Determination of the Babuska-Aziz Constant for the Linear Triangular Finite Element

Fumio Kikuchi and Liu Xuefeng

Graduate School of Mathematical Sciences, University of Tokyo,
3-8-1 Komaba, Meguro, Tokyo 153-8914, Japan

Received March 30, 2005
Revised September 16, 2005

We explicitly determine the Babuska-Aziz constant, which plays an essential role in the interpolation error estimation of the linear triangular finite element. The equation for determination is the transcendental equation $t + \tan t = 0$, so that the solution can be numerically obtained with desired accuracy and verification. Such highly accurate approximate values for the constant can be widely used for a priori and a posteriori error estimations in adaptive computation and/or numerical verification.

Key words: linear triangular finite element, Babuska-Aziz constant, interpolation error estimation

1. Introduction

The finite element method (FEM) is now recognized as a powerful numerical method for wide classes of partial differential equations. Furthermore, it also has sound mathematical bases such as highly refined a priori and a posteriori error estimations. In the classical a priori error analysis of FEM, interpolation errors are essential to derive final error estimates [4, 5]. In this process, there appear various positive constants besides the standard discretization parameter h and norms (or seminorms), but it has been very difficult to evaluate such constants explicitly. For quantitative purposes, however, it is indispensable to evaluate or bound them as accurate as possible. Thus such evaluation has been attempted for adaptive finite element calculations based on a posteriori error estimation as well as for numerical verification by FEM. In this paper, we will determine the so-called Babuska-Aziz constant [2], which appears in the estimation of the interpolation errors of the linear (P_1) triangular finite element.

More specifically, we derive a transcendental equation for the above constant. To this end, we use the reflection (or symmetry) method to solve analytically an eigenvalue problem for 2D Laplace operator. As we will see later, this constant (C_1) gives an upper bound to the optimal constant (C_0) appearing in the H^1-error estimation of P_1-interpolation functions for H^2-functions over the unit isosceles right triangle. The P_1-finite element is the most classical and fundamental one, but still in frequent use. Thus precise estimation of C_1 as well as C_0 is very important, and a number of researchers have given bounds for these two constants using various approximation methods including numerical verification, see e.g. [1, 7, 8, 9, 10, 11]. The relation between C_0 and C_1 was pointed out by Babuska and

Aziz in conjunction with the maximum angle condition [2], and later discussed in [8, 11].

The transcendental equation for the required eigenvalue $\lambda\,(= 1/C_1^2)$ is a very simple one given by $\sqrt{\lambda}+\tan\sqrt{\lambda} = 0$. Thus the constant can be easily obtained with sufficient accuracy, and can be effectively used in the quantitative error estimation of finite element solutions by the P_1-element.

2. Preliminaries

Let T be a unit right-angled isosceles triangle defined by $T = \{x = (x_1, x_2) \in \mathbf{R}^2;\ x_1 > 0,\ x_2 > 0,\ x_1 + x_2 < 1\}$, the vertices of which are denoted by $Q_1(0,0)$, $Q_2(1,0)$ and $Q_3(0,1)$. Let us define the sets V_1 and V_2 by

$$V_1 = \{v \in H^2(T);\ v(Q_1) = v(Q_2) = v(Q_3) = 0\}, \tag{1}$$

$$V_2 = \left\{v \in H^1(T);\ \int_0^1 v(x_1, 0)\, dx_1 = 0\right\}, \tag{2}$$

where $H^1(T)$ and $H^2(T)$ are respectively the first- and second-order Sobolev spaces of real square integrable functions over T. Furthermore, the space $L_2(T)$ equipped with norm $\|\cdot\|_T$ will be also used. Notice here that $\partial v/\partial x_1 \in V_2$ for $\forall v \in V_1$.

For $v \in H^2(T)$, the linear interpolation function Πv on T is the (at most) linear polynomial such that $(\Pi v)(Q_i) = v(Q_i)$ for $1 \le i \le 3$. Then $v - \Pi v \in V_1$ for $v \in H^2(T)$, so that a popular form of the interpolation error for v is ([4, 5])

$$|v - \Pi v|_{1,T} \le C_0 |v|_{2,T}, \tag{3}$$

where $|\cdot|_{1,T}$ and $|\cdot|_{2,T}$ are the usual seminorms of $H^1(T)$ and $H^2(T)$, respectively: $|v|_{1,T}^2 = \sum_{i=1}^2 \|\partial v/\partial x_i\|_T^2$, $|v|_{2,T}^2 = \sum_{i,j=1}^2 \|\partial^2 v/\partial x_i \partial x_j\|_T^2$. Moreover, C_0 is the (optimal) positive constant well-defined by the relation [1, 2, 8]

$$C_0 = \sup_{v \in V_1 \setminus \{0\}} \frac{|v|_{1,T}}{|v|_{2,T}}. \tag{4}$$

Estimation (3) is effectively used for error analysis for triangular elements of more general shape by introducing appropriate coordinate transformations [2, 4, 5, 6].

It is actually difficult to decide C_0 exactly. An upper bound for C_0 was first given by Natterer [10]. By numerical computations without verification, it is now known that $C_0 \approx 0.489$, cf. [1, 7, 11]. Moreover, C_0 has an upper bound C_1 given by

$$C_1 = \sup_{v \in V_2 \setminus \{0\}} \frac{\|v\|_T}{|v|_{1,T}}. \tag{5}$$

This constant was introduced by Babuska and Aziz in [2] to prove the maximum angle condition for the P_1-element, so that we call it the Babuska-Aziz constant here. The relation between C_0 and C_1 was fully discussed in [11] and [8], and in certain cases C_1 is more essential than C_0 itself (see below). In particular, C_1 was verified numerically in [8, 9], and it is now known that $0.492 \leq C_1 \leq 0.494$. Thus 0.493 or so is a nice approximation to C_0 for most of practical purposes: In fact, 0.5 is recommended in [11] for use as an upper bound for C_0.

Furthermore, C_1 plays its own role of enabling estimation of each partial derivative:

$$\|\partial(v - \Pi v)/\partial x_i\|_T \leq C_1 \, |\partial v/\partial x_i|_{1,T} \qquad (i = 1, 2), \tag{6}$$

which is in a sense sharper than (3), cf. [6].

It is easily seen that C_1 is determined as $C_1 = \sqrt{1/\lambda_0}$, where $\lambda_0 > 0$ is the minimum eigenvalue of the eigenvalue problem [1, 8]: *Find* λ *and* $u \in V_2 \backslash \{0\}$ *that satisfy*

$$(\nabla u, \nabla v)_T = \lambda (u, v)_T \qquad (\forall v \in V_2). \tag{7}$$

Here, $(\cdot,\cdot)_T$ denotes the inner products of both $L_2(T)$ and $L_2(T)^2$, and ∇ is the gradient operator. The present eigenvalue problem is also expressed as follows in terms of a partial differential equation, the linear constraint for V_2 and boundary conditions [1, 8]:

$$-\Delta u = \lambda u \ \text{ in } T, \ \int_0^1 u(x_1, 0)\, dx_1 = 0, \ \frac{\partial u}{\partial n} = \begin{cases} 0 & \text{on edges } Q_1Q_3 \text{ and } Q_2Q_3, \\ c & \text{on edge } Q_1Q_2 \end{cases} \tag{8}$$

where $\frac{\partial}{\partial n}$ denotes the outward normal derivative on edges, and c an unknown constant. More specifically, c must finally satisfy the relation $c + \lambda \iint_T u(x_1, x_2)\, dx_1 dx_2 = 0$. Thus, $c = 0$ if $\iint_T u(x_1, x_2)\, dx_1 dx_2 = 0$. Otherwise, c can be equated to 1 at the expense of imposing the condition $\iint_T u(x_1, x_2)\, dx_1 dx_2 = -1/\lambda$ on the magnitude of u [8].

3. Determination of the Constant

Our problem of determining C_1 now reduces to obtaining the minimum eigenvalue of (7) or (8). Since T is a triangular domain, it has been not necessarily easy to solve the associated eigenvalue problem. However, we have the following main results.

THEOREM 1. *The minimum eigenvalue* λ_0 *of* (7) *is equal to the minimum positive solution of the transcendental equation for* λ:

$$\sqrt{\lambda} + \tan \sqrt{\lambda} = 0. \tag{9}$$

The concrete value of λ_0 *can be obtained numerically with verification. For example,* $\sqrt{\lambda_0}$ *lies in the interval* $2.0287 < \sqrt{\lambda_0} < 2.0291$, *and hence* $C_1 = 1/\sqrt{\lambda_0}$ *is bounded as*

$$0.49282 < C_1 < 0.49293. \tag{10}$$

(*Numerical computation without verification gives* $C_1 = 0.49291245\cdots$.)

Proof. We will prove in several steps, each of which is based on rather well-known arguments and techniques, and is sometimes described concisely.

1°. Let Ω be a unit square domain: $\Omega = \{x = (x_1, x_2) \in \mathbf{R}^2; 0 < x_1,\ x_2 < 1\}$. Let $\{\lambda, u\} \in \mathbf{R} \times V_2 \backslash \{0\}$ be an arbitrary eigenpair of (7), and define the (symmetric) extension $\tilde{u}$ of u to Ω by reflection with respect to the line $x_1 + x_2 = 1$:

$$\begin{aligned} \tilde{u}(x_1, x_2) &= u(x_1, x_2) \ \text{ if } \ x = (x_1, x_2) \in T, \\ \tilde{u}(x_1, x_2) &= u(1 - x_2, 1 - x_1) \ \text{ if } \ x \in \Omega \backslash T. \end{aligned}$$

It is easy to see that $\tilde{u}$ belongs to $H^1(\Omega)$. Moreover, $\{\lambda, \tilde{u}\}$ becomes an eigenpair of the eigenvalue problem for Ω:

$$\tilde{u} \in \tilde{V}_2 \backslash \{0\} \ \text{ and } \ (\nabla \tilde{u}, \nabla \tilde{v})_\Omega = \lambda (\tilde{u}, \tilde{v})_\Omega \quad (\forall \tilde{v} \in \tilde{V}_2), \tag{a}$$

where $(\cdot, \cdot)_\Omega$ denotes the inner products of $L_2(\Omega)$ and $L_2(\Omega)^2$, and $\tilde{V}_2$ is defined by

$$\tilde{V}_2 = \left\{ \tilde{v} \in H^1(\Omega);\ \int_0^1 \tilde{v}(x_1, 0)\, dx_1 = 0,\ \int_0^1 \tilde{v}(1, x_2)\, dx_2 = 0 \right\}. \tag{b}$$

Conversely, any eigenpair of (a) with $\tilde{u}$ restricted to T satisfies (7), if $\tilde{u}$ is symmetric with respect to the line $x_1 + x_2 = 1$. Notice here the orthogonal decomposition of $\tilde{V}_2$ in $H^1(\Omega)$ as well as in $L_2(\Omega)$:

$$\tilde{V}_2 = \tilde{V}_2^s \oplus \tilde{V}_2^a, \quad \begin{cases} \tilde{V}_2^s = \text{subspace of symmetric functions in } \tilde{V}_2, \\ \tilde{V}_2^a = \text{subspace of antisymmetric functions in } \tilde{V}_2. \end{cases}$$

Consequently, for the present purpose, it suffices to deal with (a) in $\tilde{V}_2^s$.

2°. As is well known, a complete system of functions for $H^1(\Omega)$ is given by the totality of (orthogonal) eigenfunctions of (a):

$$\phi_{mn}(x_1, x_2) = \cos m\pi x_1 \cos n\pi x_2 \quad (m, n = 0, 1, 2, 3, \ldots).$$

Since we are interested in symmetric eigenfunctions only, we should make a complete system of symmetric functions in $H^1(\Omega)$ from the above: for $m \geq n$; $m, n = 0, 1, 2, 3, \ldots$,

$$\begin{aligned} \psi_{mn}(x_1, x_2) &= \phi_{mn}(x_1, x_2) + \phi_{mn}(1 - x_2, 1 - x_1) \\ &= \phi_{mn}(x_1, x_2) + (-1)^{m+n} \phi_{nm}(x_1, x_2). \end{aligned}$$

When restricted to T, these make a complete system of functions for $H^1(T)$. Furthermore, these are orthogonal in $L_2(\Omega)$, and also orthogonal with respect to the bilinear form $(\nabla \cdot, \nabla \cdot)_\Omega$ (and in $H^1(\Omega)$).

3°. From (b), the condition for a symmetric $\tilde{v} \in H^1(\Omega)$ to belong to $\tilde{V}_2^s$ is expressed by

$$2a_{00} + \sum_{m=1}^{\infty} (-1)^m a_{m0} = 0 \quad \text{for} \quad \tilde{v} = \sum_{m \geq n \geq 0}^{\infty} a_{mn}\psi_{mn}$$

$$\text{with} \quad \sum_{m \geq n \geq 0}^{\infty} (1 + m^2 + n^2) a_{mn}^2 < +\infty,$$

where a_{mn}'s are real coefficients and the series $\sum_{m=1}^{\infty}(-1)^m a_{m0}$ is shown to be absolutely convergent. Eliminating a_{00} by this condition, we can express $\forall \tilde{v} \in \tilde{V}_2^s$ by

$$\tilde{v} = \sum_{m=1}^{\infty} a_{m0}[\psi_{m0} - (-1)^m] + \sum_{m \geq n \geq 1}^{\infty} a_{mn}\psi_{mn}. \tag{c}$$

Clearly, ψ_{mn}'s for $m \geq n \geq 1$ are eigenfunctions of (a) with completely homogeneous Neumann's boundary condition, and the minimum of the associated eigenvalues is $2\pi^2$.

4°. Taking notice of (c), $\tilde{V}_2^s$ is expressed by the direct sum

$$\tilde{V}_2^s = W_1 \oplus W_2, \begin{cases} W_1 = \text{closure of linear combinations of } \psi_{m0} - (-1)^m \\ \hfill (m = 1, 2, 3, \ldots), \\ W_2 = \text{closure of linear combinations of } \psi_{mn} \\ \hfill (m \geq n \geq 1). \end{cases}$$

Here, W_1 and W_2 are orthogonal to each other in both $L_2(\Omega)$ and $H^1(\Omega)$, and moreover, from the observation in 3°, all the eigenfunctions in W_2 are known. Consequently, our aim will be attained if we obtain the minimum of eigenvalues associated with eigenfunctions in W_1: If it is smaller than $2\pi^2$, the obtained one is nothing but λ_0.

5°. Let us now solve the eigenvalue problem (a) in W_1 by expressing $\tilde{u} \in W_1 \backslash \{0\}$ as

$$\tilde{u} = \sum_{m=1}^{\infty} a_m\varphi_m \quad \text{with} \quad \sum_{m=1}^{\infty} m^2 a_m^2 < +\infty\,; \quad \varphi_m = \psi_{m0} - (-1)^m \quad (m \in \mathbf{N}). \tag{d}$$

More specifically, $\varphi_m(x_1, x_2) = \cos m\pi x_1 + (-1)^m(\cos m\pi x_2 - 1)$. Substituting (d) into (a) and equating $\tilde{v}$ to each of φ_m's, we have the equations for coefficients a_m's:

$$(m^2\pi^2 - \lambda)a_m = \lambda(-1)^m \sum_{n=1}^{\infty} (-1)^n a_n \quad (m \in \mathbf{N}). \tag{e}$$

The series above is absolutely convergent from (d). In addition,

$$\sum_{n=1}^{\infty}(-1)^n a_n \neq 0,\ \lambda \neq m^2\pi^2 \ \text{ and } \ a_m \neq 0 \ \ (\forall m \in \mathbf{N}),$$

so that $a_m = \lambda(-1)^m \sum_{n=1}^{\infty}(-1)^n a_n/(m^2\pi^2 - \lambda)$. Multiplying $(-1)^m$ to both sides of this equation and summing up for $m \in \mathbf{N}$, we find that

$$\sum_{m=1}^{\infty}(-1)^m a_m = \left[\sum_{n=1}^{\infty}(-1)^n a_n\right] \times \sum_{m=1}^{\infty} \lambda/(m^2\pi^2 - \lambda),$$

that is,

$$1 = \sum_{m=1}^{\infty} \frac{\lambda}{m^2\pi^2 - \lambda}, \quad \text{or} \quad \sum_{m=0}^{\infty} \frac{1}{m^2(\pi/\sqrt{\lambda})^2 - 1} = 0, \tag{f}$$

where we have used the fact that $\lambda > 0$, and the series is absolutely convergent at least for $0 < \lambda < \pi^2$. We have thus shown that the considered eigenvalue λ must satisfy (f). Conversely, for each positive solution λ of (f), we can prove in view of (e) the existence of $\tilde{u} \in W_1\backslash\{0\}$ that satisfies (a), although we omit the proof.

6°. Notice here the formula for non-integer $a \in \mathbf{R}$:

$$\sum_{m=1}^{\infty} \frac{1}{(m/a)^2 - 1} = \frac{1}{2} - \frac{\pi a}{2\tan \pi a},$$

which is for example derived from the Fourier cosine expansion of $\cos at$ on the interval $[-\pi, \pi]$ for t. Comparing the above with (f) and taking a as $\sqrt{\lambda}/\pi$, we have (9). Clearly, the minimum positive solution of (9) lies in the interval $\left(\pi^2/4, \pi^2\right)$, and is the unique solution there. It is surely smaller than $2\pi^2$, and is exactly λ_0.

7°. To obtain $\sqrt{\lambda_0} \in (\pi/2, \pi)$ numerically with verification, we can use various methods. For example, we can directly use the series in (f). Here we just give another method based on modification of the equation $t + \tan t = 0$ for $t > 0$: Let us find the minimum positive zero of

$$f(t) := \frac{\cos t}{2} + \frac{\sin t}{2t} = \sum_{m=0}^{\infty} \frac{(-1)^m (m+1) t^{2m}}{(2m+1)!} \quad (t > 0).$$

The series appearing above is an alternating one, and the absolute value of each term for fixed t converges to 0 as $m \to \infty$, monotonically for sufficiently large m. Moreover, $f(t)$ is monotonically decreasing for $0 < t < \pi$. Thus, as is well known in elementary calculus, we can compute upper and lower bounds for the minimum zero t_0 by utilizing appropriate partial sums: $f_n(t)$:= partial sum up to the term of $m = n$. It should be noted here that, at least in principle, all the computations can be performed in the finite-digit binary arithmetic without computer errors, provided that t is a rational number. For example, by taking $n = 4, 5$, we can bound $t_0 = \sqrt{\lambda_0}$ as $2.0287 < t_0 < 2.0291$, since $f(2.0291) < f_4(2.0291) < 0$ (even n) and $f(2.0287) > f_5(2.0287) > 0$ (odd n). □

Remark. Eq. (9) can be also derived as follows. The function φ_m in (d) can be also expressed by $\varphi_m(x_1, x_2) = \cos m\pi x_1 + \cos m\pi(1 - x_2) - (-1)^m$, so that $\tilde{u} \in W_1$ in (d) must be of the form, for an unknown single-variable function $g = g(t)$,

$$\tilde{u}(x_1, x_2) = g(x_1) + g(1 - x_2).$$

Substituting the above into (a), we have

$$-g''(t) = \lambda g(t) \quad \text{for} \quad 0 < t < 1, \quad g'(0) = 0, \quad g(1) + \int_0^1 g(t)\, dt = 0. \tag{11}$$

Notice in this derivation that $\tilde{v}$ in (a) can be taken from whole $\tilde{V}_2$, since W_1 is orthogonal to W_2 and V_2^a both in $L_2(\Omega)$ and $H^1(\Omega)$. Solving this eigenvalue problem, we obtain (9). Moreover, an eigenfunction associated to λ_0 is $\tilde{u}(x_1, x_2) = \cos\sqrt{\lambda_0}x_1 + \cos\sqrt{\lambda_0}(1 - x_2)$. Eq. (9) is popular in vibration analysis of a string with one end fixed and the other supported elastically, where the governing differential equation is the same as in (11).

4. Concludig Remarks

We have succeeded in determining the Babuska-Aziz constant from a very simple equation. We can effectively utilize this constant to give upper bounds of the P_1-interpolation error constants for triangles of more general shape. That is, we can derive some explicit relations for the dependence of such constants on the geometry (such as the maximum interior angle and the minimum edge length) of triangles by simple coordinate transformations. It is to be noted that they are consistent with the so-called maximum angle condition in [2]. The detailed results will be reported separately in due course.

Acknowledgements. The authors would like to express their deepest appreciation to Prof. M.T. Nakao of Graduate School of Mathematical Sciences, Kyushu University and Prof. N. Yamamoto of Department of Computer Science, The University of Electro-Communication for acquainting them with the importance of the present problem and a number of references. Some of the exact computations in this paper were performed with the aid of Mathematica® 5. This work was partially supported by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (C) (2) 16540096, and also by Yoshida Scholarship Foundation for the second author.

References

[1] P. Arbenz, Computable finite element error bounds for Poisson's equation. IMA J. Numer. Anal., **2** (1982), 475–479.

[2] I. Babuska and A.K. Aziz, On the angle condition in the finite element method. SIAM J. Numer. Anal., **13** (1976), 214–226.

[3] R.E. Bahnhill, J.H. Brown and A.R. Mitchell, A comparison of finite element error bounds for Poisson's equation. IMA J. Numer. Anal., **1** (1981), 95–103.

[4] S.C. Brenner and L.R. Scott, The Mathematical Theory of Finite Element Methods (2nd edn.). Springer, 2002.

[5] P.-G. Ciarlet, The Finite Element Method for Elliptic Problems. SIAM, 2002.

[6] P. Knabner and L. Angermann, Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Springer, 2003.

[7] R. Lehmann, Computable error bounds in finite-element method. IMA J. Numer. Anal., **6** (1986), 265–271.

[8] M.T. Nakao and N. Yamamoto, A guaranteed bound of the optimal constant in the error estimates for linear triangular element. Computing [Supplementum], **15** (2001), 163–173.

[9] M.T. Nakao and N. Yamamoto, A guaranteed bound of the optimal constant in the error estimates for linear triangular elements, Part II: Details. Perspectives on Enclosure Methods (eds. U. Kulisch et al.), the Proceedings Volume for Invited Lectures of SCAN2000, Springer-Verlag, Vienna, 2001, 265–276.

[10] F. Natterer, Berechenbare Fehlerschranken für die Methode der finite Elemente. International Series of Numerical Mathematics, **28**, Birkhäuser, 1975, 109–121.

[11] G.L. Siganevich, On the optimal estimation of error of the linear interpolation on a triangle of functions from $W_2^2(T)$ (in Russian). Doklady Akademii Nauk SSSR, **300**, No.4 (1988), 811–814.