Applications of Mathematics

Josef Dalík

A Petrov-Galerkin approximation of convection-diffusion and reaction-diffusion problems

Applications of Mathematics, Vol. 36 (1991), No. 5, 329–354

Persistent URL: http://dml.cz/dmlcz/104471

36 (1991) APPLICATIONS OF MATHEMATICS No. 5, 329—354

A PETROV-GALERKIN APPROXIMATION OF CONVECTION-DIFFUSION AND REACTION-DIFFUSION PROBLEMS

JOSEF DALÍK

(Received October 26, 1988)

Summary. A general construction of test functions in the Petrov-Galerkin method is described. Using this construction, algorithms for an approximate solution of the Dirichlet problem for the differential equation $-\varepsilon u'' + pu' + qu = f$ are presented and analyzed theoretically. The positive number ε is supposed to be much less than the discretization step and the values of $|p|$, q. An algorithm for the corresponding two-dimensional problem is also suggested and results of numerical tests are introduced.

Keywords: convection-diffusion problem with dominated convection, Petrov-Galerkin method.

AMS Subject Classification: 65L99, 65N99.

INTRODUCTION

Let a normed function space $\mathscr{H}$, a continuous bilinear form $a: \mathscr{H} \times \mathscr{H} \to \mathbb{R}$, a continuous linear form $f: \mathscr{H} \to \mathbb{R}$ and a problem

$$\text{(1)} \qquad \text{find} \quad u \in \mathscr{H}: \ a(u, v) = f(v) \quad \forall v \in \mathscr{H}$$

be given. Let us choose subspaces $\mathscr{V}$ (with a basis $\Phi_1, \ldots, \Phi_n$), $\mathscr{W}$ (with a basis $\Psi_1, \ldots, \Psi_k$) in $\mathscr{H}$ such that $n < k$, and a bilinear form $A: \mathscr{V} \times \mathscr{V} \to \mathbb{R}$. Let a linear operator $\mathscr{L}: \mathscr{V} \to \mathscr{W}$ transform the given form a into the chosen form A in the following manner:

$$a(u, \mathscr{L}v) = A(u, v) \quad \forall u, v \in \mathscr{V} .$$

We call the problem

$$\text{(2)} \qquad \text{find} \quad u^h \in \mathscr{V}: \ a(u^h, \mathscr{L}v) = f(\mathscr{L}v) \quad \forall v \in \mathscr{V}$$

a (*Petrov-Galerkin*) *$\mathscr{L}$-discretization of* (1)[1]) and the function u^h an *$\mathscr{L}$-discrete solution of* (1)[1]). If we denote $u^h = u_1\Phi_1 + \ldots + u_n\Phi_n$ then (2) can be written in

[1]) Or of the corresponding classically formulated problem.

the form

$$\text{(3)} \qquad \text{find} \quad u_1, \ldots, u_n: \sum_{j=1}^{n} u_j a(\Phi_j, \mathscr{L}\Phi_i) = f(\mathscr{L}\Phi_i) \quad \text{for} \quad i = 1, \ldots, n\,.$$

The matrix of the system (3) will be called and $\mathscr{L}$-*matrix of* $(1)^1$). In this situation the elements of $\mathscr{V}$ and $\mathscr{L}(\mathscr{V})$ are referred to as *shape functions* and *test functions* respectively.

Let us consider the problem

$$\text{(4)} \qquad -\varepsilon u'' + pu' + qu = f \quad \text{on} \quad (a, b)\,, \quad u(a) = 0 = u(b)\,,$$

where ε is a positive real number, $p \in \mathbb{C}^1(a, b)$, $q \in \mathbb{L}^\infty(a, b), 0 \leqq q - 0.5p'$ on (a, b) and $f \in \mathbb{L}^2(a, b)$. Using the Lax-Milgram theorem, one can see that (4) has exactly one week solution u in $\mathbb{H}_0^1(a, b)$ and the supposition $q \in \mathbb{L}^\infty(a, b)$ implies $u \in \mathbb{H}^2(a, b)$. But if $\varepsilon \ll |p|$ or $\varepsilon \ll q$ then there usually exist some small subintervals in (a, b), called *boundary* or *internal layers*, in which values of u change extremely quickly. This fact together with the use of the discretization step greater than ε are the reasons why the classical finite difference method and the Galerkin method, both having an optimal order of convergence, offer approximate solutions debased by oscillations on the entire interval. In order to remove these oscillations, one can simply use a suitably decentralized approximation of u in the finite difference method or substitute a suitable major value for ε in the Galerkin method. Unfortunately, both these modifications reduce the order of convergence. For example, the Galerkin approximation by linear splines reduces the order of convergence from two to one in the $\mathbb{L}^2$-norm and from one to zero in the $\mathbb{H}^1$-norm. Several numerical methods giving nonoscillating approximate solutions of various special cases of (4) have been published. For some of them higher orders of convergence than those of the modifications have been proved. See for example [2], [8], [17] and surveys in [3], [10]. There also exist algorithms producing nonoscillating approximate solutions of the two-dimensional convection-diffusion problem with good convergence properties. See [5], [9–16], [18].

The aim of this paper is to show that the Petrov-Galerkin $\mathscr{L}$-discretization with a suitable bilinear form A can be successfully applied to the problem (4) and also to its two-dimensional analogue. The analyzed algorithms search for a solution in the space of linear splines on an equidistant net on (a, b) in the following cases 1, 2.

1. $q \equiv 0$. If $p \geqq 1$ on (a, b) then it is proved that the $\mathbb{L}^2$-norm of the error is of order 1.5 and its $\mathbb{H}^1$-norm is of order 1. If one admits a first-order zero of p at one of the given nodes then the order of error decreases by 0.5 in both norms.

2. $p \equiv 0, 0 < q_0 \leqq q$ on (a, b) and q is uniformly continuous. The $\mathbb{L}^2$-, $\mathbb{H}^1$-norm of the error is of order 2, 1, respectively.

These error estimates are ε-uniform in the sense that they contain only constants independent of ε. This property is shared also by the so-called local error estimates, which are of the same order as the estimates of the corresponding global errors. All approximate solutions obtained are shown not to oscillate.

In $\mathbb{R}^n$, the symbol $\mathbf{o}$ will be reserved for the null-vector, $\mathbf{PQ}$ for a vector determined by an ordered pair $\mathbf{P}$, $\mathbf{Q}$ of points and $|\mathbf{PQ}|$ for its Euclidean norm. Notation for function spaces on an open interval (a, b) in $\mathbb{R}$ and in a bounded open subset Ω in $\mathbb{R}^2$ with a polygonal boundary Γ are used in the sense of [6]. Also the symbols $(\ ,\ )$ for the scalar product both in $\mathbb{L}^2(a, b)$ and $\mathbb{L}^2(\Omega)$, $\|\ \|$, $\|\ \|$ $\|\ \|_\infty$ for the norms in $\mathbb{L}^2(a, b)$, $\mathbb{L}^\infty(a, b)$, respectively, and $|\ |_1$, $|\ |_2$ for the seminorms in $\mathbb{H}^1(a, b)$, $\mathbb{H}^2(a, b)$, respectively, are taken from [6]. By $\|\ \|_\infty$ the max-norm in $\mathbb{R}^n$ is denoted, too. The symbol $g|_D$ stands for a restriction of a real function g from (a, b) to an interval $D \subseteq (a, b)$. If $g \in \mathbb{H}^1(a, b)$ and E is a positive piecewise constant function on (a, b) then we put $|g|_{E,1} = (Eg', g')^{1/2}$. Any generic constant in this text depends neither on ε nor on the step length.

The following problem is a weak formulation of (4).

(5) Find $u \in \mathbb{H}_0^1(a, b)$: $\alpha(u, v) = (f, v)$ $\forall v \in \mathbb{H}_0^1(a, b)$. Here $\alpha(u, v) = \int_a^b (\varepsilon u'v' + pu'v + quv)\,\mathrm{d}x$.

1.1. **Definition.** Let n be a positive integer, $m = n + 1$ and let $a = x_0 < x_1 < \ldots$ $\ldots < x_m = b$, $a = x_0^* < x_1^* < \ldots < x_{2m}^* = b$ be equidistant nodes with step length h, h^*, respectively. Further, let us put $x_{-1} = a - h$, $x_{-1}^* = a - h^*$, $x_{2m+1}^* = b + h^*$, $x_{m+1} = b + h$.

1.2. **Definition.** For each interval $D = (x_k, x_l) \subseteq (a, b)$ let us define the extension D^e of D by $D^e = (x_{k-1}, x_{l+1}) \cap (a, b)$.

1.3. **Definition.** Let us define scalar-valued functions

$$\varphi_i(x) = \begin{cases} 1 - |x - x_i| \dfrac{1}{h} & \text{for } x \in (a, b) \cap (x_{i-1}, x_{i+1}), \\ 0 & \text{for } x \in (a, b) - (x_{i-1}, x_{i+1}), \end{cases}$$

$i = 0, \ldots, m$,

$$\psi_j(x) = \begin{cases} 1 - |x - x_j^*| \dfrac{1}{h^*} & \text{for } x \in (a, b) \cap (x_{j-1}^*, x_{j+1}^*), \\ 0 & \text{for } x \in (a, b) - (x_{j-1}^*, x_{j+1}^*), \end{cases}$$

$j = 0, \ldots, 2m$, vector-valued functions

$$\Psi_i(x) = [\psi_{2i-1}(x), \psi_{2i}(x), \psi_{2i+1}(x)]^T,$$

$i = 1, \ldots, n$, and linear spaces

$$V_h(a, b) = \operatorname{span}\{\varphi_1, \ldots, \varphi_n\}, \quad V_{h^*}(a, b) = \operatorname{span}\{\psi_1, \ldots, \psi_{2n+1}\}.$$

If L: $V_h(a, b) \to V_{h*}(a, b)$ is a linear operator then an *L-discretization of* (5) is the problem

(6) find $u^h \in V_h(a, b)$: $\alpha(u^h, Lv) = (f, Lv) \quad \forall v \in V_h(a, b)$.

1.4. **Notation.** For every function $v \in V_h(a, b)$ we put

$$v = \sum_{i=1}^{n} v_i \varphi_i \quad \text{and} \quad v_0 = 0 = v_m .$$

1.5. **Lemma.** *If v is an arbitrary function from $V_h(a, b)$ then the following assertions* (a), (b), (c) *are true.*

(a) $|v|_1^2 = \frac{1}{h} \sum_{i=1}^{m} (v_i - v_{i-1})^2$.

(b) $\|v\|^2 = \frac{h}{3} \sum_{i=1}^{m} (v_{i-1}^2 + v_{i-1} v_i + v_i^2)$.

(c) $\frac{h}{6} |v|_1^2 + \frac{1}{h} \|v\|^2 = \sum_{i=1}^{n} v_i^2$.

Proof. The equations (a), (b) can be verified by a direct computation; (c) follows immediately by (a), (b).

1.6. **Definition.** Let Q be a constant such that $0 \leqq Q \leqq q - 0.5p'$ on (a, b) and let E be a positive piecewise constant function on (a, b). We define a norm

$$[v] = (|v|_{E,1}^2 + Q\|v\|^2)^{1/2}$$

in $\mathring{H}_0^1(a, b)$.

1.7. **Proposition.** *Let us suppose that a linear operator L: $V_h(a, b) \to V_{h*}(a, b)$, real numbers k, l with the property $1 \leqq k \leqq l$ and parameters ε, h satisfy the following conditions.*

(c1) *There exist positive constants C_1, C_2 such that $\max\{\varepsilon, C_1 h^l\} \leqq E \leqq C_2 h^k$ on (a, b),*

(c2) $[v]^2 \leqq C\alpha(v, Lv)$,

(c3) $|Lv|_1 \leqq Ch^{-l/2}[v]$,

(c4) $\|Lv - v\| \leqq Ch^{1-l/2}[v]$

for all $v \in V_h(a, b)$. Then the following assertions (a1), (a2) *hold for the solution u of* (5), *the L-discrete solution u^h of* (5) *and the interpolation $\tilde{u}$ of u in $V_h(a, b)$.*

(a1) $[u - \tilde{u}] \leqq Ch^{\min\{1+k/2, 2\}} |u|_2$.

(a2) $[u^h - \tilde{u}]^2 \leqq C_1(h^{1+k/2} + \|p\|_\infty h^{2-l/2} + \|q\|_\infty h^{3-k/2}) |u|_2 [u^h - \tilde{u}] +$

$+ C_2(\|p'\|_\infty + \|q\|_\infty) h^2 |u|_2 \|u^h - \tilde{u}\|$.

Proof. Let us put $\eta = u - \tilde{u}$ and $\theta = u^h - \tilde{u}$.

(i) $h|\eta|_1 + \|\eta\| \leqq Ch^2|u|_2$ is true by [15], Theorem 3.2.1.

(ii) $\alpha(\theta, L\theta) = \alpha(\eta, L\theta)$ is a consequence of the equation $\alpha(u - u^h, L\theta) = 0$. This one follows by (5) and (6).

Proof of (a1). The condition (c1) and statement (i) imply $[\eta]^2 \leqq Ch^k|\eta|_1^2 + Q\|\eta\|^2 \leqq Ch^{\min\{2+k,4\}}|u|_2^2$.

Proof of (a2).

1° $|\varepsilon(\eta', (L\theta)')| \leqq Ch^{1+k-l/2}|u|_2[\theta]$ by (i), (c1), (c3).

2° $|(p\eta', L\theta - \theta)| \leqq C\|p\|_\infty h^{2-l/2}|u|_2[\theta]$ by virtue of (i), (c4).

3° $|(p'\eta, \theta)| \leqq C\|p'\|_\infty h^2|u|_2\|\theta\|$ by (i).

4° $|(p\eta, \theta')| \leqq C\|p\|_\infty h^{2-l/2}|u|_2[\theta]$ follows by (i), (c1).

5° $|(q\eta, L\theta - \theta)| \leqq C\|q\|_\infty h^{3-l/2}|u|_2[\theta]$ is a consequence of (i), (c4).

6° $|(q\eta, \theta)| \leqq C\|q\|_\infty h^2|u|_2\|\theta\|$ follows immediately by (i).

The conditions (c1)–(c4) and facts (ii), 1°–6° give (a2).

Linear operators L will be always constructed so that

$$\operatorname{supp} Lv \subseteq \operatorname{supp} v \quad \forall v \in V_h(a, b).$$

This condition is fulfilled if and only if L can be defined as in

1.8. **Lemma.** *Let a linear operator $L: V_h(a, b) \to V_{h*}(a, b)$ be given by*

$$L\varphi_i = \varphi_i + \mathbf{x}_i^T \Psi_i,$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, x_{i3}]^T$ is a real vector for $i = 1, \ldots, n$. Then

$$\text{(a)}\quad |Lv|_1^2 = \frac{2}{h}\sum_{i=1}^{m}[v_{i-1}^2(2x_{i-1,3}^2 - 2x_{i-1,3}x_{i-1,2} + x_{i-1,2}^2 + x_{i-1,2} + 0.5) + v_{i-1}v_i(4x_{i-1,3}x_{i1} - 2x_{i-1,3}x_{i2} - 2x_{i-1,2}x_{i1} - x_{i-1,2} - x_{i2} - 1) + v_i^2(2x_{i1}^2 - 2x_{i1}x_{i2} + x_{i2}^2 + x_{i2} + 0.5)]$$

and

$$\text{(b)}\quad \|Lv - v\|^2 = \frac{h}{6}\sum_{i=1}^{m}[v_{i-1}^2(x_{i-1,2}^2 + 2x_{i-1,3}^2 + x_{i-1,3}x_{i-1,2}) + v_{i-1}v_i(x_{i-1,2}x_{i1} + 4x_{i-1,3}x_{i1} + x_{i-1,3}x_{i2}) + v_i^2(x_{i2}^2 + 2x_{i1}^2 + x_{i2}x_{i1})]$$

for an arbitrary function $v \in V_h(a, b)$.

Proof. Both statements can be proved by a direct computation.

The L-discrete solution of a problem does not oscillate whenever its L-matrix A is *monotone* (i.e. A^{-1} does exist and is non-negative). For an explanation consult [19].

1.9. **Lemma.** (Bramble, Hubbard [4]) *A real square matrix* $M = (m_{ij})$ *of order* n *is monotone whenever the following assertions* (a), (b), (c) *are true.*

(a) $i \neq j \Rightarrow m_{ij} \leqq 0$.

(b) *There is a non-empty set* $I \subseteq \{1, \ldots, n\}$ *such that*

$$\sum_{j=1}^{n} m_{ij} > 0 \Leftrightarrow i \in I \quad \text{and} \quad \sum_{j=1}^{n} m_{ij} = 0 \Leftrightarrow i \notin I .$$

(c) *For every* $i \in \{1, \ldots, n\}$ *one can find indices* $j \in I$ *and* $k_1, \ldots, k_s$ *in such a way that each of the numbers* $m_{ik_1}, m_{k_1k_2}, \ldots, m_{k_sj}$ *is non-zero.*

1.10. **Lemma.** (Raviart [19]) *If* $M\boldsymbol{x} = \boldsymbol{b}$ *is a system with a monotone matrix* $M = (m_{ij})$ *of order* n *then*

$$\sum_{j=1}^{n} m_{ij} \geqq \alpha_* > 0 \quad \text{for} \quad i = 1, \ldots, n \quad \Rightarrow \quad \|\boldsymbol{x}\|_\infty \leqq \alpha_*^{-1} \|\boldsymbol{b}\|_\infty .$$

2. THE CASE $q \equiv 0$

Let us consider the problem (4) provided with restrictions $q = 0$ and $p \geqq 1$ on (a, b). A weak formulation of this problem is

(7) find $u \in \mathcal{H}_0^1(a, b)$: $\alpha_1(u, v) = (f, v) \quad \forall v \in \mathcal{H}_0^1(a, b)$. Here $\alpha_1(u, v) = \int_a^b (\varepsilon u'v' + pu'v)\, dx$.

2.1. **Definition.** Let us put

$$E_{1i} = \max\{\varepsilon, \textstyle\int_{x_{i-1}}^{x_i} p\varphi_{i-1}\, dx\},$$

$$E_1(x) = E_{1i} \quad \text{for} \quad x \in \langle x_{i-1}, x_i), \quad i = 1, \ldots, m \quad \text{and}$$

$$A_1(u, v) = \textstyle\int_a^b (E_1 u'v' + pu'v)\, dx \quad \forall u, v \in \mathcal{H}^1(a, b) .$$

2.2. **Remark.** Obviously, the following assertions (a)–(c) hold for $i = 1, \ldots, m$.

(a) $A_1(\varphi_{i-1}, \varphi_i) \leqq 0$,

(b) $A_1(\varphi_{i-1} + \varphi_i + \varphi_{i+1}, \varphi_i) = 0$,

(c) $\varepsilon \leqq \int_{x_i}^{x_{i+1}} p\varphi_i\, dx \Rightarrow A_1(\varphi_{i+1}, \varphi_i) = 0$.

2.3. **Definition.** Let us define a linear operator $L_1 : V_h(a, b) \to V_{h*}(a, b)$ by

$$L_1\varphi_i = \varphi_i + [a_i, 0, -a_{i+1}]\, \Psi_i \quad \text{for} \quad i = 1, \ldots, n, \quad \text{where}$$

$$a_i = (E_{1i} - \varepsilon)/(p, \psi_{2i-1}) \quad \text{for} \quad i = 1, \ldots, m .$$

2.4. **Lemma.** *We have* $\alpha_1(u, L_1 v) = A_1(u, v)\ \forall u, v \in V_h(a, b)$.

Proof. Since α_1 and A_1 are bilinear, it is sufficient to prove

(i) $\alpha_1(\varphi_j, L_1\varphi_i) = A_1(\varphi_j, \varphi_i)$ for $i, j = 1, \ldots, n$:

$$\alpha_1(\varphi_{i-1}, L_1\varphi_i) = \alpha_1(\varphi_{i-1}, \varphi_i) + a_i\alpha_1(\varphi_{i-1}, \psi_{2i-1}) =$$

$$= \alpha_1(\varphi_{i-1}, \varphi_i) - (E_{1i} - \varepsilon)\frac{1}{h} = \alpha_1(\varphi_{i-1}, \varphi_i) + \int_a^b (E_1 - \varepsilon)\varphi'_{i-1}\varphi'_i\,dx =$$

$$= A_1(\varphi_{i-1}, \varphi_i)\,.$$

If $j = i, i + 1$ then (i) can be proved analogously and for the other values of j it holds obviously.

2.5. **Lemma.** *There exists a constant* K *such that*

$$|a_i| \leqq K \quad \textit{for} \quad i = 1, \ldots, m$$

holds for all positive ε, h.

Proof. Let us put $p_i = p(x_i)$ for $i = 1, \ldots, m$ and denote by c, C constants satisfying $0 \leqq c \leqq -p' \leqq C$ on (a, b). Then

$$p_i + c(x_i - x) \leqq p(x) \leqq p_i + C(x_i - x)$$

on $(x_{i-1}, x_i\rangle$. Hence

$$E_{1i} - \varepsilon < E_{1i} \leqq \int_{x_{i-1}}^{x_i} [p_i + C(x_i - x)]\,\varphi_{i-1}\,\mathrm{d}x = p_i h/2 + Ch^2/3$$

and

$$p_i h/2 + ch^2/4 = p_i h/2 + c(x_i - x, \psi_{2i-1}) \leqq (p, \psi_{2i-1})\,.$$

These two inequalities give $|a_i| \leqq (6p_i + 4Ch)/(6p_i + 3ch)$. This together with $p_i \geqq 1$ yields $|a_i| \leqq K$ for $K = 1 + \frac{2}{3}C$.

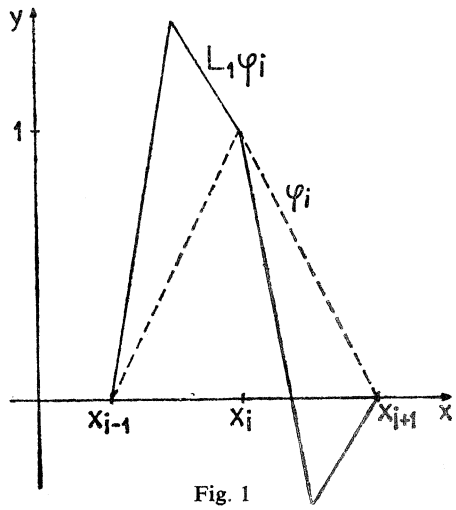

Fig. 1

2.6. **Remark.** If $p = 1$ on (a, b) and $\varepsilon < h/2$ then $E_{1i} = h/2$ and $a_i = 1 - 2/h$. The graph of $L_1\varphi_i$ can be seen in Fig. 1. Thus, up to the right-hand side, the classical upwind scheme is obtained by using the test functions $L_1\varphi_i$ in this case.

2.7. **Theorem.** *The L_1-matrix of problem* (7) *is monotone.*

Proof. If $M = (m_{ij})$ stands for the L_1-matrix of (7) then M is tridiagonal and

$$m_{i,i-1} = A_1(\varphi_{i-1}, \varphi_i) \quad \text{for} \quad i = 2, \ldots, n,$$
$$m_{ii} = A_1(\varphi_i, \varphi_i) \quad \text{for} \quad i = 1, \ldots, n,$$
$$m_{i,i+1} = A_1(\varphi_{i+1}, \varphi_i) \quad \text{for} \quad i = 1, \ldots, n-1$$

according to 2.4. Now using $p' \leqq 0$ and 2.2(a), (b), (c), one can easily verify 1.9(a), (b), (c).

2.8. **Definition.** Let Q be a constant such that $0 \leqq Q \leqq -0.5p'$ on (a, b). We put

$$[v]_1 = (|v|^2_{E_1,1} + Q\|v\|^2)^{1/2}$$

for arbitrary $(c, d) \subseteq (a, b)$ and $v \in \mathcal{H}^1(c, d)$.

2.9. **Theorem.** *Suppose that $\varepsilon < h$, u is an exact solution of the problem* (7) *and u^h is an L_1-discrete solution of* (7). *Then*

$$[u - u^h]_1 \leqq Ch^{3/2}|u|_2 .$$

Proof. Let us denote by $\tilde{u}$ an interpolate of u in $V_h(a, b)$ and put $\eta = u - \tilde{u}$, $\theta = u^h - \tilde{u}$.

(i) Obviously there exist positive constants C_1, C_2 such that $\max\{\varepsilon, C_1 h\} \leqq E_1 \leqq C_2 h$ on (a, b).

(ii) $[v]_1^2 \leqq \alpha_1(v, L_1 v)\ \forall v \in V_h(a, b)$ is a consequence of $[v]_1^2 \leqq A_1(v, v)$ and 2.4.

(iii) $|L_1 v|_1 \leqq Ch^{-1/2}[v]_1\ \forall v \in V_h(a, b)$: By 1.8(a) it follows that $|L_1 v|_1^2 = (1/h)\sum_{i=1}^{m}(4a_i^2 + 1)(v_i - v_{i-1})^2$. This, 2.5 and 1.5(a) give $|L_1 v|_1 \leqq C|v|_1$. The last inequality together with (i) implies (iii).

(iv) $\|L_1 v - v\| \leqq Ch^{1/2}[v]_1\ \forall v \in V_h(a, b)$: By 1.8(b) we get $\|L_1 v - v\|^2 = (h/3)\sum_{i=1}^{m} a_i^2(v_i - v_{i-1})^2$. With respect to this equality and 2.5, 1.5(a) we obtain $\|L_1 v - v\| \leqq Ch|v|_1$. Now, (iv) holds by (i).

By virtue of (i)–(iv) and 1.7, the following assertions (v), (vi) are true.

(v) $[\eta]_1 \leqq Ch^{3/2}|u|_2$,

(vi) $[\theta]_1^2 \leqq C_1 h^{3/2}|u|_2\,[\theta]_1 + C_2 h^2 |u|_2\,\|\theta\|$.

It follows by (i) and the Friedrichs inequality that $\|\theta\| \leqq Ch^{-1/2}[\theta]_1$. This and (vi) imply

(vii) $[\theta]_1 \leqq Ch^{3/2}|u|_2$

and the statement of the theorem is an immediate consequence of (v), (vii).

2.10. **Remark.** If the exact solution u of (7) has a boundary or internal layer then the norm $|u|_2$ is proportional to $\varepsilon^{-1.5}$ thanks to the behaviour of u within the layers; see [15], Theorems 2.1, 2.3. Hence, Theorem 2.9 does not give any information concerning exactness of L_1-discrete solutions of such problems. The following local error estimate is much more valuable.

2.11. **Theorem.** *Suppose that $\varepsilon < h$, u is an exact solution of the problem* (7) *and u^h is an L_1-discrete solution of* (7). *Let a subinterval $D = (a, x_l)$ in (a, b) have the property*

$$l < m \;\Rightarrow\; \varepsilon \leqq \int_{x_l}^{x_{l+1}} p\varphi_l \,\mathrm{d}x \,.$$

Then

$$[(u - u^h)|_D]_1 \leqq Ch^{3/2}|u|_{D^e}|_2 \,.$$

Proof. The function $u^h = u_1\varphi_1 + \ldots + u_n\varphi_n$ is a solution of the equations

$$\sum_{j=1}^{n} u_j A_1(\varphi_j, \varphi_i) = (f, L_1\varphi_i) \quad \text{for} \quad i = 1, \ldots, n \tag{8}$$

by 2.4. The case $l = m$ being trivial, we suppose that $l < m$ and

(i) $\varepsilon \leqq \int_{x_l}^{x_{l+1}} p\varphi_l \,\mathrm{d}x$.

If we put $c = x_{l+1}$ then $D^e = (a, c)$. Let us consider the problem

$$\text{find} \quad \bar{u} \in \mathcal{H}^1(a, c): \; \beta_1(\bar{u}, v) = (f, v) \;\; \forall v \in \mathcal{H}_0^1(a, c) \quad \text{and} \tag{9}$$

$$\bar{u}(a) = 0 \,, \quad \bar{u}(c) = u(c) \,,$$

$$\text{where} \quad \beta_1(\bar{u}, v) = \int_a^c (\varepsilon\bar{u}'v' + p\bar{u}'v) \,\mathrm{d}x \,.$$

(ii) $\bar{u} = u|_{D^e}$ is true obviously.

Let $\bar{\varphi}_i = \varphi_i|_{D^e}$ for $i = 0, \ldots, l+1$, let $R_1: V_h(a, c) \to V_{h^*}(a, c)$ be the linear operator from 2.3 and $\bar{u}^h = \bar{u}_1\bar{\varphi}_1 + \ldots + \bar{u}_{l+1}\bar{\varphi}_{l+1}$ an R_1-discrete solution of (9).

(iii) $\bar{u}_i = u_i$ for $i = 1, \ldots, l$: Clearly, $\bar{u}_{l+1} = u(c)$ and $\bar{u}_1, \ldots, \bar{u}_l$ satisfy

$$\sum_{j=1}^{l+1} \bar{u}_j \beta_1(\bar{\varphi}_j, R_1\bar{\varphi}_i) = (f, R_1\bar{\varphi}_i) \quad \text{for} \quad i = 1, \ldots, l \,. \tag{10}$$

Using 2.4, one can easily see that $R_1\bar{\varphi}_i = L_1\varphi_i|_{D^e}$ and $\beta_1(\bar{\varphi}_j, R_1\bar{\varphi}_i) = A_1(\varphi_j, \varphi_i)$ for $j = 1, \ldots, l+1$, $i = 1, \ldots, l$. Hence (10) can be written in the form

$$\sum_{j=1}^{l+1} \bar{u}_j A_1(\varphi_j, \varphi_i) = (f, L_1\varphi_i) \quad \text{for} \quad i = 1, \ldots, l \,. \tag{11}$$

We have $A_1(\varphi_{l+1}, \varphi_l) = 0$ according to (i) and 2.2(c). Thus, one can easily see that the values $u_1, \ldots, u_l$ ($\bar{u}_1, \ldots, \bar{u}_l$) do not depend on $u_{l+1}, \ldots, u_n$ (on $\bar{u}_{l+1}$, respectively). This and the fact that equations (10) are exactly the first l equations from (8) imply (iii).

Now

$$[(u - u^h)|_D]_1 \overset{\text{(ii),(iii)}}{=} [(\bar{u} - \bar{u}^h)|_D]_1 \leqq [\bar{u} - \bar{u}^h]_1 \overset{2.9,\text{(ii)}}{\leqq} Ch^{3/2}|u|_{D^e}|_2 \,.$$

2.12. **Remark.** If one computes an L_1-discrete solution u^h from Fig. 2 then one uses a bilinear form A_1 with $E_1 = 0.05$. In Fig. 2, besides u^h a Galerkin solution $u_G \in V_h(0, 1)$ of the problem

$$-0.05u'' + u' = x \quad \text{on} \quad (0, 1), \quad u(0) = 0 = u(1)$$

is shown. The reader can observe an essential difference in the exactness of u^h and u_G in the whole interval.

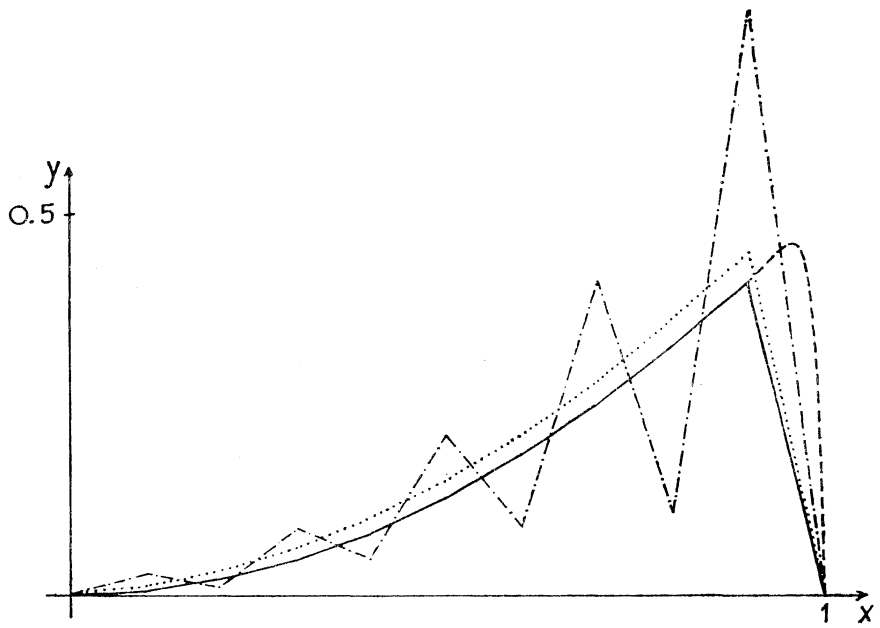

Fig. 2

$-0.01u'' + u' = x$ on $(0, 1)$, $u(0) = 0 = u(1)$
— — — — exact solution
———— L_1-discrete solution with step 0.1
— · — · — · Galerkin solution with step 0.1
$-0.05u'' + u' = x$ on $(0, 1)$, $u(0) = 0 = u(1)$
. . . . . Galerkin solution with step 0.1

2.13. **Remark.** Theorems 2.9 and 2.11 can be reformulated for the case $p \leqq -1$ on (a, b) in an obvious way. If we admit that p has a first-order zero at one of the nodes $x_0, \ldots, x_m$ then Theorems 2.9, 2.11 remain valid with the following modification: $h^{1.5}$ has to be substituted by h.

3. THE CASE $p \equiv 0$

Let us consider the problem (4) provided with restrictions $p = 0$, $0 < q_0 \leqq q$ and q is uniformly continuous on (a, b). A weak formulation of this problem is

(12) find $u \in \mathring{H}^1_0(a, b)$: $\alpha_2(u, v) = (f, v) \quad \forall v \in \mathring{H}^1_0(a, b)$. Here $\alpha_2(u, v) = \int_a^b (\varepsilon u'v' + quv)\,\mathrm{d}x$.

3.1. **Definition.** Let us put

$$E_{2i} = \max\{\varepsilon, h \int_{x_{i-1}}^{x_i} q\varphi_{i-1}\varphi_i\,\mathrm{d}x\},$$
$$E_2(x) = E_{2i} \quad \text{for} \quad x \in \langle x_{i-1}, x_i), \quad i = 1, \ldots, m \quad \text{and}$$
$$A_2(u, v) = \int_a^b (E_2 u'v' + quv)\,\mathrm{d}x \quad \forall u, v \in \mathring{H}^1(a, b).$$

3.2. **Remark.** Obviously, the following assertions (a)–(c) hold for $i = 1, \ldots, n$.

(a) $A_2(\varphi_{i-1}, \varphi_i) = A_2(\varphi_i, \varphi_{i-1}) \leqq 0$.

(b) $A_2(\varphi_{i-1} + \varphi_i + \varphi_{i+1}, \varphi_i) = (q, \varphi_i) \geqq q_0 h > 0$.

(c) $\varepsilon \leqq h \int_{x_{i-1}}^{x_i} q\varphi_{i-1}\varphi_i\,\mathrm{d}x \Rightarrow A_2(\varphi_{i-1}, \varphi_i) = 0 = A_2(\varphi_i, \varphi_{i-1})$.

3.3. **Definition.** Let us define a linear operator L_2: $V_h(a, b) \to V_{h*}(a, b)$ by

$$L_2\varphi_i = \varphi_i + \mathbf{c}_i^T \Psi_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $\mathbf{c}_i = \mathbf{o}$ in the case $E_{2i} = \varepsilon = E_{2,i+1}$ and $\mathbf{c}_i$ is a solution of the equations

(13) $$\alpha_2(\varphi_j, L_2\varphi_i) = A_2(\varphi_j, \varphi_i) \quad \text{for} \quad j = i - 1, i, i + 1$$

in the case $E_{2i} > \varepsilon$ or $E_{2,i+1} > \varepsilon$.

3.4. **Lemma.** *We have* $\alpha_2(u, L_2 v) = A_2(u, v)$ $\forall u, v \in V_h(a, b)$.

3.5. **Lemma.** *There exist positive constants* C *and* h_0 *such that*

$$\|\mathbf{c}_i\|_\infty \leqq C \quad \textit{for} \quad i = 1, \ldots, n$$

is true for all positive ε *and all* $h \in (0, h_0)$.

Proof. Using the uniform continuity of q, we define h_0 as a positive number satisfying

(i) $|x - y| < 2h_0 \Rightarrow |q(x) - q(y)| < \frac{2}{7} q_0$ for all $x, y \in (a, b)$.

Let i be a fixed index, $\mathbf{c}_i = [c_1, c_2, c_3]^T$, m_j the minimum and M_j the maximum of q on $\langle x_{j-1}, x_j\rangle$ for $j = i, i + 1$.

Obviously, it is sufficient to consider the case $E_{2i} > \varepsilon$ or $E_{2,i+1} > \varepsilon$. Then

(ii) $\varepsilon < (h^2/6) \max\{M_i, M_{i+1}\}$.

By a simple modification of (13) we get

$$(14)\qquad \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \frac{1}{h} \begin{bmatrix} -E_{2i} + \varepsilon \\ 2E_{2i} + 2E_{2,i+1} - 4\varepsilon \\ -E_{2,i+1} + \varepsilon \end{bmatrix}, \quad \text{where}$$

$$a_{11} = \alpha_2(\varphi_{i-1}, \psi_{2i-1}) \geqq (h/4)\, m_i\,,$$

$$a_{22} = \alpha_2(\varphi_i - \varphi_{i-1} - \varphi_{i+1}, \psi_{2i}) \geqq 4\varepsilon/h + (h/6)\,(m_i + m_{i+1})\,,$$

$$a_{33} = \alpha_2(\varphi_{i+1}, \psi_{2i+1}) \geqq (h/4)\, m_{i+1}\,,$$

$$a_{21} = \alpha_2(\varphi_i - \varphi_{i-1}, \psi_{2i-1}) \leqq (h/12)\,(M_i - m_i) \quad \text{and}$$

$$a_{23} = \alpha_2(\varphi_i - \varphi_{i+1}, \psi_{2i+1}) \leqq (h/12)\,(M_{i+1} - m_{i+1})\,.$$

It remains to find upper estimates of $a_{12} = \alpha_2(\varphi_{i-1}, \psi_{2i})$ and $a_{32} = \alpha_2(\varphi_{i+1}, \psi_{2i})$. Taking into account

$$-\frac{\varepsilon}{h} + \frac{h}{24}\, m_i \leqq a_{12} \leqq -\frac{\varepsilon}{h} + \frac{h}{24}\, M_i\,,$$

we investigate the following cases $1° - 3°$.

$1°$ $-\dfrac{\varepsilon}{h} + \dfrac{h}{24}\, M_i \leqq 0$. Then $|a_{12}| \leqq \dfrac{\varepsilon}{h} - \dfrac{h}{24}\, m_i$ and we obtain

$|a_{12}| \leqq \dfrac{h}{6} \max\{M_i, M_{i+1}\} - \dfrac{h}{24}\, m_i$ by (ii).

$2°$ $-\dfrac{\varepsilon}{h} + \dfrac{h}{24}\, m_i \geqq 0$. In this case $|a_{12}| \leqq -\dfrac{\varepsilon}{h} + \dfrac{h}{24}\, M_i$.

$3°$ $-\dfrac{\varepsilon}{h} + \dfrac{h}{24}\, m_i < 0 < -\dfrac{\varepsilon}{h} + \dfrac{h}{24}\, M_i$. Then $|a_{12}| \leqq \dfrac{h}{24}\,(M_i - m_i)$.

It follows by these estimates that

$$a_{11} - |a_{12}| \geqq \min\left\{\frac{h}{24}\,(7m_i - 4\max\{M_i, M_{i+1}\})\,,\right.$$

$$\left.\frac{\varepsilon}{h} + \frac{h}{24}\,(6m_i - M_i)\,,\ \frac{h}{24}\,(7m_i - M_i)\right\} = \frac{h}{24}\,(7m_i - 4\max\{M_i, M_{i+1}\})\,.$$

Analogously,

$$a_{33} - |a_{32}| \geqq \frac{h}{24}\,(7m_{i+1} - 4\max\{M_i, M_{i+1}\}) \quad \text{and}$$

$$a_{22} - |a_{21}| - |a_{23}| \geqq 4\frac{\varepsilon}{h} + \frac{h}{12}\,[3(m_i + m_{i+1}) - M_i - M_{i+1}]$$

is true. Using the last three inequalities and (i), one arrives at

$$\min_{1 \leqq i \leqq 3} \left(a_{ii} - \sum_{j \neq i} |a_{ij}|\right) \geqq C_1 h \quad \text{for} \quad C_1 = \frac{q_0}{24}.$$

At the same time, there is an upper estimate $C_2 h$ of the absolute values of all the right-hand sides in (14) according to (ii) and the definition of E_2. Hence

$$\|\mathbf{c}_i\|_\infty \leqq C \quad \text{for} \quad C = \frac{C_2}{C_1}.$$

3.6. **Remark.** (a) If $q(x) = 1$ on (a, b) and $\varepsilon < h^2/6$ then $E_{2i} = h^2/6 = E_{2,i+1}$ and $\mathbf{c}_i = (h^2 - 6)/(h^2 + 12)\,[-1, 2, -1]^T$. See Fig. 3.

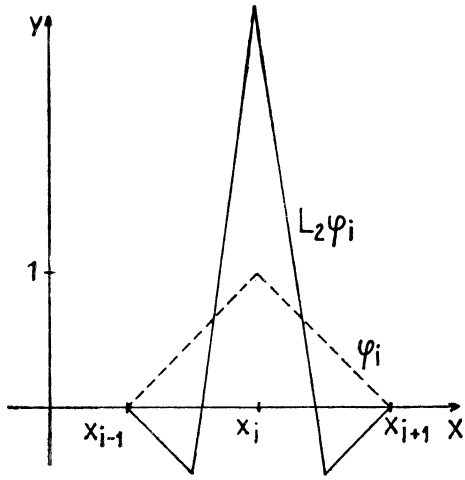

Fig. 3

(b) Whenever $\varepsilon \leqq h \int_{x_{i-1}}^{x_i} q\varphi_{i-1}\varphi_i \,\mathrm{d}x$ for $i = 1, \ldots, n$, the L_2-matrix of (12) is diagonal by 3.2 and it can be derived from the stiffness matrix appearing in the Galerkin method by lumping. See for example [1].

3.7. **Theorem.** *The L_2-matrix of problem* (12) *is monotone and*

$$\|u^h\|_\infty \leqq C \frac{\|f\|_\infty}{q_0}$$

for the L_2-discrete solution u^h of (12).

Proof. The statement immediately follows by 3.2(a), (b) ,1.9 and 1.10.

3.8. **Definition.** Let us put

$$[v]_2 = (|v|^2_{E_2,1} + q_0\|v\|^2)^{1/2}$$

for arbitrary $(c, d) \subseteq (a, b)$ and $v \in \mathcal{H}^1(c, d)$.

3.9. **Theorem.** *Let u be an exact solution of the problem* (12) *and u^h an L_2-discrete solution of* (12). *There exist positive constants C and h_0 such that*

$$[u - u^h]_2 \leqq Ch^2|u|_2$$

holds for all $h \in (0, h_0)$ and $\varepsilon < h^2$.

Proof. Let us denote by $\tilde{u}$ an interpolate of u in $V_h(a, b)$.

(i) Obviously there exist positive constants C_1, C_2 such that $\max\{\varepsilon, C_1h^2\} \leqq \leqq E_2 \leqq C_2h^2$ on (a, b).

(ii) $[v]_2^2 \leqq \alpha_2(v, L_2v)$ $\forall v \in V_h(a, b)$ is a consequence of $[v]_2^2 \leqq A_2(v, v)$ and 3.4.

(iii) $|L_2v|_1 \leqq (C/h)[v]_2$ $\forall v \in V_h(a, b)$: By means of 1.8(a) and the constant C from 3.5 the following estimate can be obtained:

$$|L_2v|_1^2 \leqq \frac{48}{h} \max\{1, C^2\} \sum_{i=1}^{n} v_i^2 .$$

Hence $|L_2v|_1^2 \leqq C(|v|_1^2 + (1/h^2)\|v\|^2)$ with respect to 1.5(c). This and (i) imply (iii).

(iv) $\|L_2v - v\| \leqq C[v]_2$ $\forall v \in V_h(a, b)$: Using 1.8(b) and the constant C from 3.5, the following estimate can be derived:

$$\|L_2v - v\|^2 \leqq 3h \max\{1, C^2\} \sum_{i=1}^{n} v_i^2 .$$

This inequality and 1.5(c) imply $\|L_2v - v\|^2 \leqq C(h^2|v|_1^2 + \|v\|^2)$ and one gets (iv) by (i).

Assertions (i)–(iv) and Proposition 1.7 imply $[u - \tilde{u}]_2 \leqq Ch^2|u|_2$ and $[u^h - \tilde{u}]_2 \leqq \leqq Ch^2|u|_2$; the statement follows immediately.

Now, we illustrate the need for a local error estimate.

3.10. **Example.** If u is a solution of the problem

$$-\varepsilon u'' + u = 1 \quad \text{on} \quad (0, 1), \quad u(0) = 0 = u(1),$$

then it has boundary layers in neighborhoods of 0 and 1. One can easily see that $|u|_2 > \varepsilon^{-3/4}$ and, at the same time,

$$-\sqrt{(\varepsilon)} \ln \varepsilon < t_1 < t_2 < 1 + \sqrt{(\varepsilon)} \ln \varepsilon \;\Rightarrow\; |u/_{(t_1,t_2)}|_2 < 1 .$$

3.11. **Theorem.** *Let u be an exact solution of the problem* (12), *u^h an L_2-discrete solution of* (12), *and let a subinterval $D = (x_k, x_l)$ in (a, b) have the property*

$$0 < k \;\Rightarrow\; \varepsilon \leqq h \int_{x_{k-1}}^{x_k} q\varphi_{k-1}\varphi_k \, dx \quad \textit{and} \quad l < m \;\Rightarrow\; \varepsilon \leqq h \int_{x_l}^{x_{l+1}} q\varphi_l\varphi_{l+1} \, dx .$$

Then there exist positive constants C *and* h_0 *such that*

$$[(u - u^h)|_D]_2 \leqq Ch^2|u|_{D^e}|_2$$

is true for all $h \in (0, h_0)$ *and* $\varepsilon < h^2$.

Proof. Let us denote $D^e = (c, d)$ and consider the problem

(15) find $\bar{u} \in H^1(c, d)$: $\beta_2(\bar{u}, v) = (f, v) \quad \forall v \in H^1(c, d)$ and $\bar{u}(c) = u(c)$, $\bar{u}(d) = u(d)$, where

$$\beta_2(\bar{u}, v) = \int_c^d (\varepsilon \bar{u}'v' + q\bar{u}v)\,dx .$$

Let R_2: $V_h(c, d) \to V_{h*}(c, d)$ be a linear operator from 3.3 and $\bar{u}^h$ an R_2-discrete solution of (15).

Similarly as in 2.11 one can see that $\bar{u} = u|_{D^e}$ and $\bar{u}^h|_D = u^h|_D$. These facts and 3.9 give

$$[(u - u^h)|_D]_2 = [(\bar{u} - \bar{u}^h)|_D]_2 \leqq [\bar{u} - \bar{u}^h]_2 \leqq Ch^2|\bar{u}|_2 \leqq Ch^2|u|_{D^e}|_2 .$$

3.12. **Remark.** (a) If one computes the L_2-discrete solution u^h of the problem from Fig. 4 then one uses a bilinear form A_2 with $E_2 = 0.0016$. In Fig. 4, besides u^h a Galerkin solution $u_G \in V_h(0, 1)$ of the problem

$$-0.0016u'' + u = 1 + \sin 2\pi x \quad \text{on} \quad (0, 1), \quad u(0) = 0 = u(1)$$

is shown. An essential difference in the exactness of u^h and u_G in the whole interval can be observed.

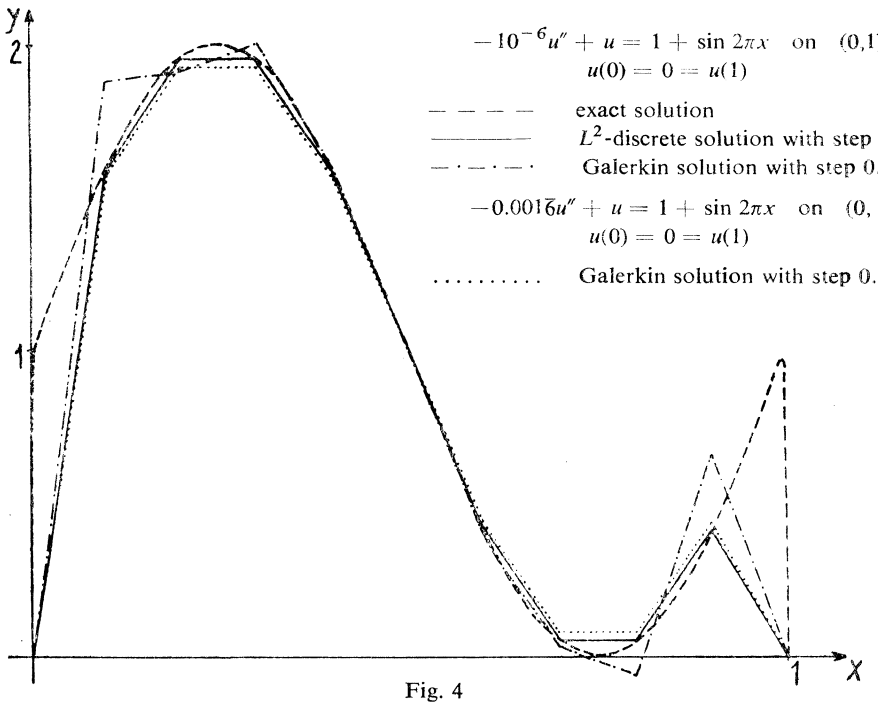

Fig. 4

(b) L_2-discrete solutions of the problem from Fig. 5 with steps 0.1, 0.05 and 0.025 have been computed. The mutual differences of their values at each of the points 0.1, ..., 0.9 are less than $4 \cdot 10^{-6}$.

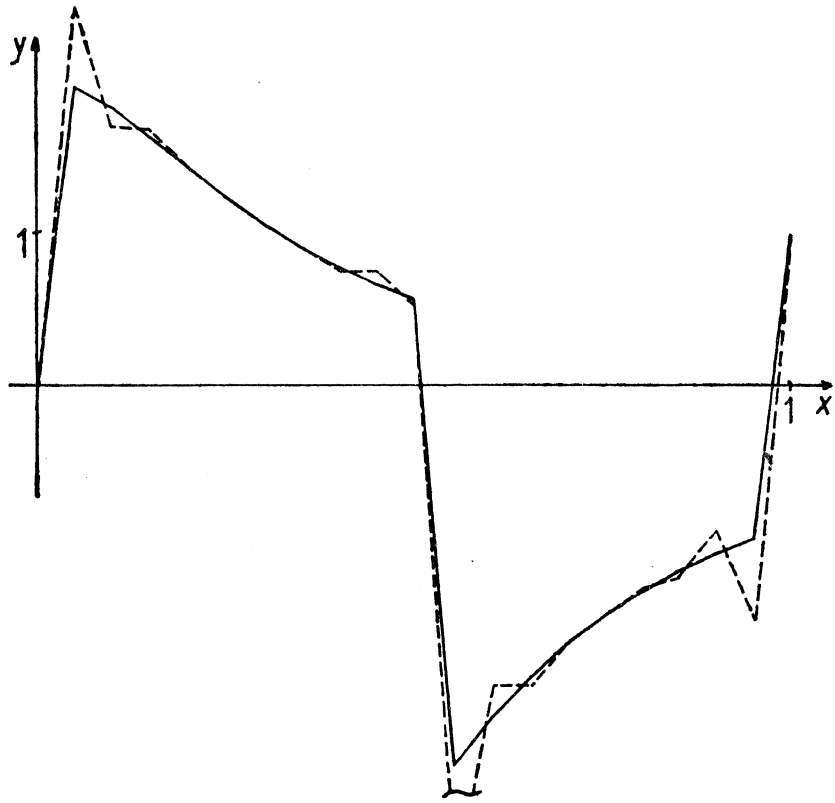

Fig. 5

$$-10^{-6}u'' + (1 + 10x^2)\,u = \begin{cases} 2 & \text{for } x \leqq 0.5\,, \\ -10 & \text{for } x > 0.5 \end{cases} \qquad u(0) = 0\,, \quad u(1) = 1$$

———— L_2-discrete solution with step 0.05
— — — — Galerkin solution with step 0.05

4. A MORE GENERAL ONE-DIMENSIONAL CASE

Let us consider the problem (4) satisfying $0 < Q \leqq q - 0.5p'$ on (a, b) for a constant Q. A weak formulation of this problem is

(16) find $u \in \mathcal{H}_0^1(a, b)$: $\alpha_3(u, v) = (f, v) \quad \forall v \in \mathcal{H}_0^1(a, b)$. Here $\alpha_3(u, v) = \int_a^b (\varepsilon u'v' + pu'v + quv)\,\mathrm{d}x$.

4.1. **Definition.** Let us define a linear operator $L_3 : V_h(a, b) \to V_{h*}(a, b)$ by

$$L_3\varphi_i = \varphi_i + \boldsymbol{d}_i^T \Psi_i \quad \text{for} \quad i = 1, \ldots, n ,$$

where the vectors $\boldsymbol{d}_i$ satisfy

$$\alpha_3(\varphi_j, L_3\varphi_i) \leqq 0 ,$$

$$\alpha_3(\varphi_j, \varphi_i) \geqq 0 \Rightarrow \alpha_3(\varphi_j, L_3\varphi_i) = 0$$

for $j = i - 1, i + 1$ and

$$d_{i1}^2 + \frac{1}{h} d_{i2}^2 + d_{i3}^2 \quad \text{is minimal} .$$

4.2. **Remark.** In [7] an apriori error estimate is proved illustrating that the size of the $\mathbb{H}_0^1$- and $\mathbb{L}^2$-norms of the error of the L_3-discrete solution is of the same order as stated in Theorem 2.9, 3.9.

4.3. **Remark.** In Fig. 7, one can see graphs of the test functions used in the computation of u_8 from Fig. 6. The accuracy of L_3-discrete solutions u_8, u_{16}, u_{32} ($h = 0.03125$) from Fig. 6 is compared in Tab. 1. Approximate solutions of a problem which is no special case of (16) are given in Fig. 8.

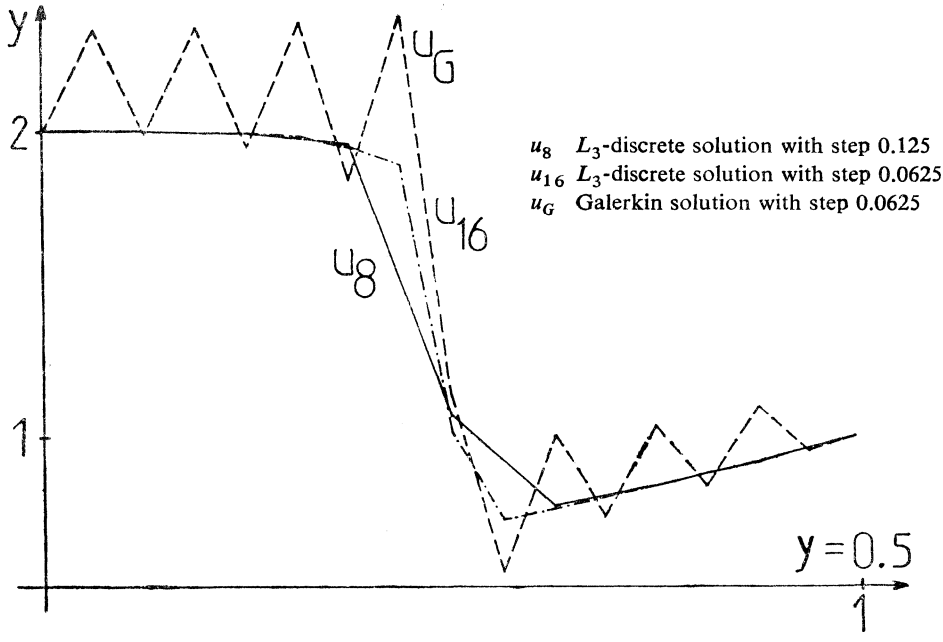

Fig. 6

$-10^{-6}u'' + 10\cos(\pi x)u' + (10x^2 + 0.1)\,u = x$ on $(0, 1)$, $u(0) = 2$, $u(1) = 1$

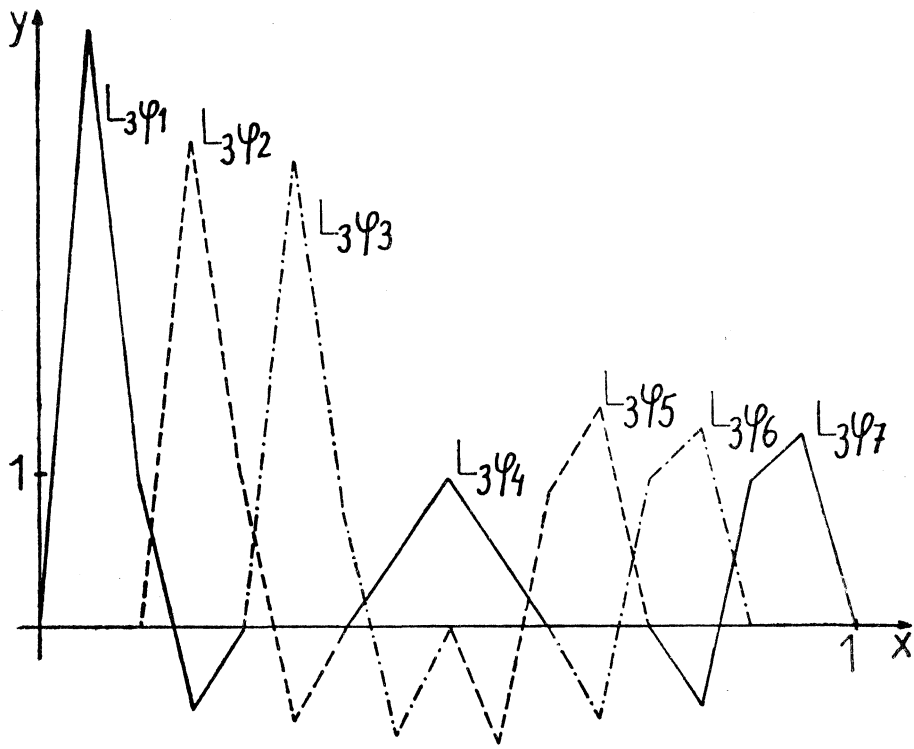

Fig. 7

Tab. 1.

x	$u_{16} - u_8$	$u_{32} - u_{16}$
0.125	−0.000472	−0.000153
0.250	−0.001798	−0.000533
0.375	−0.009081	−0.0027
0.500	−0.042566	−0.0433035
0.625	−0.006124	−0.0013585
0.750	−0.0022208	−0.0005918
0.875	−0.0012405	−0.0003525

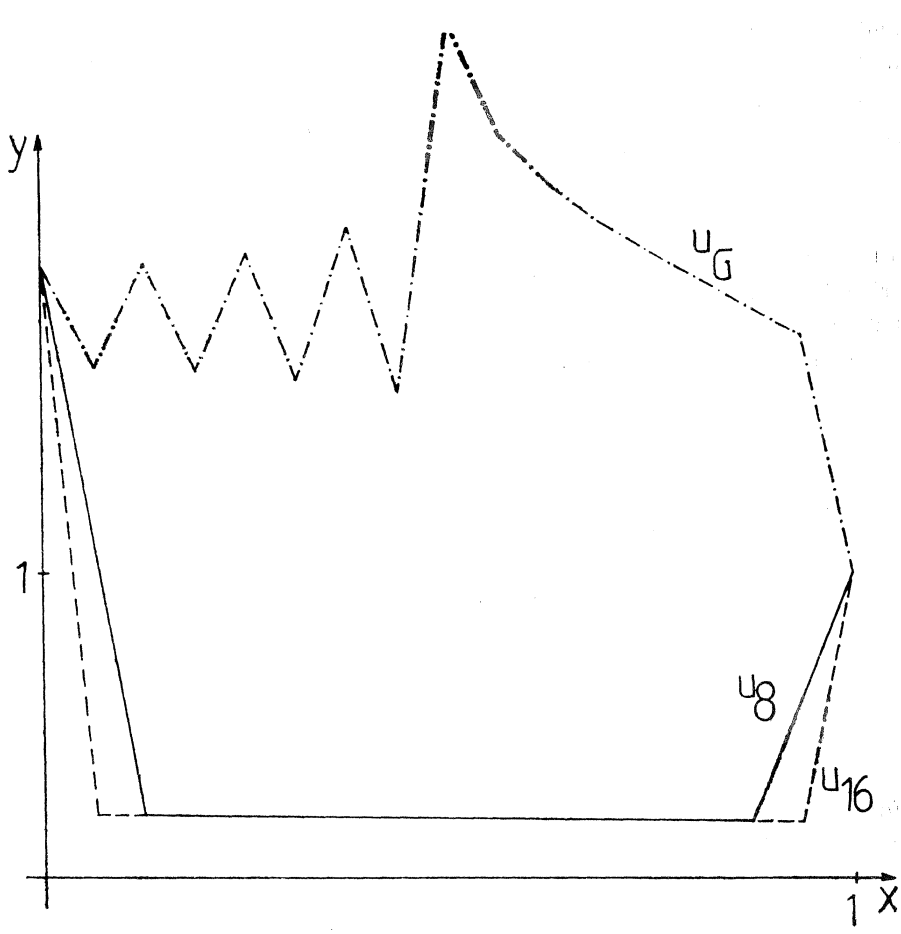

Fig. 8

$-10^{-6}u'' - 10\cos(\pi x)u' + (10x^2 + 0.1)\,u = x$ on $(0,1)$, $u(0) = 2$, $u(1) = 1$

u_8 L_3-discrete solution with step 0.125

u_{16} L_3-discrete solution with step 0.0625

u_G Galerkin solution with step 0.0625

5. A TWO-DIMENSIONAL PROBLEM

Let us apply the basic ideas of the method used in Sections 2 and 3 to the problem

$$-\varepsilon \Delta u + \boldsymbol{p} \,.\, \mathbf{grad}\, u + qu = f \quad \text{on} \quad \Omega\,, \quad u|_\Gamma = 0\,. \tag{17}$$

Here ε is a positive real number, $\boldsymbol{p}(q)$ is a sufficiently smooth vector (scalar) valued function on Ω and $f \in \mathbb{L}^2(\Omega)$. The functions $\boldsymbol{p}$, q are supposed to satisfy either $\operatorname{div} \boldsymbol{p} \leqq \leqq 0$, $q = 0$ on Ω or $\boldsymbol{p} = \mathbf{o}$, $0 < q_0 \leqq q$ on Ω.

Only a brief description of this application is presented. A theoretical analysis is not complete yet.

A weak formulation of (17) is

(18) find $u \in \mathbb{H}_0^1(\Omega)$: $a(u, v) = (f, v) \quad \forall v \in \mathbb{H}_0^1(\Omega)$, where

$$a(u, v) = \int_\Omega [\varepsilon \, \mathbf{grad}\, u \,.\, \mathbf{grad}\, v + (\mathbf{p} \,.\, \mathbf{grad}\, u + qu)\, v] \, \mathrm{d}\mathbf{x} \,.$$

5.1. **Definition.** Let τ be an arbitrary triangulation of Ω. We denote by $\mathscr{T}_\tau$ the set of triangles, by $\mathscr{N}_\tau$ the set of nodes of τ and by $\mathscr{I}_\tau$ the set of $\mathbf{P} \in \mathscr{N}_\tau$ satisfying $\mathbf{P} \notin \Gamma$.

Let us construct a new triangulation τ^* by dividing each triangle from $\mathscr{T}_\tau$ into four equal parts in a way illustrated in Fig. 9.

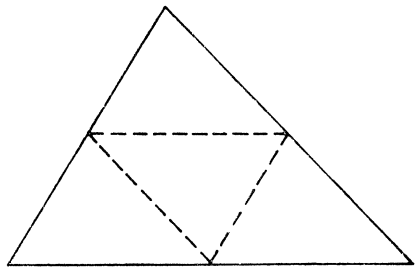

Fig. 9

For each vertex $\mathbf{P} \in \mathscr{N}_\tau$ ($\mathbf{P} \in \mathscr{N}_{\tau^*}$) define a real function $\varphi_\mathbf{P}(\psi_\mathbf{P})$ continuous on Ω, linear on each triangle from $\mathscr{T}_\tau$ (from $\mathscr{T}_{\tau^*}$) and such that

$$\varphi_\mathbf{P}(\mathbf{Q}) = \begin{cases} 1 \text{ for } \mathbf{Q} = \mathbf{P} \\ 0 \text{ for } \mathbf{Q} \in \mathscr{N}_\tau - \{\mathbf{P}\} \end{cases} .$$

$$\left(\psi_\mathbf{P}(\mathbf{Q}) = \begin{cases} 1 \text{ for } \mathbf{Q} = \mathbf{P} \\ 0 \text{ for } \mathbf{Q} \in \mathscr{N}_{\tau^*} - \{\mathbf{P}\} \end{cases} . \right)$$

Let us denote $V_\tau = \operatorname{span}\{\varphi_\mathbf{P};\ \mathbf{P} \in \mathscr{I}_\tau\}$ and $V_{\tau^*} = \operatorname{span}\{\psi_\mathbf{P};\ \mathbf{P} \in \mathscr{I}_{\tau^*}\}$.

5.2. **Definition.** For arbitrary $u, v \in \mathbb{H}^1(\Omega)$ we put

$$b(u, v) = \sum_{T \in \mathscr{T}_\tau} b_T(u, v) \,,$$

where

$$b_T(u, v) = a_T(u, v) + \delta_T(u, v) \,,$$

$$a_T(u, v) = \int_T [\varepsilon \, \mathbf{grad}\, u \,.\, \mathbf{grad}\, v + (\mathbf{p} \,.\, \mathbf{grad}\, u + qu)\, v] \, \mathrm{d}\mathbf{x} \,,$$

$$\delta_T(u, v) = \int_T \left[T_x \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + T_y \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} + T_s \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right) \right] \mathrm{d}\mathbf{x}$$

and the constants T_x, T_y, T_s satisfy the system of equations

$$\begin{aligned}
\delta_T(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}) &= r_1 = \min\{0, -a_T(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}), -a_T(\varphi_{\mathbf{Q}}, \varphi_{\mathbf{P}})\} \\
\delta_T(\varphi_{\mathbf{Q}}, \varphi_{\mathbf{R}}) &= r_2 = \min\{0, -a_T(\varphi_{\mathbf{Q}}, \varphi_{\mathbf{R}}), -a_T(\varphi_{\mathbf{R}}, \varphi_{\mathbf{Q}})\} \\
\delta_T(\varphi_{\mathbf{R}}, \varphi_{\mathbf{P}}) &= r_3 = \min\{0, -a_T(\varphi_{\mathbf{R}}, \varphi_{\mathbf{P}}), -a_T(\varphi_{\mathbf{P}}, \varphi_{\mathbf{R}})\}
\end{aligned} \tag{19}$$

for each $T \in \mathscr{T}_\tau$ with nodes $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$.

Lemma 5.3 can be proved by a direct computation and Lemma 5.4 is true obviously.

5.3. **Lemma.** *Let τ be a triangulation of Ω. The determinant of the matrix of system* (19) *equals* -0.125 *for each* $T \in \mathscr{T}_\tau$.

5.4. **Lemma.** *The following assertions* (a), (b) *hold for an arbitrary triangulation τ of Ω.*

(a) $b(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}) \leqq 0 \quad \forall \mathbf{P}, \mathbf{Q} \in \mathscr{N}_\tau, \quad \mathbf{P} \neq \mathbf{Q}$.

(b) $\sum\limits_{\mathbf{Q} \in \mathscr{N}_\tau} b(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}) = \sum\limits_{\mathbf{Q} \in \mathscr{N}_\tau} a(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}) \quad \forall \mathbf{P} \in \mathscr{I}_\tau$.

5.5. **Lemma.** *If τ is a triangulation of Ω then the eigenvalues of the matrix* $\begin{bmatrix} T_x & T_s \\ T_s & T_y \end{bmatrix}$ *are non-negative for each* $T \in \mathscr{T}_\tau$.

Proof. Let $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$ be the vertices and s the area of an arbitrary triangle $T \in \mathscr{T}_\tau$. Denote $a = |\mathbf{PQ}|$, $b = |\mathbf{PR}|$, $\boldsymbol{a} = (1/a)\,\mathbf{PQ} = [a_1, a_2]^T$, $\boldsymbol{b} = (1/b)\,\mathbf{PR} = [b_1, b_2]^T$ and $M = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$. If we put

$$\begin{bmatrix} T_x & T_s \\ T_s & T_y \end{bmatrix} = M^T \begin{bmatrix} Z_x & Z_s \\ Z_s & Z_y \end{bmatrix} M \quad \text{and} \tag{20}$$

$$\delta_T^*(u, v) = \int_T \left[Z_x \frac{\partial u}{\partial \boldsymbol{a}} \frac{\partial v}{\partial \boldsymbol{a}} + Z_y \frac{\partial u}{\partial \boldsymbol{b}} \frac{\partial v}{\partial \boldsymbol{b}} + Z_s \left(\frac{\partial u}{\partial \boldsymbol{a}} \frac{\partial v}{\partial \boldsymbol{b}} + \frac{\partial u}{\partial \boldsymbol{b}} \frac{\partial v}{\partial \boldsymbol{a}} \right) \right] \mathrm{d}\boldsymbol{x},$$

then $\delta_T^*(u, v) = \delta_T(u, v)$ for all $u, v \in \mathbb{C}^1(T)$. Thus instead of (19), one can solve the system of equations

$$\begin{aligned}
\delta_T^*(\varphi_{\mathbf{P}}, \varphi_{\mathbf{Q}}) &= r_1 \\
\delta_T^*(\varphi_{\mathbf{Q}}, \varphi_{\mathbf{R}}) &= r_2 \\
\delta_T^*(\varphi_{\mathbf{R}}, \varphi_{\mathbf{P}}) &= r_3
\end{aligned} \quad \text{, in detail} \quad \frac{s}{(ab)^2} \begin{bmatrix} -b^2 & 0 & -ab \\ 0 & 0 & ab \\ 0 & -a^2 & -ab \end{bmatrix} \begin{bmatrix} Z_x \\ Z_y \\ Z_s \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}.$$

Clearly, $Z_x = -(r_1 + r_2)\,a^2/s$, $Z_y = -(r_2 + r_3)\,b^2/s$ and $Z_s = r_2 ab/s$. This together with

(i) $r_i \leqq 0$ for $i = 1, 2, 3$ yields

(ii) $\det \begin{bmatrix} Z_x & Z_s \\ Z_s & Z_y \end{bmatrix} \geqq 0$. This result and (20) imply

(iii) $\det \begin{bmatrix} T_x & T_s \\ T_s & T_y \end{bmatrix} \geqq 0$. At the same time, we have

(iv) $T_x + T_y \geqq 0$: $T_x + T_y = Z_x + Z_y + 2\boldsymbol{a}^T\boldsymbol{b}\, Z_s \geqq Z_x + Z_y - 2\sqrt{(Z_x Z_y)} \geqq 0$ by virtue of (20), (ii), (i).

The statement follows by (iii) and (iv).

Hence the eigenvalues of the tensor $\begin{bmatrix} T_x + \varepsilon & T_s \\ T_s & T_y + \varepsilon \end{bmatrix}$ of the so-called artificial diffusion are greater than those of the tensor $\begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}$ of diffusion. Another way how to increase the diffusion coefficients in the bilinear form a in order to arrive at a bilinear form satisfying Lemma 5.4 is presented in [19]. Now, we construct a linear operator $L_\tau: V_\tau \to V_{\tau^*}$ such that

$$a(u, L_\tau v) = b(u, v) \quad \forall u, v \in V_\tau ,$$
$$\operatorname{supp} L_\tau v \subseteq \operatorname{supp} v \quad \forall v \in V_\tau .$$

5.6. **Definition.** Let us put

$$\mathcal{N}_\tau(\boldsymbol{P}) = \{\boldsymbol{Q} \in \mathcal{N}_\tau;\ \boldsymbol{Q} \in \operatorname{supp} \varphi_{\boldsymbol{P}}\} , \quad \mathcal{N}_{\tau^*}(\boldsymbol{P}) = \{\boldsymbol{R} \in \mathcal{N}_{\tau^*};\ \boldsymbol{R} \in \operatorname{supp} \psi_{\boldsymbol{P}}\}$$

for an arbitrary node $\boldsymbol{P} \in \mathcal{J}_\tau$.

In Fig. 10, the sets $\operatorname{supp} \varphi_{\boldsymbol{K}}$ and $\operatorname{supp} \psi_{\boldsymbol{K}}$ are sketched. Obviously, $\mathcal{N}_\tau(\boldsymbol{K}) = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{K}\}$ and $\mathcal{N}_{\tau^*}(\boldsymbol{K}) = \{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{I}, \boldsymbol{J}, \boldsymbol{K}\}$.

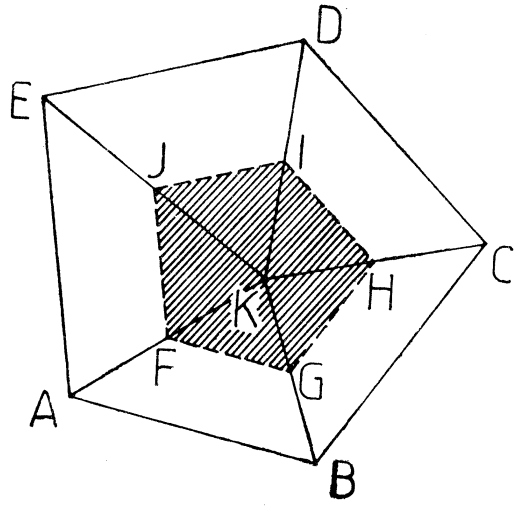

Fig. 10

5.7. **Definition.** Let us define a linear operator $L_\tau: V_\tau \to V_{\tau^*}$ such that

$$L_\tau \varphi_{\boldsymbol{P}} = \sum_{\boldsymbol{R} \in \mathcal{N}_{\tau^*}(\boldsymbol{P})} x_{\boldsymbol{P},\boldsymbol{R}} \psi_{\boldsymbol{R}}$$

and the coefficients $x_{P,R}$ satisfy the system of equations

$$a(\varphi_Q, L_\tau \varphi_P) = b(\varphi_Q, \varphi_P) \quad \forall Q \in \mathcal{N}_\tau(P) \tag{21}$$

for any node $P \in \mathcal{T}_\tau$.

Regarding the analysis of the one-dimensional cases, the systems (21) are solved by a weighted least squares method, making coefficients $x_{P,R}$ as small as possible. Hence in general, (21) need not be satisfied exactly.

In 5.8 and 5.9, the following notation will be used:

$$\Omega = (0, 1) \times (0, 1),$$

g is a linear spline on the boundary Γ of Ω, related to the standard equidistant net with step 0.05 such that

$$g(x, y) = \begin{cases} 1 & \text{for } y = 0 \vee (x = 1 \wedge y \leqq 0.2) \\ 0 & \text{for } y \geqq 0.25 \vee (x = 0 \wedge y \geqq 0.05) \end{cases}$$

τ = the uniform triangulation shown in Fig. 11.

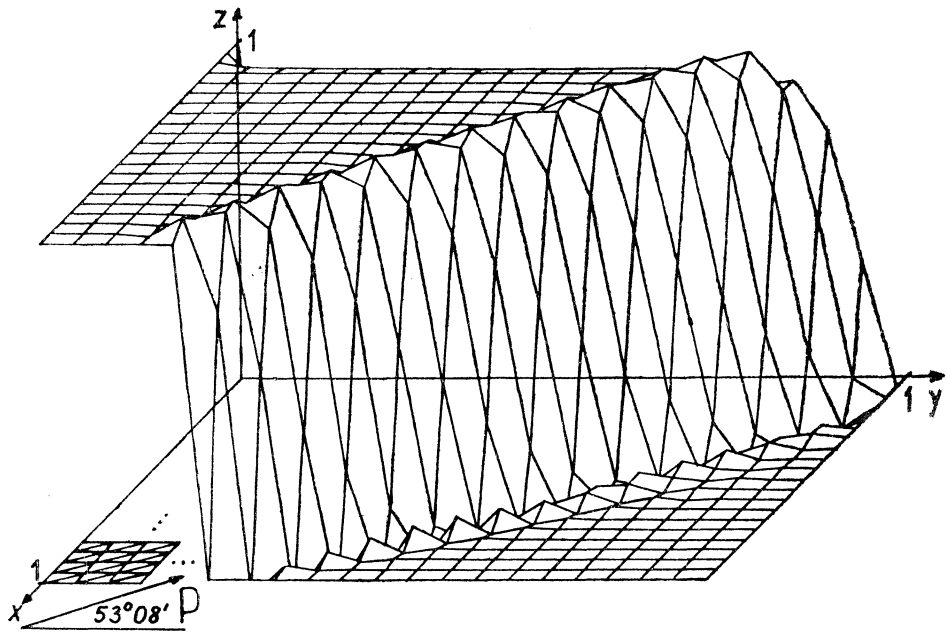

Fig. 11

5.8. **Example.** An approximate solution of the problem

$$-10^{-6}\Delta u - 0.8\frac{\partial u}{\partial x} + 0.6\frac{\partial u}{\partial y} = 0 \quad \text{on} \quad \Omega, \tag{22}$$

$$u = g \quad \text{on} \quad \Gamma$$

has been computed in the space V_τ. From its graph in Fig. 11, one can see that the inner layer along the segment AB ($A = (1, 0.2)$ and $B = (0, 0.95)$) is preserved. This result is comparable with a solution of the same problem, published in [5]. Since the coefficients in (22) are constant and τ is a uniform triangulation, the shape of all the test functions $L_\tau \varphi_P$, $P \in \mathcal{I}_\tau$, is the same. See Fig. 12. Likewise, the artificial diffusion on triangles is of two types only. It is shown in Fig. 13.

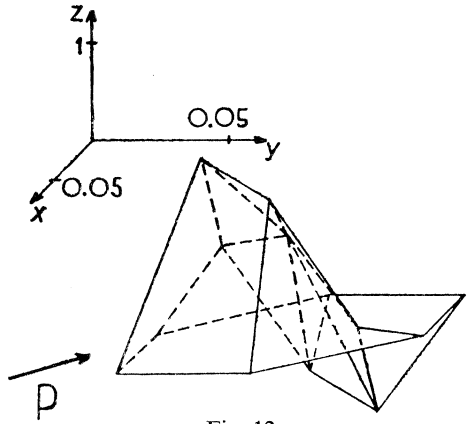

Fig. 12

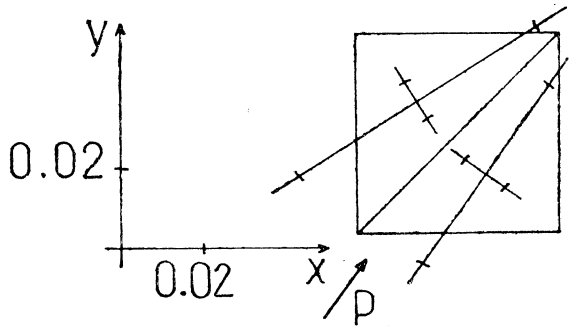

Fig. 13

5.9. **Example.** Two approximate solutions of the problem

$$-10^{-6}\Delta u + u = xy \quad \text{on } \Omega, \quad u = g \quad \text{on } \Gamma \tag{23}$$

have been computed. One, u^h, for the triangulation τ and the other, u^{2h}, for the triangulation τ_0 defined by $\tau_0^* = \tau$. It is well-known that the exact solution u coincides with the function $u_0 = xy$ outside the boundary layers. Let us denote

$$\|v\|_t = \max\{|v(P)|\,; \quad P \in \mathcal{I}_t\}$$

for each continuous function v on Ω and $t = \tau, \tau_0$. We have

$$\|u^{2h} - u_0\|_{\tau_0} = 0.0007229 \text{ and } \|u^h - u_0\|_\tau = 0.0001811\,.$$

Hence, the seminorm of error seems to be proportional to h^2. The shape of test functions $L_{\tau_0}\varphi_P$, $P \in \mathscr{J}_{\tau_0}$, can be seen in Fig. 14.

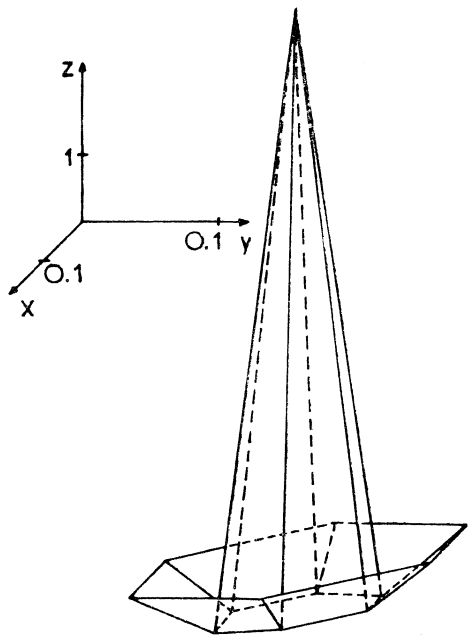

Fig. 14

References

[1] *J. E. Akin:* Application and implementation of finite element methods. Academic Press, London, New York, 1982.

[2] *J. W. Barret, K. W. Morton:* The mathematics of finite elements and applications IV. Academic Press, London, New York (1982), 403—411.

[3] *P. Bar-Yoseph, M. Israeli:* An asymptotic finite element method for improvement of solutions of boundary layer problems. Numer. Math. Vol. 49, 4 (1986), 425—438.

[4] *J. H. Bramble, B. E. Hubbard:* New monotone type approximations for elliptic problems. Math. Comp. 18 (1964), 349—367.

[5] *A. N. Brooks, T. J. R. Hughes:* Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Computer Math. in Appl. Mech. and Eng. 32 (1982), 199—259.

[6] *P. Ciarlet:* The finite element method for elliptic problems. North-Holland, Amsterdam, 1978.

[7] *J. Dalík:* An apriori error estimate of an approximation of a two-point boundary value problem by the Petrov-Galerkin method (Czech). Knižnice obd. a věd. spisů VUT Brno, Sv. A-35 (1988), 19—28.

[8] *E. P. Doolan, J. J. H. Miller, W. H. A. Schilders:* Uniform numerical methods for problems with initial and boundary layers. Boole Press, Dublin, 1980.

[9] *R. Glowinski:* Numerical nethods for nonlinear variational problems. Appendix II. Springer-Verlag, New York, Berlin, 1984.

[10] *P. W. Hemker, P. M. De Zeeuw:* Defect correction for the solution of a singular perturbation problem (preprint). Math. centrum, 1982.
[11] *P. W. Hemker:* Numerical aspects of singular perturbation problems (preprint). Math. centrum, Amsterdam, 1982.
[12] *T. Ikeda:* Maximum principle in finite element models for convection-diffusion phenomena. North-Holland, Amsterdam, New York, Oxford, 1983.
[13] *C. Johnson, U. Nävert:* Analysis of some finite element methods for advection-diffusion problems (research report). Chalmers Univ. of Techn., Göteborg, 1980.
[14] *C. Johnson, U. Nävert, J. Pitkäranta:* Finite elements method for linear hyperbolic problems (research report). Chalmers Univ. of Techn., Göteborg, 1982.
[15] *U. Nävert:* A finite element method for convection-diffusion problems (thesis). Chalmers Univ. of Techn., Göteborg, 1982.
[16] *U. Nävert:* The streamline diffusion method for timedependent convection-diffusion problems with small diffusion (research report). Chalmers Univ. of Techn., Göteborg, 1981.
[17] *E. O'Riordan:* Singularly perturbed finite element methods. Numer. Math. Vol. 44, 3 (1984), 425—434.
[18] *G. D. Raithby:* Skew upstream differencing schemes for problems involving fluid flow. Comp. Meth. Appl. Mech. Eng. Vol 9 (1976), 153—164.
[19] *P. A. Raviart:* Les méthodes d'élements finis en mécanique des fluides II. 3. Edditions Eyrolles, Paris, 1981.

Souhrn

PETROVOVA-GALERKINOVA APROXIMACE PROBLÉMŮ TYPU KONVEKCE-DIFÚZE A REAKCE-DIFÚZE

JOSEF DALÍK

V předloženém článku je prezentována nová obecná konstrukce testovacích funkcí jako varianta Petrovovy-Galerkinovy metody. Je využita při tvorbě a teoretické analýze nových algoritmů pro numerické řešení Dirichletovy úlohy pro diferenciální rovnici $-\varepsilon u'' + pu' + qu = f$ na intervalu (0, 1). Pozornost je soustředěna na případ, kdy kladné číslo ε je podstatně menší než hodnoty funkcí $|p|$ a q. Je navržen rovněž algoritmus pro numerické řešení odpovídající rovinné úlohy a jsou uvedeny výsledky numerických experimentů.

Author's address: Dr. *Josef Dalík*, katedra matematiky a deskriptivní geometrie stavební fakulty VUT, Barvičova 85, 662 37 Brno.