Aplikace matematiky

Jitka Segethová

Reducing the bandwidth in solving linear algebraic systems arising in the finite element method

Aplikace matematiky, Vol. 25 (1980), No. 4, 286–304

Persistent URL: http://dml.cz/dmlcz/103862

REDUCING THE BANDWIDTH IN SOLVING LINEAR ALGEBRAIC SYSTEMS ARISING IN THE FINITE ELEMENT METHOD

JITKA SEGETHOVÁ

(Received June 1, 1978)

1. INTRODUCTION

The matrix of the system of linear algebraic equations, arising in the application of the finite element method to one-dimensional problems, is a bandmatrix. In approximation of high order, the band is very wide but the elements situated far from the diagonal of the matrix are negligibly small as compared with the diagonal elements. The situation is similar but much more complex for two or three-dimensional problems.

In numerical practice, the outer diagonals of the matrix influence the solution of the system very little. The aim of this paper is to show that it is possible to work with a matrix the bandwidth of which is reduced, i.e., some non-zero diagonals of the original matrix are replaced by zeros. For a sufficiently high order of approximation (i.e., for a sufficiently wide band) the error caused by this replacement may be negligible in comparison with the rounding errors involved in the numerical solution of the system.

We choose the hill functions ω_N [1], [5] for the trial functions in solving a model problem by the finite element method since we can study their asymptotic behavior for $N \to \infty$ by means similar to those used in [4]. This as well as the formulation of the model problem is the contents of Section 2.

In Section 3 we investigate the behavior of the elements of the matrix of the system as $N \to \infty$. This study gives us a key to the introduction of a matrix the band of which is reduced. In conclusion, the difference of the original and the reduced matrices is estimated.

Section 4 is devoted to the analysis of the rounding errors, which is based on [6], [7]. In this part of the paper we obtain error bounds for numerical solution of the original and the reduced systems. These bounds are quantities of the same order. Moreover, a simple numerical example illustrates the statements of Section 4.

2. ASYMPTOTIC BEHAVIOR OF THE HILL FUNCTIONS. THE MODEL PROBLEM

We employ the following particular type of hill functions (called also B-splines) as trial functions.

Definition 2.1. *Let $\mathbb{R}$ be a one-dimensional Euclidean space. Let us put*

$$\omega_1(t) = 1\,, \quad |t| \leqq \tfrac{1}{2}\,, \tag{2.1}$$
$$= 0\,, \quad |t| > \tfrac{1}{2}\,,$$

$$\omega_N = \omega_{N-1} * \omega_1 \tag{2.2}$$

where $*$ denotes the convolution.

The above introduced hill function ω_N is a piecewise polynomial function of degree $N-1$, it is continuous together with its derivatives up to the order $N-2$ and its support is the interval $\langle -\frac{1}{2}N, \frac{1}{2}N\rangle$ (see [1], [5]). Further,

$$\omega_N(-t) = \omega_N(t)\,, \quad \omega_N'(-t) = -\omega_N'(t)\,. \tag{2.3}$$

The asymptotic behavior of the hill functions ω_N for $N \to \infty$ is studied in [4]. Let us perform a more detailed analysis necessary for our further purposes.

Theorem 2.1. *Let ω_N be given by* (2.1), (2.2). *Then there exists a positive constant C independent of N such that*

$$\left|\omega_N(t\sqrt{N})\sqrt{N} - \sqrt{(6\pi^{-1})}\, e^{-6t^2}\right| \leqq CN^{-1}\,, \tag{2.4}$$

$$\left|\omega_N'(t\sqrt{N})N + 12\sqrt{(6\pi^{-1})}\, te^{-6t^2}\right| \leqq CN^{-1} \tag{2.5}$$

uniformly with respect to all $t \in \mathbb{R}$.

Proof. The convolution formula (2.2) is analogous to the convolution formula of the probability theory describing the probability density of the sum of N independent random variables with the same density. The bound (2.4) follows immediately from the central limit theorem ([2] Ch. XV, Sec. 5, Theorem 2, and Ch. XVI, Sec. 2, Theorem 2, cf. also [4] Theorem 3.2).

Moreover, we obtain (2.5) in the same way as (2.4) by substituting ω_N' for ω_N in the proofs of the corresponding theorems of [2].

We confine ourselves to a simple model problem in order to be able to present the main idea of our investigation of reducing the bandwidth. The analysis can be done in an analogous way for a very wide class of problems.

Definition 2.2. *Let us put*

$$Lu = -u'' + cu\,, \quad c > 0\,, \tag{2.6}$$

and consider the boundary-value problem

$$Lu = f \quad on \quad \left(-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi\right), \tag{2.7}$$

$$u'\left(-\tfrac{1}{2}\pi\right) = u'\left(\tfrac{1}{2}\pi\right) = 0 \tag{2.8}$$

where $f \in L_2\left(-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi\right)$.

Let N, M *be positive integers,*

$$2M > N\,. \tag{2.9}$$

Let us put

$$Q = \left[\tfrac{1}{2}(N-1)\right]^*), \tag{2.10}$$

$$h = \pi/(2M)\,, \tag{2.11}$$

$$P = 2M + 2Q + 1\,. \tag{2.12}$$

The function

$$u_N(t) = \sum_{j=-M-Q}^{M+Q} x_{N,h,j}\omega_{N,h,j}(t)$$

where

$$\omega_{N,h,j}(t) = \omega_N(t/h - j) \tag{2.13}$$

is said to be the approximate Ritz-Galerkin solution of the problem (2.7), (2.8) *if the coefficients* $x_{N,h,j}$; $j = -M - Q, \ldots, M + Q$, *are the solution of the system*

$$A_{N,h}x_{N,h} = s_{N,h} \tag{2.14}$$

of P *linear algebraic equations, where*

$$A_{N,h} = \{a_{N,h,j,k}\}_{j,k=-M-Q}^{M+Q}\,,$$

$$x_{N,h} = \{x_{N,h,j}\}_{j=-M-Q}^{M+Q}\,,$$

$$s_{N,h} = \{s_{N,h,k}\}_{k=-M-Q}^{M+Q}\,.$$

The elements of the matrix $A_{N,h}$ *and of the vector* $s_{N,h}$ *are given by the relations*

$$a_{N,h,j,k} = I'_{N,h,j,k} + cI_{N,h,j,k} = (\omega'_{N,h,j}, \omega'_{N,h,k}) + c(\omega_{N,h,j}, \omega_{N,h,k})\,; \tag{2.15}$$
$$j, k = -M - Q, \ldots, M + Q\,,$$

$$s_{N,h,k} = (f, \omega_{N,h,k})\,; \quad k = -M - Q, \ldots, M + Q\,, \tag{2.16}$$

where $(\cdot, \cdot)$ *is the scalar product in* $L_2\left(-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi\right)$.

We omit the subscripts N, h *and write* I_{jk}, I'_{jk}, a_{jk}, x_j, s_k, A, x, s *wherever it is not ambiguous.*

The following theorem shows some basic properties of the elements of the matrix A.

*) $[z]$ denotes the largest integer which fulfils $[z] \leqq z$.

Theorem 2.2. *Let a_{jk} be given by* (2.15) *of Definition* 2.2. *Then*

$$I_{jk} = I_{kj} = I_{-j,-k} = I_{-k,-j}, \tag{2.17}$$

$$I'_{jk} = I'_{kj} = I'_{-j,-k} = I'_{-k,-j}, \tag{2.18}$$

$$a_{jk} = a_{kj} = a_{-j,-k} = a_{-k,-j} \tag{2.19}$$

for $j, k = -M - Q, \ldots, M + Q$. *Further,*

$$I_{jk} = I'_{jk} = a_{jk} = 0 \quad \text{for} \quad |j - k| \geqq N; \quad j, k = -M - Q, \ldots, M + Q, \tag{2.20}$$

i.e., A is a symmetric bandmatrix of bandwidth $2N - 1$.

Proof. The symmetry relations (2.17), (2.18), (2.19) are an easy consequence of the substitution of the equalities (2.3) into (2.15). Since the support of ω_N is finite we obtain (2.20). ∎

The relations (2.17), (2.18), (2.19) allow us to study only one quarter of the elements of the matrix A. In the following, we choose those pairs of indices j, $k = -M - Q, \ldots, M + Q$ for which $0 \leqq k - j$, $0 \leqq k + j$. With respect to (2.20) we introduce the following four sets of pairs of indices j, k.

Definition 2.3. *Let N, M be positive integers satisfying* (2.9), *let Q be given by* (2.10). *We put*

$$S^{++} = \{(j, k);\ |j| \leqq M + Q,\ |k| \leqq M + Q,\ 0 \leqq k - j < N,\ 0 \leqq k + j\},$$

$$S^{+-} = \{(j, k);\ |j| \leqq M + Q,\ |k| \leqq M + Q,\ 0 \leqq k - j < N,\ k + j \leqq 0\},$$

$$S^{-+} = \{(j, k);\ |j| \leqq M + Q,\ |k| \leqq M + Q,\ -N < k - j \leqq 0,\ 0 \leqq k + j\},$$

$$S^{--} = \{(j, k);\ |j| \leqq M + Q,\ |k| \leqq M + Q,\ -N < k - j \leqq 0,\ k + j \leqq 0\}.$$

In the notation introduced in Definition 2.3 we confine ourselves to the analysis of only the elements of the set S^{++} in the following. The properties of the elements of the other three sets can be established according to Theorem 2.2 in an apparent way.

Let us mention that all the four sets S^{++}, S^{+-}, S^{-+}, S^{--} depend on the integers N, M. Therefore we sometimes write $S^{++}(N, M)$ or $S^{++}(N)$ instead of S^{++} etc.

The knowledge of the asymptotic behavior of the hill functions enables us to obtain asymptotic formulae for the elements a_{jk}, s_k as $N \to \infty$.

Theorem 2.3. *Let a_{jk}, s_k be given by* (2.15), (2.16) *of Definition* 2.2. *Then*

$$I_{jk} = I^*_{jk}(1 + O(N^{-1})), \tag{2.21}$$

$$I'_{jk} = I'^*_{jk}(1 + O(N^{-1})), \tag{2.22}$$

$$a_{jk} = (I'^*_{jk} + cI^*_{jk})(1 + O(N^{-1})) \tag{2.23}$$

for $N \to \infty$ *and* $(j, k) \in S^{++}$ *where**)

$$I^*_{jk} = (6N^{-1})^{3/2}\,\pi^{-1}h^{-1}\,\tfrac{1}{6}Nh^2\,\sqrt{(\tfrac{1}{2}\pi)}\exp\left(-3N^{-1}(j-k)^2\right)\times \tag{2.24}$$
$$\times\,\Phi\left(\sqrt{(3N^{-1})}\,(N+j-k)\right),$$

$$I'^*_{jk} = (6N^{-1})^{3/2}\,\pi^{-1}h^{-1}\exp\left(-3N^{-1}(j-k)^2\right)\times \tag{2.25}$$
$$\times\left(\sqrt{(\tfrac{1}{2}\pi)}\,(1-6N^{-1}(j-k)^2)\,\Phi\left(\sqrt{(3N^{-1})}\,(N+j-k)\right)-\right.$$
$$\left.-\sqrt{(6N^{-1})}\,(N+j-k)\times\exp\left(-3N^{-1}(N+j-k)^2\right)\right)$$

for $j < M - Q$ *and*

$$I^*_{jk} = \tfrac{1}{2}(6N^{-1})^{3/2}\,\pi^{-1}h^{-1}\,\tfrac{1}{6}Nh^2\,\sqrt{(\tfrac{1}{2}\pi)}\exp\left(-3N^{-1}(j-k)^2\right)\times \tag{2.26}$$
$$\times\left(\Phi\left(\sqrt{(3N^{-1})}\,(2M-j-k)\right)-\Phi\left(\sqrt{(3N^{-1})}\,(-N-j+k)\right)\right),$$

$$I'^*_{jk} = \tfrac{1}{2}(6N^{-1})^{3/2}\,\pi^{-1}h^{-1}\exp\left(-3N^{-1}(j-k)^2\right)\left(\sqrt{(\tfrac{1}{2}\pi)}\times\right. \tag{2.27}$$
$$\times(1-6N^{-1}(j-k)^2)\left(\Phi\left(\sqrt{(3N^{-1})}\,(2M-j-k)\right)-\Phi\left(\sqrt{(3N^{-1})}\times\right.\right.$$
$$\left.\times\,(-N-j+k)\right)+\sqrt{(6N^{-1})}\left((-N-j+k)\times\right.$$
$$\left.\left.\times\exp\left(-3N^{-1}(-N-j+k)^2\right)-(2M-j-k)\times\right.\right.$$
$$\left.\left.\times\exp\left(-3N^{-1}(2M-j-k)^2\right)\right)\right)$$

for $M - Q \leqq j$.

Further,

$$|s_k|^2 \leqq \|f\|^2_{L_2}\,I_{kk} \quad \textit{for} \quad k = -M-Q, \ldots, M+Q\,. \tag{2.28}$$

Proof. Considering that $(j, k) \in S^{++}$ and that (2.9) holds, we distinguish two possibilities in the proof, $j < M - Q$ and $j \geqq M - Q$, which correspond to different limits of integration in the calculation of (2.15). In both the cases we substitute the expressions

$$\omega_N(t) = \sqrt{(6\pi^{-1}N^{-1})}\exp\left(-6N^{-1}t^2\right) + O(N^{-3/2}),$$
$$\omega'_N(t) = 12N^{-1}\,\sqrt{(6\pi^{-1}N^{-1})}\,t\exp\left(-6N^{-1}t^2\right) + O(N^{-2}),$$

following from (2.4), (2.5) of Theorem 2.1, into (2.15). Taking into account that the functions $\Phi(t)$ and $\exp(-t^2)$ are bounded in $\mathbb{R}$ and considering that the length w of the interval of integration in (2.15) fulfils the inequality

$$w \leqq Nh\,,$$

we obtain (2.21) and (2.22) as well as (2.23) by direct calculation. The terms of higher degree in N^{-1} are neglected in these formulae. Applying the Schwarz inequality to (2.16), we obtain (2.28).

*) We use the usual notation

$$\Phi(t) = \frac{2}{\sqrt{\pi}}\int_0^t \exp(-u^2)\,\mathrm{d}u\,.$$

3. SYSTEM WITH REDUCED MATRIX

In this section we investigate the asymptotic behavior of the elements a_{jk}, s_k as $N \to \infty$. We introduce the matrix B of a reduced bandwidth as well as the associate approximate solution of the model problem making use of the knowledge of this asymptotic behavior.

In conclusion, we estimate the difference $B - A$ of the original and the reduced matrices and the difference $t - s$ of the original and the reduced right-hand parts of the system.

Let us start with the study of the elements a_{jk} of the matrix A.

Definition 3.1. *Let N, M be positive integers satisfying* (2.9). *Moreover, let there exist positive constants C, D independent of N such that*

$$M = M(N) \leqq CN^D . \tag{3.1}$$

Let Q be given by (2.10). *Let us choose $\frac{1}{2} > \varepsilon > 0$ and $Z > 0$ (independently of N) and denote by K the least odd integer not less than $ZN^{1/2+\varepsilon}$. Further let us put*

$$R = \tfrac{1}{2}(K - 1) . \tag{3.2}$$

Let us introduce the following five sets of pairs of indices j, k:

$$S_1^{+\,+}(N) = \{(j,k);\ j = -Q, \ldots, M - Q - 1;\ k = j + K, \ldots, j + N - 1\} \cap \cap S^{+\,+}(N) ,$$

$$S_2^{+\,+}(N) = \{(j,k);\ j = M - Q, \ldots, M - R - 1;\ k = j + K, \ldots, M + Q\} ,$$

$$S_3^{+\,+}(N) = \{(j,k);\ j = M - R, \ldots, M;\ k = M + R + 1, \ldots, M + Q\} ,$$

$$S_4^{+\,+}(N) = \{(j,k);\ j = M + 1, \ldots, M + R - 1;\ k = M + R + 1, \ldots, M + Q\} ,$$

$$S_5^{+\,+}(N) = \{(j,k);\ j = M + R, \ldots, M + Q - 1;\ k = j + 1, \ldots, M + Q\} .$$

Moreover, we write

$$\tilde{S}^{+\,+}(N) = \bigcup_{i=1}^{5} S_i^{+\,+}(N) .$$

In an analogous way we can introduce the sets $\tilde{S}^{+\,-}(N)$, $\tilde{S}^{-\,+}(N)$, $\tilde{S}^{-\,-}(N)$ (cf. Theorem 2.2 and Definition 2.3) and put

$$\tilde{S}(N) = \tilde{S}^{+\,+}(N) \cup \tilde{S}^{+\,-}(N) \cup \tilde{S}^{-\,+}(N) \cup \tilde{S}^{-\,-}(N) .$$

Apparently, $\tilde{S}^{+\,+}(N) \subset S^{+\,+}(N)$. The sets $S_i^{+\,+}(N)$ are mutually disjoint for N sufficiently large by definition.

In the following we always suppose that N, M are positive integers satisfying (2.9), (3.1). Moreover, P, Q always denote the integers given by (2.12), (2.10) of Definition

2.2 while R is given by (3.2) of Definition 3.1. With respect to (2.11), (3.1) we also have

$$h = h(N) = \pi/(2M(N)) . \tag{3.3}$$

Theorem 3.1. *Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence of pairs of integers. Let N_0 be a positive integer such that*

$$(j(N), k(N)) \in \tilde{S}^{++}(N) \quad \textit{for} \quad N \geqq N_0 . \tag{3.4}$$

Let us put

$$\alpha_{j(N),k(N)} = \pi h(\tfrac{1}{6}N)^{3/2} a_{j(N),k(N)} . \tag{3.5}$$

Let J be an arbitrary real number. Then there exists a positive constant C_J independent of N such that

$$|\alpha_{j(N),k(N)}| \leqq C_J N^{-J} \quad \textit{for} \quad N \geqq N_0 . \tag{3.6}$$

Proof. 1. First we will prove that, choosing a fixed i and considering a sequence $\{j(N), k(N)\}_{N=1}^{\infty}$ such that

$$(j(N), k(N)) \in S_i^{++}(N) \quad \text{for} \quad N \geqq N_i \tag{3.7}$$

with some positive N_i, we find a positive constant C_{iJ} independent of N such that

$$|\alpha_{j(N),k(N)}| \leqq C_{iJ} N^{-J} \quad \text{for} \quad N \geqq N_i . \tag{3.8}$$

Having proved (3.8) for $i = 1, \ldots, 5$, we obtain the statement of the theorem by the following argument.

Let us consider the sequence (3.4). Let us split its subsequence $\{j(N), k(N)\}_{N=N_0'}^{\infty}$ with some $N_0' \geqq N_0$ into five sets U_i; $i = 1, \ldots, 5$, such that

$$U_i = \{(j(N), k(N)) ; \ (j(N), k(N)) \in S_i^{++}(N)\} .$$

Apparently

$$\{j(N), k(N)\}_{N=N_0'}^{\infty} = \bigcup_{i=1}^{5} U_i . \tag{3.9}$$

For a sufficiently large N_0', each of the sets U_i is either empty or infinite and at least one of them is infinite with respect to (3.4). Let us put

$$I = \{i; U_i \neq \emptyset\} .$$

Then $I \neq \emptyset$ and if $i \in I$ we can write

$$U_i = \{j(N_n^i), k(N_n^i)\}_{n=1}^{\infty} , \quad N_0' \leqq N_n^i , \quad N_n^i < N_{n+1}^i .$$

For each $i \in I$ we obtain from (3.8) that

$$|\alpha_{j(N_n^i),k(N_n^i)}| \leqq C_{iJ}(N_n^i)^{-J} \quad \text{for} \quad N_n^i \geqq N_i .$$

Putting $C'_J = \max_{i \in I} C_{iJ}$, we have

$$\left|\alpha_{j(N),k(N)}\right| \leqq C'_J N^{-J} \quad \text{for} \quad N \geqq N''_0$$

where $N''_0 = \max(N'_0, \max_{i \in I} N_i)$, from which (3.6) follows immediately.

Therefore it remains to show (3.8) for $i = 1, \ldots, 5$. We will write (j, k) instead of $(j(N), k(N))$ wherever it is not ambiguous.

2. Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence satisfying (3.7) for $i = 1$. From Definition 3.1 and Theorem 2.3 we find that a_{jk} is given by (2.23), (2.24), (2.25) in this case, i.e.,

$$\begin{aligned}(3.10)\quad \alpha_{jk} = \exp\left(-3N^{-1}(j-k)^2\right)\left(\sqrt{\left(\tfrac{1}{2}\pi\right)}\left(\tfrac{1}{6}ch^2N + 1 - 6N^{-1}(j-k)^2\right) \times\right.\\ \times \Phi\left(\sqrt{(3N^{-1})}\,(N + j - k)\right) - \sqrt{(6N^{-1})}\,(N + j - k) \times\\ \left.\times \exp\left(-3N^{-1}(N + j - k)^2\right)\right)\left(1 + O(N^{-1})\right).\end{aligned}$$

Since $j + K \leqq k \leqq j + N - 1$ in $S_1^{+\,+}(N)$ we have

$$k - j \geqq K \geqq ZN^{1/2+\varepsilon},$$

i.e.,

$$(3.11) \qquad (k-j)^2 N^{-1} \geqq Z^2 N^{2\varepsilon} \to \infty \quad \text{for} \quad N \to \infty.$$

Therefore there exists a positive constant C'_{1J} independent of N such that

$$(3.12) \qquad \exp\left(-3N^{-1}(j-k)^2\right) < C'_{1J} N^{-J}$$

for an arbitrary real J. The functions $\Phi(t)$ and $t \exp(-t^2)$ are bounded in $\mathbb{R}$. For any sequence the elements of which lie in $S^{+\,+}(N)$, we have $0 \leqq k - j < N$ from Definition 2.3 and, moreover,

$$(3.13) \qquad (k-j)^2 N^{-1} < N.$$

Therefore we finally obtain

$$(3.14) \qquad \left|\alpha_{jk}\right| \exp\left(3N^{-1}(j-k)^2\right) \leqq D_1 N$$

with some positive constant D_1 from (2.9) and (3.3). Considering (3.12) and (3.14), we find a positive constant C_{1J} independent of N such that (3.8) holds with $i = 1$.

3. Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence satisfying (3.7) for $i = 2$. From Definition 3.1 and Theorem 2.3 we find that a_{jk} is given by (2.23), (2.26), (2.27) in this case, i.e.,

$$\begin{aligned}(3.15)\quad \alpha_{jk} = \tfrac{1}{2}\exp\left(-3N^{-1}(j-k)^2\right)\left(\sqrt{\left(\tfrac{1}{2}\pi\right)}\left(\tfrac{1}{6}ch^2N + 1 - 6N^{-1}(j-k)^2\right) \times\right.\\ \times \left(\Phi\left(\sqrt{(3N^{-1})}\,(2M - j - k)\right) - \Phi\left(\sqrt{(3N^{-1})}\,(-N - j + k)\right)\right) +\\ + \sqrt{(6N^{-1})}\left((-N - j + k)\exp\left(-3N^{-1}(-N - j + k)^2\right) -\right.\\ \left.\left.- (2M - j - k)\exp\left(-3N^{-1}(2M - j - k)^2\right)\right)\right)\left(1 + O(N^{-1})\right).\end{aligned}$$

Since $j + K \leqq k$ in $S_2^{+\,+}(N)$ we obtain (3.11) also in this case. Obviously (3.13) also holds because $S_2^{+\,+}(N) \subset S^{+\,+}(N)$. Therefore we can repeat the argument of part 2 of the proof in order to find a positive constant C_{2J} independent of N such that (3.8) holds with $i = 2$.

4. Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence satisfying (3.7) for $i = 3$. From Definition 3.1 and Theorem 2.3 we find that α_{jk} is given by (3.15). Further we obtain

$$k - j \geqq R + 1 > \tfrac{1}{2} Z N^{1/2+\varepsilon}$$

with respect to (3.2) and the inequality

$$K \geqq Z N^{1/2+\varepsilon} \tag{3.16}$$

Therefore

$$(k - j)^2 N^{-1} \geqq \tfrac{1}{4} Z^2 N^{2\varepsilon} \to \infty \quad \text{for} \quad N \to \infty\,. \tag{3.17}$$

Now we can repeat the argument of part 2 in order to find a positive constant C_{3J} independent of N such that (3.8) holds with $i = 3$.

5. Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a seuqence satisfying (3.7) for $i = 4$. We find that α_{jk} is given by (3.15). Further we obtain

$$k - j \leqq Q - 1 < \tfrac{1}{2} N$$

with respect to (2.10). Hence we finally have

$$(-N + k - j) N^{-1/2} < -\tfrac{1}{2} N^{1/2} \to -\infty \quad \text{for} \quad N \to \infty\,.$$

Using an asymptotic expansion for $\Phi(t)$ at $+\infty$ we obtain (see e.g. [3])

$$\Phi(t) = 1 - \exp(-t^2)\left((t\sqrt{\pi})^{-1} + O(t^{-3})\right), \quad t \to \infty\,. \tag{3.18}$$

Therefore there exists a positive constant $\tilde{C}'_{4J}$ independent of N such that

$$0 < 1 + \Phi\left(\sqrt{(3N^{-1})}\,(-N - j + k)\right) \leqq \tilde{C}'_{4J} N^{-J} \tag{3.19}$$

for an arbitrary real J. Analogously there exists a constant $\tilde{C}''_{4J}$ such that

$$0 < (N + j - k) \exp\left(-3N^{-1}(-N - j + k)^2\right) \leqq \tilde{C}''_{4J} N^{-J}\,. \tag{3.20}$$

From Definition 3.1 we further obtain that

$$k + j - 2M \geqq R + 2 > \tfrac{1}{2} Z N^{1/2+\varepsilon}$$

with respect to (3.2), (3.16). Then

$$(k + j - 2M) N^{-1/2} \geqq \tfrac{1}{2} Z N^{\varepsilon} \to \infty \quad \text{for} \quad N \to \infty \tag{3.21}$$

and we find positive constants $\hat{C}'_{4J}$, $\hat{C}''_{4J}$ idenpendent of N such that

$$0 < \Phi\left(\sqrt{(3N^{-1})}\,(2M - j - k)\right) + 1 \leqq \hat{C}'_{4J} N^{-J}\,, \tag{3.22}$$

$$(3.23) \qquad 0 < (-2M + k + j)\exp(-3N^{-1}(2M - j - k)^2) \leqq \hat{C}''_{4J}N^{-J}$$

by the same argument as above. Since the function $\exp(-t^2)$ is bounded in $\mathbb{R}$ and

$$\tfrac{1}{6}ch^2N + 1 - 6N^{-1}(j - k)^2 \leqq C'_{4J}N$$

with some constant C'_{4J} we can find a positive constant C_{4J} independent of N such that (3.8) holds with $i = 4$ if we take into account the inequalities (3.19), (3.20,) (3.22), (3.23).

6. Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence satisfying (3.7) for $i = 5$. From Definition 3.1 we obtain

$$k - j \leqq Q - R < \tfrac{1}{2}N - \tfrac{1}{2}ZN^{1/2+\varepsilon},$$

$$k + j - 2M \geqq 2R + 1 \geqq ZN^{1/2+\varepsilon}$$

with respect to (2.10), (3.2), (3.16). Repeating the argument of part 5 we finally find a positive constant C_{5J} independent of N such that (3.8) holds with $i = 5$.

The proof of the theorem is complete.

Corollary. *Let $\{j(N), k(N)\}_{N=1}^{\infty}$ be a sequence of pairs of integers. Let N_0 be a positive integer such that*

$$(3.24) \qquad (j(N), k(N)) \in \tilde{S}(N) \quad \textit{for} \quad N \geqq N_0 .$$

Let $\alpha_{j(N),k(N)}$ be given by (3.5), let J be an arbitrary real number. Then there exists a positive constant $\tilde{C}_J$ independent of N such that

$$|\alpha_{j(N),k(N)}| \leqq \tilde{C}_J N^{-J} \quad \textit{for} \quad N \geqq N_0 .$$

Proof. Theorem 3.1 treats the case when (3.4) holds for the sequence $\{j(N), k(N)\}$. From Theorem 2.2 we obtain the statement of the corollary for any sequence such that $(j(N), k(N)) \in \tilde{S}^{+-}(N)$ (or $\tilde{S}^{-+}(N)$ or $\tilde{S}^{--}(N)$). Proceeding in the way analogous to that of part 1 of the proof of Theorem 3.1, we obtain the statement for an arbitrary sequence satisfying (3.24). ∎

A similar statement holds for the elements s_k of the right-hand part s.

Theorem 3.2. *Let $\{k(N)\}_{N=1}^{\infty}$ be a sequence of integers. Let N_0 be a positive integer such that*

$$M + R + 1 \leqq |k(N)| \leqq M + Q \quad \textit{for} \quad N \geqq N_0 .$$

Let us put

$$\sigma_{k(N)} = \pi h(\tfrac{1}{6}N)^{3/2} s_{k(N)} .$$

Let J be a real number. Then there exists a positive constant C_J^* independent of N such that

$$|\sigma_{k(N)}| \leqq C_J^* \|f\|_{L_2} N^{-J} \quad \textit{for} \quad N \geqq N_0 .$$

Proof. According to (2.28) of Theorem 2.3, it is sufficient to show that

$$|\pi^2 h^2 (\tfrac{1}{6}N)^3 I_{kk}| \leqq C_J^* N^{-J} \quad \text{for} \quad N \geqq N_0 . \tag{3.25}$$

We have $I_{kk} = I_{-k,-k}$ from (2.17) of Theorem 2.2, which allows us to confine ourselves only to the study of sequences $\{k(N)\}_{N=1}^{\infty}$ such that

$$M + R + 1 \leqq k(N) \leqq M + Q \quad \text{for} \quad N \geqq N_0 . \tag{3.26}$$

The result obtained for these sequences can be generalized in the way indicated in the proof of Corollary.

Thus let $\{k(N)\}_{N=1}^{\infty}$ be a sequence satisfying (3.26). From Theorem 2.3 we find that I_{kk} is given by (2.21), (2.26) in this case, i.e.,

$$\pi^2 h^2 (\tfrac{1}{6}N)^3 I_{kk}^* = (\tfrac{1}{2}\pi)^{3/2} h^3 (\tfrac{1}{6}N)^{5/2} \left(\Phi(2\sqrt{(3N^{-1})}(M - k)) - \Phi(-\sqrt{(3N)})\right) .$$

Moreover, we obtain that

$$k - M \geqq R + 1 > \tfrac{1}{2} Z N^{1/2+\varepsilon}$$

with respect to (3.2), (3.16), (3.26). Therefore

$$(k - M) N^{-1/2} \geqq \tfrac{1}{2} Z N^{\varepsilon} \to \infty \quad \text{for} \quad N \to \infty . \tag{3.27}$$

Using the expansion (3.18) for $\Phi(t)$ and proceeding in the way analogous to part 5 of the proof of Theorem 3.1, we find a positive constant C_J^* independent of N such that (3.25) is fulfilled for $k(N)$ satisfying (3.26). This implies the statement of the theorem immediately. ∎

We showed in Theorem 3.1, Corollary, and Theorem 3.2 that some of the elements a_{jk} of the matrix A and some of the elements s_k of the vector s (premultiplied by the factor $\pi h(\frac{1}{6}N)^{3/2}$) converge rapidly to 0 as $N \to \infty$.

Working with the quantities α_{jk}, σ_k we obtain that $\alpha_{00} = O(1)$ as $N \to \infty$. We will employ this fact in the proof of Theorem 3.3. Apparently, $\alpha_{jk} \to 0$ implies $a_{jk} \to 0$ and $\sigma_k \to 0$ implies $s_k \to 0$. This is the behavior of e.g. all the elements a_{jk} lying outside the band of width $2K - 1$.

Let us note that the statement is based on the condition (3.16). In general, if we put $K \geqq Z N^{\beta}$ with $\beta \leqq \frac{1}{2}$ then none of the relations (3.11), (3.17), (3.21), (3.27) can be fulfilled. In this sense, the condition (3.16) determines the narrowest band with the property that the elements a_{jk} situated outside the band tend to 0.

We will study a matrix and a vector formed from the original matrix A and vector s by substituting zeros for the elements tending to zero. First we will introduce the necessary notation.

Definition 3.2. *Let*

$$B_{N,h} = B = \{b_{jk}\}_{j,k=-M-Q}^{M+Q}$$

be a square matrix with the elements

$$b_{jk} = 0 \quad \textit{for} \quad (j, k) \in \tilde{S}(N) ,$$

$$= a_{jk} \quad \textit{otherwise}\,.$$

Moreover, let

$$t_{N,h} = t = \{t_k\}_{k=-M-Q}^{M+Q}$$

be the vector the elements of which are given by the relations

$$t_k = 0 \quad \textit{for} \quad M + R + 1 < |k|\,,$$

$$= s_k \quad \textit{otherwise}\,.$$

The function

$$\hat{u}_N(t) = \sum_{j=-M-Q}^{M+Q} \hat{x}_{N,h,j}\omega_{N,h,j}(t) \tag{3.28}$$

where $\omega_{N,h,j}$ is given by (2.13) *is said to be the associate approximate Ritz-Galerkin solution of the problem* (2.7), (2.8) *if the coefficients $\hat{x}_{N,h,j}$; $j = -M - Q, \ldots, M + Q$, are the solution of the system*

$$B_{N,h}\hat{x}_{N,h} = t_{N,h} \tag{3.29}$$

of P linear algebraic equations where

$$\hat{x}_{N,h} = \hat{x} = \{\hat{x}_{N,h,j}\}_{j=-M-Q}^{M+Q} = \{\hat{x}_j\}_{j=-M-Q}^{M+Q}\,.$$

In addition, we write

$$C = B - A\,, \quad u = t - s\,. \tag{3.30}$$

The matrix B is introduced in Definition 3.2 in such a way that

$$B = \begin{bmatrix} B' & O & O \\ O & B^* & O \\ O & O & B'' \end{bmatrix} \tag{3.31}$$

where B^* is a square matrix of order

$$P^* = 2M + 2R + 1\,, \tag{3.32}$$

B', B'' are diagonal matrices of order $Q - R$ and the O's denote zero matrices of appropriate types.

Similarly

$$t^T = (O, t^{*T}, O) \tag{3.33}$$

where $t^* = \{t_k\}_{k=-M-R}^{M+R}$ and the superscript T denotes the transpose. Let us write

$$\hat{x}^T = (\hat{x}'^T, \hat{x}^{*T}, \hat{x}''^T) \tag{3.34}$$

where $\hat{x}^* = \{\hat{x}_j\}_{j=-M-R}^{M+R}$. In this notation we obtain

$$\hat{x}' = \hat{x}'' = O \tag{3.35}$$

from (3.29). Therefore (3.28) is equivalent to

$$\hat{u}_N(t) = \sum_{j=-M-R}^{M+R} \hat{x}_j \omega_{N,h,j}(t)$$

and $\hat{x}_j$ are determined from the system

$$B^* \hat{x}^* = t^* \tag{3.36}$$

of P^* linear algebraic equations. Moreover, the matrix B^* is a symmetric bandmatrix of bandwidth $2K - 1$ according to Definition 3.2 and

$$b_{jk} = b_{kj} = b_{-j,-k} = b_{-k,-j}\,; \quad j, k = -M - Q, \ldots, M + Q\,.$$

We will study the relation between the approximate and the associate approximate solution in the next section. First we will estimate the differences $A - B$, $s - t$ making use of the matrix and vector norms.

Definition 3.3. *Let $D = \{d_{jk}\}_{j,k=1}^{n}$ be a square matrix, $r = \{r_k\}_{k=1}^{n}$ a vector. We write*

$$\|D\| = \max_{1 \leqq j \leqq n} \sum_{k=1}^{n} |d_{jk}|\,, \tag{3.37}$$

$$\|r\| = \max_{1 \leqq k \leqq n} |r_k|\,. \tag{3.38}$$

The expressions (3.37), (3.38) are the well-known matrix and vector norms (cf. e.g. [7]).

Theorem 3.3. *Let A, B, C, s, t, u be the matrices and vectors introduced in Definitions 2.2 and 3.2, let ε_0 be an arbitrary positive number independent of N. Then there exists an integer $N_0 > 0$ such that*

$$\frac{\|C\|}{\|B\|} < \varepsilon_0\,, \quad \frac{\|u\|}{\|B\|} < \varepsilon_0 \tag{3.39}$$

for all $N \geqq N_0$ (and any M, h satisfying (2.9), (3.1), (3.3)).

Proof. 1. Let us estimate the ratio

$$\frac{\|C\|}{\|B\|} = \frac{\pi h(\tfrac{1}{6}N)^{3/2}\,\|B - A\|}{\pi h(\tfrac{1}{6}N)^{3/2}\,\|B\|}\,. \tag{3.40}$$

We have $\pi h(\frac{1}{6}N)^{3/2}\,\|B\| \geqq |\alpha_{00}|$ from (3.5). Further, we obtain that

$$\alpha_{00} = \left(\sqrt{(\tfrac{1}{2}\pi)}\,(\tfrac{1}{6}ch^2N + 1)\,\Phi(\sqrt{(3N)}) - \sqrt{(6N)}\exp(-3N)\right)(1 + O(N^{-1}))$$

from Definition 3.1 and Theorem 2.3. Since there exist positive constants $\bar{C}_{1J}$, $\bar{C}_{2J}$ independent of N such that

$$1 - \Phi(\sqrt{(3N)}) \leqq \bar{C}_{1J} N^{-J}$$

(cf. (3.18)) and

$$\sqrt{(6N)} \exp(-3N) \leqq \bar{C}_{2J} N^{-J}$$

for any real J we conclude that there exists a positive constant C_3 independent of N such that

$$|\alpha_{00}| \geqq C_3 \tag{3.41}$$

for a sufficiently large N.

2. With respect to Theorem 3.1 and Definition 3.2, we have

$$|\gamma_{jk}| \leqq C_j N^{-J}$$

for all the elements γ_{jk} of the matrix $\pi h(\frac{1}{6}N)^{3/2} C$, any real J, and a sufficiently large N. From (3.40), (3.41) we finally obtain the first inequality in (3.39).

The second inequality in (3.39) follows in an analogous way from Theorem 3.2 and Definition 3.2.

4. A NUMERICAL EXAMPLE. DISCUSSION

In this section we are concerned with the study of the difference of the approximate and the associate approximate Ritz-Galerkin solution of the problem (2.7), (2.8). Keeping the notation of the previous sections, it is sufficient to estimate the difference $x - \hat{x}$ for this purpose.

However, we will consider the situation from the practical, i.e., numerical point of view. Following the ideas of Wilkinson [6], [7], let us distinguish between the true solution x_t of the system (2.14) and the computed solution x_c of the same system; let us introduce the vectors $\hat{x}_t$, $\hat{x}_c$ with an analogous meaning with respect to the system (3.29).

Let us solve the system of linear algebraic equations by the Gaussian elimination (based on the triangular decomposition of the matrix of the system) with partial pivoting. Let us consider the floating-point arithmetic with the mantissa of t binary digits. (The analysis is analogous and follows also [6], [7] in the case of decimal computation.)

The detailed analysis of the rounding errors made in the arithmetic operations on digital computers in the floating-point computation is a very complex and deep problem. There are phenomena observed widely in practice but rigorously demonstrated only in certain special cases. In view of these rather unusual conditions we use several statements of [6], [7] the assumptions of which cannot be rigorously verified. Similarly we make some assumptions of this nature in the following. These assumptions are justified by a numerical example presented in the conclusion of this section.

Let us return to the error analysis. The computed solution x_c of the system (2.14) is the true solution of the perturbed system

$$(A + E)\, x_c = s$$

where

$$\|E\| \leqq 2^{-t} g q(P)\,, \tag{4.1}$$

$$q(P) = 2{\cdot}14\left(\tfrac{1}{2}P + 1\right)(P - 1)\,,$$

P is the order of the matrices A, E, and

$$g = \max_{1 \leqq r \leqq P} \max_{-M-Q \leqq j,k \leqq M+Q} |a_{jk}^{(r)}|$$

with $A^{(r)} = \{a_{jk}^{(r)}\}_{j,k=-M-Q}^{M+Q}$ defined as the matrix appearing in the rth step of the elimination (i.e., the original system is $A^{(1)}x = s$). The error expressed by E originates in the triangular decomposition and represents the prevailing part of the total error. The perturbations resulting from the rounding errors made in the backsubstitution are negligible as compared with E. Further, the bound (4.1) for E may be attained in practice. All these statements are taken from [7] (Ch. 3, Secs. 14–33).

It is

$$\|(T + V)^{-1}\| \leqq \frac{\|T^{-1}\|}{1 - \|T^{-1}V\|} \leqq \frac{\|T^{-1}\|}{1 - \|T^{-1}\|\,\|V\|}$$

provided that $\|T^{-1}\|\,\|V\| < 1$ (cf. e.g. [7]). Considering that $Ax_t = s$ and supposing that $\|A^{-1}\|\,\|E\| < 1$ (this assumption is quite natural; if it is not true the rounding errors are so large that their analysis is impossible), we arrive at the inequality

$$\frac{\|x_c - x_t\|}{\|x_t\|} \leqq \frac{\varkappa(A)\,\|E\|/\|A\|}{1 - \varkappa(A)\,\|E\|/\|A\|} \tag{4.2}$$

where $\varkappa(A) = \|A\|\,\|A^{-1}\|$ is the condition number of the matrix A.

Taking into account the relations (3.31), (3.33), (3.34), (3.35), we concluded in Section 3 that it is sufficient to solve the system (3.36) for the vector $\hat{x}^*$ in order to obtain the solution $\hat{x}$ of the system (3.29). Let $\hat{x}_c^*$ be the computed solution of the system (3.36), let

$$\hat{x}_c^T = (O, \hat{x}_c^{*T}, O)\,.$$

By the same argument as above we see that $\hat{x}_c$ is the true solution of the perturbed system

$$(B + \hat{E})\, \hat{x}_c = t\,,$$

i.e., with respect to (3.30),

$$(A + C + \hat{E})\, \hat{x}_c = s + u \tag{4.3}$$

where

$$\hat{E} = \begin{bmatrix} O & O & O \\ O & \hat{E}^* & O \\ O & O & O \end{bmatrix},$$

$\hat{E}^*$ is a square matrix of order P^* and the O's denote zero matrices of appropriate types. According to [7] we have

$$\|\hat{E}\| = \|\hat{E}^*\| \leqq 2^{-t}\hat{g}^* q(P^*) \tag{4.4}$$

where

$$\hat{g}^* = \max_{1 \leqq r \leqq P^*} \max_{-M-R \leqq j,k \leqq M+R} |b_{jk}^{(r)}|$$

and $B^{*(r)} = \{b_{jk}^{(r)}\}_{j,k=-M-R}^{M+R}$ is the matrix appearing in the rth step of the elimination.

Considering the matrix $C + \hat{E}$ and the vector u in (4.3) to be the perturbations of the system (2.14), we obtain

$$\|\hat{x}_c - x_t\| \leqq \frac{\varkappa(B)\,(\|C\|\,\|x_t\|/\|B\| + \|\hat{E}\|\,\|x_t\|/\|B\| + \|u\|/\|B\|)}{1 - \varkappa(B)\,\|\hat{E}\|/\|B\|} \tag{4.5}$$

provided that $\|B^{-1}\|\,\|\hat{E}\| < 1$.

Let us suppose that there exists a positive constant G independent of P^* such that

$$\hat{g}^* \leqq G\|B^*\|\,. \tag{4.6}$$

Numerical results confirm this assumption in the whole practically investigated range of values of P^*. Further, let G be chosen in such a way that the bound (4.6) for $\hat{g}^*$ may be attained.

From (4.4), (4.6) we thus obtain

$$\|\hat{E}\| \leqq 2^{-t}G\|B^*\|\,q(P^*) \leqq 2^{-t}G\|B\|\,q(P^*)\,.$$

In particular,

$$\|\hat{E}\|/\|B\| \leqq 2^{-t}Gq(P^*) \tag{4.7}$$

where the bound is attainable and tends to infinity as $N \to \infty$ (and $P \to \infty$, $P^* \to \infty$ as well). From Theorem 3.3 we have $\|u\|/\|B\| \to 0$ as $N \to \infty$. Let us now choose a sufficiently large N and consider such x_t that

$$\|\hat{E}\|\,\|x_t\|/\|B\| \gg \|u\|/\|B\|\,. \tag{4.8}$$

Numerical results in the following correspond apparently to this situation. Neglecting the term $\|u\|/\|B\|$ in (4.5) with respect to (4.8), we arrive at

$$\frac{\|\hat{x}_c - x_t\|}{\|x_t\|} \leqq \frac{\varkappa(B)\,(\|C\|/\|B\| + \|\hat{E}\|/\|B\|)}{1 - \varkappa(B)\,\|\hat{E}\|/\|B\|}\,. \tag{4.9}$$

Moreover, we have $\|C\|/\|B\| \to 0$ as $N \to \infty$ from Theorem 3.2. Neglecting now also the term $\|C\|/\|B\|$ in (4.9) with respect to (4.7), we finally obtain

$$\frac{\|\hat{x}_c - x_t\|}{\|x_t\|} \leqq \frac{\varkappa(B)\,\|\hat{E}\|/\|B\|}{1 - \varkappa(B)\,\|\hat{E}\|/\|B\|}\,. \tag{4.10}$$

Comparing the estimates (4.2), (4.10), we conclude that both the bounds are of the same nature as long as we suppose $\varkappa(B) \sim \varkappa(A)$. In fact, numerical results show that rather $\varkappa(B) < \varkappa(A)$. Furthermore, it is $P^* < P$ and therefore we can expect that $\hat{x}_c$ (obtained from $\hat{x}_c^*$) gives better results than x_c for a sufficiently large P.

The following simple numerical example illustrates the above statements. Let us seek the approximate Ritz-Galerkin solution of the problem (2.7), (2.8) with the operator L given in (2.6). We put

$$f(x) = -\sin(dx)\,, \quad d > 0\,.$$

The exact solution of the problem is

$$u(x) = -\frac{1}{d^2 + c}\sin(dx) + \frac{d}{(d^2 + c)\sqrt{c}}\,\frac{\cos(\frac{1}{2}\pi d)}{\operatorname{ch}(\frac{1}{2}\pi\sqrt{c})}\operatorname{sh}(x\sqrt{c})\,.$$

The approach of the whole paper is asymptotic; the statements always assume that N is sufficiently large. Under such conditions, Z in Definition 3.1 may be chosen arbitrarily. In numerical computation we are forced to work only with some limited range of values of N. Our experiments included $N \leqq 15$. With regard to this fact, we computed the associate approximate Ritz-Galerkin solution (Definition 3.2) for all the values R given (instead of (3.2) of Definition 3.1) by $R = [\frac{1}{2}(K - 1)]$ with $K = N - 1, N - 2, \ldots, 2$. The solution corresponding to $K = N$ is the approximate solution (cf. Definition 2.2). In this way, we obtained a series of results by solving systems of linear algebraic equations with matrices of bandwidth $2N - 1, 2N - 3, \ldots, 3$.

Table 1.

N	$\eta(N, N)$	$\eta(N, N-1)$	$\eta(N, N-2)$	$\eta(N, N-3)$	$\eta(N, N-4)$
2	0·587 −3				
3	0·139 −4	0·185 −2			
4	0·451 −5	0·287 −3	0·201 −2		
5	0·348 −6	0·151 −4	0·907 −3	0·164 −2	
6	0·240 −6	0·401 −6	0·108 −3	0·191 −2	0·119 −2
7	0·231 −6	0·232 −6	0·735 −5	0·303 −3	0·246 −2
8	0·231 −6	0·231 −6	0·353 −6	0·844 −4	0·590 −3
9	0·231 −6	0·231 −6	0·356 −6	0·188 −5	0·475 −3
10	0·231 −6	0·231 −6	0·231 −6	0·307 −5	0·898 −5
11	0·233 −6	0·231 −6	0·231 −6	0·239 −6	0·325 −4
12	0·231 −6	0·231 −6	0·231 −6	0·249 −6	0·615 −6
13	0·324 −5	0·232 −6	0·232 −6	0·231 −6	0·970 −6

The results given in Table 1 correspond to $c = 0{\cdot}25$, $d = 7$, $M = 16$ (i.e., $h \doteq 0{\cdot}098$). The results obtained with various other values of these parameters have very similar character.

The computation was carried out on an IBM computer in double precision. The resulting system of linear algebraic equations was solved by the Gaussian elimination. In Table 1 we present the quantity

$$\eta(N, K) = \left((4M + 1)^{-1} \sum_{m=-2M}^{2M} |\hat{u}_N(\tfrac{1}{2}mh) - u(\tfrac{1}{2}mh)|^2\right)^{1/2}$$

where the associate approximate solution $\hat{u}_N$ depends on K in the sense of the above remarks and Definitions 3.1 and 3.2. The format of the table entries is a mantissa and a decimal exponent. Let us note that

$$\max_{x \in \langle -\pi/2, \pi/2 \rangle} |u(x)| \doteq 0{\cdot}020$$

for $c = 0{\cdot}25$, $d = 7$.

The values of $\eta(N, N)$ decrease (for fixed $M = 16$) with increasing N up to $N = 5$. In double precision (the mantissa of $t = 56$ binary digits), the rounding errors prevail in the total error for $N > 5$ and $\eta(N, N)$ is almost constant up to $N = 12$; then it increases. This implies that it is necessary to use more precision in order to obtain results of higher accuracy.

On the other hand, it enables us to present a true illustration for our statements. The rounding errors are so large that the result is not affected by narrowing the band of the matrix of the system up to some width. This "critical bandwidth" depends on N. For $N = 7$ the critical bandwidth is 11 (i.e., 1 + 1 diagonals may be deleted), for $N = 12$ the critical bandwidth is 17 (i.e., 3 + 3 diagonals may be deleted).

In general, the presented numerical results confirm the statements of this section concerning the bounds (4.2) and (4.10).

References

[1] *I. Babuška:* Approximation by hill functions. Comment. Math. Univ. Carolinae 11 (1970), 787–811.

[2] *W. Feller:* An introduction to probability theory and its applications. Vol. 2. Wiley, New York 1966.

[3] *I. S. Gradšteĭn, I. M. Ryžik:* Tables of integrals, sums, series, and products (Russian). 5th edition. Nauka, Moskva 1971.

[4] *K. Segeth:* Universal approximation by hill functions. Czechoslovak Math. J. 22 (1972), 612–640.

[5] *J. Segethová:* Numerical construction of the hill functions. SIAM J. Numer. Anal. 9 (1972), 199–204.

[6] *J. H. Wilkinson:* Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach. 8 (1961), 281–330.

[7] *J. H. Wilkinson:* Rounding errors in algebraic processes. HMSO, London 1963.

Souhrn

ZUŽOVÁNÍ PÁSU PŘI ŘEŠENÍ SOUSTAV LINEÁRNÍCH ALGEBRAICKÝCH ROVNIC VZNIKAJÍCÍCH Z METODY KONEČNÝCH PRVKŮ

JITKA SEGETHOVÁ

Matice soustavy lineárních algebraických rovnic, vznikající při řešení jednorozměrných úloh metodou konečných prvků, je pásová. Při aproximaci vysokého řádu je její pás velmi široký, avšak prvky na okraji pásu (daleko od diagonály) jsou zanedbatelně malé vzhledem k diagonálním prvkům.

V práci se ukazuje, že je prakticky možné pracovat s maticí soustavy, jejíž pás je zúžen, tj. některé nenulové diagonály původní matice jsou nahrazeny nulami. Nepřesnost pramenící z tohoto zúžení pásu může být při dostatečně vysokém řádu aproximace (tj. při dostatečně širokém pásu původní matice) zanedbatelná ve srovnání se zaokrouhlovacími chybami vznikajícími při numerickém řešení soustavy.

Celý problém se studuje na modelové okrajové úloze a je doplněn numerickým příkladem.

Author's address: RNDr. *Jitka Segethová,* CSc., Matematicko-fyzikální fakulta Karlovy univerzity, Malostranské nám. 25, 118 00 Praha 1.