VALIDATED COMPUTATION FOR A LINEAR ELLIPTIC PROBLEM WITH A PARAMETER

Nobito Yamamoto, Mitsuhiro T. Nakao & Yoshitaka Watanabe

Abstract

We propose an efficient method for validated computing of the solution of linear elliptic problem with a parameter. A numerical method to obtain the k-th eigenvalue of given symmetric matrices is also developed. We present some numerical examples which concern with the problem to determine a constant appearing in error estimation of the Finite Element Method (FEM).

1. Introduction

We consider the following linear elliptic equation defined on a bounded convex polygonal domain $\Omega \subset \mathbf{R}^2$.

$$(1) \qquad \begin{cases} -\triangle u &= \lambda(u+g) \qquad \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial n} &= 0 \qquad\qquad\quad \text{on} \quad \partial\Omega. \end{cases}$$

Our aim is to obtain numerically a weak solution $u \in X$ with guaranteed accuracy for a given $g \in L_0^2(\Omega)$ and a parameter λ, where $X := H^1(\Omega) \bigcap L_0^2(\Omega)$, $L_0^2(\Omega) := \{v \in L^2(\Omega) | \int_\Omega v dx = 0\}$ and $0 < \lambda < a$. The positive constant a is taken so that λ should not be an eigenvalue of the Laplacian $-\triangle$ with Neumann boundary condition. Then the operator $-\triangle - \lambda$ is one of the Fredholm operators with index=0, and the equation (1) has a unique solution within $0 < \lambda < a$.

This problem appears in computation of the constant related to the error estimates of FEM with linear triangular elements.

Since (1) is linear, validated computaion will be done rather easily, e.g. [6],[4]. But in the case where the parameter λ continuously varies in some interval, the cost of computation may considerably grow. In this paper, we propose a method of efficient computation for changing parameters.

2. Outline of the validated computation for the problem (1)

First, we give an outline of our method for validated computation of the solution of (1):

1. Let $S_h \subset H^1(\Omega)$ be a finite element space and P_h be a projection from X to $S_h \bigcap L_0^2(\Omega)$ as follows:

$$\begin{aligned} (\nabla P_h u, \nabla v_h) &= (\nabla u, \nabla v_h), \quad \forall v_h \in S_h, \\ (P_h u, 1) &= 0, \end{aligned}$$

where $(\cdot, \cdot)$ is the L^2 inner product. Moreover, let us take $(\nabla\cdot, \nabla\cdot)$ as an inner product of X, and put $S_h^\perp$ as the orthogonal complement of $S_h \bigcap L_0^2(\Omega)$ with respect to that inner product.

We assume that S_h has the following approximation property for the Poisson equation with $f \in L_0^2(\Omega)$ in the right-hand side. Namely, for the solution v of

$$\left\{\begin{array}{rcll} -\triangle v &=& f & \text{in} \quad \Omega, \\ \dfrac{\partial v}{\partial n} &=& 0 & \text{on} \quad \partial\Omega, \end{array}\right.$$

and its projection $P_h v$, it holds that

$$\|\nabla(I - P_h)v\| \le C_0 h \|f\|, \tag{2}$$

where $\|\cdot\|$ means the L^2-norm, h a parameter of S_h concerning the mesh size, and C_0 a constant independent of h.

2. Let $(-\triangle)^{-1}$ be an inverse operator of $-\triangle$ with the Neumann boundary condition. Then the equation (1) is represented of the following fixed point form:

$$u \;=\; (-\triangle)^{-1}(\lambda(u+g)).$$

We can write this as follows :

$$\left\{\begin{array}{lcl} P_h u &=& P_h((-\triangle)^{-1}(\lambda(u+g))) \\ (I - P_h)u &=& (I - P_h)((-\triangle)^{-1}(\lambda(u+g))) \end{array}\right. \tag{3}$$

The first equation of the above is equivalent to :

$$\begin{array}{rcl} (\nabla P_h u, \nabla v_h) - \lambda(P_h u, v_h) &=& \lambda((I - P_h)u + g, v_h), \\ && \forall v_h \in S_h. \end{array}$$

Then we define a mapping

$$R_h \;:\; L_0^2(\Omega) \ni f \mapsto R_h f \in S_h \bigcap L_0^2(\Omega)$$

by

$$\begin{array}{rcl} (\nabla R_h f, \nabla v_h) - \lambda(R_h f, v_h) &=& (f, v_h), \\ && \forall v_h \in S_h. \end{array} \tag{4}$$

Notice that we can verify that $R_h f$ is well defined through showing in actual computation that the corresponding matrix is nonsingular. Using $P_h u = \lambda R_h((I - P_h)u + g)$ and (3), we define the operator T by:

$$T(u) \;:=\; \lambda R_h((I - P_h)u + g) + (I - P_h)(-\triangle)^{-1}(\lambda(u+g)).$$

Then (1) turns to be equivalent to a fixed point equation on X, that is, $u = T(u)$.

3. Since the operator T is a bounded continuous affine mapping on H^1, and moreover it is compact, if we find a bounded closed convex set $U \subset X$ such that $T(U) = \{T(u) | u \in U\} \subset U$ holds, then there exists a solution $u \in T(U)$ of $u = T(u)$ by Schauder's fixed point theorem. Taking convex sets $U_h \subset S_h \bigcap L_0^2$ and $U_h^\perp \subset S_h^\perp (\subset X)$, we define the set U by $U = U_h \oplus U_h^\perp$, which is usually referred as the candidate set. Then it is sufficient for our purpose to show that

$$\begin{cases} P_h T(U) & \subset & U_h \\ (I - P_h)T(U) & \subset & U_h^\perp, \end{cases}$$

which we call the verification condition.

4. We define the set U_h and $U_h^\perp$ as follows:
For a given $\alpha > 0$,

$$\begin{aligned} U_h^\perp &:= \{u_h^\perp \in S_h^\perp \,|\, \|\nabla u_h^\perp\| \le \alpha\} && (5) \\ U_h &:= \{u_h \in S_h \textstyle\bigcap L_0^2 \,|\, u_h = \lambda R_h(v + g), v \in U_h^\perp\}. && (6) \end{aligned}$$

From this definition, if

$$(I - P_h)T(U) \;\subset\; U_h^\perp \tag{7}$$

holds, then $P_h T(U) \subset U_h$ also holds. Applying (2) to (7), we obtain a sufficient condition for the verification.

Theorem 1
For the candidate set U which is constituted of $U_h^\perp$ and U_h satisfying (5) and (6), respectively, if

$$C_0 h \lambda \sup_{u \in U} \|u + g\| \;\le\; \alpha \tag{8}$$

holds, then there exists a solution $u \in U$ of $u = T(u)$.

In this case, the solution is unique (globally) because the problem (1) has a unique solution for $0 < \lambda < a$.

3. Validated computaion for the operator R_h

In order to define α and the set U so that the above condition (8) holds, we have to calculate the image of R_h with guaranteed accuracy. In the following, we propose a method to calculate $R_h f$ without so much cost when the parameter λ varies continuously in some interval.

Take $\{\phi_i\}_{i=1,\cdots,n}$ as a basis of the finite elelment subspace S_h. Then, from (4), we represent $R_h f = \sum_{i=1,\cdots,n} x_i \phi_i$ by using the solution $\vec{x} = (x_1, x_2, \cdots, x_n)^T$ of the following linear system:

$$G_\lambda \vec{x} \;=\; \vec{f},$$

where

$$\begin{aligned} G_\lambda &= D - \lambda L, \\ D &= ((\nabla\phi_i, \nabla\phi_j))_{i,j=1,\cdots,n}, \\ L &= ((\phi_i, \phi_j))_{i,j=1,\cdots,n} \end{aligned}$$

are $n \times n$ matrices, and $\vec{f}$ is a vector such that

$$\vec{f} \;=\; ((f, \phi_1), (f, \phi_2), \cdots, (f, \phi_n))^T.$$

To solve this system with validated computation, here we adopt the method by Rump [5] in which the smallest singular value σ_λ of G_λ is used. Note that σ_λ is equal to the smallest absolute value of the eigenvalues of G_λ because of the symmetry of G_λ.

We assume that λ belongs to an interval Λ, with the center $\breve{\lambda}$ and the width ν, and that the vector $\vec{f}$ in the right-hand side also belongs to an interval valued vector, which has the center vector $\breve{f}$ and the width vector $[\vec{f}]$.

Let $\tilde{x}$ be an approximate solution of $G_{\breve{\lambda}}^{-1}\breve{f}$, obtained by floating point arithmetic. Since the diffrernce of $\vec{x}$ from $\tilde{x}$ is written as

$$\vec{x} - \tilde{x} \;=\; G_\lambda^{-1}((\vec{f} - \breve{f}) + \breve{f} - G_{\breve{\lambda}}\tilde{x} + (\lambda - \breve{\lambda})L\tilde{x}),$$

we have

$$\begin{aligned} \|\vec{x} - \tilde{x}\| \;&\leq\; \|G_\lambda^{-1}\|_2(\frac{1}{2}\|[\vec{f}]\| + \|\breve{f} - G_{\breve{\lambda}}\tilde{x}\| + \frac{\nu}{2}\|L\tilde{x}\|) \\ &\leq\; \frac{1}{\sigma_0}(\frac{1}{2}\|[\vec{f}]\| + \|\breve{f} - G_{\breve{\lambda}}\tilde{x}\| + \frac{\nu}{2}\|L\tilde{x}\|), \end{aligned}$$

where

$$\sigma_0 \;=\; \inf_{\lambda \in \Lambda} \sigma_\lambda.$$

Here, $\|\cdot\|$ stands for the usual Euclidian norm, and $\|\cdot\|_2$ for 2-norm of matrices induced by the Euclidian norm.

In this way, an error of the approximate solution $\tilde{x}$ can be obtained with guaranteed accuracy through rigorous calculation of the right-hand side of the above inequality.

4. Estimation of the smallest singular value of G_λ

Since the smallest singular value σ_λ equals the smallest absolute value of the eigenvalues, we have to compute the minimum absolute eigenvalue for λ over the interval Λ, which takes considerable costs. Therefore, we try to estimate σ_λ by some explicit functions of λ.

Let μ_1 and μ_2 be the first and the second smallest eigenvalue of G_λ, respectively. Since (1) is a Neumann problem, the matrix D which corresponds to the Laplacian is nonnegative and 0 is the smallest eigenvalue. Actually,

$$D\vec{1} \;=\; 0, \tag{9}$$

where $\vec{1} = (1, 1, \cdots, 1)^T$.

In what follows, we restrict λ within a range $0 < \lambda < b$ in which the matrix G_λ is not singular. Then, from (9) and the positive definiteness of L,

$$\mu_1 \;<\; 0 \;<\; \mu_2$$

holds. Thus one of μ_1 and μ_2 which has the smallest absolute value gives the smallest singular value.

First, we estimate an upper bound of μ_1. Since this is the smallest eigenvalue of G_λ, from (9) and the symmetry of G_λ , we have

$$
\begin{aligned}
\mu_1 &\leq -\frac{\vec{1}^T L \vec{1}}{\|\vec{1}\|^2}\lambda \\
&=: -\chi_1(\lambda).
\end{aligned} \tag{10}
$$

Next, we estimate a lower bound of μ_2 using the following lemma obtained by Weyl :

Lemma 1

Let A, B and C be real symmetric matrices with the size n such that $A = B+C$ holds. Define $\lambda_i(A)$, $\lambda_i(B)$ and $\lambda_i(C)$, $(i = 1, \cdots, n)$ as the eigenvalues of A, B and C, respectively, where the index i means the order of magnitude (λ_1 is the smallest). Then

$$
\lambda_i(B) + \lambda_1(C) \leq \lambda_i(A) \leq \lambda_i(B) + \lambda_n(C) \tag{11}
$$

and

$$
|\lambda_i(A) - \lambda_i(B)| \leq \|C\|_2 \tag{12}
$$

hold.

The proof is given by an elementary consideration on linear algebra. See [2], for example.

If, in the first inequality (11), we take $i = 2$, $A = D$, $B = G_\lambda$, and $C = \lambda L$, then we have a lower bound of μ_2 by

$$
\mu_2 \geq \rho_2 - \lambda\|L\|_2. \tag{13}
$$

Here, ρ_2 denotes the second smallest eigenvalue of the matrix D. In the actual calculations, $\|L\|_2$ is overestimated by $\|L\|_\infty$, the infinity norm of the matrix L. Thus defining

$$
\chi_2(\lambda) := \rho_2 - \lambda\|L\|_\infty,
$$

we obtain a lower bound of the smallest singular value by

$$
\sigma_\lambda \geq min(\chi_1(\lambda), \chi_2(\lambda)).
$$

5. Validated computation of the second eigenvalue of the matrix D

In order to apply the argument in the previous section to our problems, we have to calculate the second eigenvalue of the matrix D with guaranteed accuracy. In this section, we show a new method to obtain a bound of an eigenvalue of a symmetric matrix as well as to decide the index of the eigenvalue in order of magnitude. First, a lemma concerning the number of nonnegative eigenvalues is described.

Lemma 2

Let A be an arbitrary real symmetric matrix and can be decomposed as

$$
A = M^T B M,
$$

with a symmetric matrix B and a nonsingular matrix M. Then the matrices A and B have the same numbers of nonnegative eigenvalues.

The proof is omitted because the lemma is easily derived from Sylvester's law of inertia.

Using Lemma 2 and (12) in Lemma 1, we have a numerical method to estimate eigenvalues and to decide the orders of them as follows :

Theorem 2

Let A be an arbitrary symmetric matrix and $\tilde{\rho}$ be an approximation to an eigenvalue of A. Taking positive numbers δ_1 and δ_2, define

$$\begin{aligned} Y_1 &:= A - (\tilde{\rho} - \delta_1) I \\ &\textit{and} \\ Y_2 &:= A - (\tilde{\rho} + \delta_2) I, \end{aligned}$$

where I is the identity matrix. For Y_i, $i = 1, 2$, take a diagonal matrix B_i and a nonsingular matrix M_i, and compute the following quantities rigorously:

$$\begin{aligned} \varepsilon_1 &:= \|Y_1 - M_1^T B_1 M_1\|_2 \\ &\textit{and} \\ \varepsilon_2 &:= \|Y_2 - M_2^T B_2 M_2\|_2. \end{aligned}$$

Let B_1 and B_2 have $k-1$ and $k+r$ negative elements, respectively, with $k > 0$ and $r \geq 0$. Then there exist from the k-th to the $(k+r)$-th eigenvalue within an interval

$$[\tilde{\rho} - \delta_1 - \varepsilon_1 \quad , \quad \tilde{\rho} + \delta_2 + \varepsilon_2].$$

Proof

From Lemma 2, $M_1^T B_1 M_1$ and $M_2^T B_2 M_2$ have $k-1$ and $k+r$ negative eigenvalues, respectively. Let μ_k be the k-th eigenvalue of Y_1. If $\mu_k < -\varepsilon_1$, then $M_1^T B_1 M_1$ should have more than k negative eigenvalues because of Lemma 1. Thus it is necessary that $\mu_k \geq -\varepsilon_1$ holds. From $\mu_k = \rho_k - (\tilde{\rho} - \delta_1)$, we obtain a lower bound of ρ_k as follows :

$$\tilde{\rho} - \delta_1 - \varepsilon_1 \leq \rho_k. \tag{14}$$

Let ν_{k+r} be the $k+r$-th eigenvalue of Y_2. We know $\nu_{k+r} < \varepsilon_2$ from Lemma 1, and using $\nu_{k+r} = \rho_{k+r} - (\tilde{\rho} + \delta_2)$, obtain an upper bound of ρ_{k+r}:

$$\rho_{k+r} \leq \tilde{\rho} + \delta_2 + \varepsilon_2. \tag{15}$$

□

In the actual calculations, we use, for example, LDL^T-decomposition of Y_i, and ∞-norm instead of 2-norm. Using this method, we can do validated calculation of the second eigenvalue of the matrix D.

6. Numerical examples

In this section, we show some numerical examples on the following problem which appears in the computation of the constant in the error estimation of FEM with linear triangular elements:

$$\left\{ \begin{array}{rcll} -\triangle u &=& \lambda(u+g) & \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial n} &=& 0 & \text{on} \quad \partial\Omega \\ g &=& \dfrac{1}{2}((1-x)^2 + y^2) - \dfrac{1}{3}, & \end{array} \right. \tag{16}$$

where Ω is the standard triangle with vertices $(0,0)$, $(1,0)$ and $(0,1)$. We use a uniform triangular mesh with linear elements for the finite element subspace S_h. The parameter h is taken as $\frac{1}{n}$, where n means the number of partition of the edge. In the examples here, we take $n = 40$. The range of λ for which the problem (16) has a unique solution is $0 < \lambda < \pi^2$, and our calculation is done for $1.5 \le \lambda \le 4.0$. We show results for $n = 40$ with taking $C_0 = 4.9389492 \times 10^{-1}$ and $\delta_1 = \delta_2 = 1.0 \times 10^{-6}$.

The second engenvalue of the matrix D is obtained as

$$\rho_2 \in [0.5448499001 \times 10^{-2}, 0.5450499007 \times 10^{-2}].$$

The infinity norm of L is:

$$\|L\|_\infty \in [0.625000000000102 \times 10^{-3}, 0.625000000000104 \times 10^{-3}],$$

and

$$\frac{\vec{1}^T L \vec{1}}{\|\vec{1}\|^2} \in [0.580720092914863 \times 10^{-3}, 0.580720092915615 \times 10^{-3}].$$

From these, we can take χ_1 as a lower bound of the smallest singular value of G_λ for $0 < \lambda < 8.717598$.

Table 1. Validated results for (16)

Λ	σ_0	α	relative error of the solution
$[1.5, 2.0]$	0.8710802×10^{-3}	4.2093907×10^{-3}	$17.6579209 \times 10^{-2}$
$[2.0, 2.5]$	1.1614402×10^{-3}	5.2870586×10^{-3}	$13.3700798 \times 10^{-2}$
$[2.5, 3.0]$	1.4518003×10^{-3}	6.3905212×10^{-3}	$10.8107669 \times 10^{-2}$
$[3.0, 3.5]$	1.7421603×10^{-3}	7.5237493×10^{-3}	9.1117521×10^{-2}
$[3.5, 4.0]$	2.0325204×10^{-3}	8.6927330×10^{-3}	7.9009187×10^{-2}
$[3.0, 3.2]$	1.7421603×10^{-3}	6.8084311×10^{-3}	3.8701039×10^{-2}
$[3.2, 3.4]$	1.8583043×10^{-3}	7.2644782×10^{-3}	3.6672089×10^{-2}
$[3.4, 3.6]$	1.9744484×10^{-3}	7.7261107×10^{-3}	3.4885805×10^{-2}
$[3.6, 3.8]$	2.0905924×10^{-3}	8.1938205×10^{-3}	3.3299602×10^{-2}
$[3.8, 4.0]$	2.2067364×10^{-3}	8.6681660×10^{-3}	3.1879869×10^{-2}
$[3.5, 3.6]$	2.0325204×10^{-3}	7.7183901×10^{-3}	1.9088204×10^{-2}
$[3.6, 3.7]$	2.0905925×10^{-3}	7.9517124×10^{-3}	1.8733604×10^{-2}
$[3.7, 3.8]$	2.1486644×10^{-3}	8.1866387×10^{-3}	1.8398176×10^{-2}
$[3.8, 3.9]$	2.2067364×10^{-3}	8.4232425×10^{-3}	1.8080145×10^{-2}
$[3.9, 4.0]$	2.2648084×10^{-3}	8.6616026×10^{-3}	1.7777932×10^{-2}

The relative errors are estimated by $\|u - u_h\|_2 / \|u_h\|_2$, where $u_h := \check{\lambda} R_h g$.

We used `INTLIB_90`[3] in the numerical experiments, a library for interval arithmetic with consideration of the influence of rounding error.

Used machines are Sun Ultra Enterprise 450(single CPU). The details are shown in Table 2.

Table 2. Specification of numerical environment

	Ultra Enterprise 450
OS	SunOS 5.5.1
software	WorkShop Compiler Fortran 90 1.2
CPU	UltraSPARC-II 300MHz

7. Conclution

We proposed a new verification method for parametrized elliptic problems, and showed an application to a problem which appears in the computation of the constant in the error estimation of FEM. Moreover, we developed a new method to obtain a range of the k-th eigenvalue of a symmetric matrix.

As concerns the latter method, though it works in this case, there may be critical cases where the LDL^T-decomposition causes some considerable error and the obtained range of the eigenvalue is too large. We are now improving the method in order that it can be applied to arbitrary symmetric matrices with sufficient accuracy.

References

[1] Behnke, H., Inclusion of Eigenvalues of General Eigenvalue Problems for Matrices, in U. Kulisch and H. J. Stetter (Editors), Scientific Computation with Automatic Result Verification, *Computing, Supplement*, Vol.6 (1988) pp.69–78.

[2] Chatelin,F., *Valeurs propres de matrices*, Masson, Paris,1988.

[3] Kearfott, R. B., and Kreinovich, V., *Applications of Interval Computations*, Kluwer Academic Publishers, Netherland, 1996. (`http://interval.usl.edu/kearfott.html`)

[4] Nakao, M.T. and Yamamoto, N., Numerical verification of solutions for nonlinear elliptic problems using L^∞ residual method, *Journal of Mathematical Analysis and Applicatons*, Vol.217 (1998) 246-262.

[5] Rump, S. M., Verification Methods for Dense and Sparse Systems of Equations, in Jürgen Herzberger (Editor), *Topics in Validated Computations* (Proceedings of the IMACS-GAMM International Workshop on Validated Computation, Oldenburg, Germany, August 30th–September 3rd, 1993), Elsevier Science, North Holland,

[6] Yamamoto, N. and Nakao, M.T., Numerical verifications for solutions to elliptic equations using residual iterations with higher order finite element, *Journal of Computational and Applied Mathematics*, Vol.60 (1995) pp.271-279.

[7] Yamamoto, N., Watanabe, Y. and Nakao, M.T., Verification Methods of Generalized Eigenvalue Problems, preprint

Nobito Yamamoto & Mitsuhiro T. Nakao :
Graduate School of Mathematics, Kyushu University 33,Fukuoka 812-8581, Japan
Yoshitaka Watanabe :
Computer Center, Kyushu University 33,Fukuoka 812-8581, Japan