1}	6.863×10^{-1}	Failed
5	4.469×10^{-2}	2.251×10^{-1}	1.842	5.329×10^{-1}	1.433×10^{-1}	2.589×10^{-1}	1.159
6	2.235×10^{-2}	2.251×10^{-1}	1.728	2.665×10^{-1}	1.433×10^{-1}	1.141×10^{-1}	4.901×10^{-1}
7	1.118×10^{-2}	2.251×10^{-1}	1.676	1.333×10^{-1}	1.433×10^{-1}	5.359×10^{-2}	2.296×10^{-1}
8	5.586×10^{-3}	2.251×10^{-1}	2.715	6.663×10^{-2}	1.433×10^{-1}	7.033×10^{-2}	1.877×10^{-1}

5.4. Several convex domains

In this section, we will present numerical results on several convex domains. All computations are carried out on Cent OS (Linux), Quad-Core AMD Opteron(tm) Processor 8376, 2.30 GHz with 512GB RAM by using MATLAB 2010a with a toolbox for verified computations, INTLAB [**29**]. We use Gmsh [**6**] (`http://geuz.org/gmsh/`) to obtain triangular mesh. Let us treat the following model problem.

$$\begin{cases} -\Delta u = u^2 + 10, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{84}$$

Here, Ω is assumed to be convex polygonal domains.

5.4.1. Rectangular domain. Let us draw V_h as the piecewise linear finite element subspace (P_1 elements). Firstly, we consider (84) on $\Omega = (0,2) \times (0,1)$: rectangular domain. An approximate solution $\hat{u} \in V_h$ of (84) is appeared on Figure 5.9 with the mesh size 1/16. The verification result is shown in Table 5.9. Here, $C_{e,2}$ is respect to a verified lower bound of the minimal spectrum of Laplacian $-\Delta$ given by (48). The residual bound is presented by $C_{2,h}$ using the newest smoothing technique in (73). Based on Newton-Kantorovich theorem, there is an unique solution in the error bound ρ.

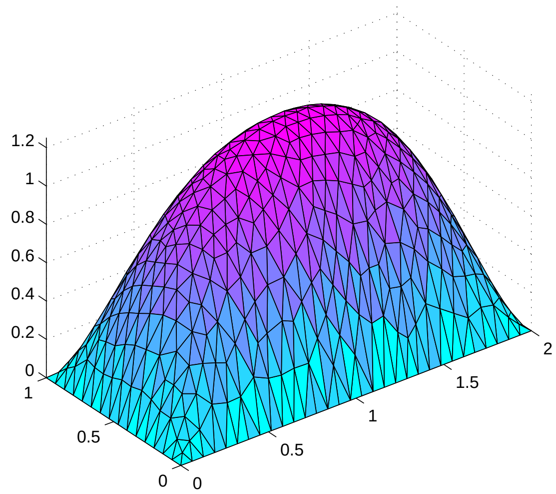

Figure 5.9: $\hat{u}$ of (84) on rectangular domain

Table 5.9: Verified results on rectangular domain

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ
3	7.712×10^{-2}	2.931×10^{-1}	1.351	6.491×10^{-1}	1.215×10^{-1}	1.438×10^{-1}	9.507×10^{-1}
4	3.527×10^{-2}	2.864×10^{-1}	1.286	3.236×10^{-1}	1.161×10^{-1}	6.203×10^{-2}	4.298×10^{-1}
5	1.756×10^{-2}	2.852×10^{-1}	1.264	1.617×10^{-1}	1.151×10^{-1}	2.966×10^{-2}	2.073×10^{-1}
6	8.714×10^{-3}	2.849×10^{-1}	1.253	8.021×10^{-2}	1.148×10^{-1}	1.445×10^{-2}	1.013×10^{-1}
7	4.889×10^{-3}	2.848×10^{-1}	1.249	3.953×10^{-2}	1.147×10^{-1}	7.064×10^{-3}	4.953×10^{-2}

5.4.2. Hexagonal domain. Secondly, let Ω be the hexagonal domain. There are two approximate solutions $\hat{u}_1, \hat{u}_2 \in V_h$ given by finite element method. These are displayed in Figure 5.10 and Figure 5.11 with the mesh size 1/16. For the first approximate solution $\hat{u}_1$, verification results are shown in Table 5.10 and Table 5.11. Here, we use P_1 (piecewise linear) and P_2 (piecewise quadratic) elements for getting $\hat{u}_1$. We adopt RT_0 space for P_1-element and RT_1 space for P_2-element. Comparing two cases in Table 5.10 and Table 5.11, we can observe that higher order elements yield improved result .

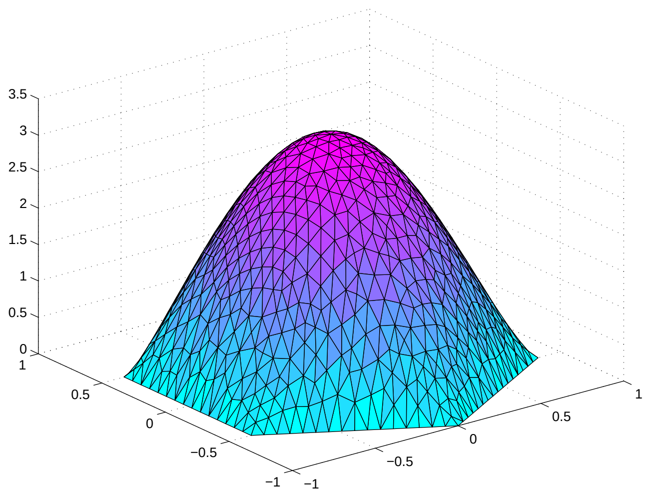

Figure 5.10: Shape of $\hat{u}_1$ for (84)

Table 5.10: Verified results for $\hat{u}_1$: P_1 with $p_h \in RT_0$

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ_1
3	6.288×10^{-2}	3.777×10^{-1}	3.667	9.003×10^{-1}	2.018×10^{-1}	2.441	Failed
4	3.231×10^{-2}	3.749×10^{-1}	3.477	4.609×10^{-1}	1.988×10^{-1}	1.107	Failed
5	1.886×10^{-2}	3.743×10^{-1}	3.404	2.248×10^{-1}	1.981×10^{-1}	5.155×10^{-1}	Failed
6	8.745×10^{-3}	3.741×10^{-1}	3.334	1.131×10^{-1}	1.978×10^{-1}	2.482×10^{-1}	4.403×10^{-1}
7	4.819×10^{-3}	3.739×10^{-1}	3.308	5.662×10^{-2}	1.977×10^{-1}	1.224×10^{-1}	2.004×10^{-1}

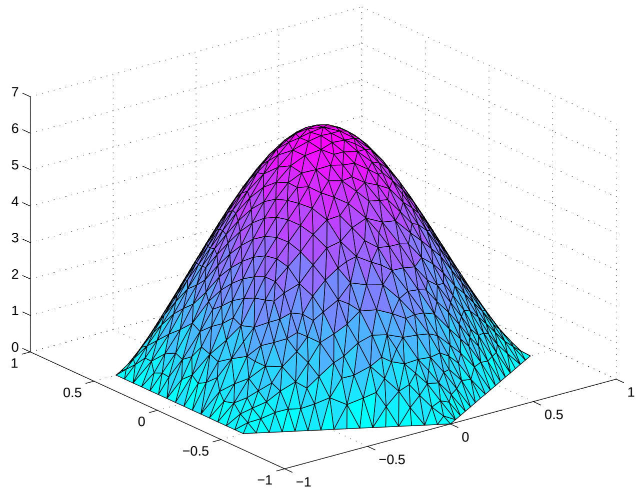

Figure 5.11: Shape of $\hat{u}_2$ for (84)

Table 5.11: Verified results for $\hat{u}_1$: P_2 with $p_h \in RT_1$

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ_1
3	5.776×10^{-2}	3.779×10^{-1}	3.821	1.255×10^{-1}	2.019×10^{-1}	3.699×10^{-1}	6.351×10^{-1}
4	2.823×10^{-2}	3.749×10^{-1}	3.496	4.479×10^{-2}	1.987×10^{-1}	1.088×10^{-1}	1.662×10^{-1}
5	2.138×10^{-2}	3.745×10^{-1}	3.437	1.491×10^{-2}	1.983×10^{-1}	3.491×10^{-2}	5.216×10^{-2}

Next, we present results with respect to $\hat{u}_2$ which is from P_2 finite element space. Table 5.12 presents verified results for $\hat{u}_2$. Moreover, the comparison of each evaluation (67), (70) and (73) implies our proposed one works well in Table 5.13. Numeric values on last column in Table 5.13 express upper bound of absolute error ρ_2 using (73) residual bounds. Based on Newton-Kantorovich theorem, we prove that there is a solution in $\overline{B}(\hat{u}_2, \rho_2)$.

Table 5.12: Verified results for $\hat{u}_2$: P_2 with $p_h \in RT_1$

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ_2
3	5.776×10^{-2}	3.779×10^{-1}	5.167	1.966×10^{-1}	2.019×10^{-1}	1.061	Failed
4	2.823×10^{-2}	3.749×10^{-1}	3.805	6.715×10^{-2}	1.987×10^{-1}	1.931×10^{-1}	2.865×10^{-1}
5	2.138×10^{-2}	3.745×10^{-1}	3.642	2.204×10^{-2}	1.983×10^{-1}	5.794×10^{-2}	8.272×10^{-2}

Table 5.13: $\hat{u}_2$: P_2 with several evaluation

$1/2^\gamma$	(67)	(70)	(73)	$\|u - \hat{u}_2\|_V \le \rho_2$
3	11.18	0.7509	0.1966	Failed
4	5.467	0.3796	0.0672	0.2865
5	4.141	0.2776	0.0221	0.0828

5.5. Nonconvex domain

Another example is the case that Ω is assumed to be noncovex domain. Let us consider (84) on $\Omega = (0,2)^2 \setminus (1,2)^2$: L-shape domain. An approximate solution $\hat{u} \in V_h$ of (84) is appeared on Figure 5.12 with the mesh size 1/16. Verification results are shown in Table 5.14. Here, C_M is calculated using Raviart-Thomas mixed finite element method by the procedure in Subsection 3.2.2. The computational cost of C_M becomes larger than that of the case on convex domain. Further, $C_{2,h}$ uses the method in (73) with P_1-RT_0 smoothing technique. Based on Newton-Kantorovich theorem, there are unique solution in the error bound in ρ.

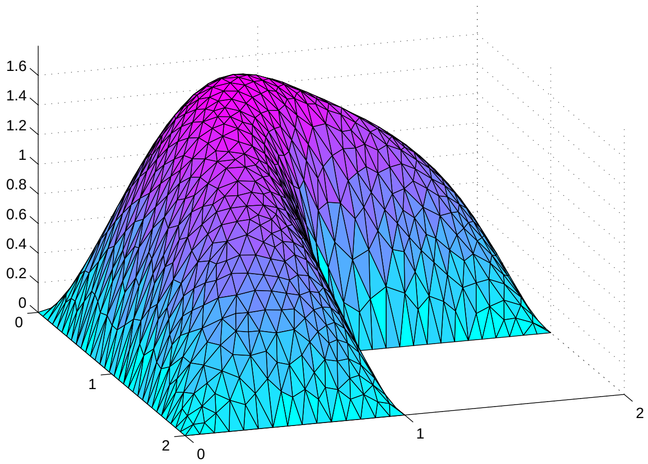

Figure 5.12: $\hat{u}$ of (84) on L-shape domain

Table 5.14: Verification for L-shape domain

$1/2^\gamma$	C_M	$C_{e,2}$	C_1	$C_{2,h}$	C_3	$C_1^2 C_{2,h} C_3$	ρ
3	1.031×10^{-1}	3.341×10^{-1}	1.872	1.221	1.578×10^{-1}	6.739×10^{-1}	Failed
4	5.558×10^{-2}	3.257×10^{-1}	1.693	6.415×10^{-1}	1.501×10^{-1}	2.758×10^{-1}	1.301
5	3.229×10^{-2}	3.234×10^{-1}	1.617	3.601×10^{-1}	1.479×10^{-1}	1.392×10^{-1}	6.295×10^{-1}

5.6. GUI toolbox

In order to illustrate that our proposed method can treat arbitrary polygonal domain, we develop GUI toolbox on MATLAB for the model problem (84). Users can define the concerning domain by using GUI. Arbitrary domain including with hole is treatable in this toolbox. This toolbox is designed for verified computations to Semilinear elliptic PDEs (7). Here, we just put the interface of our toolbox as a graphic.

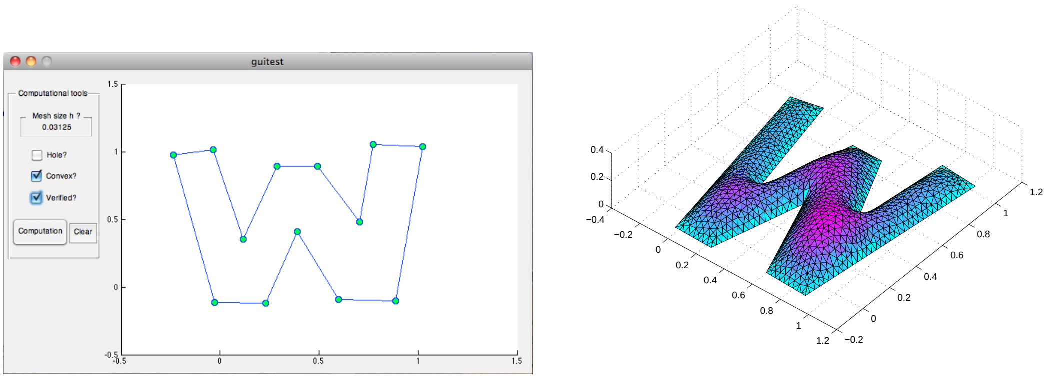

Figure 5.13: Example of GUI toolbox

Appendix A

Notes of Raviart-Thomas elements on triangle

In this part, we would like to note representations of the lowest (RT_0) and 1st order (RT_1) Raviart-Thomas element on a triangle element K_h. Vertices of K_h are numbered as 1, 2, 3. Their coordinates are (x_1, y_1), (x_2, y_2), (x_3, y_3). Let us denote $a_i = x_j y_k - x_k y_j$, $b_i = y_j - y_k$, $c_i = x_k - x_j$ where (i, j, k) are even permutation of $(1, 2, 3)$. Here, we put subtense of each vertex as e_i with direction from j to k. See K_h in Figure 1.1. Then it follows

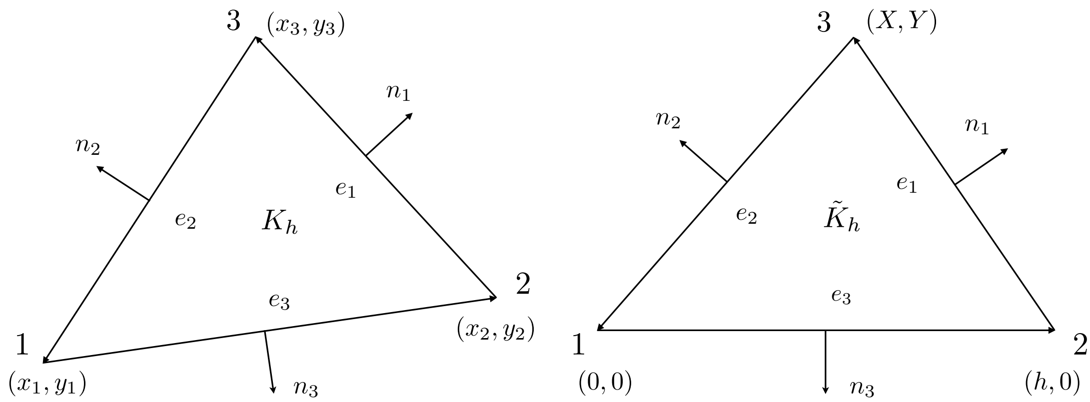

Figure 1.1: Triangle elements K_h and $\tilde{K}_h$

$$|e_i| = (b_i^2 + c_i^2)^{1/2}, \ D = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = b_j c_k - b_k c_j.$$

Furthermore, the unit normal vector n_i on each side is given by

$$n_i = \begin{pmatrix} n_1^{(i)} \\ n_2^{(i)} \end{pmatrix} = \frac{-\sigma}{|e_i|} \begin{pmatrix} b_i \\ c_i \end{pmatrix},$$

where $\sigma = D/|D|$ is corresponding to the direction of numbering. Namely,

$$\sigma = \begin{cases} 1, & (i, j, k \ : \ \text{counter clockwise rotation}), \\ -1, & (i, j, k \ : \ \text{clockwise rotation}). \end{cases}$$

For $q \in RT_k(K_h)$, degrees of freedom are given by

$$\int_{\partial K_h} q \cdot n \ \varphi_k ds, \ \ \varphi_k \in R_k(\partial K_h), \ \text{ for } k \geq 0, \tag{85}$$

$$\int_{K_h} q \cdot q_{k-1} ds, \ \ q_{k-1} \in (P_{k-1}(K_h))^2, \ \text{ for } k \geq 1. \tag{86}$$

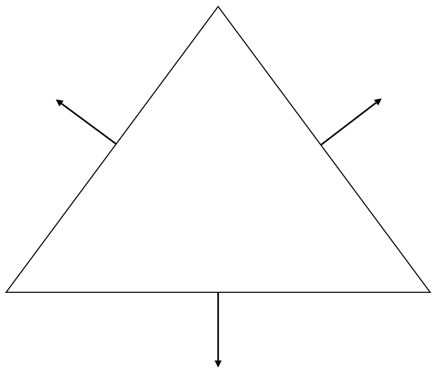

Figure 1.2: $RT_0(K_h)$

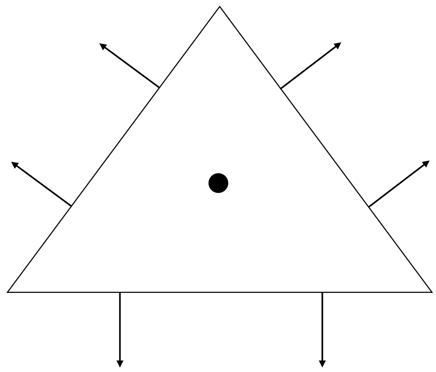

Figure 1.3: $RT_1(K_h)$

A.1. RT_0 ELEMENTS

For $p_h \in RT_0$, the representation of RT_0 element p_h on a triangle K_h is given by

$$p_h|_{K_h} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \alpha_3 \begin{pmatrix} x \\ y \end{pmatrix}$$

Let us explain how to determine coefficients α_i. Three freedoms are given by the following form, which is equivalent to (85) in case of $k = 0$.

$$\gamma_i = |e_i| \ p_h \cdot n_i$$

Notice that $p_h \cdot n_i = p_h|_{(x_j,y_j)} \cdot n_i$, we have

$$\begin{bmatrix} n_1^{(1)} & n_2^{(1)} & x_2 n_1^{(1)} + y_2 n_2^{(1)} \\ n_1^{(2)} & n_2^{(2)} & x_3 n_1^{(2)} + y_3 n_2^{(2)} \\ n_1^{(3)} & n_2^{(3)} & x_1 n_1^{(3)} + y_1 n_2^{(3)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \gamma_1/|e_1| \\ \gamma_2/|e_2| \\ \gamma_3/|e_3| \end{bmatrix}$$

$$\Longleftrightarrow \sigma \begin{bmatrix} -b_1 & -c_1 & a_1 \\ -b_2 & -c_2 & a_2 \\ -b_3 & -c_3 & a_3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}.$$

Using facts for $i = 1, 2, 3$,

$$\begin{cases} \sum a_i = D, & \sum b_i = \sum c_i = 0, \\ \sum b_i x_i = D, & \sum a_i x_i = \sum c_i x_i = 0, \\ \sum c_i y_i = D, & \sum a_i y_i = \sum b_i y_i = 0, \end{cases}$$

and $\sigma D = |D|$, we have

$$\alpha_1 = -\frac{\sum \gamma_i x_i}{|D|},\ \alpha_2 = -\frac{\sum \gamma_i y_i}{|D|},\ \alpha_3 = \frac{\sum \gamma_i}{|D|}.$$

Therefore, RT_0 element on K_h can be expressed with freedoms γ_i

$$p_h|_{K_h} = \sum_{i=1}^{3} \frac{\gamma_i}{|D|} \begin{pmatrix} x - x_i \\ y - y_i \end{pmatrix} = \sum_{i=1}^{3} \gamma_i \psi_i,$$

where ψ_i are base functions of RT_0 finite element space.

Remark A.1. *The image of $RT_0(K_h)$ is given in Figure 1.2. Further for $q \in (L^2(\Omega))^2$, let us define a linear functional, $F_i(q) = |e_i|\{q(x_j, y_j) \cdot n_i\}$ $(i = 1, 2, 3)$. It follows*

$$F_i(\psi_j) = \delta_{ij} = \begin{cases} 1, & (i = j), \\ 0, & (i \neq j), \end{cases} \quad 1 \leq i, j \leq 3.$$

A.2. RT_1 ELEMENTS

Next let us consider 1st order Raviart-Thomas finite element. Degrees of freedom are denoted by $\gamma_i \in \mathbb{R}$ $(i = 1, ..., 8)$. For simplicity, we will transform triangle K_h to $\tilde{K}_h$, which has vertices $(0,0)$, $(h,0)$, (X,Y) in Figure 1.1.

$$h = (b_3^2 + c_3^2)^{1/2}, \quad \begin{pmatrix} X \\ Y \end{pmatrix} = \frac{1}{h} \begin{pmatrix} c_3 & -b_3 \\ b_3 & c_3 \end{pmatrix} \begin{pmatrix} -c_2 \\ b_2 \end{pmatrix}, \quad D = hY,$$

$$n_1 = \frac{\sigma}{|e_1|} \begin{pmatrix} Y \\ -(X-h) \end{pmatrix}, \quad n_2 = \frac{\sigma}{|e_2|} \begin{pmatrix} -Y \\ X \end{pmatrix}, \quad n_3 = \frac{\sigma}{|e_3|} \begin{pmatrix} 0 \\ -h \end{pmatrix}.$$

In the following, we would like to explain RT_1 element on $\tilde{K}_h$. RT_1 element p_h is represented on $\tilde{K}_h$,

$$p_h|_{\tilde{K}_h} = \begin{pmatrix} \alpha_1 + \alpha_2 x + \alpha_3 y \\ \alpha_4 + \alpha_5 x + \alpha_6 y \end{pmatrix} + (\alpha_7 x + \alpha_8 y) \begin{pmatrix} x \\ y \end{pmatrix}.$$

Coefficients α_i are obtained by the following method of determination with respect to γ_i. For $i = 1, 2, 3$, degrees of freedom are given by (85) and (86),

$$\begin{aligned} \gamma_i &= \int_{e_i} p_h \cdot n_i \phi_j ds, \\ \gamma_{i+3} &= \int_{e_i} p_h \cdot n_i \phi_k ds, \\ \gamma_7 &= \int_{\tilde{K}_h} p_h \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} ds, \\ \gamma_8 &= \int_{\tilde{K}_h} p_h \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} ds, \end{aligned}$$

where ϕ_j, ϕ_k denote piecewise linear functions on e_i, satisfying

$$\phi_j(x_j, y_j) = \phi_k(x_k, y_k) = 1, \ \phi_j(x_k, y_k) = \phi_k(x_j, y_j) = 0.$$

So that we have

$$\frac{\sigma}{6}\begin{bmatrix} 3Y & Y(X+2h) & Y^2 & -3(X-h) & -(X-h)(X+2h) & -(X-h)Y & hY(X+2h) & hY^2 \\ -3Y & -2XY & -2Y^2 & 3X & 2X^2 & 2XY & 0 & 0 \\ 0 & 0 & 0 & -3h & -h^2 & 0 & 0 & 0 \\ 3Y & Y(2X+h) & 2Y^2 & -3(X-h) & -(X-h)(2X+h) & -2(X-h)Y & hY(2X+h) & 2hY^2 \\ -3Y & -XY & -Y^2 & 3X & X^2 & XY & 0 & 0 \\ 0 & 0 & 0 & -3h & -2h^2 & 0 & 0 & 0 \\ 6 & 2(X+h) & 2Y & 0 & 0 & 0 & h^2+hX+X^2 & (2X+h)Y/2 \\ 0 & 0 & 0 & 6 & 2(X+h) & 2Y & (2X+h)Y/2 & Y^2 \end{bmatrix}$$

$$*\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \\ 2\gamma_7/D \\ 2\gamma_8/D \end{bmatrix}.$$

Solving above linear system, we have the value of each coefficients. Then, RT_1 element is described on $\tilde{K}_h$,

$$p_h|_{\tilde{K}_h} = \sum_{i=1}^{8} \gamma_i \psi_i,$$

where ψ_i is base functions as following

$$\psi_1 = \frac{2}{|D|}\begin{pmatrix} -2x + \frac{X}{Y}y + \frac{4}{h}(x^2 - \frac{X}{Y}xy) \\ -y + \frac{4}{h}(xy - \frac{X}{Y}y^2) \end{pmatrix},$$

$$\psi_2 = \frac{2}{|D|}\begin{pmatrix} h - x - (\frac{X+3h}{Y})y + \frac{4}{Y}xy \\ -2y + \frac{4}{Y}y^2 \end{pmatrix},$$

$$\psi_3 = \frac{2}{|D|}\begin{pmatrix} -2X + 3(\frac{X+3h}{h})x - \frac{3X}{D}(X-h)y + \frac{4}{h}(-x^2 + (\frac{X-h}{Y})xy) \\ -2Y + \frac{3Y}{h}x - \frac{3}{h}(X-2h)y + \frac{4}{h}(-xy + (\frac{X-h}{Y})y^2) \end{pmatrix},$$

$$\psi_4 = \frac{2}{|D|}\begin{pmatrix} -x - \frac{X}{Y}y + \frac{4}{Y}xy \\ -2y + \frac{4}{Y}y^2 \end{pmatrix},$$

$$\psi_5 = \frac{2}{|D|}\begin{pmatrix} -2h + 6x - 3(\frac{X-h}{Y})y + \frac{4}{h}(-x^2 + (\frac{X-h}{Y})xy) \\ 3y + \frac{4}{h}(-xy + (\frac{X-h}{Y})y^2) \end{pmatrix},$$

$$\psi_6 = \frac{2}{|D|}\begin{pmatrix} X - (\frac{3X+2h}{h})x + \frac{X}{D}(3X+h)y + \frac{4}{h}(x^2 - \frac{X}{Y}xy) \\ Y - \frac{3Y}{h}x + (\frac{3X-h}{h})y + \frac{4}{h}(xy - \frac{X}{Y}y^2) \end{pmatrix},$$

$$\psi_7 = \frac{8}{h|D|}\begin{pmatrix} 2x - \frac{X}{Y}y - \frac{2}{h}x^2 + (\frac{2X-h}{D})xy \\ y - \frac{2}{h}xy + (\frac{2X-h}{D})y^2 \end{pmatrix},$$

$$\psi_8 = \frac{8}{D|D|}\begin{pmatrix} -(2X-h)x + \frac{X}{Y}(X+h)y + (\frac{2X-h}{h})x^2 - 2(\frac{X^2-Xh+h^2}{D})xy \\ -(X-2h)y + (\frac{2X-h}{h})xy - 2(\frac{X^2-Xh+h^2}{D})y^2 \end{pmatrix}.$$

Remark A.2. *See Figure 1.3 for degrees of freedom to* $RT_1(\tilde{K}_h)$. *A linear functional is defined by* $F_i(q)$, $(i = 1, ..., 8)$ *for* $q \in (L^2(\Omega))^2$, *such that*

$$
\begin{aligned}
F_l(q) &= \int_{e_l} q \cdot n_l \phi_m ds, \\
F_{l+3}(q) &= \int_{e_l} q \cdot n_l \phi_n ds, \\
F_7(q) &= \int_{\tilde{K}_h} q \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} dx, \\
F_8(q) &= \int_{\tilde{K}_h} q \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} dx
\end{aligned}
$$

where (l, m, n) *are even permutation of* $(1, 2, 3)$. *Then, we have* $F_i(\psi_j) = \delta_{ij}$, $(1 \leq i, j \leq 8)$.

Remark A.3. *For the inner product of each base functions* ψ_i *such as* (ψ_i, ψ_j), $(i, j = 1, ..., 8)$, *we need to integrate base functions on* $\tilde{K}_h$ *instead of* K_h. *For instance,*

$$
(\psi_i, \psi_j) := \int_{\tilde{K}_h} \psi_i \cdot \psi_j dx, \ (i, j = 1, ..., 8).
$$

Bibliography

1. R.A. Adams, *Sobolev spaces*, Academic Press, New York, 1975 (English).

2. D. Braess, *Finite elements*, third ed., Cambridge University Press, Cambridge, 2007.

3. S.C. Brenner and R. Scott, *The mathematical theory of finite element methods*, Texts in Applied Mathematics Series, Springer, 2010.

4. H. Brézis, *Functional analysis, sobolev spaces and partial differential equations*, Universitext Series, Springer, 2010.

5. F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer Series in Computational Mathematics, vol. 15, Springer-Verlag, New York, 1991.

6. G. Christophe and R. Jean-François, *Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities*, http://www.geuz.org/gmsh/.

7. P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for newton's method and extensions to related methods*, SIAM Journal on Numerical Analysis **16** (1979), no. 1, 1–10.

8. P. Grisvard, *Elliptic problems in nonsmooth domains*, Monographs and studies in mathematics, Pitman Advanced Pub. Program, 1985.

9. C. Johnson, *Numerical solution of partial differential equations by the finite element method*, vol. 32, Cambridge University Press, 1987.

10. L.V. Kantorovich and G.P. Akilov, *Functional analysis in normed spaces*, International series of monographs in pure and applied mathematics, Pergamon Press, 1964.

11. G. Kedem, *A posteriori error bounds for two-point boundary value problems*, SIAM Journal on Numerical Analysis **18** (1981), no. 3, 431–448.

12. F. Kikuchi and X. Liu, *Estimation of interpolation error constants for the p0 and p1 triangular finite elements*, Computer Methods in Applied Mechanics and Engineering **196** (2007), no. 37-40, 3750 – 3758.

13. F. Kikuchi and H. Saito, *Remarks on a posteriori error estimation for finite element solutions*, J. Comput. Appl. Math. **199** (2007), no. 2, 329–336.

14. K. Kobayashi, *A constructive a priori error estimation for finite element discretizations in a non-convex domain using singular functions*, Japan Journal of Industrial and Applied Mathematics **26** (2009), 493–516.

15. K. Kobayashi, *On the interpolation constants over triangular elements*, RIMS Kokyuroku **1733** (2011-03), 58–77.

16. X. Liu and S. Oishi, *Verified eigenvalue evaluation for elliptic operator on arbitrary polygonal domain*, Prepare to publication.

17. M.A. McCarthy and R.A. Tapia, *Computable a posteriori l_∞-error bounds for the approximate solution of two-point boundary value problems*, SIAM Journal on Numerical Analysis **12** (1975), no. 6, 919–937.

18. M.T. Nakao, *A numerical approach to the proof of existence of solutions for elliptic problems*, Japan Journal of Industrial and Applied Mathematics **5** (1988), 313–332.

19. M.T. Nakao, K. Hashimoto, and Y. Watanabe, *A numerical method to verify the invertibility of linear elliptic operators with applications to nonlinear problems*, Computing **75** (2005), no. 1, 1–14.

20. M.T. Nakao and T. Kinoshita, *Some remarks on the behaviour of the finite element solution in nonsmooth domains*, Appl. Math. Lett. **21** (2008), no. 12, 1310–1314.

21. M.T. Nakao and Y. Watanabe, *Numerical verification methods for solutions of semilinear elliptic boundary value problems*, Nonlinear Theory and Its Applications, IEICE **2** (2011), no. 1, 2–31.

22. M.T. Nakao and N. Yamamoto, *Numerical verification*, Nihonhyouron-sya, 1998.

23. S. Oishi, *Numerical verification of existence and inclusion of solutions for nonlinear operator equations*, J. Comput. Appl. Math. **60** (1995), 171–185.

24. S. Oishi and S.M. Rump, *Fast verification of solutions of matrix equations*, Numerische Mathematik **90** (2002), 755–773.

25. M. Plum, *Computer-assisted existence proofs for two-point boundary value problems*, Computing **46** (1991), 19–34.

26. M. Plum, *Computer-assisted proofs for semilinear elliptic boundary value problems*, Japan J. Indust. Appl. Math. **26** (2009), no. 2-3, 419–442.

27. W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math. **5** (1947), 241–269.

28. P. Raviart and J. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, Mathematical Aspects of Finite Element Methods (Ilio Galligani and Enrico Magenes, eds.), Lecture Notes in Mathematics, vol. 606, Springer Berlin / Heidelberg (1977), 292–315.

29. S.M. Rump, *INTLAB - INTerval LABoratory*, Developments in Reliable Computing (Tibor Csendes, ed.), Kluwer Academic Publishers, Dordrecht (1999), 77–104, http://www.ti3.tu-harburg.de/rump/.

30. S.M. Rump, *Verification of positive definiteness*, BIT **46** (2006), no. 2, 433–452.

31. S.M. Rump, *Verification methods: rigorous results using floating-point arithmetic*, Acta Numer. **19** (2010), 287–449.

32. S.M. Rump, *Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse*, Bit Numerical Mathematics **51** (2011), no. 2, 367–384.

33. G. Strang and G.J. Fix, *An analysis of the finite element method*, Wellesley-Cambridge Press, 2008.

34. T. Sunaga, *Theory of an interval algebra and its application to numerical analysis*, Japan Journal of Industrial and Applied Mathematics **26** (2009), 125–143.

35. A. Takayasu and S. Oishi, *A method of computer assisted proof for nonlinear two-point boundary value problems using higher order finite elements*, Nonlinear Theory and Its Applications, IEICE **2** (2011), no. 1, 74–89.

36. A. Takayasu, S. Oishi, and T. Kubo, *Numerical existence theorem for solutions of two-point boundary value problems of nonlinear differential equations*, Nonlinear Theory and Its Applications, IEICE **1** (2010), no. 1, 105–118.

37. M. Urabe, *Galerkin's procedure for nonlinear periodic systems*, Archive for Rational Mechanics and Analysis **20** (1965), 120–152.

38. N. Yamamoto and M.T. Nakao, *Numerical verifications for solutions to elliptic equations using residual iterations with a higher order finite element*, Journal of Computational and Applied Mathematics **60** (1995), no. 1-2, 271 – 279.

39. T. Yamamoto, *Error bounds for approximate solutions of systems of equations*, Japan Journal of Industrial and Applied Mathematics **1** (1984), 157–171.