DEMONSTRATIO MATHEMATICA
Vol. XXX No 4 1997

Ewa Lewińska

L^2-ANALYSIS OF FINITE ELEMENT EIGENFUNCTION APPROXIMATION WITH NUMERICAL INTEGRATION

1. Introduction

The effect of numerical integration on the finite element approximation of eigenfunctions of the eigenvalue problem

$$L\varphi = \begin{cases} -\sum_{i,j=1}^{2} \partial_j(a_{ij}(x)\partial_i\varphi) = \lambda\varphi & \text{in } \Omega, \\ \varphi = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a convex polygonal domain in $\Re^2$, will be considered.

Similar problems have been studied by Fix [10], Babuska and Osborn [2], Banerjee and Osborn [3], [4], Banerjee and Suri [5], Lewińska [11], Vanmaele and Van Keer [12], [13], Andreev, Kascieva and Vanmaele [1] and others. In [3], [4] there were obtained the optimal estimates for eigenvalues and for eigenfunctions of (1.1) in the space $H_0^1(\Omega)$. Another approach using the maximum norm estimates for FEM was presented in [11].

This paper is devoted to the eigenfunction estimates in the space $L^2(\Omega)$. We show for the finite element space of piecewise polynomials of a degree $k \geq 2$ that, if the precision of the numerical integration is like that for the source linear boundary value problem, the eigenfunction estimates in $L^2(\Omega)$ are optimal (like those for the finite element approximation without numerical integration). In the case of piecewise linear finite elements ($k = 1$) a slight loss in the order of convergence for eigenfunctions takes place despite the assumed increased accuracy of the quadrature rule. Our method of proof makes use of the L^∞ estimates for FEM and in that it differs from the approaches of other authors. In Sec.2 notation and the problem are set up. In Sec.3 the convergence results are established. Sec.4 contains our main result and final remarks.

1991 *Mathematics Subject Classification*: 65N25, 65N30.

2. Problem setting and notations

The variational formulation of (1.1) is:

(2.1) Find $(\lambda, \varphi) \in \Re \times H_0^1(\Omega)$ such that $\varphi \neq 0 \wedge a(\varphi, v) = \lambda b(\varphi, v)$ $\forall v \in H_0^1(\Omega)$ with $a(u,v) = \int_\Omega \sum_{i,j=1}^2 a_{i,j}(\partial_i u)(\partial_j v)\,dx$, $b(u,v) = \int_\Omega uv\,dx$.

We assume that the coefficients a_{ij} are sufficiently smooth, that $a_{ij} = a_{ji}$ $\forall i,j$ and

$$\exists \tilde{a} > 0 : \forall x \in \Omega \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \geq \tilde{a} \sum_{i=1}^{2} \xi_i^2 \quad \forall (\xi_1, \xi_2) \in \Re^2.$$

These assumptions imply the symmetry of the form a and its $H_0^1(\Omega)$-ellipticity, i.e.

$$\exists a > 0 : a(v,v) \geq a\|v\|_1^2 \quad \forall v \in H_0^1(\Omega). \tag{2.2}$$

The approximate space for $V = H_0^1(\Omega)$ will be

$$V_h = \{v_h \in C(\overline{\Omega}) : v_h|_{\partial\Omega} = 0 \wedge (v_h|_K \in P_k(K)\ \forall K \in \tau_h)\},$$

where $P_k(K)$ are polynomials of a degree k on a triangle K of a uniformly regular triangularization τ_h of Ω. Thus about τ_h we assume

$$\exists \nu > 0 : \nu h \leq \varrho_K \leq h_K \leq h\ \forall K \in \tau_h\ \forall h \leq h_0, \tag{2.3}$$

where $h_K = \operatorname{diam} K$; $\varrho_K = \sup\{\operatorname{diam} S : S \text{is a ball contained in } K\}$; $h = \max h_K$. Also as it is usual we will require that no vertex of any closed triangle K belonging to τ_h lies on the interior of a side of another triangle and that the union of all the triangles gives Ω.

The quadrature rule is first defined on a reference element $\tilde{K}$ as

$$\int_{\tilde{K}} \tilde{f}(\tilde{x})\,d\tilde{x} \approx \sum_{l=1}^{L} \tilde{\omega}_l \tilde{f}(\tilde{b}_l)$$

with weights $\tilde{\omega}_l > 0$ and knots $\tilde{b}_l$.

Let $F_K\tilde{x} = B_K\tilde{x} + b_K$ be an affine mapping from $\tilde{K}$ onto K. The quadrature rule is transferred onto each element $K \in \tau_h$ by

$$\int_K f(x)dx \approx \sum_{l=1}^{L} \omega_{l,K} f(b_{l,K}) \quad \text{with} \quad \omega_{l,K} = |\det B_K|\tilde{\omega}_l, b_{l,K} = F_K(\tilde{b}_l) \in K.$$

We denote by $E_K(f)$ an error of the quadrature rule for the element $K \in \tau_h$, i.e.

$$E_K(f) = \int_K f(x)\,dx - \sum_{l=1}^{L} \omega_{l,K} f(b_{l,K}).$$

The approximate forms $a_h : V_h \times V_h \to \Re$ and $b_h : C(\overline{\Omega}) \times C(\overline{\Omega}) \to \Re$ are obtained by applying the above quadrature to the forms a, b respectively, i.e.

$$a_h(u_h, v_h) = \sum_{K \in \tau_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{2} (a_{ij}(\partial_i u_h)(\partial_j v_h))(b_{l,K}) \quad \forall u_h, v_h \in V_h$$

and

$$b_h(u, v) = \sum_{K=\tau_h} \sum_{l}^{L} \omega_{l,K}(uv)(b_{l,K}) \quad \forall u, v \in C(\overline{\Omega}).$$

Throughout the text we assume that the quadrature is exact for the polynomials of a degree $2k - 2$, i.e.

$$E_{\widetilde{K}}(\widetilde{f}) = 0 \quad \forall \widetilde{f} \in P_{2k-2}(\widetilde{K}). \tag{2.4}$$

By [7, Th. 4.1.2] the assumption (2.4) implies the uniform V_h-ellipticity of the forms a_h, i.e.

$$\exists \beta > 0 : a_h(v_h, v_h) \geq \beta \|v_h\|_1^2 \quad \forall h \leq h_0 \ \forall v_h \in V_h. \tag{2.5}$$

Thus the approximate problem will be

$$\begin{aligned} &\text{Find } (\lambda, \varphi_h) \in \Re \times V_h \text{ such that } \varphi_h \neq 0 \wedge a_h(\varphi_h, v_h) \\ &= \lambda b_h(\varphi_h, v_h) \ \forall v_h \in V_h. \end{aligned} \tag{2.6}$$

The eigenvalue problems (2.1), (2.6) can be transformed to an operator form with the help of the solution operators $T : L^2(\Omega) \to V$ and $\widetilde{T}_h : C(\overline{\Omega}) \to V_h$ defined as

$$a(Tu, v) = b(u, v) \ \forall v \in V \quad \forall u \in L^2(\Omega), \tag{2.7}$$

$$a_h(\widetilde{T}_h u, v_h) = b_h(u, v_h) \quad \forall v_h \in V_h \ \forall u \in C(\overline{\Omega}). \tag{2.8}$$

By the Lax–Milgram theorem, T belongs to $L(L^2, V)$ but there arises the problem of existence of $\widetilde{T}_h$. However let us observe that for each $u \in C(\overline{\Omega})$ the mapping $b_h(u, \cdot) : V_h \to \Re$ is a linear functional on a finite dimensional space V_h so by (2.5) and the Lax—Milgram theorem, $\widetilde{T}_h u$ exists. At this stage we leave open the question of boundedness of $\widetilde{T}_h : C(\overline{\Omega}) \to V_h$. This problem will be tackled later.

With the introduction of the solution operators $T, \widetilde{T}_h$ the eigenvalue problems (2.1), (2.6) are equivalent to

$$\mu\varphi = T\varphi, \quad \varphi \in V = H_0^1, \tag{2.9}$$

$$\mu\varphi_h = \widetilde{T}_h\varphi_h, \quad \varphi_h \in V_h, \quad \mu = 1/\lambda. \tag{2.10}$$

Let us observe that in fact the operator T, which has a domain $H = L^2(\Omega)$, can be treated as an operator from H into H (while in (2.9) T has been

treated as an operator from V into V). The eigenvalue problem

$$\mu\varphi = T\varphi, \quad \varphi \in H = L^2 \tag{2.11}$$

has the same eigenvalues and eigenfunctions as (2.9) except for $\mu = 0$ since Range $T \subseteq V$.

From now onward we will examine the relations between the eigenvalue problem (2.11) and its approximation (2.10).

In our text we will often refer to the fact that the operator $T : L^2 \to L^2$ is compact and that the boundary value problem (2.7) is "regular" in the sense (see [7])

$$\forall u \in L^2(\Omega) \quad Tu \in V \cap H^2(\Omega) \wedge T \in L(L^2, H^2). \tag{2.12}$$

At the end of this section let us complete setting up notations. Thus for $A \subseteq \Re^2$ the norms and the seminorms in $H^k(A), W^{k,q}(A)$ will be denoted by $\| \|_{k,A}$, $\| \|_{k,q,A}$, $| |_{k,A}$, $| |_{k,q,A}$. Sometimes subscripts will be omitted if it causes no confusion. The norm in L^2 will be denoted either by $\| \|_0$ or by $| |_0$. Also the norm

$$\|v_h\|_{j,\tau_h} = \sqrt{\sum_{K \in \tau_h} \|v_h\|^2_{j,K}} \quad \forall v_h \in V_h$$

will be frequently used as well as the a-projectors $\prod_h : V \to V_h$ defined by the formula

$$a(u - \prod_h u, v_h) = 0 \quad \forall u \in V \ \forall v_h \in V_h. \tag{2.13}$$

Throughout the text C will stand for a generic constant.

3. Convergence results

Like Banerjee, Osborn in [4], we will apply classical results of Descloux, Nassif, Rappaz [8]. The following conditions will thus have to be checked:

$$\lim_{h \to 0} \inf_{v_h \in V_h} \|u - v_h\|_0 = 0 \quad \forall u \in H = L^2, \tag{3.1}$$

$$\lim_{h \to 0} \|(T - \widetilde{T}_h|_{V_h}\|_{L(H)} = 0. \tag{3.2}$$

The convergence (3.1) obviously takes place, since (3.1) does hold for every $u \in V = H_0^2$ and V is dense in H.

In order to establish (3.2) we will prove

LEMMA 1. *If* $E_{\widetilde{K}}(\widetilde{f}) = 0 \forall \widetilde{f} \in P_k(\widetilde{K})$, *then*

$$|(b - b_h)(u_h, v_h)| \leq Ch^2 |u_h|_{1,\Omega} |v_h|_{1,\Omega} \quad \forall u_h, v_h \in V_h. \tag{3.3}$$

Proof. The technique presented in [9, pp.345–6] will be followed. Let $\widetilde{u}, \widetilde{v}$ belong to $P_k(\widetilde{K})$ and let $u_0, v_0 \in \Re$ be their mean values, i.e.

$$u_0 = \frac{1}{\mu(\widetilde{K})} \int_{\widetilde{K}} \widetilde{u}\, d\widetilde{x}, \quad v_0 = \frac{1}{\mu(\widetilde{K})} \int_{\widetilde{K}} \widetilde{v}\, d\widetilde{x}.$$

The following identity is obvious: $\widetilde{u}\widetilde{v} = (\widetilde{u} - u_0)(\widetilde{v} - v_0) + u_0(\widetilde{v} - v_0) + \widetilde{u}v_0$. Since the functions $u_0(\widetilde{v} - v_0)$, $\widetilde{u}v_0$ are polynomials of a degree k, by the assumption of Lemma, $E_{\widetilde{K}}(u_0(\widetilde{v} - v_0)) = E_{\widetilde{K}}(\widetilde{u}v_0) = 0$ and

$$\begin{aligned} E_{\widetilde{K}}(\widetilde{u}\widetilde{v}) &= E_{\widetilde{K}}((\widetilde{u} - u_0)(\widetilde{v} - v_0)) \\ &\leq \Big| \int_{\widetilde{K}} (\widetilde{u} - u_0)(\widetilde{v} - v_0) \Big| + \Big| \sum_{l=1}^{L} \widetilde{\omega}_l((\widetilde{u} - u_0)(\widetilde{v} - v_0))(\widetilde{b}_l) \Big| \\ &\leq |\widetilde{u} - u_0|_{0,\widetilde{K}} |\widetilde{v} - v_0|_{0,\widetilde{K}} + \mu(\widetilde{K}) |\widetilde{u} - u_0|_{0,\infty,\widetilde{K}} |\widetilde{v} - v_0|_{0,\infty,\widetilde{K}}. \end{aligned}$$

In the finite dimensional space $P_k(\widetilde{K})$ all the norms are equivalent so

$$|E_{\widetilde{K}}(\widetilde{u}\widetilde{v})| \leq C |\widetilde{u} - u_0|_{0,\widetilde{K}} |\widetilde{v} - v_0|_{0,\widetilde{K}}.$$

Taking into account the fact that

$$|\widetilde{u} - u_0|_{0,\widetilde{K}} \leq C |\widetilde{u}|_{1,\widetilde{K}}, \quad |\widetilde{v} - v_0|_{0,\widetilde{K}} \leq C |\widetilde{v}|_{1,\widetilde{K}},$$

we get

$$|E_{\widetilde{K}}(\widetilde{u}\widetilde{v})| \leq C |\widetilde{u}|_{1,\widetilde{K}} |\widetilde{v}|_{1,\widetilde{K}} \forall \widetilde{u}, \widetilde{v} \in P_k(\widetilde{K}). \tag{3.4}$$

Assuming that $u, v \in P_k(K)$ by the standard reference element technique and by (3.4), we arrive at

$$|E_K(uv)| \leq Ch^2 |u|_{1,K} |v|_{1,K} \quad \forall u, v \in P_k(K)\ \forall K \in \tau_h.$$

The last result implies (3.3) since:

$$(b - b_h)(u_h, v_h) = \sum_K E_K(u_h|_K v_h|_K) \quad \forall u_h, v_h$$

and $u_h|_K, v_h|_K \in P_k(K)$. ∎

Now we are in a position to prove (3.2).

THEOREM 1. *Suppose* $E_{\widetilde{K}}(\widetilde{f}) = 0 \forall \widetilde{f} \in P_{2k-2}(\widetilde{K}) \cup P_k(\widetilde{K})$. *Then*

$$\|(T - \widetilde{T}_h)v_h\|_0 \leq Ch \|v_h\|_0 \quad \forall v_h \in V_h. \tag{3.5}$$

Proof. Making use of the a-projectors defined in (2.13) we get

$$\|(T - \widetilde{T}_h)v_h\|_0 \leq \|f_{1h}\|_0 + \|f_{2h}\|_0 \tag{3.6}$$

with $f_{1h} = (I - \prod_h)Tv_h$, $f_{2h} = (\prod_h T - \widetilde{T}_h)v_h$. By classical results and by (2.12), $\|f_{1h}\|_0 \leq Ch^2 |Tv_h|_2 \leq Ch^2 \|v_h\|_0$. By (2.5), (2.13), (2.7), (2.8):

$$\beta \|f_{2h}\|_1^2 \leq a_h((\prod_h T - \widetilde{T}_h)v_h, f_{2h}) \tag{3.7}$$

$$= (a_h - a)(\prod_h Tv_h, f_{2h}) + a(\prod_h Tv_h, f_{2h}) - a_h(\widetilde{T}_h v_h, f_{2h})$$
$$= (a_h - a)(\prod_h Tv_h, f_{2h}) + (b - b_h)(v_h, f_{2h}).$$

Lemma 1 and inverse inequalities allow us to write

$$|(b - b_h)(v_h, f_{2h})| \leq Ch^2 |v_h|_1 |f_{2h}|_1 \leq Ch\|v_h\|_0 |f_{2h}|_1. \tag{3.8}$$

From [7, Th. 4.1.4], inverse inequalities and the uniform boundedness of the a-projectors $\prod_h$ in $L(H_0^1)$ there follows

$$\begin{aligned} |(a_h - a)(\prod_h Tv_h, f_{2h})| &\leq Ch^k \|\prod_h Tv_h\|_{k,\tau_h} \|f_{2h}\|_1 \\ &\leq Ch^k h^{1-k} \|\prod_h Tv_h\|_1 \|f_{2h}\|_1 \\ &\leq Ch\|Tv_h\|_1 \|f_{2h}\|_1 \leq \|v_h\|_0 \|f_{2h}\|_1. \end{aligned} \tag{3.9}$$

Combining (3.7)–(3.9) we arrive at: $\|f_{2h}\|_0 \leq \|f_{2h}\|_1 \leq Ch\|v_h\|_0$, which together with (3.6) yields (3.5). ∎

4. Estimates for eigenfunctions

Let the spaces H, V_h and the operators $T, \widetilde{T}_h, \widetilde{T}_h \prod_h$ be complexified in the usual manner. Having established (3.1) and (3.2), we can apply the classical results on eigenvalue approximation from [8]. Let $\mu_0 \neq$ be an eigenvalue of T with a finite multiplicity m. Let us observe that the operator $T : L^2 \to L^2$ is selfadjoint with respect to the scalar product a, because for any $u, v \in H$ we have $a(Tu, v) = b(u, v) = b(v, u) = a(Tv, u) = a(u, Tv)$. Thus the invariant subspace M for T and $\mu_0 \neq 0$ is equal to the eigenspace, so $M = \mathrm{Ker}(\mu_0 - T)$.

Let $\Gamma = O(\mu_0, r)$ be a circle with a centre μ_0 and a radius $r < 0$. Let Γ lie in the resolvent set $\varrho(T)$ and enclose no other points of the spectrum $\sigma(T)$. By [8], inside Γ there lie exactly m eigenvalues $\widetilde{\mu}_{1h}, \widetilde{\mu}_{2h}, \ldots, \widetilde{\mu}_{mh}$ of $\widetilde{T}_h|_{V_h}$ counting their multiplicities. Also for any compact set $F \subseteq \varrho(T)$

$$\|R_z(\widetilde{T}_h)v_h\|_0 \leq \mathrm{Const}\|v_h\|_0 \quad \forall v_h \in V_h \ \forall z \in F \tag{4.1}$$

for h sufficiently small with $R_z(\widetilde{T}_h) = (z - \widetilde{T}_h|_{V_h})^{-1}$ denoting the resolvent operator for $\widetilde{T}_h|_{V_h}$.

Let $\widetilde{M}_h$ be an algebraic sum of invariant subspaces for $\widetilde{\mu}_{ih}$ and $\widetilde{T}_h$. Let us observe that $\widetilde{T}_h$ is selfadjoint with respect to the scalar product a_h, because

$$a_h(\widetilde{T}_h u_h, v_h) = b_h(u_h, v_h) = b_h(v_h, u_h) = a_h(\widetilde{T}_h v_h) = a_h(u_h, \widetilde{T}_h v_h).$$

Therefore invariant subspaces are equal to eigenspaces, so

$$\begin{cases} \widetilde{M}_h = \mathrm{Ker}(\widetilde{\mu}_{1h} - \widetilde{T}_h) + \mathrm{Ker}(\widetilde{\mu}_{2h} - \widetilde{T}_h) + \ldots \mathrm{Ker}(\widetilde{\mu}_{mh} - \widetilde{T}_h), \\ M = \mathrm{Ker}(\mu_0 - T). \end{cases} \tag{4.2}$$

It is known that the so-called gap $\tilde{\delta}_H(M, \widetilde{M}_h)$ is a good measure of the eigenfunction approximation. Let us recollect that

$$(4.3)\quad \begin{cases} \tilde{\delta}_H(M, \widetilde{M}_h) = \max\{\delta_H(M, \widetilde{M}_h), \delta_H(\widetilde{M}_h, M)\}, \\ \delta_H(M, \widetilde{M}_h) = \sup\{\inf_{\varphi_h \in M_h} \|\varphi - \varphi_h\|_0 : \varphi \in M \wedge (\|\varphi\|_0 = 1)\}, \\ \delta_H(\widetilde{M}_h, M) = \sup\{\inf_{\varphi \in M} \|\varphi_h - \varphi\|_0 : \varphi_h \in \widetilde{M}_h \wedge (\|\varphi_h\|_0 = 1)\}. \end{cases}$$

By (3.1), (3.2) and [8], we get $\tilde{\delta}_H(M, \widetilde{M}_h) \to 0$. Hence for h sufficiently small

$$(4.4)\qquad \tilde{\delta}_H(M, \widetilde{M}_h) = \delta_H(M, \widetilde{M}_h) = \delta_H(\widetilde{M}_h, M).$$

Now we will pass to the estimate of $\delta_H(M, \widetilde{M}_h)$ which seems much easier than an evaluation of $\delta_H(\widetilde{M}_h, M)$, because in the definition (4.3) of $\delta_H(M, \widetilde{M}_h)$ the supremum is taken over $\varphi \in M$ and we can assume that the functions $\varphi \in M$ are sufficiently smooth. Therefore a high convergence rate can be expected. In the definition (4.3) of $\delta_H(\widetilde{M}_h, M)$ the supremum is taken over $\varphi_h \in \widetilde{M}_h$ being continuous and no more. Thus, let us introduce spectral projections

$$P : V \to V \quad \text{for } T : V \to V, \qquad P = \frac{1}{2\pi i} \int_\Gamma R_z(T)\, dz,$$

$$P_h : V \to V \quad \text{for } \tilde{T}_h \prod_h : V \to V, \qquad P_h = \frac{1}{2\pi i} \int_\Gamma R_z(\tilde{T}_h \prod_h)\, dz,$$

$$\tilde{P}_h : V_h \to V_h \text{ for } \tilde{T}_h|_{V_h} : V_h \to V_h, \quad \tilde{P}_h = \frac{1}{2\pi i} \int_\Gamma R_z(\tilde{T}_h)\, dz.$$

In [6, pp. 184, 238—9] the following is proved

$$(4.5)\qquad R_z(\tilde{T}_h \prod_h) = R_z(\prod_h \tilde{T}_h \prod_h) = R_z(\tilde{T}_h)\prod_h + \frac{1}{z}(I - \prod_h),$$

$$(4.6)\qquad \text{Range } P_h V = \text{Range } \tilde{P}_h V_h = \widetilde{M}_h.$$

Therefore $P_h \varphi \in \widetilde{M}_h$ for any $\varphi \in M$ and we can write

$$(4.7)\quad \begin{aligned} \delta_H(M, \widetilde{M}_h) &= \sup\{\inf_{\varphi_h \in \widetilde{M}_h} \|\varphi - \varphi_h\|_0 : \varphi \in M \wedge (\|\varphi\|_0 = 1)\} \\ &\le \sup\{\|\varphi - P_h \varphi\|_0 : \varphi \in M \wedge (\|\varphi\|_0 = 1)\}. \end{aligned}$$

Examining the quantity $\|\varphi - P_h \varphi\|_0$ we will establish:

LEMMA 1. *By the assumptions of Theorem* 3.1

$$(4.8)\quad \begin{aligned} \tilde{\delta}_H(M, \widetilde{M}_h) \le C(&\sup\{W_1(\varphi) : \varphi \in M \wedge \|\varphi\|_0 = 1\} \\ &+ \sup\{W_2(\varphi) : \varphi \in M \wedge \|\varphi\|_0 = 1\}) \end{aligned}$$

with

$$W_1(\varphi) = \|(I - \prod_h)\varphi\|_0, \quad W_2(\varphi) = \|(\prod_h T - \tilde{T}_h \prod_h)\varphi\|_0.$$

Proof. Since $M = PV$, then

$$\varphi - P_h\varphi = P\varphi - P_h\varphi = \left\{\frac{1}{2\pi i}\int_\Gamma [R_z(T) - R_z(\tilde{T}_h\prod_h)]\,dz\right\}\varphi$$

$$= \frac{1}{2\pi i}\int_\Gamma R_z(\tilde{T}_h\prod_h)(T - \tilde{T}_h\prod_h)R_z(T)\varphi\,dz.$$

By the formula (4.5)

$$\varphi - P_h\varphi = \frac{1}{2\pi i}\int_\Gamma R_z(\tilde{T}_h)\prod_h(T - \tilde{T}_h\prod_h)R_z(T)\varphi\,dz \tag{4.9}$$

$$+ (I - \prod_h)(T - \tilde{T}_h\prod_h)\frac{1}{2\pi i}\int_\Gamma \frac{1}{z}R_z(T)\varphi\,dz.$$

Since $R_z(T)\varphi = \frac{1}{z-\mu_0}\varphi$, for $\varphi \in M$, by the property (4.1) we have

$$\|R_z(\tilde{T}_h)(\prod_h T - \tilde{T}_h\prod_h)R_z(T)\varphi\|_0$$

$$\leq \|R_z(\tilde{T}_h)\|_{L(H)}\left\|(\prod_h T - \tilde{T}_h\prod_h)\frac{1}{z-\mu_0}\varphi\right\|_0$$

$$\leq C\max_{z\in\Gamma}\frac{1}{|z-\mu_0|}\|(\prod_h T - \tilde{T}_h\prod_h)\varphi\|_0 \leq CW_2(\varphi).$$

Thus the first term in (4.9) is bounded by $CW_2(\varphi)$. As to the second term is concerned, let us observe that $(I - \prod_h)(T - \tilde{T}_h\prod_h) = (I - \prod_h)T$, while

$$\frac{1}{2\pi i}\int_\Gamma \frac{1}{z}R_z(T)\varphi\,dz = \frac{1}{2\pi i}\int_\Gamma \frac{1}{z}\cdot\frac{1}{z-\mu_0}\varphi\,dz = \frac{1}{\mu_0}\varphi.$$

Therefore the second term in (4.9) is equal to $(I-\prod_h)T(\frac{1}{\mu_0}\varphi) = (I-\prod_h)\varphi$. The above analysis together with (4.9), (4.7) and (4.4) prove (4.8). ∎

LEMMA 2. *Suppose* $E_{\tilde{K}}(\tilde{f}) = 0\ \forall \tilde{f} \in P_{2k-k}(\tilde{K})$. *Then*

$$\exists C \geq 0 : \|\tilde{T}_h\|_{L(C(\overline{\Omega}),V_h)} \leq C \quad \forall h \leq h_0. \tag{4.10}$$

Proof. Since $\sum_l \omega_{l,K} = \mu(K)$ and since by (2.3), $\mu(K) \leq Ch^2$, we have for any $u \in C(\overline{\Omega}), v_h \in V_h$:

$$|b_h(u, v_h)| \leq \|u\|_{C(\overline{\Omega})}\sum_K\sum_l \omega_{l,K}\|v_h\|_{0,\infty K} \leq Ch^2\|u\|_{C(\overline{\Omega})}\sum_K \|v_h\|_{0,\infty,K}.$$

From the Hölder inequality $\sum_K \|v_h\|_{0,\infty,K} \leq \sqrt{\sum_K 1^2}\sqrt{\sum_K \|v_h\|^2_{0,\infty,K}}$. The sum $\sum_K 1^2$ is equal to the number of triangles K in τ_h. According

to the assumption (2.3), it is not greater than $\frac{C}{h^2}$. By inverse inequalities $\|v_h|_{0,\infty,K} \leq \frac{c}{h}\|v_h\|_{0,2K}$ and consequently

$$|b_h(u, v_h)| \leq Ch^2\|u\|_{C(\overline{\Omega})}\sqrt{\frac{C}{h^2}}\sqrt{\sum_K \frac{C^2}{h^2}\|v_h\|^2_{0,2,K}} \leq C\|u\|_{C(\overline{\Omega})}\|v_h\|_0. \tag{4.11}$$

By (2.5), we have $\beta\|\widetilde{T}_h u\|_1^2 \leq a_h(\widetilde{T}_h u, \widetilde{T}_u) = b_h(u, \widetilde{T}_h u)$ for any $u \in C(\overline{\Omega})$. Now a quick glance at (4.11) is enough to draw the conclusion (4.10). ∎

Lemmas 1, 2 enable us to prove our main result.

THEOREM 1. *Suppose* $E_{\widetilde{K}}(\widetilde{f}) = 0\ \forall \widetilde{f} \in P_{2k-2}(\widetilde{K}) \cup P_k(\widetilde{K})$ *and* $M = \mathrm{Ker}(\mu_0 - T) \subseteq H_0^1(\Omega) \cap W^{r,\infty})\Omega)$, *where* $r = \max(k+1, 2k-1)$. *Then*

$$\widetilde{\delta}_H(M, \widetilde{M}_h) \leq Ch^p \tag{4.12}$$

where $p = k+1$ *for* $k \leq 2$ *and* $p = 2 - \varepsilon$ *for* $k = 1$ *with any* $\varepsilon > 0$.

Proof. Throughout the proof we will consider $\varphi \in M$ such that $\|\varphi\|_0 = 1$. Since in the finite dimensional space M all the norms are equivalent, we will be able to bound $\|\varphi\|_{k+1}$, $\|\varphi\|_{k+1,\infty}$, $\|\varphi\|_{2k-1,\infty}$, $\|\varphi\|_{1,3}$ by a constant C. Lemma 1, the property (2.12) and the Aubin–Nitsche lemma imply

$$W_1(\varphi) = \|(I - \prod_h)\varphi\|_0 \leq Ch^{k+1}|\varphi|_{k+1} \leq Ch^{k+1}. \tag{4.13}$$

Thus by (4.8), it is enough to estimate the term $W_2(\varphi)$. Let $W_2(\varphi)$ be splitted up into:

$$W_2(\varphi) \leq W_2^1(\varphi) + W_2^2(\varphi) + W_2^3(\varphi) \tag{4.14}$$

with

$$W_2^1(\varphi) = \|(\prod_h T - T)\varphi\|_0, \quad W_2^2(\varphi)\|(T - \widetilde{T}_h)\varphi\|_0,$$
$$W_2^3(\varphi) = \|\widetilde{T}_h(\varphi - \prod_h \varphi)\|_0.$$

The term $W_2^1(\varphi)$ may be estimated again by the Aubin–Nitsche lemma in the following way

$$W_2^1(\varphi) \leq Ch^{k+1}|T\varphi|_{k+1} = Ch^{k+1}|\mu_0\varphi|_{k+1} \leq Ch^{k+1}. \tag{4.15}$$

From Lemma 2 the approximate solution operators $\widetilde{T}_h$ are uniformly bounded in $L(C(\overline{\Omega}), H_0^1)$, so $W_2^3(\varphi) \leq \|\widetilde{T}_h(\varphi - \prod_h \varphi\|_1 \leq C\|\varphi - \prod_h \varphi\|_{0,\infty}$. The classical L^∞-estimates in FEM (see [7]) account for

$$W_2^3(\varphi) \leq C\|\varphi - \prod_h \varphi\|_{0,\infty} \tag{4.16}$$
$$\leq \begin{cases} Ch^{k+1}|\varphi|_{k+1,\infty} \leq Ch^{k+1} & \text{for } k \geq 2, \\ Ch^{2-\varepsilon}|\varphi|_{2,\infty} \leq Ch^{2-\varepsilon} & \text{for } k = 1. \end{cases}$$

The only point remaining concerns the behaviour of the term $W_2^2(\varphi)$. By [7, Exc. 4.1.3]

$$W_2^2(\varphi) \leq \sup_{g\in L^2} \frac{1}{\|g\|_0} \inf_{v_h\in V_h} \{C\|T\varphi - \widetilde{T}_h\varphi\|_1 \|Tg - v_h\|_1 + |(a-a_h)(\widetilde{T}_h\varphi, v_h)| + |(b-b_h)(\varphi, v_h)|\}. \tag{4.17}$$

Again there are three terms to be estimated. At first let us remind that by [7, Th. 4.1.6]:

$$\|T\varphi - \widetilde{T}_h\varphi\|_{1,\Omega} \leq Ch^k(\|T\varphi\|_{k+1} + \|\varphi\|_{k,q}) \quad \text{with } q \geq 2,\ k > 2/q.$$

Setting $q = 2$ for $k \geq 2$ and $q = 3$ for $k = 1$ and making use of the equivalence of norms in M, we see that

$$\|T\varphi - \widetilde{T}_h\varphi\|_{1,\Omega} \leq Ch^k. \tag{4.18}$$

For any $g \in H = L^2$ we choose $v_h = J_h(Tg)$ in (4.17) with $J_h(Tg)$ being an interpolant of Tg. The interpolant is well-defined since by (2.12), $Tg \in H^2 \subset C(\overline{\Omega})$. Thus $\|Tg - J_h(Tg)\|_1 \leq Ch|Tg|_2 \leq Ch|g|_0$. The last inequality together with (4.18) yield

$$\|T\varphi - \widetilde{T}_h\varphi\|_1 \|Tg - v_h\|_1 \leq Ch^{k+1}|g|_0. \tag{4.19}$$

As the second term in (4.17) is concerned, we get from [3, Lemma 3.2, p. 149]

$$|(a-a_h)(\widetilde{T}_h\varphi, v_h)| \leq Ch^{2k-1}\|\widetilde{T}_h\varphi\|_{k,\tau_h}\|v_h\|_{k,\tau_h}.$$

By inverse inequalities $\|v_h\|_{k,\tau_h} \leq Ch^{2-k}\|v_h\|_{2,\tau_h}$. By the classical interpolation theory and (2.12): $\|v_h\|_{2,\tau_h} = \|J_h(Tg)\|_{2,\tau_h} \leq \|J_h(Tg) - Tg\|_{2,\tau_h} + \|Tg\|_{2,\Omega} \leq C\|Tg\|_{2,\Omega} \leq C|g|_0$. Hence

$$|(a-a_h)(\widetilde{T}_h\varphi, v_h)| \leq Ch^{k+1}\|\widetilde{T}_h\varphi\|_{k,\tau_h}|g|_0. \tag{4.20}$$

To establish the uniform boundedness of the term $\|\widetilde{T}_h\varphi\|_{k,\tau_h}$, we can write

$$\begin{aligned}
\|\widetilde{T}_h\varphi\|_{k,\tau_h} &\leq \|\widetilde{T}_h - \textstyle\prod_h T\varphi\|_{k,\tau_h} + \|\textstyle\prod_h T\varphi\|_{k,\tau_h} \\
&\leq Ch^{1-k}\|(\widetilde{T}_h - \textstyle\prod_h T)\varphi\|_1 + \|\textstyle\prod_h(\mu_0\varphi)\|_{k,\tau_h} \\
&\leq Ch^{1-k}(\|(\widetilde{T}_h - T)\varphi\|_1 + \|(I - \textstyle\prod_h)T\varphi\|_1) \\
&\quad + \mu_0\|(\textstyle\prod_h - I)\varphi\|_{k,\tau_h} + \mu_0\|\varphi\|_k \\
&\leq Ch^{1-k}(\|(\widetilde{T}_h - T)\varphi\|_1 + h^k|T\varphi|_{k+1}) + \mu_0 Ch|\varphi|_{k+1} + \mu_0\|\varphi\|_k.
\end{aligned}$$

Bearing in mind (4.18), we actually state that $\|\widetilde{T}_h\varphi\|_{k,\tau_h} \leq C$, which together with (4.20) gives

$$|(a-a_h)(\widetilde{T}_h\varphi, v_h)| \leq Ch^{k+1}|g|_0. \tag{4.21}$$

To determine the third term in (4.17), once again from Banerjee [3, Lemma 3.2, p.149], we get

$$\begin{aligned}|(b-b_h)(\varphi, v_h)| &= |\sum_K E_K(\varphi|_K \cdot 1|_K \cdot J_h(Tg)|_K)| \\ &\leq Ch^{2k-1}\|\varphi\|_{2k-1,\infty}\|1\|_{k,\tau_h}\|J_h(Tg)\|_{k,\tau_h} \\ &\leq Ch^{2k-1}\|J_h(Tg)\|_{k,\tau_h}.\end{aligned}$$

In the application of the result of Banerjee there is a certain trick. Since two piecewise polynomials are needed while we have only one $v_h = J_h(Tg)$ (φ is not), we choose the constant function 1 to be the second piecewise polynomial. Above φ stands for a coefficient which should belong to $W^{2k-1,\infty}(\Omega)$. Hence our assumption is: $M \subset W^{2k-1,\infty}(\Omega)$. From inverse inequalities, the classical interpolation theory and (2.12), we get

$$\|J_h(Tg)\|_{k,\tau_h} \leq Ch^{2-k}\|J_h(Tg)\|_{2,\tau_h} \leq Ch^{2-k}|Tg|_2 \leq Ch^{2-k}|g|_0.$$

Thus

$$|(b-b_h)(\varphi, v_h)| \leq Ch^{k+1}|g|_0. \tag{4.22}$$

The formulae (4.19), (4.21), (4.22) applied in (4.17) yield

$$W_2^2(\varphi) \leq Ch^{k+1}. \tag{4.23}$$

Combining (4.23) with (4.14)–(4.16), (4.13) and (4.8) we obtain our assertion (4.12). ∎

Final remarks:

1) About the smoothness of eigenfunctions in general case it is only known that the eigenspace M is a subset of $H_0^1 \cap H^2$. No better regularity result is available for Ω-a polygon. However in a specific case M may be a set of very smooth functions. For example for the square $\Omega = (0,\pi)\times(0,\pi)$ and laplacian L, the operator $T = L^{-1}$ has eigenvalues $\mu_{nm} = 1/(n^2+m^2)$ with eigenfunctions $\varphi = (\sin nx)(\sin mx)$ of class C^∞.

2) For $k \geq 2$, $\max(2k-2, k) = 2k-2$, so our assumption about the precision of the quadrature is like that for the corresponding source linear boundary value problem. Besides that for $k \geq 2$ the estimate of the gap $\widetilde{\delta}$ is optimal in the sense that it does also hold for the classical finite element approximation (i.e. without numerical integration). For $k = 1$, $\max(2k-2, k) = 1$, so we assume that the quadrature is exact for all the linear functions. For the corresponding source BVP the quadrature is assumed to be exact for constans only. Simultaneously for $k = 1$ the thesis of Theorem 1 states a slight loss in the order of convergence.

References

[1] A. B. Andreev, V. A. Kascieva, M. Vanmaele, *Some results in lumped mass finite element approximation of eigenvalue problems using numerical quadrature formulas*, J. Comp. Appl. Math. 43, (1992), 291–311.

[2] I. Babuska, J. E. Osborn, *Eigenvalue error estimates for elliptic problems. Handbook of numerical analysis*, Vol.II. Finite element methods. Part I,(eds.: Ciarlet P. G., Lions J.-L.). North Holland. Amsterdam, (1991), 641–787.

[3] U. Banerjee, *A note on the effect of numerical quadrature in finite element eigenvalue approximation*, Num. Math. 61, (1992), 145–152.

[4] U. Banerjee, J. E. Osborn, *Estimation of the effect of numerical integration in finite element eigenvalue approximation*, Num. Math. 56, (1990), 735–762.

[5] U. Banerjee, M. Suri, *Analysis of numerical integration in p-version finite element eigenvalue approximation*, Numer. Methods for PDEs. 8, (1992), 381–394.

[6] F. Chatelin, *Spectral Approximation of Linear Operators. Computer Science and Applied Mathematics*, Academic Press (1983).

[7] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North Holland. Amsterdam. (1978).

[8] J. Descloux, N. Nassif, J. Rappaz, *On spectral approximations*, Part I. The problem of convergence, RAIRO Anal. Numer. 12, (1978), 97–112.

[9] J. Descloux, J. Rappaz, *Approximation of solution branches of nonlinear equations*, RAIRO Num. Anal. 16, (1982) 319–349.

[10] G. J. Fix, *Effects of quadrature errors in finite element approximation of steady state, eigenvalue and parabolic problems*, The Mathematical Foundations of the FEM with Applications to PDEs (ed. Aziz A. K.). Academic Press 1972, 525–557.

[11] E. Lewinska, *Approximation of the eigenvalue problem for elliptic operator by finite element method with numerical integration*, Matem. Stos. 37, (1994), 3–14.

[12] M. Vanmaele, R. Van Keer, *Convergence and error estimates for a finite element method with numerical quadrature for a second order elliptic eigenvalue problem*, Numerical Treatment of Eigenvalue Problems. Vol.5 (eds. Albrecht J., Collatz L., Hagedorn P., Vette W.). International Series of Numerical Mathematics, 96 Birkhaü ser Verlag. Basel, (1991), 225–236.

[13] M. Vanmaele, R. Van Keer, *Error estimates for a FEM with numerical quadrature for a class of elliptic eigenvalue problems*, Numerical Methods (eds. Greenspan D., Ró zsa P.). North Holland, (1991), 267–282.

INSTITUTE OF MATHEMATICS
WARSAW UNIVERSITY OF TECHNOLOGY
Pl. Politechniki 1
00-660 WARSAW, POLAND
E-mail: ewalew@plwatu2l.bitnet

Received April 23, 1997.