Publ. RIMS, Kyoto Univ.
Vol. 6 (1970), 323-341

A posteriori Error Estimates of Galerkin-approximate Solutions to the Third Boundary Value Problem for Elliptic Differential Equations*

By

Tetsuhiko MIYOSHI

§1. Introduction

In the present paper we shall show that in the third boundary value problem for elliptic partial differential equations one can get some formulas giving a posteriori error estimates to the approximate solutions obtained by Galerkin's method.

Trefftz has proposed in [12] an approximation method which can be used also for getting error estimates to the approximate solutions obtained by Galerkin's method. His method is based on the use of trial functions satisfying the given differential equations. For details of his method, see [3], [5], [12]. In practical applications, however, his method is not so convenient, because in general it is not easy to find functions satisfying the conditions requested for the trial functions.

On the other hand, in [4] and [9], Bramble, Payne and Weinberger gave integral inequalities which can be used for getting error estimates by the use of arbitrary functions satisfying only some smoothness conditions. However the error estimation based on their integral inequalities are not valid, say, when the coefficients of the given differential equation are not continuous.

Our method also is based on the use of some trial functions. They

Received May 19, 1970.

* This paper is based on author's lecture in the 'Symposium on Applied Mathematics' held Dec. 22-23, 1969 at Mikuruma Hall, Kyoto.

do not need to satisfy any condition except smoothness conditions. Error estimates obtained by our method, nevertheless, is valid when the coefficients are not continuous.

In our method, in order to get the formulas for error estimation, we use some integral inequalities connected with so-called 'equivalent norms' in Sobolev space $\mathring{W}_2^1(\Omega)$. Since our aim is to get practical error estimates of approximate solutions, the constants appearing in the integral inequalities must be estimated practically. In §2, following the method of Friedrichs [6], we shall derive these inequalities with the constants which can be evaluated practically.

In §3 we shall derive formulas for a posteriori error estimations for Galerkin-approximate solutions to self-adjoint and positive definite problems. Some numerical results obtained by our method will be shown in §4.

As well known, for self-adjoint and positive definite problems, if a suitable coordinate system is employed, Galerkin's method is, in a certain sense, an optimal approximation method in accuracy. In the appendix it will be shown that, although in a little weakened sense, the above property is preserved even when the problems lose the self-adjointness or the positive definiteness.

Throughout the present paper, a piecewise continuous function is meant only a function that is smooth on the closure of each related subinterval. When a piecewise continuous function is continuous on the whole interval, it is called a piecewise smooth function. The boundary of a domain in (x_1, x_2)-plane is called to be piecewise smooth if for any point p on the boundary there is a disk V centered at p such that the portion of the boundary inside V can be described in a suitable local coordinate system by an equation $x_2=f(x_1)$ with piecewise smooth $f(x_1)$. The terms piecewise continuous and piecewise smooth are used in a similar way for functions of two variables under the assumption that the boundaries of related subregions are piecewise smooth.

§2. Integral Inequalities

Let Ω be a bounded domain in (x_1, x_2)-plane with piecewise smooth

boundary Γ' and Γ^* be a portion of Γ. Then in many cases we have

$$(2.1) \qquad \|u\|^2 \leqslant C_1\|u\|^2_{\Gamma^*} + C_2 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2$$

and

$$(2.2) \qquad \|u\|^2_{\Gamma'} \leqslant C_3\|u\|^2_{\Gamma^*} + C_4 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2$$

for any function $u \in W_2^1(\Omega)$ ($W_2^k(\Omega)$ denotes the Sobolev space of functions). In (2.1) and (2.2) $\|u\|$, $\|u\|_{\Gamma^*}$ and $\|u\|_{\Gamma'}$ denote respectively $L^2(\Omega)$-, $L^2(\Gamma^*)$- and $L^2(\Gamma')$-norm of u and $\{C_i\}$ are some definite constants independent of function u.

If these inequalities are valid for smooth functions in $\bar{\Omega}$, then they are valid for any $u \in W_2^1(\Omega)$, because the class $C^1(\bar{\Omega})$ is dense in $W_2^1(\Omega)$ and the imbedding operator $W_2^1(\Omega) \to L^2(\Gamma')$ is continuous by the assumption of piecewise smoothness of the boundary Γ' (see, for example, [1], [11]).

First consider the closed region Ω_0 bounded by two smooth arcs

$$S_0^{(i)}: \ x_1 = g_i(x_2) \ (i=1, 2), \ \alpha \leqslant x_2 \leqslant \beta, \ g_1(x_2) < g_2(x_2)$$

and by two line segments

$$S_0^{(3)}: \ x_2 = \alpha, \ g_1(\alpha) \leqslant x_1 \leqslant g_2(\alpha),$$

$$S_0^{(4)}: \ x_2 = \beta, \ g_1(\beta) \leqslant x_1 \leqslant g_2(\beta).$$

Put

$$L = \max_{\alpha \leqslant x_2 \leqslant \beta} [g_2(x_2) - g_1(x_2)],$$

$$l = \min_{\alpha \leqslant x_2 \leqslant \beta} [g_2(x_2) - g_1(x_2)]$$

and suppose that

$$\left| \frac{d g_i(x_2)}{d x_2} \right| \leqslant T_0 \qquad (i=1, 2).$$

Then making use of the equality

$$(2.3) \qquad u(x_1, x_2) = u[g_1(x_2), x_2] + \int_{g_1(x_2)}^{x_1} \frac{\partial u}{\partial x_1}(t, x_2)\, dt,$$

by double integration we easily get

$$(2.4)\qquad \|u\|^2_{\Omega_0} \leqslant k_0^{(1)}\|u\|^2_{S_0^{(1)}} + k_0^{(2)}\left\|\frac{\partial u}{\partial x_1}\right\|^2_{\Omega_0} \quad (k_0^{(1)}=2L,\ k_0^{(2)}=L^2)$$

and

$$\text{(i)}\qquad \|u\|^2_{S_0^{(1)}} \leqslant \frac{2\sqrt{1+T_0^2}}{l}\|u\|^2_{\Omega_0} + L\sqrt{1+T_0^2}\left\|\frac{\partial u}{\partial x_1}\right\|^2_{\Omega_0}.$$

If we use instead of (2.3) the equality

$$u(x_1, x_2) = u[g_2(x_2), x_2] - \int_{x_1}^{g_2(x_2)} \frac{\partial u}{\partial x_1}(t, x_2)\,dt,$$

then we easily get

$$\text{(ii)}\qquad \|u\|^2_{S_0^{(2)}} \leqslant \frac{2\sqrt{1+T_0^2}}{l}\|u\|^2_{\Omega_0} + L\sqrt{1+T_0^2}\left\|\frac{\partial u}{\partial x_1}\right\|^2_{\Omega_0}.$$

Further, if we put

$$u[g_1(x_2)+s\{g_2(x_2)-g_1(x_2)\},\ x_2] = U(s, x_2) \qquad (0 \leqslant s \leqslant 1)$$

and make use of the equalities

$$U(s, x_2) = U(s, \alpha) + \int_\alpha^{x_2} \frac{\partial U}{\partial x_2}(s, t)\,dt$$

and

$$U(s, x_2) = U(s, \beta) - \int_{x_2}^{\beta} \frac{\partial U}{\partial x_2}(s, t)\,dt,$$

then by double integration we easily get

$$\text{(iii)}\qquad \left.\begin{matrix}\|u\|^2_{S_0^{(3)}} \\ \|u\|^2_{S_0^{(4)}}\end{matrix}\right\} \leqslant \frac{2L}{l(\beta-\alpha)}\|u\|^2_{\Omega_0} + \frac{L(\beta-\alpha)(1+T_0^2)}{l}\sum_{i=1}^{2}\left\|\frac{\partial u}{\partial x_i}\right\|^2_{\Omega_0}.$$

Adding (i), (ii) and (iii) and substituting (2.4) into the result we have

$$(2.5)\qquad \|u\|^2_{\Gamma_0} \leqslant k_0^{(3)}\|u\|^2_{S_0^{(1)}} + k_0^{(4)}\sum_{i=1}^{2}\left\|\frac{\partial u}{\partial x_i}\right\|^2_{\Omega_0},$$

where $\Gamma_0 = \bigcup_{i=1}^{4} S_0^{(i)}$ and $k_0^{(3)}$, $k_0^{(4)}$ are definite constants.

Now, suppose that, for example, $S_0^{(1)} = \Gamma^*$, then the inequalities (2.4) and

(2.5) are the desired inequalities of the type (2.1) and (2.2) for the region Ω_0.

Starting from the inequalities (2.4) and (2.5) we can obtain the desired inequalities for more complicated regions. For example, suppose that the region Ω_0 is connected with another closed region Ω_1 with piecewise smooth boundary Γ_1 in such a way that the intersection of Ω_0 and Ω_1 is only a curve S_1 and for any smooth function u in Ω_1 hold the following inequalities.

$$\|u\|^2_{\Omega_1} \leqslant \gamma_1^{(1)}\|u\|^2_{S_1} + \gamma_1^{(2)} \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2_{\Omega_1}, \tag{2.6}$$

$$\|u\|^2_{\Gamma_1} \leqslant \gamma_1^{(3)}\|u\|^2_{S_1} + \gamma_1^{(4)} \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2_{\Omega_1}, \tag{2.7}$$

where $\{\gamma_1^{(i)}\}$ are given definite constants. Then the inequalities (2.4), (2.5) and (2.6) imply the validity of the inequality of the type

$$\|u\|^2_{\Omega_0 \cup \Omega_1} \leqslant k_1^{(1)}\|u\|^2_{S_0^{(1)}} + k_1^{(2)} \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2_{\Omega_0 \cup \Omega_1}, \tag{2.8}$$

and (2.7) implies with (2.5)

$$\|u\|^2_{\Gamma_{01}} \leqslant k_1^{(3)}\|u\|^2_{S_0^{(1)}} + k_1^{(4)} \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2_{\Omega_0 \cup \Omega_1} \tag{2.9}$$

for any smooth function u in $\Omega_0 \cup \Omega_1$, where $\{k_1^{(i)}\}$ are definite constants and Γ_{01} denotes the boundary of $\Omega_0 \cup \Omega_1$. Therefore, if the closure of the domain Ω under consideration is constructed by finite number of subregions $\{\Omega_i\}$ which are obtained by continuing the above procedure, then it is evident that the inequalities of the type (2.1) and (2.2) hold for domain Ω and, of course, the constants are explicitly determined.

Remark 1. Let $\mathfrak{S}$ be a piecewise smooth arc lying in Ω. Then, by the above discussion we see that an inequality of the type

$$\|u\|^2_{\mathfrak{S}} \leqslant C_5\|u\|^2_{\Gamma^*} + C_6 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2 \tag{2.10}$$

will be easily obtained.

In fact, it is sufficient if we can get an inequality of the type

$$\|u\|_{\mathfrak{S}}^2 \leqslant C_7\|u\|^2 + C_8 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2 .$$

But this inequality has just the same type as inequalities (i), (ii) and (iii). Therefore the evaluation of these constants are not difficult for many cases.

§3. Error Estimates

Let us consider the equation

$$Lu \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) = f(x_1, x_2) \qquad \text{in } \Omega \tag{3.1}$$

under the boundary condition

$$\left[\sum_{i,j=1}^{2} a_{ij}\frac{\partial u}{\partial x_j}\cos(n, x_i) + \sigma u\right]_{\Gamma} = 0. \tag{3.2}$$

Here Ω is a bounded domain with piecewise smooth boundary Γ, and n is outward normal to Γ. The function $a_{ij}(a_{ij}=a_{ji})$ is piecewise continuous in Ω and for any $\xi \in R^2$

$$\sum_{i,j=1}^{2} a_{ij}\xi_i\xi_j \geqslant \delta_0 \sum_{i=1}^{2} \xi_i^2 \qquad (\delta_0 = \text{const.} > 0). \tag{3.3}$$

The function $f(x_1, x_2)$ is square summable over Ω. We assume for σ that it is non-negative and piecewise continuous on Γ, and there is a positive constant σ_0 and a portion Γ^* of Γ consisting of piecewise smooth arcs such that

$$\sigma \geqslant \sigma_0 > 0 \qquad \text{on } \Gamma^*. \tag{3.4}$$

Furthermore we assume that for any $u \in W_2^1(\Omega)$ and its trace to Γ hold the inequalities of the type (2.1) and (2.2).

Put

$$B[u, \phi] \equiv \sum_{i,j=1}^{2} \left(a_{ij}\frac{\partial u}{\partial x_j}, \frac{\partial \phi}{\partial x_i}\right) + \int_{\Gamma} \sigma u\phi \, ds \tag{3.5}$$

for $u, \phi \in W_2^1(\Omega)$, where $(u, v) \equiv \int_{\Omega} u \cdot v \, dx_1 \, dx_2$. Then the inequality (2.1) and the continuity of the imbedding operator $W_2^1(\Omega) \to L^2(\Gamma)$ implies that

the norm introduced by

$$\|u\|_H^2 \equiv B[u, u] \tag{3.6}$$

is equivalent to the norm in $W_2^1(\Omega)$. Therefore by Riesz's representation theorem the problem (3.1) with (3.2) has always a unique weak solution for any $f \in L^2(\Omega)$, that is, there exists a function $u \in W_2^1(\Omega)$ satisfying the equation

$$B[u, \phi] = (f, \phi) \qquad \text{for any } \phi \in W_2^1(\Omega) \quad (f \in L^2(\Omega)). \tag{3.7}$$

Let $\boldsymbol{\phi}_n = (\phi_1, \phi_2, \cdots, \phi_n)$ $(\phi_i \in W_2^1(\Omega))$ be a vector consisting of linearly independent functions. In Galerkin's method we seek the approximate solution of order n in the form

$$u_n = \sum_{i=1}^{n} \alpha_i \phi_i \tag{3.8}$$

and determine the coefficients $\{\alpha_i\}$ by solving the system of equations

$$B[u_n, \phi_i] = (f, \phi_i) \qquad i = 1, 2, \cdots, n. \tag{3.9}$$

(**a**) First we consider such case that all of the coefficients $\{a_{ij}\}$ are smooth. Let us introduce two functional $F(\phi)$ and $G(\psi)$ by defining

$$F(\phi) = \|\phi\|_H^2 - 2(f, \phi) \qquad \text{for } \phi \in W_2^1(\Omega) \tag{3.10}$$

and

$$G(\psi) = F(\psi) - k_1 \|f - L\psi\|^2 - k_2 \left\| \frac{\partial \psi}{\partial \nu} + \sigma \psi \right\|_\Gamma^2 \qquad \text{for } \psi \in W_2^2(\Omega), \tag{3.11}$$

where $\dfrac{\partial \psi}{\partial \nu}$ denotes the co-normal derivative $\sum a_{ij} \dfrac{\partial \psi}{\partial x_j} \cos(n, x_i)$ and

$$k_1 = \max\left(\frac{2C_1}{\sigma_0}, \frac{2C_2}{\delta_0} \right), \quad k_2 = \max\left(\frac{2C_3}{\sigma_0}, \frac{2C_4}{\delta_0} \right). \tag{3.12}$$

Theorem 1. *Let $u \in W_2^1(\Omega)$ be the weak solution of the boundary value problem* (3.1), (3.2). *Then hold the following estimates*:

(I) $$\min_{W_2^1(\Omega)} F(\phi) = -(f, u) \geqslant \sup_{W_2^2(\Omega)} G(\psi),$$

(II) $$\|u - \phi\|_H^2 = F(\phi) + (f, u) \leqslant F(\phi) - G(\psi)$$

for any $\phi\in W_2^1(\Omega)$ *and* $\psi\in W_2^2(\Omega)$. *If the solution* u *belongs to* $W_2^2(\Omega)$, *then we have*

(I)′ $$\min_{W_2^2(\Omega)} F(\phi)=-(f,\ u)=\max_{W_2^2(\Omega)} G(\psi).$$

Proof. By generalized Green's identity, for any $\psi\in W_2^2(\Omega)$

$$\begin{aligned}\|u-\psi\|_H^2&=\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial}{\partial x_j}(u-\psi),\ \frac{\partial}{\partial x_i}(u-\psi)\right)+\int_\Gamma\sigma(u-\psi)^2\,ds\\&=\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial u}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right)+\int_\Gamma\sigma u(u-\psi)\,ds\\&\quad-\left\{\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial\psi}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right)+\int_\Gamma\sigma\psi(u-\psi)\,ds\right\}\\&=(f,\ u-\psi)-\left\{(L\psi,\ u-\psi)+\int_\Gamma\left(\frac{\partial\psi}{\partial\nu}+\sigma\psi\right)(u-\psi)\,ds\right\}.\end{aligned}$$

Therefore, by Schwarz's inequality

$$\|u-\psi\|_H^2\leqslant\|f-L\psi\|\cdot\|u-\psi\|+\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma\cdot\|u-\psi\|_\Gamma. \tag{3.13}$$

On the other hand, by (2.1) and (2.2) we have

$$\|u-\psi\|\leqslant\sqrt{\frac{k_1}{2}}\|u-\psi\|_H,\quad \|u-\psi\|_\Gamma\leqslant\sqrt{\frac{k_2}{2}}\|u-\psi\|_H.$$

Substituting these inequalities into (3.13) we have

$$\|u-\psi\|_H^2\leqslant k_1\|f-L\psi\|^2+k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma^2. \tag{3.14}$$

Therefore, for any $\psi\in W_2^2(\Omega)$ it holds that

$$\begin{aligned}-(f,\ u)&=F(\psi)-\|u-\psi\|_H^2\\&\geqslant F(\psi)-k_1\|f-L\psi\|^2-k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma^2=G(\psi).\end{aligned} \tag{3.15}$$

The theorem follows immediately from these relations.

(**b**) Now, if one of the coefficients $\{a_{ij}\}$ is only piecewise continuous in Ω, the above estimates lose it's validity because the function $L\psi$ is not integrable. We propose another formula valid for such case too.

Let $\mathfrak{S}$ be the set consisting of all discontinuous points of $\{a_{ij}\}$ and

$\Omega=\sum_{k=1}^{N}\oplus R_k$ be a direct sum of subregions, in each of which $\{a_{ij}\}$ are smooth (we can assume that the boundary of R_k is piecewise smooth). Let us assume that

$$(3.16)\qquad \|u\|_{\mathfrak{S}}^2 \leqslant C_5\|u\|_{\Gamma^*}^2+C_6\sum_{i=1}^{2}\left\|\frac{\partial u}{\partial x_i}\right\|^2 \qquad \text{for any } u\in \mathring{W}_2^1(\Omega),$$

where C_5 and C_6 are explicitly given constants (see the remark in §2). Of course, we assume the validity of the inequalities (2.1) and (2.2) for the domain Ω.

For any $\psi\in\{\psi\in W_2^1(\Omega);\ \psi\in W_2^2(\Omega-\overline{\mathfrak{S}})\}$ holds the equality

$$(3.17)\qquad \|u-\psi\|_H^2=(f,\ u-\psi)-\left\{\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial\psi}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right)+\int_\Gamma \sigma\psi(u-\psi)\,ds\right\},$$

where u is the exact solution. By generalized Green's identity

$$(3.18)\qquad \begin{aligned}&\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial\psi}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right)\\&=\sum_{i,j=1}^{2}\left[\sum_{k=1}^{N}\left(a_{ij}\frac{\partial\psi}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right)_{R_k}\right]\\&=\sum_{k=1}^{N}\left[(L\psi,\ u-\psi)_{R_k}+\int_{\partial R_k}\left(\sum_{i,j=1}^{2}a_{ij}\frac{\partial\psi}{\partial x_j}\cos(n,\ x_i)\right)(u-\psi)\,ds\right]\\&=\sum_{k=1}^{N}(L\psi,\ u-\psi)_{R_k}+\int_\Gamma\frac{\partial\psi}{\partial\nu}(u-\psi)\,ds\\&\quad+\sum_{k=1}^{N}\int_{\partial R_k-\Gamma}\left(\sum_{i,j=1}^{2}a_{ij}\frac{\partial\psi}{\partial x_j}\cos(n,\ x_i)\right)(u-\psi)\,ds,\end{aligned}$$

where ∂R_k denotes the boundary of region R_k.

Here we may write that

$$(3.19)\qquad \begin{aligned}&\int_{\partial R_k-\Gamma}\left(\sum_{i,j=1}^{2}a_{ij}\frac{\partial\psi}{\partial x_j}\cos(n,\ x_i)\right)(u-\psi)\,ds\\&\quad+\int_{\partial(\Omega-R_k)-\Gamma}\left(\sum_{i,j=1}^{2}a_{ij}\frac{\partial\psi}{\partial x_j}\cos(n,\ x_i)\right)(u-\psi)\,ds\end{aligned}$$

$$=\int_{\partial R_k-\Gamma}\left\{\sum_{i,j=1}^{2}\left[a_{ij}^{(+k)}\left(\frac{\partial\psi}{\partial x_j}\right)^{(+k)}-a_{ij}^{(-k)}\left(\frac{\partial\psi}{\partial x_j}\right)^{(-k)}\right]\times\right.$$

$$\left.\times\cos(n,\ x_i)\right\}(u-\psi)\,ds,$$

where $(+k)$ and $(-k)$ are symbols to denote the traces of functions to $\partial R_k-\Gamma$ from inside of R_k and to $\partial(\Omega-R_k)-\Gamma(=\partial R_k-\Gamma$ as a set) from inside of $\Omega-R_k$ respectively. For brevity, let us introduce two functions $\widetilde{L\psi}$ and $\left[\frac{\partial\psi}{\partial\nu}\right]'$ by defining

$$\widetilde{L\psi}\equiv\{L\psi \text{ in } R_k-\partial R_k \qquad (k=1,\ 2,\ \cdots,\ N)\} \tag{3.20}$$

and

$$\left[\frac{\partial\psi}{\partial\nu}\right]'\equiv\left\{\sum_{i,j=1}^{2}\left[a_{ij}^{(+k)}\left(\frac{\partial\psi}{\partial x_j}\right)^{(+k)}-a_{ij}^{(-k)}\left(\frac{\partial\psi}{\partial x_j}\right)^{(-k)}\right]\cos(n,\ x_i)\right. \tag{3.21}$$

$$\left.\text{on } \partial R_k-\Gamma \quad (k=1,\ 2,\ \cdots,\ N)\right\}$$

for $\psi\in\{\psi\in W_2^1(\Omega);\ \psi\in W_2^2(\Omega-\overline{\mathfrak{S}})\}$. By suitably defining its value on the set of measure zero, each of these functions can be regarded as a function belonging to $L^2(\Omega)$ and $L^2(\sum_k\partial R_k-\Gamma)$ respectively. By (3.18) and (3.19) we have

$$\sum_{i,j=1}^{2}\left(a_{ij}\frac{\partial\psi}{\partial x_j},\ \frac{\partial}{\partial x_i}(u-\psi)\right) \tag{3.22}$$

$$=(\widetilde{L\psi},\ u-\psi)+\int_\Gamma\frac{\partial\psi}{\partial\nu}(u-\psi)\,ds+\int_{\mathfrak{S}}\left[\frac{\partial\psi}{\partial\nu}\right]'(u-\psi)\,ds.$$

Substituting this into (3.17) and using Schwarz's inequality we have

$$\|u-\psi\|_H^2\leqslant\|f-\widetilde{L\psi}\|\cdot\|u-\psi\|+\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma\cdot\|u-\psi\|_\Gamma \tag{3.23}$$

$$+\left\|\left[\frac{\partial\psi}{\partial\nu}\right]'\right\|_{\mathfrak{S}}\cdot\|u-\psi\|_{\mathfrak{S}}.$$

Therefore, by (2.1), (2.2) and (3.16) we have

$$\|u-\psi\|_H^2\leqslant\frac{3}{2}\left(k_1\|f-\widetilde{L\psi}\|^2+k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma^2+k_3\left\|\left[\frac{\partial\psi}{\partial\nu}\right]'\right\|_{\mathfrak{S}}^2\right), \tag{3.24}$$

where

$$k_3=\max\left(\frac{2C_5}{\sigma_0}, \frac{2C_6}{\delta_0}\right). \tag{3.25}$$

From (3.24), we then get the following theorem in a similar way as for theorem 1.

Theorem 2. *Let us set*

$$\tilde{G}(\psi)\equiv F(\psi)-\frac{3}{2}\left(k_1\|f-\widetilde{L}\psi\|^2+k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_{\Gamma}^2+k_3\left\|\left[\frac{\partial\psi}{\partial\nu}\right]'\right\|_{\mathfrak{S}}^2\right), \tag{3.26}$$

where k_1, k_2 and k_3 are the constants defined by (3.12) *and* (3.25). *Then for any $\phi\in W_2^1(\Omega)$ and $\psi\in\{\psi\in W_2^1(\Omega);\ \psi\in W_2^2(\Omega-\overline{\mathfrak{S}})\}$ holds the estimate*

$$\|u-\phi\|_H^2\leqq F(\phi)-\tilde{G}(\psi), \tag{3.27}$$

where u is the exact solution of the problem (3.7).

Now, let u_n be the solution of (3.9). Then

$$F(u_n)=\|u_n\|_H^2-2(f,\ u_n)=-(f,\ u_n).$$

Therefore, for example, in case (b), the error of Galerkin-approximate solution u_n is estimated by

$$\|u-u_n\|_H^2\leqq-(f,\ u_n)-\tilde{G}(\psi),$$

where ψ is arbitrary function satisfying the condition in Theorem 2. Note that u_n minimize the functional F in the linear manifold spanned by the functions $\{\phi_i\}$ $(i=1, 2, \cdots, n)$.

Remark 2. If the boundary condition (3.2) is inhomogeneous,

$$\left[\sum_{i,j=1}^{2}a_{ij}\frac{\partial u}{\partial x_j}\cos(n, x_i)+\sigma u\right]=b \qquad b\in L^2(\Gamma), \tag{3.2}'$$

then the functional F and $\tilde{G}$ in Theorem 2 take the following expressions respectively.

$$F(\phi)=\|\phi\|_H^2-2(f,\ \phi)-2\int_{\Gamma}b\phi\,ds, \tag{3.10}'$$

$$\tilde{G}(\psi)=F(\psi)-\frac{3}{2}\left(k_1\|f-\widetilde{L}\psi\|^2+k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi-b\right\|_{\Gamma}^2\right. \tag{3.26}'$$

$$+k_3\left\|\left[\frac{\partial\psi}{\partial\nu}\right]'\right\|_{\mathfrak{S}}^2\Big).$$

Remark 3. Let us put $f=0$ in (3.7). Then, clearly $u=0$. Therefore by (3.14) we have

$$\|\psi\|_H^2\leqslant k_1\|L\psi\|^2+k_2\left\|\frac{\partial\psi}{\partial\nu}+\sigma\psi\right\|_\Gamma^2 \qquad \text{for any } \psi\in W_2^2(\Omega),$$

which is similar to the results in [4], [9]. By using this inequality we can again estimate the error of approximate solutions. But this inequality can not be used when the exact solution belongs merely to $W_2^1(\Omega)$, because the boundary value of $\partial u/\partial x_i$ can not be well defined for such case and the expression $\partial u/\partial\nu$ has no positive meaning.

§4. Numerical Example

Let Ω be the unit square $0<x_1, x_2<1$ and Γ be its boundary. We approximate the solution of the problem

$$-\frac{\partial}{\partial x_1}\left(a\frac{\partial u}{\partial x_1}\right)-\frac{\partial}{\partial x_2}\left(a\frac{\partial u}{\partial x_2}\right)=1 \qquad \text{in } \Omega$$

$$\left[\frac{\partial u}{\partial\nu}+\sigma u\right]_\Gamma=0,$$

where the function a and σ satisfy the following conditions.

$$a=\begin{cases}\tilde{a} & \text{in } R_1\\ 1 & \text{in } R_2\end{cases} \qquad R_1=\{(x_1, x_2);\ 0\leqslant x_1, x_2\leqslant 0.5\} \qquad R_2=\bar{\Omega}-R_1,$$

$$\sigma=\begin{cases}1 & \text{on } \Gamma_1\\ \tilde{\sigma} & \text{on } \Gamma_2\end{cases} \qquad \Gamma_1=\{(x_1, x_2)\in\Gamma;\ x_1=0 \text{ or } x_2=0\} \qquad \Gamma_2=\Gamma-\Gamma_1.$$

The approximate solution is sought in the form

$$u_{28}(x_1, x_2)=\sum_{i+j\leqslant 6} a_{ij}x_1^i x_2^j$$

determining $\{a_{ij}\}$ by solving the system of equations (3.9).

The constants necessary for error estimation are evaluated as follows. By (2.4) and (i) in §2 it is easy to see that

$$\|u\|_S^2 \leqq \|u\|_{\Gamma_1}^2 + 0.5 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2 \tag{5.1}$$

and

$$\|u\|_{\Gamma_2}^2 \leqq 4\|u\|_{\Gamma_1}^2 + 3 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2. \tag{5.2}$$

The later inequality implies inequality

$$\|u\|_{\Gamma}^2 \leqq 5\|u\|_{\Gamma_1}^2 + 3 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2. \tag{5.3}$$

Put $\mathfrak{S} = \partial R_1 - \Gamma$ and $S = \{(x_1, x_2) \in \Omega;\ x_1 = 0.5\}$. Then by (i) in §2 we see that

$$\|u\|_S^2 \leqq 2\|u\|^2 + 0.25 \left\| \frac{\partial u}{\partial x_1} \right\|^2, \tag{5.4}$$

thus

$$\begin{aligned} \|u\|_{\mathfrak{S}}^2 &\leqq 4\|u\|^2 + 0.25 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2 \\ &\leqq 4\|u\|_{\Gamma_1}^2 + 2.25 \sum_{i=1}^{2} \left\| \frac{\partial u}{\partial x_i} \right\|^2. \end{aligned} \tag{5.5}$$

Therefore the constants are estimated $C_1 = 1.0$, $C_2 = 0.5$, $C_3 = 5.0$, $C_4 = 3.0$, $C_5 = 4.0$ and $C_6 = 2.25$.

Let $\bar{u}_{28}$ be the Galerkin-approximate solution and $\underline{u}_{28}$ be the function maximizing the functional $\tilde{G}(u_{28})$. We computed the values $F(\bar{u}_{28})$, $\tilde{G}(\underline{u}_{28})$ and $\tilde{G}(\bar{u}_{28})$ for various $\tilde{a}$ and $\tilde{\sigma}$. The result is shown in Table I. For $\tilde{a} = 1$, $\tilde{\sigma} = 1$ the bound of error of the approximate solution $\bar{u}_{28}$ is given by

$$\begin{aligned} \|u - \bar{u}_{28}\|_H^2 &\equiv \sum_{i=1}^{2} \left\| \frac{\partial}{\partial x_i}(u - \bar{u}_{28}) \right\|^2 + \int_{\Gamma} (u - \bar{u}_{28})^2 ds \\ &\leqq F(\bar{u}_{28}) - \tilde{G}(\underline{u}_{28}) \\ &= -0.2905227 + 0.2905229 \leqq 3.0 \times 10^{-7}. \end{aligned}$$

Since polynomials are used as coordinate functions, the estimates become poor with the decreasing of $\tilde{a}$ or $\tilde{\sigma}$. This is caused by the

discontinuity of the derivatives of the exact solution. Therefore, if we want more accurate estimates for such cases, another suitable coordinate functions, for example, suitable functions belonging to the class $C(\bar{\Omega}) \cap C^2(\bar{R}_1) \cap C^2(\bar{R}_2)$, must be employed.

Table I. Estimates of $F(\bar{u}_{28})$, $\tilde{G}(u_{28})$ and $\tilde{G}(\bar{u}_{28})$.

$\tilde{\sigma}$ \ $\tilde{a}$	1.0	0.8	0.6
1.0	−0.2905227 −0.2905229 −0.2905256	−0.2928 −0.2944 −0.2990	−0.2960 −0.3025 −0.3357
0.8	−0.3193 −0.3199 −0.3201	−0.3221 −0.3295 −0.3309	−0.3260 −0.3862 −0.3954
0.6	−0.3584 −0.3617 −0.3627	−0.3620 −0.3725 −0.3767	−0.3669 −0.4345 −0.4537

The computation has been carried out by the use of TOSBAC 3400 at R.I.M.S., Kyoto University.

APPENDIX: A Remark on the Optimality of Galerkin's Method

Let H be denoted the space $\mathring{W}_2^1(\Omega)$ renormed by the norm $\|\cdot\|_H$ defined by (3.6). Let $u \in \mathring{W}_2^1(\Omega)$ be the solution of the problem (3.7). Then the operator $G: f(\in L^2(\Omega)) \to u(\in H)$ is linear and continuous. We use the notation $\|G\|_H$ to denote the operator-norm of $G: L^2(\Omega) \to H$. Note that G can be regarded as a homeomorphism of H' (dual space of H) on to H. When we approximate the solution of problem (3.7) in the form (3.8), the error of best approximation (in the norm of H) is given obviously by

$$u - P_{\phi_n} u = (I - P_{\phi_n}) Gf,$$

where P_{ϕ_n} denotes the orthogonal projection of H onto the linear manifold spanned by the functions $\{\phi_i\}$ $(i=1, 2, \cdots, n)$.

Definition 1. A set of functions $\boldsymbol{\xi}_n = (\xi_1, \xi_2, \cdots, \xi_n) \in H$ is called an optimal n-dimensional coordinate system for problem (3.7) if

(1) $$\|(I-P_{\xi_n})G\|_H \leqslant \|(I-P_{\phi_n})G\|_H \qquad \text{for any } \boldsymbol{\phi}_n \in H.$$

Now consider the eigenvalue problem

(2) $$G\phi=\lambda\phi \qquad \phi \in L^2(\Omega).$$

The operator G is self-adjoint, positive definite and completely continuous as an operator $L^2(\Omega)\to L^2(\Omega)$, since the imbedding $H\to L^2(\Omega)$ is completely continuous [11]. Therefore the eigenvalues of (2) can be numbered so that

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant \cdots > 0$$

and the corresponding eigenfunctions $e_1, e_2, \cdots, e_n, \cdots$ make a complete system both in $L^2(\Omega)$ and H. It is then easily seen that

(3) $$\|(I-P_{e_n})G\|_H=\sqrt{\lambda_{n+1}}.$$

On the other hand, for an arbitrary $\boldsymbol{\phi}_n=(\phi_1, \phi_2, \cdots, \phi_n)\in H$ consider a function v such that $v=\sum_{i=1}^{n+1} a_i e_i$, $\|v\|=1$ and $B[Gv, \phi_i]=0\,(i=1, 2, \cdots, n)$. Then we can easily prove that $\|(I-P_{\phi_n})Gv\|_H \geqslant \sqrt{\lambda_{n+1}}$ [2]. By (3) this implies that

$$\|(I-P_{e_n})G\|_H \leqslant \|(I-P_{\phi_n})G\|_H,$$

that is, $\boldsymbol{e}_n=(e_1, e_2, \cdots, e_n)$ is an optimal coordinate system.

For an approximate solution u_n obtained by Galerkin's method, we have by (3.9)

(4) $$u_n=P_{\phi_n}u=P_{\phi_n}Gf.$$

Therefore, if an optimal coordinate system is employed we have

(5) $$\|u-u_n\|_H \leqslant \inf_{\phi_n\in H}\{\|(I-P_{\phi_n})G\|_H\}\cdot\|f\|.$$

This inequality shows that, if an optimal coordinate system, say, the system consisting of eigenfunctions of operator G, is used, the procedure in Galerkin's method is optimal in accuracy for general $f\in L_2(\Omega)$.

Our purpose in this appendix is to extend these results to more general problem

(6) $$B[u, \phi]+(Ku, \phi)=(f, \phi) \qquad \text{for any } \phi\in H(f\in L^2(\Omega)),$$

where K is a differential operator of first order with bounded, measurable coefficients. We assume that this problem has always a unique solution $u \in H$ for any $f \in L^2(\Omega)$.

By Riesz's representation theorem, equation (6) can be represented by the equation

$$(u, \phi)_H + (GKu, \phi)_H = (Gf, \phi)_H \qquad \text{for any } \phi \in H$$

or

$$u + GKu = Gf. \tag{7}$$

Since $GK: H \to H$ is completely continuous (see, e.g., [8]), by Riesz-Schauder's theory the assumption of unique solvability of problem (6) implies the unique solvability of the equation

$$u + GKu = v \qquad v \in H$$

and thus the operator $(I+GK)$ has bounded inverse $(I+GK)^{-1}: H \to H$ by Banach's theorem. Therefore, the solution of (7) is given by $u = (I+GK)^{-1}Gf = [G^{-1}(I+GK)]^{-1}f = G(I+KG)^{-1}f$.

Let us put

$$G' = G(I+KG)^{-1}. \tag{8}$$

Note that the operator $(I+KG)$ is not only a homeomorphism of H' onto H', but of $L^2(\Omega)$ onto $L^2(\Omega)$. Because, the equation

$$u + KGu = f \qquad f \in L^2(\Omega)$$

has always a unique solution u belonging at least to H', but this implies $u = f - KGu \in L^2(\Omega)$.

Corresponding to definition 1 we put

Definition 2. A set of functions $\boldsymbol{\xi}_n = (\xi_1, \xi_2, \cdots, \xi_n) \in H$ is called a quasi-optimal n-dimensional coordinate system for problem (6) if

$$\|(I - P_{\xi_n})G'\|_H \leqslant C\|(I - P_{\phi_n})G'\|_H \qquad \text{for any } \boldsymbol{\phi}_n \in H, \tag{9}$$

where C is a constant independent of n and function $\boldsymbol{\xi}_n$.

Theorem (A). *The vector $\boldsymbol{e}_n$ of the first n eigenfunctions of the*

eigenvalue problem (2) *is a quasi-optimal coordinate system for problem* (6) *and*

$$c_1\sqrt{\lambda_{n+1}} \leqq \|(I-P_{e_n})G'\|_H \leqq c_2\sqrt{\lambda_{n+1}} \tag{10}$$

$$c_1=\|I+KG\|^{-1} \quad c_2=\|(I+KG)^{-1}\|,$$

where λ_{n+1} *is the* $(n+1)$*-th eigenvalue of the eigenvalue problem* (2).

Proof. The theorem can be proved by almost similar way to the self-adjoint, positive definite case [2].

Let $\lambda_1 \geqq \lambda_2 \geqq \cdots > 0$ be the eigenvalues of the problem (2) and $\{e_i\}$ $(i=1, 2, \cdots)$ be the system of orthonormalized eigenfunctions corresponding to these eigenvalues. Since $\{e_i\}$ is complete both in $L^2(\Omega)$ and in H, for any $f \in L^2(\Omega)$

$$\begin{aligned} G'f &= \sum_{i=1}^{\infty} (G'f, e_i)\, e_i = \sum_{i=1}^{\infty} ((I+KG)^{-1}f, Ge_i)\, e_i \\ &= \sum_{i=1}^{\infty} \sqrt{\lambda_i}\,((I+KG)^{-1}f, e_i)\sqrt{\lambda_i}\, e_i, \end{aligned} \tag{11}$$

and by Parseval's equality,

$$\begin{aligned} \|G'f-P_{e_n}G'f\|_H^2 &= \sum_{i=n+1}^{\infty} \lambda_i((I+KG)^{-1}f, e_i)^2 \\ &\leqq \lambda_{n+1} \sum_{i=n+1}^{\infty} ((I+KG)^{-1}f, e_i)^2 \\ &\leqq \lambda_{n+1}\|(I+KG)^{-1}f\|^2 \leqq \lambda_{n+1}\, c_2^2 \|f\|^2 \end{aligned}$$

which establishes the second inequality in (10). On the other hand, for $f=(I+KG)e_{n+1}$ in (11) we have $G'f=\lambda_{n+1}\, e_{n+1}$, therefore

$$\begin{aligned} \lambda_{n+1} &= \|G'f-P_{e_n}G'f\|_H^2 \leqq \|(I-P_{e_n})G'\|_H^2 \|f\|^2 \\ &\leqq \|(I-P_{e_n})G'\|_H^2 \|I+KG\|^2 \end{aligned}$$

from which follows the first inequality in (10).

Quasi-optimality of e_n: Let $\phi_n=(\phi_1, \phi_2, \cdots, \phi_n) \in H$ be arbitrary vector and $v=\sum_{i=1}^{n+1} a_i e_i$ be the vector satisfying $\|v\|=1$ and $B[Gv, \phi_i]=0$ $(i=1, 2, \cdots, n)$. Then we have

$$\|(I-P_{\phi_n})G'(I+KG)v\|_H^2$$

$$=\|(I-P_{\phi_n})Gv\|_H^2=\|Gv\|_H^2=\sum_{i=1}^{n+1}\lambda_i a_i^2\geqslant\lambda_{n+1}.$$

Therefore we have

$$\sqrt{\lambda_{n+1}}\leqslant\|(I-P_{\phi_n})G'\|_H\|I+KG\|,$$

that is,

$$c_1\sqrt{\lambda_{n+1}}\leqslant\|(I-P_{\phi_n})G'\|_H.$$

Then from (10) we readily get

$$\|(I-P_{e_n})G'\|_H\leqslant\frac{c_2}{c_1}\|(I-P_{\phi_n})G'\|_H$$

which completes the proof of the theorem.

Now, let us apply Galerkin's method to the problem (6). Let $\{\phi_i\}$ be a complete system of linearly independent functions. We seek the approximate solution of order n in the form $u_n=\sum_{i=1}^{n}\alpha_i\phi_i$, and determine the coefficients $\{\alpha_i\}$ by solving the system of equations

$$B[u_n,\phi_i]+(Ku_n,\phi_i)=(f,\phi_i)\qquad(i=1,2,\cdots,n). \tag{12}$$

Clearly this system of equations is equivalent to the equation

$$u_n+P_{\phi_n}GKu_n=P_{\phi_n}Gf \tag{13}$$

which approximate the original equation (7). Hence the well known theory of approximation method is applicable and we can verify that for sufficiently large n the system of equations (12) has always a unique solution u_n and holds the following asymptotic error estimate.

$$\|u-u_n\|_H\leqslant\text{const}\cdot\|u-P_{\phi_n}u\|_H \tag{14}$$

(for more detail proof, see [7], [10]). Therefore, if we employ a quasi-optimal coordinate system, then we have

$$\|u-u_n\|\leqslant\text{const}\cdot\{\inf_{\phi_n\in H}\|(I-P_{\phi_n})G'\|\}\|f\|. \tag{15}$$

This inequality shows that, if we use a quasi-optimal coordinate system, say, the system consisting of eigenfunctions of operator G, the procedure

in Galerkin's method is quasi-optimal in accuracy for problem (6).

Remark. The above discussion is applicable to first and second boundary value problems without essential modification.

Acknowledgement

The author wishes to thank Professor M. Urabe for his criticisms and help in this work. He is also grateful to Professor M. Yamaguti for valuable advices.

References

[1] Agmon, S., Lectures on Elliptic Boundary Value Problems, Van Nostrand, 1965, §2.
[2] Babuska, I., M. Prager and E. Vitasek, Numerical Processes in Differential Equations, Interscience, 1966.
[3] Bittner, L., Abschätzungen bei Variationsmethoden mit Hilfe von Dualitätssatzen I, Numer. Math. **11** (1968), 129-143.
[4] Bramble, J. I. and L. E. Payne, Some integral inequalities for uniformly elliptic operators, Contributions to Diff. Eqs. **1** (1963), 129-135.
[5] Collatz, L., The Numerical Treatment of Differential Equations, 3rd ed., Springer, 1966, Chap. V §6.
[6] Friedrichs, K., Die Randwert- und Eigenwertprobleme aus der Theorie der elastischen Platten (Anwendungen der direkten Methoden der Variationsrechnung), Math. Ann. **98** (1928), 205-247.
[7] Kantorovich, L. V. and G. P. Akilov, Functional Analysis in Normed Spaces, Pergamon, 1964, Chap. XIV.
[8] Mikhlin, S. G., Variational Methods in Mathematical Physics, Pergamon, 1964, Chap. IX.
[9] Payne, L. E. and H. F. Weinberger, New bounds for solutions of second order elliptic partial differential equations, Pacific J. Math. **8** (1958), 551-573.
[10] Polsky, N. I., On the convergence of certain approximate methods of analysis, Ukrain. Mat. Ž. **7** (1955), 41-56.
[11] Smirnov, V. I., A Course of Higher Mathematics, Pergamon, 1964, Vol. V, Chap. IV.
[12] Trefftz, E., Konvergenz und Fehlerschätzung beim Ritzschen Verfahren, Math. Ann. **100** (1928), 503-521.