ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
20, 4 (1990), pp. 607–624

H. MARCINKOWSKA and A. SZUSTALEWICZ (Wrocław)

SOME REMARKS ON THE GALERKIN APPROXIMATION OF PARABOLIC EQUATIONS

1. Introduction. This paper is a continuation of an earlier work [5] of the first-named author. In [5] the initial-boundary value problem for the parabolic equation

$$Au+u_t = f \tag{1}$$

was studied in a time-dependent domain $\Omega_{(t)} \subset R^n$, $0 < t < T$. The problem was reduced to that in a constant domain Ω of R^n by means of a diffeomorphism. Using two equivalent weak formulations, we proposed two approximate methods of solution.

The aim of the present paper is to study deeply these methods and the related error estimates.

2. Basic notation and assumptions. Preliminary lemmas. For convenience of the reader we recall some notation used in [5].

For $x, y \in R^N$ we denote by $|x|$ the Euclidean norm and by $\langle x, y\rangle$ the scalar product. Given an $(N \times N)$-matrix C, we write

$$|C| = \sup_{|x|=1} |Cx| \quad (x \in R^N)$$

for its spectral norm.

All the derivatives in the sequel are understood in the distributional sense.

Let $\Omega \subset R^n$ be a bounded domain having the segment property. We put $\Delta_T = \Omega \times (0, T)$ and consider the operator A in divergence form

$$A = -\sum_{j,k=1}^{n} D_k a_{jk}(x, t) D_j + \sum_{j=1}^{n} a_j(x, t) D_j + a(x, t)$$

assuming that

(a_1) the coefficients a_{jk}, a_j, a are bounded in Δ_T;

(a_2) there is a constant $c>0$ such that

$$\sum_{j,k=1}^{n} a_{jk}(x,t)\xi_j\xi_k \geqslant c|\xi|^2$$

for $(x,t)\in \Delta_T$, $\xi\in R^n$ (this means the uniform ellipticity of A).

For any linear normed space X we denote by $L^2(0,T;X)$ the set of all functions $[0,T]\ni t\to u(t)\in X$ such that

$$\int_0^T \|u(t)\|_X^2 dt < \infty .$$

In what follows we use the following Hilbert spaces:

(i) the space $L^2(\Omega)$ with the scalar product $(\ ,\ )_\Omega$ and the norm $\|\ \|_\Omega$;

(ii) the closed subspace V of the Sobolev space $H_1(\Omega)$ satisfying $C_0^\infty(\Omega)\subset V\subset H_1(\Omega)$, equipped with the induced norm $\|\ \|_{1,\Omega}$ (see [1]);

(iii) the space $H(V)=\{u\in L^2(0,T;V)\colon u_t\in L^2(\Delta_T)\}$ with the scalar product induced by $H_1(\Delta_T)$;

(iv) the space $L^{2,N}(0,T)=L^2(0,T;R^N)$ with the scalar product

$$(u,v)=\int_0^T \langle u(t),v(t)\rangle dt$$

and the norm $\|u\|=(u,u)^{1/2}$.

The product of N copies of $H_k(0,T)$ is denoted by $H_k^N(0,T)$.

Remark. It is easy to prove that $H(V)$ is a closed subspace of $H_1(\Delta_T)$, and therefore a Hilbert space. Suppose namely that

$$\|u_n-u\|_{1,\Delta_T}\to 0 \quad \text{as } n\to\infty,\ u_n\in H(V).$$

This means that the sequence $p_n(t)=\|u_n(\cdot,t)-u(\cdot,t)\|_{1,\Omega}$ tends to zero in $L^2(0,T)$, and therefore it contains a subsequence $\{p_{n_k}\}$ which tends to zero almost everywhere in $(0,T)$. This yields $u(\cdot,t)\in V$ for almost every $t\in(0,T)$, so $u\in H(V)$.

We relate with the operator A two bilinear forms:

$$b(t;\varphi,\psi)=\sum_{j,k=1}^{n}(a_{jk}D_j\varphi, D_k\psi)_\Omega+\sum_{j=1}^{n}(a_jD_j\varphi,\psi)_\Omega+(a\varphi,\psi)_\Omega$$

for $\varphi,\psi\in H_1(\Omega)$ and

$$B(u,v)=\int_0^T b(t;u,v)dt-\int_0^T (u,v_t)_\Omega dt+\big(u(\cdot,T),v(\cdot,T)\big)_\Omega \tag{2}$$

for $u,v\in H(V)$. For any $v\in H(V)$ we put (see [5], Lemma 3)

$$\bar v=[v,v(\cdot,0)]\in H(V)\times L^2(\Omega).$$

It was proved in [5] that after a suitable change of the unknown function in (1) the inequality

(a_3) $b(t; v, v) \geqslant 2\varkappa \|v\|_{1,\Omega}^2$

holds for any $v \in V$, $t \in (0, T)$ with a positive constant $\varkappa$. Hence

LEMMA 1. *The following inequality holds for any* $v \in H(V)$:

$$B(v, v) \geqslant \varkappa\Big(\int_0^T \|v\|_{1,\Omega}^2 dt + \|v(\cdot, 0)\|_\Omega^2 + \|v(\cdot, T)\|_\Omega^2\Big).$$

For the proof it is sufficient to integrate by parts the integral

$$\int_0^T (v, v_t)_\Omega dt.$$

Suppose we are given two Hilbert spaces H_0, H_+ with the scalar products $(\, , \,)_0$, $(\, , \,)_+$ and the corresponding norms $\| \; \|_0$, $\| \; \|_+$, respectively. We assume that $H_+ \subset H_0$, the inclusion is dense and continuous. Then every $f \in H_0$ defines a linear continuous functional $l_f(g) = (f, g)$ over H_+. It is easy to prove that $\{l_f\colon f \in H_0\}$ is a complete set of functionals and that $\|l_f\| \leqslant \|f\|_0$. Therefore, identifying l_f with f and putting $H_- = (H_+)^*$ (with the usual norm which is denoted by $\| \; \|_-$), we have the continuous and dense imbedding $H_0 \mathrel{\acute{\subset}} H_-$. Putting $l(g) = (l, g)_0$ we extend the scalar product in H_0 to the bilinear form over $H_- \otimes H_+$. The generalized Schwarz inequality

$$|(f, g)_0| \leqslant \|f\|_- \|g\|_+ \tag{3}$$

holds for any $f \in H_-$, $g \in H_+$.

Let us assume now that $Z \subset H_+$ is a finite-dimensional linear space and let P be the operator of orthogonal projection in H_0 on Z. We have

$$\|Pu\|_+ \leqslant c\|P_u\|_0 \leqslant c\|u\|_0 \leqslant c\|u\|_+$$

for $u \in H_0$, and therefore

$$\|Pu\|_- = \sup_{v \in H_+} \frac{|(Pu, v)_0|}{\|v\|_+} = \sup_{v \in H_+} \frac{|(u, Pv)_0|}{\|v\|_+} \leqslant \|u\|_- \sup_{v \in H_+} \frac{\|Pv\|_+}{\|v\|_+} \leqslant c\,\|u\|_-.$$

Thus P may be extended to a continuous linear operator $P\colon H_- \to H_-$; we denote this extension also by P.

LEMMA 2. *For any* $u \in H_-$, $v \in H_+$ *we have* $Pu \in Z$ *and*

$$(Pu, v)_0 = (u, Pv)_0. \tag{4}$$

Proof. Let $u_n \to u$ in H_-, $u_n \in H_0$. Then

$$(Pu_n, v)_0 = (u_n, Pv)_0,$$

and passing to the limit we get (4) in view of (3). Moreover, $Pu_n \to Pu$ in H_- and Z is closed with respect to each norm. Therefore $Pu \in Z$.

LEMMA 3. *Let $u \in H_-$. Then $Pu = 0$ is equivalent to the identity*

(5) $$\forall_{z \in Z} (u, z)_0 = 0.$$

Proof. Let $u_n \to u$ in H_-, $u_n \in H_0$. We have the orthogonal decomposition $u_n = Pu_n + u_n^{\perp}$, and therefore for any $z \in Z$ we obtain

$$(u_n, z)_0 = (Pu_n, z)_0.$$

Passing to the limit we get

(6) $$(u, z)_0 = (Pu, z)_0.$$

Putting now $z = Pu$ in (5) we obtain $\|Pu\|_0 = 0$, and therefore $Pu = 0$. The converse statement is obvious in virtue of (6).

LEMMA 4. *Suppose $L: H_+ \to H_-$ is linear and continuous. Then $T := PL|_Z$ is a linear continuous mapping in Z equipped with the norm $\| \ \|_0$.*

Proof. As all the norms are equivalent on Z, we have for $z \in Z$

$$\|PLz\|_0 \leqslant c_1 \|PLz\|_- \leqslant c_2 \|Lz\|_- \leqslant c_3 \|z\|_+ \leqslant c_4 \|z\|_0$$

with some positive constants c_j.

We consider now some examples of the triplet $H_+ \subset H_0 \subset H_-$.

EXAMPLE 1. $H_+ = H(V)$, $H_0 = L^2(\Delta_T)$. The inclusion $H(V) \subset L^2(\Delta_T)$ is obviously continuous and dense because $C_0^\infty(\Delta_T) \subset H(V)$. The space H_- is denoted by $H^*(V)$ and the extended scalar products by $(\, , \,)_0$, $(\, , \,)_{\Delta_T}$, respectively.

EXAMPLE 2. $H_+ = H(V) \times L^2(\Omega)$, $H_0 = L^2(\Delta_T) \times L^2(\Omega)$. The scalar products are defined as

$$(\, , \,)_0 = (\, , \,)_{\Delta_T} + (\, , \,)_\Omega.$$

Every linear functional $l \in H_-$ is of the form

$$l(v, \varphi) = (f, v)_{\Delta_T} + (\varphi, \psi)_\Omega \quad \text{for } v \in H(V), \ \psi \in L^2(\Omega)$$

with some $f \in H^*(V)$, $\varphi \in L^2(\Omega)$. So $H_- = H^*(V) \times L^2(\Omega)$ and the following equality holds:

$$\|l\|^2 = \|f\|^2_{H^*(V)} + \|\varphi\|^2_\Omega.$$

A special case of the approximate methods of solving (1), considered in [5] and in this paper, is the finite element method. Using the notation of [3] we suppose that $\{T_h\}$ is a family of triangulations of the considered domain $\Omega \subset R^n$ (which is assumed to be a polyhedron) with

$$h = \max_{K \in T_h} \operatorname{diam} K$$

and that the following conditions hold:

(f_1) the family $\{T_h\}$ is regular; this means that there is a constant σ such that

$$\forall_h \forall_{K \in T_h} \frac{h_K}{\varrho_K} \leqslant \sigma,$$

where

$$h_K = \operatorname{diam} K \quad \text{and} \quad \varrho_K = \sup\{\operatorname{diam} B\colon\ B \text{ is a ball in } K\};$$

(f_2) (K, P_K, Σ_K) with $K \in \bigcup_h T_h$ is the family of finite elements of class C^0;

(f_3) (K, P_K, Σ_K) is affinely equivalent to a pattern finite element $(\hat{K}, \hat{P}_K, \hat{\Sigma}_K)$;

(f_4) $P_r \subset \hat{P} \subset H_1(\hat{K})$, where P_r is the set of all polynomials of degree $\leqslant r$;

(f_5) the set $\hat{\Sigma}$ is defined by means of the derivations of order $\leqslant s$ (so $s = 0$ in the case of finite elements of Lagrange type);

(f_6) there is a positive constant α such that

$$\forall_h \forall_{K \in T_h} \alpha h \leqslant h_K.$$

3. Simultaneous space-time Galerkin approximation. For given $u_0 \in L^2(\Omega)$, $g \in H^*(V)$ (in particular, for $g \in L^2(\Delta_T)$) and any arbitrary $v \in H(V)$ let us put

$$l_{g,u_0}(v) = (g, v)_{\Delta_T} + \big(u_0, v(\cdot, 0)\big)_\Omega,$$

where $v(\cdot, 0)$ is the trace according to [5], Lemma 3. We consider the initial-boundary value problem for equation (1) in the following weak form:

(P_1) Find $u \in H(V)$ satisfying the identity

$$B(u, v) = l_{g,u_0}(v) \tag{7}$$

for any $v \in H(V)$.

Let $Z_d \subset H(V)$ be a finite-dimensional linear subspace. In [5] the following Galerkin method of solving (P_1) was proposed:

(R_d) Find a function $u_d \in Z_d$ such that

$$B(u_d, z) = l_{g,u_0}(z) \tag{8}$$

for any $z \in Z_d$.

Following [6] we are going to prove the stability of this method. For this purpose we write identity (8) in a slightly different form. Applying the Schwarz inequality in $L^2(\Delta_T)$, for a fixed $u \in H(V)$ and arbitrary $v \in H(V)$ we get

$$\Big|\int_0^T b(t; u, v)dt\Big| \leqslant c\|u\|_{1,\Delta_T}\|v\|_{1,\Delta_T},$$

where c is a constant depending on the upper bounds of the coefficients of the operator A. Thus

$$l\colon v \to \int_0^T b(t; u, v)dt$$

is a linear functional over $H(V)$, and therefore (see Example 1) there is an $\tilde{A}u \in H^*(V)$ such that

$$(9)\qquad \int_0^T b(t; u, v)dt = (\tilde{A}u, v)_{\Delta_T},$$

$$(10)\qquad \|\tilde{A}u\|_{H^*(V)} = \|l\| \leqslant c\|u\|_{1,\Delta_T}.$$

Notice that if $Au \in L^2(\Delta_T)$, then integrating by parts the left-hand side of (9) with $v \in C_0^\infty(\Delta_T)$ we obtain $\tilde{A}u = Au$. Integrating by parts with respect to t, we see that the right-hand side of (2) (in view of [5], Lemma 6) yields now, for u, $v \in H(V)$,

$$(11)\qquad B(u, v) = (\tilde{A}u, v)_{\Delta_T} + (u_T, v)_{\Delta_T} + (u(\cdot, 0), v(\cdot, 0))_\Omega.$$

In the sequel we use the triplet $H_+ \subset H_0 \subset H_-$ defined in Example 2. Defining for an arbitrary $u \in H(V)$

$$\bar{u} = [u, u(\cdot, 0)] \in H(V) \times L^2(\Omega) \quad \text{and}$$

$$\bar{L}\bar{u} = [\tilde{A}u + u_t, u(\cdot, 0)] \in H^*(V) \times L^2(\Omega)$$

we can rewrite (11) as

$$(12)\qquad B(u, v) = (\bar{L}\bar{u}, \bar{v})_0,$$

and this yields the desired form of (8), namely

$$(13)\qquad (\bar{L}\bar{u}_d, \bar{z})_0 = (g^*, \bar{z})_0,$$

with $z \in Z_d$ and $g^* = [g, u_0]$.

Let $\bar{Z}_d = \{\bar{z}\colon z \in Z_d\}$ and let P_d be the orthogonal projection in H_0 onto $\bar{Z}_d$. Obviously, $\bar{Z}_d$ is a finite-dimensional subspace of H_+ and $\bar{L}$ is a linear continuous operator $H_+ \to H_-$ in view of (10). Consequently, Lemmas 2–4 hold true. We consider $\bar{Z}_d$ as a normed space with the induced norm $\|\ \|_0$. Putting $\bar{T}_d = P_d\bar{L}|_{\bar{Z}_d}$, we can write (13) (or, equivalently, (8)) as the Galerkin operator equation

$$(14)\qquad \bar{T}_d\bar{u}_d = P_d g^*.$$

THEOREM 1. *Equation* (14) *is uniquely solvable for any* $g \in L^2(\Delta_T)$, $u \in L^2(\Omega)$. *The solution operator* S_d *is bounded by a constant not depending on the space* Z_d.

Proof. For any $\bar{z} \in \bar{Z}_d$ we have

$$(\bar{T}_d \bar{z}, \bar{z})_0 = (\bar{L}\bar{z}, P_d \bar{z})_0.$$

Since P_d is the identity on $\bar{Z}_d$, using (12) and Lemma 1 we obtain

$$(T_d \bar{z}, \bar{z})_0 = B(z, z) \geqslant \varkappa \|\bar{z}\|_0. \tag{15}$$

Moreover, $(T_d \bar{z}, \bar{z})_0 \leqslant \|T_d \bar{z}\|_0 \|\bar{z}\|_0$. Thus

$$\|T_d \bar{z}\|_0 \geqslant \varkappa \|\bar{z}\|_0,$$

which means that the continuous operator T_d: $\bar{Z}_d \to \bar{Z}_d$ is invertible and that its range Y is closed in $\bar{Z}_d$. Moreover, one can also prove that Y is dense in $\bar{Z}_d$, because if for some $\bar{z}_0 \in \bar{Z}_d$ and any $\bar{z} \in \bar{Z}_d$ the equality $(T_d \bar{z}, \bar{z}_0)_0 = 0$ holds, then putting $\bar{z} = \bar{z}_0$ we get $\bar{z}_0 = 0$ in view of (15). Thus $Y = \bar{Z}_d$, so T_d^{-1} is defined on the whole space $\bar{Z}_d$ and $\|T_d^{-1}\| \leqslant 1/\varkappa$. Therefore, equation (14) has a unique solution

$$\bar{u}_d = T_d^{-1} P_d g^*,$$

where $g^* = [g, u_0] \in H_0$ and the solution operator $S_d = T_d^{-1} P_d$ satisfies

$$\|S_d g^*\|_0 \leqslant \frac{1}{\varkappa} \|P_d g^*\|_0 \leqslant \frac{1}{\varkappa} \|g^*\|_0.$$

Hence $\|S_d\| \leqslant 1/\varkappa$ and this completes the proof.

Theorem 1 says that the approximate method (R_d) yields uniquely defined approximate solutions of (P_1), and (R_d) is stable for any $g \in L^2(\Delta_T)$, $u_0 \in L^2(\Omega)$.

Let $\{z_j\}_{j=1}^{M_d}$ be a basis in Z_d and let us put

$$u_d = \sum_{j=1}^{M_d} \xi_j z_j$$

with unknown coefficients ξ_j. Then (8) is equivalent to the linear algebraic system of equations

$$\sum_{j=1}^{M_d} \xi_j B(z_j, z_k) = l_{g,u_0}(z_k) \quad (k = 1, \ldots, M_d).$$

Suppose, in particular, that Z_d is a finite-element space connected with a triangulation T_d of the space-time domain Δ_T satisfying (f_1)–(f_5). Then M_d is of order $d^{-(n+1)}$ and the density matrix $[B(z_j, z_k)]$ is a matrix with one non-zero band of width depending on the chosen triangulation, growing with n. It was proved in [5] that if the exact solution u is in $H_{r+1}(\Delta_T)$ and $s = 0$ (so we use finite elements of Lagrange type), then the error

$$\left(\int_0^T \|u(\cdot, t) - u_d(\cdot, t)\|_{1,\Omega}^2 dt\right)^{1/2}$$

is of order d^r.

4. Two-step Galerkin approximations. Let us consider another weak formulation of the boundary value problem for equation (1):

(P_2) Given $g \in L^2(\Delta_T)$ and $u_0 \in L^2(\Omega)$, find $u \in H(V)$ such that
(i) the identity

$$(u_t, v)_\Omega + b(t; u, v) = (g, v)_\Omega \tag{16}$$

holds for any $v \in V$ and for almost all $t \in (0, T)$;
(ii) $u(\cdot, 0) = u_0$.

Problem (P_2) yields the usual (see [4]) Galerkin semidiscretization in space variables, namely:

($Q_{1,h}$) Given a finite-dimensional linear space $V_h \subset V$ with a basis $\{v_j\}_{j=1}^{N_h}$, find a $U \in H(V_h)$ such that the identities

$$(U_t, v)_\Omega + b(t; U, v) = (g, v)_\Omega, \tag{17}$$

$$(U(\cdot, 0), v)_\Omega = (u_0, v)_\Omega \tag{18}$$

hold for all $v \in V_h$, $t \in (0, T)$.

By means of the decomposition

$$U(x, t) = \sum_{j=1}^{N_h} \alpha_j(t) v_j(x) \tag{19}$$

problem ($Q_{1,h}$) is reduced to the following one:

($Q_{2,h}$) Find $\alpha \in H_1^{N_h}(0, T)$ such that

$$C\dot{\alpha} + B(t)\alpha = \beta(t), \tag{20}$$

$$C\alpha(0) = \gamma, \tag{21}$$

where

$$\begin{gathered} C_{kj} = (v_j, v_k)_\Omega, \qquad B_{kj}(t) = b(t; v_j, v_k), \\ \beta_k(t) = (g(\cdot, t), v_k)_\Omega, \qquad \gamma_k = (v_j, v_k)_\Omega \quad (j, k = 1, \ldots, N_h). \end{gathered} \tag{22}$$

In [5] the Galerkin method for the approximate solution of ($Q_{2,h}$) was proposed. Let $X_{h,\tau}$ be a finite-dimensional subspace of $H_1^{N_h}(0, T)$ with basis $\{\phi^{(m)}\}_{m=1}^{M_{h,\tau}}$, and let us put

$$d(\alpha, \phi) = (B\alpha, \phi) - (C\alpha, \dot{\phi}) + \langle C\alpha(T), \phi(T)\rangle,$$

$$p_{\beta,\gamma}(\phi) = (\beta, \phi) + \langle \gamma, \phi(0)\rangle.$$

We now formulate an approximate problem as follows:

($Q_{h,\tau}^*$) Find $\alpha^* \in X_{h,\tau}$ such that

$$d(\alpha^*, \phi) = p_{\beta,\gamma}(\phi) \tag{23}$$

holds for any $\phi \in X_{h,\tau}$.

Solving $(Q^*_{h,\tau})$ we obtain an approximate solution of (P_2) as

$$U^*(x, t) = \sum_{j=1}^{N_h} \alpha_j^*(t) v_j(x).$$

Let us put

$$(v\cdot\phi)(x, t) = \sum_{k=1}^{N_h} v_k(x)\phi_k(t).$$

It is easy to check that for any $\alpha, \phi \in H_1^{N_h}(0, T)$ and U defined by (19) we have

$$v\cdot\phi \in H(V), \quad d(\alpha, \phi) = B(U, v\cdot\phi), \quad p_{\beta,\gamma}(\phi) = l_{g,u_0}(v\cdot\phi).$$

Therefore, α^* is a solution of $(Q^*_{h,\tau})$ if and only if U^* satisfies

(24) $$B(U^*, v\cdot\phi^{(m)}) = l_{g,u_o}(v\cdot\phi^{(m)})$$

for $m = 1, \ldots, M_{h,\tau}$. The functions $\{v\cdot\phi^{(m)}\}_{m=1}^{M_{h,\tau}}$ form a linear independent system in $H(V)$. Suppose namely that

$$\sum_{m=1}^{M_{h,\tau}} c_m(v\cdot\phi^{(m)})(x, t) = \sum_{k=1}^{N_h} v_k(x) \sum_{m=1}^{M_{h,\tau}} c_m \phi_k^{(m)}(t) = 0.$$

Then

$$\sum_{m=1}^{M_{h,\tau}} c_m \phi_k^{(m)})(t) = 0$$

for any $t \in (0, T)$ and any $k = 1, \ldots, N_h$. This means that

$$\sum_{m=1}^{M_{h,\tau}} c_m \phi_k^{(m)}) = 0,$$

and therefore $c_m = 0$ for any m. Thus, the approximate problem $(Q^*_{h,\tau})$ is a particular case of the problem (R_d) with the space Z_d spanned by the system $\{v\cdot\phi^{(m)}\}_{m=1}^{M_{h,\tau}}$. According to Theorem 1, this yields the unique approximate solution of (P_2) (or, equivalently, of (P_1)) and is stable for $g \in L^2(\Delta_T)$, $u \in L^2(\Omega)$.

In the sequel we consider a special form of $X_{h,\tau}$, namely $X_{h,\tau} = (Y_\tau)^{N_h}$, where Y_τ is a subspace of $H_1(0, T)$ with a finite basis $\{Y_j\}_{j=1}^{R_\tau}$. Then the vector functions $\phi^{(j,r)}$ with $\phi_k^{(j,r)} = \delta_{kr} y_j$ $(k, r = 1, \ldots, N_h; j = 1, \ldots, R_\tau)$ form a basis in X_τ and we have

$$(v\cdot\phi^{(j,r)})(x, t) = v_r(x) y_j(t).$$

Identity (24) and, equivalently, problem $(Q^*_{h,\tau})$ reduce now to the system of $M_{h,\tau} = N_h \cdot R_\tau$ equations

(25) $$\sum_{m=1}^{N_h} \sum_{r=1}^{R_\tau} \xi_{mr} B(v_m y_r, v_k y_s) = l_{g,u_0}(v_k y_s)$$

$$(k = 1, \ldots, N_h; \; s = 1, \ldots, R_\tau)$$

if we put

$$U^*(x,t) = \sum_{m=1}^{N_h} \sum_{r=1}^{R_\tau} \xi_{mr} v_m(x) y_r(t)$$

with unknown coefficients ξ_{mr}. It is easy to verify that the matrix of the system (25) (the density matrix) consists of blocks which are $(R_\tau \times R_\tau)$-matrices. If, in particular, $n = 1$ and V_h, Y_τ are finite element spaces containing sectionally linear splines only, then the density matrix is a three-diagonal block matrix and each block is a three-diagonal matrix.

From now on we suppose that

(a_4) Ω is a polyhedron in R^n;

(a_5) V_h is a finite element space of Lagrange type connected with the triangulation T_h of Ω and (f_1)-(f_5) hold;

(a_6) Y_τ is a finite element space of Lagrange type connected with the triangulation T_τ of the segment $[0, T]$ and (f_1)-(f_5) hold with h replaced by τ and r replaced by l;

(a_7) the derivatives $D_t^j g$ are in $L^2(\Delta_T)$ $(j = 0, 1, \ldots, l)$;

(a_8) the derivatives in the classical sense

$$\frac{\partial^r}{\partial t^r} a_{jk}, \ \frac{\partial^r}{\partial t^r} a_j, \ \frac{\partial^r}{\partial t^r} a \qquad (r = 0, 1, \ldots, l)$$

exist, are bounded in Δ_T, and continuous with respect to t.

Our aim is to estimate the error $U - U^*$ in a suitably defined norm. This is easy to do using the Schwarz inequality

$$\|U - U^*\|_{\Delta_T}^2 \leqslant \|\alpha - \alpha^*\|^2 \sum_{k=1}^{N_h} \|v_k\|_\Omega^2 \tag{26}$$

and

$$\int_0^T \|U(\cdot, t) - U^*(\cdot, t)\|_{1,\Omega}^2 dt \leqslant \|\alpha - \alpha^*\|^2 \sum_{k=1}^{N_h} \|v_k\|_{1,\Omega}^2. \tag{27}$$

In [5] some estimates for the error $\alpha - \alpha^*$ were obtained with constants depending on the coefficients of (20) or, equivalently, on the space-discretization parameter h. We are now going to estimate the right-hand sides of (26) and (27) in terms of h and τ.

In the sequel we denote by $\bar{c}$, $\bar{c}_j$, etc. positive constants not depending on h. For a fixed triangulation T_h

$$\Omega = \bigcup_{s=1}^{t_h} K_s$$

we denote by y_j $(j = 1, \ldots, N_h)$ the knots and assume that $v_j(y_k) = \delta_{jk}$. We also write

$$\tau_{j,h} = \{s\colon\ K_s \in T_h,\ y_j \in K_s\}.$$

We now prove some lemmas:

Lemma 5. *There exists an integer $\hat{q}$ with the property that for every $h > 0$, $j = 1, \ldots, N_h$ the set $\tau_{j,h}$ contains at most $\hat{q}$ elements.*

Proof (a contrario). Let us assume that

$$\underset{m}{\forall}\ \underset{h}{\exists}\ \underset{j}{\exists}\ \bar{\bar{\tau}}_{j,h} > m,$$

where $\bar{\bar{\tau}}_{j,h}$ is the cardinality of $\tau_{j,h}$. Let $D(x, r)$ denote the ball with center at $x \in R^n$ and radius r. The inclusion

$$\bigcup_{s \in \tau_{j,h}} K_s \subset D(y_j, h)$$

holds for arbitrary $y_j \in \tau_{j,h}$. Thus

$$\text{(28)} \qquad \operatorname{Vol}\Big(\bigcup_{s \in \tau_{j,h}} K_s\Big) < \operatorname{Vol}(D(y_j, h)) = h^n v_n,$$

where v_n is the volume of the unit ball in R^n. Now, in view of (f_1) and (f_6), bounding the left-hand side value from below, we have

$$\operatorname{Vol}\Big(\bigcup_{s \in \tau_{j,h}} K_s\Big) = \sum_{s \in \tau_{j,h}} \operatorname{Vol} K_s > \sum_{s \in \tau_{j,h}} \varrho_{K_s}^n v_n > m(\alpha/\sigma)^n h^n v_n,$$

which is a contradiction with (28) because m may be arbitrarily large.

Lemma 6. *There exists a positive constant $\hat{c}$ such that for an arbitrary $h > 0$ the inequality $N_h \leqslant \hat{c}h^{-n}$ holds true.*

Proof. Let us denote by $C(h)$ the n-dimensional cube with the side length equal to h. Then $\operatorname{diam} C(h) = n^{1/2}h$. The inclusion

$$\Omega_C := c_1 C(1) \supset \Omega$$

holds for some positive c_1. Let us divide the cube Ω_C into $c_1 n^{n/2} h^{-n}$ cubes with the side length equal to $n^{-1/2}h$ (and diam $\leqslant h$). Using (f_1) and (f_6) we can write the inequalities

$$\varrho_K/\alpha h \geqslant \varrho_K/h_K \geqslant 1/\sigma \quad \text{or} \quad \sigma\varrho_K/\alpha \geqslant h,$$

which imply the possibility of covering the whole Ω with $c_1 n^{n/2} h^{-n} \alpha\sigma^{-1}$ cubes of diam $\leqslant \varrho_K$ for an arbitrary $K \in T_h$. Since the pattern element $\hat{K}$ has β knots, the number of all knots in Ω is less than $\beta c_1 n^{n/2} h^{-n} \alpha\sigma^{-1}$.

Lemma 7. *Let us put $K_s = F_s(\hat{K})$ with*

$$\text{(29)} \qquad F_s\colon y = A_s x + a_s \quad (s = 1, \ldots, t_h)$$

according to (f_3). *Then*

$$\bar{c}_1 h^n \leqslant |\det A_s| \leqslant \bar{c}_2 h^n \tag{30}$$

holds for any s *and* h.

Proof. For a measurable set $\Xi \subset R^n$ let us write $|\Xi|$ for its Lebesgue measure. Then the substitution $y = F_s(x)$ in the integral gives

$$|K_s| = \int_{K_s} dy = |\det A_s| \int_{\hat{K}} dx = |\det A_s|\,|\hat{K}|,$$

so

$$|\det A_s| = \frac{|K_s|}{|\hat{K}|}.$$

Since for any $K \in T_h$ we have $c\varrho_K^n \leqslant |K| \leqslant h_K^n$ with some positive c, (29) follows from (f_1) and (f_6).

Lemma 8. *Using the notation of Lemma* 7 *let us put*

$$A_s = [A_{jk,s}], \quad A_s^{-1} = [D_{jk,s}].$$

Then

$$|D_{jk,s}| \leqslant \bar{c}h^{-1}. \tag{31}$$

Proof. For a fixed p let us choose two points $\bar{x}$, $\bar{\bar{x}} \in \hat{K}$ such that

$$(\bar{x}-\bar{\bar{x}})_j = \bar{c}_1 \delta_{pj} \quad (j = 1, \ldots, n).$$

Then for any $k = 1, \ldots, n$ we have

$$h \geqslant |F_s(\bar{x}) - F_s(\bar{\bar{x}})| \geqslant \bar{c}_1 |A_{kp,s}|,$$

and this yields (31) by using (30).

Lemma 9. *Let*

$$C(\xi) = \sum_{j,k=1}^{N_h} C_{kj}\xi_k\xi_j.$$

Then

$$C(\xi) \geqslant 2h^n \hat{c}|\xi|^2. \tag{32}$$

Proof. Let us put

$$p_s(\xi) = \sum_{j,k=1}^{N_h} \xi_k\xi_j \int_{K_s} v_j(y)v_k(y)dy$$

and suppose that K_s contains the knots $y_{n_j(s)}$ ($j = 1, \ldots, r$) only. Equivalently, the only base functions not vanishing on K_s are $v_{n_j(s)}$ ($j = 1, \ldots, r$), and therefore

$$p_s(\xi) = \sum_{j,k=1}^{r} \xi_{n_k(s)}\xi_{n_j(s)} \int_{K_s} v_{n_j(s)}(y)v_{n_k(s)}(y)dy$$

or, after transforming the integral by means of (29),

$$p_s(\xi) = \sum_{j,k=1}^{r} \hat{\xi}_k \hat{\xi}_j \int_{\hat{K}} \hat{v}_j(x)\hat{v}_k(x)dx\,|\det A_s|, \tag{33}$$

where $v_{n_j(s)}(y) = \hat{v}_j(x)$, $\xi_{n_j(s)} = \hat{\xi}_j$ $(j = 1,\ldots, r)$. It follows from (33) and Lemma 7 that

$$p_s(\xi) \geqslant 2h^n \hat{c}|\hat{\xi}|^2.$$

Then summing over all s we get (32).

LEMMA 10. *Let*

$$B(t, \xi) = \sum_{j,k=1}^{N_h} B_{kj}(t)\xi_k\xi_j.$$

Then

$$B(t, \xi) \geqslant 2h^n \hat{c}|\xi|^2 \tag{34}$$

for $t \in [0, T]$.

Proof. Let us put

$$v_\xi = \sum_{j=1}^{N_h} \xi_j v_j.$$

Then

$$B(t, \xi) = b(t; v_\xi, v_\xi) \geqslant 2\varkappa C(\xi)$$

in view of (a_3), and (34) follows from (32).

We need the following special form of the theorem of Gerschgorin in further calculations (for its proof see [2]).

LEMMA 11. *Suppose λ is an eigenvalue of a symmetric matrix $A = [A_{jk}]$ and put*

$$r_j = \sum_{k\neq j} |A_{jk}|.$$

Then for some j we have

$$A_{jj} - r_j \leqslant \lambda \leqslant A_{jj} + r_j. \tag{35}$$

Using this lemma it is easy to prove

LEMMA 12. *We have*

$$|C| \leqslant \hat{c}h^n. \tag{36}$$

Proof. Since

$$\operatorname{supp} v_j = \bigcup_{s\in\tau_{j,k}} K_s,$$

we have

$$C_{jk} = \sum_{s\in\tau_{j,h}\cap\tau_{k,h}} \int_{K_s} v_j(y)v_k(y)dy, \tag{37}$$

and therefore putting

$$\varrho_j = \{k\colon\ y_k\in K_s \text{ for } s\in\tau_{j,h}\}, \tag{38}$$

we get

$$r_j \leqslant \sum_{\substack{k\in\varrho_j\\ k\neq j}} \sum_{s\in\tau_{j,h}} \Big|\int_{K_s} v_j(y)v_k(y)dy\Big|. \tag{39}$$

According to Lemma 5, the number of terms on the right-hand side is bounded by a constant $\hat{p}$. Transforming the integrals in (37) and (39) by means of (29) and using (30), we get

$$|C_{jj}|+r_j \leqslant \hat{c}h^n.$$

Since C is symmetric, we can use Lemma 11 to obtain

$$|C| = \max|\lambda| \leqslant |C_{jj}|+r_j,$$

which yields (36).

It follows from (a_8) that $B_{kj}\in C^l[0, T]$. Putting

$$B_j = \sup_{[0,T]} |B^{(j)}| \qquad (j = 0, 1, \ldots, l)$$

we have

LEMMA 13. *The inequality*

$$|B_j| \leqslant \hat{c}h^{(n/2-2)} \qquad (j = 0, 1, \ldots, l) \tag{40}$$

holds true.

Proof. According to (f_3) there is, for a fixed j, a basis function $\hat{v}_j$ over $\hat{K}$ such that $v_j(y) = \hat{v}_j(x)$ with x, y related by (29). Differentiation of both sides yields, in view of Lemma 6,

$$\left|\frac{\partial v_j}{\partial y_r}\right|^2 \leqslant \hat{c}_1 h^{-2} \sum_{s=1}^{n} \left|\frac{\partial \hat{v}_j}{\partial x_s}\right|.$$

Integrating both sides and using Lemma 7, we obtain

$$\|v_j\|^2_{1,K_s} \leqslant \hat{c}_2 h^{n-2} \tag{41}$$

for every $K_s\in T_h$. Now, in view of (41) and Lemma 5, we have

$$\|v_j\|^2_{1,\Omega} = \sum_{s\in\tau_{j,h}} \|v_j\|^2_{1,K_s} \leqslant \hat{c}_2\hat{q}h^{n-2}, \tag{42}$$

and therefore

$$|B_{kj}| \leqslant \hat{c}_3 \|v_j\|_{1,\Omega}\|v_k\|_{1,\Omega} \leqslant \hat{c}_4 h^{n-2}. \tag{43}$$

Moreover,

(44)
$$|B| \leqslant \Big(\sum_{j,k=1}^{N_h} |B_{kj}|^2\Big)^{1/2}.$$

But for a fixed j the term $|B_{kj}|$ does not vanish only for $k \in \varrho_j$ (see (38)), so the second sum contains at most $\hat{p}$ terms. Therefore using (43), (44) and Lemma 6 we get (40) for $j = 0$. The proof for $j \geqslant 1$ goes along the same lines and is omitted.

LEMMA 14. *Suppose that $B(t)$ is symmetric for $t \in [0, T]$. Then*

(45)
$$|B_j| \leqslant \hat{c}h^{n-2} \quad (j = 0, 1, \ldots, l).$$

Proof. It follows from Lemma 11 that for some j we have

$$|B(t)| \leqslant \sum_{k=1}^{N_h} |B_{kj}(t)|.$$

The sum contains at most $\hat{p}$ terms, so (43) yields (45) for $j = 0$. For $j \geqslant 1$ the proof is quite the same.

LEMMA 15. *The inequality $|C^{-1}| \leqslant \hat{c}h^{-n}$ holds true.*

The above lemma follows from Lemma 9.

LEMMA 16. *The following inequalities hold:*

(46)
$$|\gamma| \leqslant \hat{c}h^{n/2},$$

(47)
$$\|\beta^{(j)}\| \leqslant \hat{c}h^{n/2} \quad (j = 0, 1, \ldots, l).$$

Proof. Denoting the support of v_j by $\underline{v}_j$ we have

(48)
$$\Big|\int_{\underline{v}_j} u_0 v_j dy\Big|^2 \leqslant \int_{\underline{v}_j} u_0^2 dy \int_{\underline{v}_j} v_j^2 dy.$$

Since for any j we have

$$\int_{\underline{v}_j} v_j^2(y) dy = \sum_{s \in \tau_{j,h}} \int_{K_s} v_j^2(y) dy,$$

transforming the integral on the right by means of (29) and using (30) together with Lemma 5 we get

(49)
$$\int_{\underline{v}_j} v_j^2 dy \leqslant \hat{c}_1 h^n.$$

Inequalities (48) and (49) yield

(50)
$$|\gamma|^2 = \sum_{j=1}^{N_h} \Big|\int_{\underline{v}_j} u_0 v_j dy\Big|^2 \leqslant \alpha \hat{c}_1 h^n$$

with

$$\alpha = \sum_{j=1}^{N_h} \int_{\underline{v}_j} u_0^2 dy.$$

Suppose that the pattern element $\hat{K}$ constains $\hat{r}$ knots. Then each point of Ω belongs to $\underline{v}_j$ for at most $\hat{r}$ different j. Therefore

(51) $$\alpha \leqslant \hat{r} \int_{\Omega} u_0^2 \, dy$$

and (46) follows from (50) and (51). The proof of (47) goes along the same lines.

We can prove now our main result:

THEOREM 2. *Suppose that the assumptions* (a_1)–(a_8) *and* (f_1)–(f_6) *hold true. Then*

(52) $$\|U - U^*\|_{\Delta_T} \leqslant c\phi_l(\tau, h)$$

and

(53) $$\Big(\int_0^T \|U(\cdot, t) - U^*(\cdot, t)\|_{1,\Omega}^2 dt\Big)^{1/2} \leqslant ch^{-1}\phi_l(\tau, h),$$

where

(54) $$\phi_l(\tau, h) = \tau^l h^{-(n/2)(l+3)-2(l+2)}$$

and c is a constant not depending on τ and h.

If the bilinear form b is symmetric in u, v for any $t \in [0, T]$, then we can put

(55) $$\phi_l(\tau, h) = \tau^l h^{-(n/2)-2(l+2)}.$$

Proof. We denote by c_j any positive constant not depending on h and τ. It follows from (49) and Lemma 6 that

(56) $$\sum_{k=1}^{N_h} \|v_k\|_{\Omega}^2 \leqslant c_1.$$

Similarly, using (42) and Lemma 6 we get

(57) $$\sum_{k=1}^{N_h} \|v_k\|_{1,\Omega}^2 \leqslant c_2 h^{-2}.$$

In view of (26) and (27) it remains to estimate the error $\|\alpha - \alpha^*\|_0$, where α and α^* are the solutions of ($Q_{2,h}$) and ($Q^*_{h,\tau}$), respectively. It follows from the lemmas proved above and from [5] (Theorems 1, 4, 5 and estimate (27)) that $\alpha \in H_l^{N_h}(0, T)$ and the following estimates hold true:

(58) $$\|\alpha - \alpha^*\| \leqslant c_3 \mu \varkappa^{-1} \|\alpha^{(l+1)}\| \tau^l,$$

(59) $$\|\alpha\| \leqslant \varkappa^{-1}(\|\beta\| + |\gamma|),$$

where $\varkappa = h^n$ and $\mu = h^{(n/2)-2}$ in the case of an arbitrary form b, and $\mu = h^{n-2}$ when b is symmetric. It remains to estimate $\|\alpha^{(l+1)}\|$. As α satisfies (20), we have

$$\alpha^{(l+1)} = C^{-1}\Big(\beta^{(l)} + \sum_{k=0}^{l} \binom{l}{k} B^{(k)} \alpha^{(l-k)}\Big),$$

and therefore

$$\|\alpha^{(l+1)}\| \leqslant |C^{-1}|\left(\|\beta^{(l)}\| + \sum_{k=0}^{l}\binom{l}{k}B^{(k)}\|\alpha^{(l-k)}\|\right).$$

Consequently, using induction on l and estimates obtained in Lemmas 12–16 together with (59), we obtain

$$\|\alpha^{(l+1)}\| \leqslant c_4 h^{-(n/2)(l+2)-2(l+1)} \tag{60}$$

and

$$\|\alpha^{(l+1)}\| \leqslant c_5 h^{-(n/2)-2(l+1)} \tag{61}$$

in the case where β is symmetric in u and v. Thus the theorem follows from (26) and (27) by using (56)–(60).

Suppose now $\tau = h^{\alpha}$ with some $\alpha > 0$. Then using (54) we get

$$\phi_1(h^{\alpha}, h) = h^{\beta} \quad \text{with } \beta = l\alpha - 2(l+2) - \frac{n}{2}(l+3)$$

and a sufficient condition for $\phi_l(h^{\alpha}, h) \to 0$ as $h \to 0$ is

$$\alpha > \left[2(l+2) + \frac{n}{2}(l+3)\right] l^{-1}.$$

In the simplest case $n = l = 1$ this yields $\alpha > 8$, in the symmetric case using (55) we obtain

$$\alpha > \left[2(l+2) + \frac{n}{2}\right] l^{-1},$$

so in the case $n = l = 1$ the inequality $\alpha > 6.5$ is sufficient for $\phi_1(h^{\alpha}, h) \to 0$ as $h \to 0$.

It is evident that the estimates of the error (52), (53) obtained in the general case are most unsatisfactory. It seems that for some special choice of spaces V_h and $X_{h,\tau}$ the estimate of the error could be improved, but this needs further investigations.

References

[1] R. A. Adams, *Sobolev Spaces*, New York 1975.
[2] Å. Björck and G. Dahlquist, *Numerical Methods*, Prentice-Hall, 1974 (Polish translation: Warszawa 1983).
[3] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Amsterdam 1978 (Russian translation: Moscow 1980).
[4] J. Douglas, Jr., and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal. 7 (1970), pp. 575–626.

[5] H. Marcinkowska, *On Galerkin approximations of parabolic equations in time dependent domains*, Zastos. Mat. 19 (1987), pp. 289–307.
[6] W. L. Wendland, *Asymptotic convergence of boundary element methods*, Technische Hochschule Darmstadt, Preprint 611 (1981), pp. 1–42.

HANNA MARCINKOWSKA
MATHEMATICAL INSTITUTE
UNIVERSITY OF WROCŁAW
PL. GRUNWALDZKI 2/4
PL 50-384 WROCŁAW

ADAM SZUSTALEWICZ
INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
UL. PRZESMYCKIEGO 20
PL 51-151 WROCŁAW

Received on 1987.12.15