Russ. J. Numer. Anal. Math. Modelling, Vol. 16, No. 3, pp. 191–214 (2001)

On the solution of algebraic systems for serendipity FEM

O. AXELSSON,* Y. L. GURIEVA,† and V. P. IL'IN†

Abstract — The properties of algebraic systems for the serendipity FEM with the hierarchical basis functions are investigated for an anisotropic two-dimensional boundary value problem. The constants of the spectral equivalence of the matrices are obtained. Several one- and two-level iterative processes are considered, including the defect-correction approach. Theoretical estimates are confirmed by the results of numerical experiments.

Algebraic two-level and multilevel iterations are efficient approaches to the solution of large-scale linear systems of equations with special sparse matrix structure. They can provide general finite element matrices with an optimal (or nearly optimal) order of computational complexity, which is proportional (or nearly proportional) to the number of degrees of freedom, i.e. the numbers of unknowns, see [1–6]. One of the major problems here is the development of a variable-step preconditioning algorithm which presents a considerable generalization of the conjugate gradient methods based on the properties of the Krylov spaces [7, 9]. An important application of such techniques is the algebraic multigrid methods for recursively embedded grids by which one can obtain a condition number independent of the meshsize of the finest mesh.

The purpose of this paper is to investigate several two-level iterative algorithms for solving algebraic systems for serendipity finite element methods based on hierarchical quadratic basis functions for which the matrix structure is particularly advantageous and which are of considerable practical interest. The first results on this topic were obtained for the Laplace equation in [1, 2]. Our aim is to extend these results to the anisotropic diffusion equation and construct fast iterative solvers on the basis of efficient preconditioners and the defect correction approach, which have the computational cost comparable to the bilinear case.

The paper is organized as follows. In Section 1, we present the block structure and some algebraic properties of stiffness matrices for the serendipity FEM for a rectangular grid for the two-dimensional Dirichlet boundary value problem with anisotropic diffusion equation, namely the structure of local and global matrices, spectral estimates, and auxiliary inequalities. Several two- and one-level preconditioned conjugate gradient iterative processes are considered in Section 2. In

*Faculty of Natural Sciences, Mathematics and Informatics, University of Nijmegen, NL-6525ED Nijmegen, The Netherlands

†Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia

This work is supported by the grants NWO-RFBR 047.008.007 and RFBR N 99-01-00579.

particular, we consider the application of the explicit and implicit incomplete factorization methods as well as 'classical' and modified defect correction methods. In Section 3, we present the results of numerical experiments for the suggested algorithms for different mesh sizes and stopping criteria. In conclusion, we discuss the numerical results and make a comparative analysis of different algorithms in terms of the convergence rate and the computational complexity. Also, the advantage of the serendipity approach over the bilinear FEM is demonstrated.

1. ALGEBRAIC PROPERTIES OF SERENDIPITY FEM SYSTEMS

Consider a weak formulation of the model Dirichlet boundary value problem for the second order elliptic equation: for a given function $f \in H^1(\Omega)$ find the function $u \in H_0^1(\Omega)$ such that

$$\begin{aligned} a(u,v) &\equiv \int_\Omega \left(a_{1,1}\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + a_{2,2}\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) \mathrm{d}x\mathrm{d}y \\ &= \int_\Omega f\, v\, \mathrm{d}x\mathrm{d}y = (f,v) \end{aligned}$$

for any $v \in H_0^1(\Omega)$, where $\Omega = \{a < x_1 < b, c < x_2 < d\}$. In the following we assume that $u \in H^3(\Omega) \cap H_0^1(\Omega)$ and the coefficients $a_{k,l}$, for simplicity, are some piecewise positive constants with no discontinuities within the elements. We use regular rectangular elements with hierarchical basis functions of serendipity type, i.e. with bilinear basis functions corresponding to the vertex nodes and quadratic basis functions at the mid-edge points. For the reference element $\Omega_{i,j}^h = (0,1)^2$ (see Fig. 1) such basis functions are

$$\begin{array}{l} \varphi_1 = (1-x)(1-y),\ \ \varphi_2 = x(1-y),\ \ \varphi_3 = xy,\ \ \varphi_4 = (1-x)y \\ \varphi_5 = 4x(1-x)(1-y),\ \ \varphi_6 = 4xy(1-x) \\ \varphi_7 = 4(1-x)y(1-y),\ \ \varphi_8 = 4xy(1-y). \end{array} \tag{1.1}$$

In general, we consider the non-uniform rectangular grid

$$\begin{gathered} x_{i+1} = x_i + h_i^x,\ y_{j+1} = y_j + h_j^y,\ i = 0,\dots,I,\ j = 0,\dots,J \\ x_0 = a,\ x_{I+1} = b,\ y_0 = c,\ y_{J+1} = d \end{gathered}$$

and define $h = \max\{h_i^x, h_j^y\}$, but we use notations $h_i^x = h_x$, $h_j^y = h_y$ for the simpler case of the uniform grid.

The finite element solution is defined as a function $u^h \in V_h = \mathrm{Span}\{\varphi_k^{(i,j)}, k = 1,\dots,8\} \subset H_0^1(\Omega)$ which satisfies the variational equation

$$a(u^h, v^h) = (f, v^h) \quad \forall v^h \in V_h = \mathrm{Span}\{\varphi_k^{(i,j)}\} \subset H_0^1(\Omega)$$

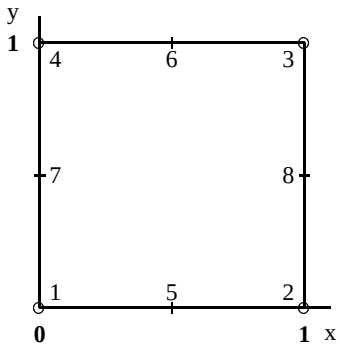

Figure 1. Reference serendipity element.

where $\varphi_k^{(i,j)}$ are the basis functions of type (1.1) defined in each grid cell (element) $\Omega_{i,j}^h = \{x_i < x < x_{i+1}, y_j < y < y_{j+1}\}$.

The local stiffness matrix for the element $\Omega_{i,j}^h$ is defined as $A_{i,j} = \{[\varphi_k, \varphi_l]_{\Omega_{i,j}^h}\}$ and can be presented as the sum

$$A_{i,j} = A_{i,j}^x + A_{i,j}^y \tag{1.2}$$

where

$$A_{i,j}^x = \left\{ a_{1,1} \int\limits_{\Omega_{i,j}^h} \frac{\partial \varphi_k}{\partial x} \frac{\partial \varphi_l}{\partial x} \mathrm{d}x\mathrm{d}y \right\}$$

$$A_{i,j}^y = \left\{ a_{2,2} \int\limits_{\Omega_{i,j}^h} \frac{\partial \varphi_k}{\partial y} \frac{\partial \varphi_l}{\partial y} \mathrm{d}x\mathrm{d}y \right\}$$

and the matrices have the entries:

$$A_{i,j}^x = \frac{a_{1,1}}{3} \frac{h_j^y}{h_i^x} \left[\begin{array}{cccc|cccc} 1 & -1 & -1/2 & 1/2 & 0 & 0 & 1 & -1 \\ -1 & 1 & 1/2 & -1/2 & 0 & 0 & -1 & 1 \\ -1/2 & 1/2 & 1 & -1 & 0 & 0 & -1 & 1 \\ 1/2 & -1/2 & -1 & 1 & 0 & 0 & 1 & -1 \\ \hline 0 & 0 & 0 & 0 & 16/3 & 8/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8/3 & 16/3 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 8/5 & -8/5 \\ -1 & 1 & 1 & -1 & 0 & 0 & -8/5 & 8/5 \end{array}\right]$$

$$A^y_{i,j} = \frac{a_{2,2}}{3}\frac{h^x_i}{h^y_j}\left[\begin{array}{cccc|cccc} 1 & 1/2 & -1/2 & -1 & 1 & -1 & 0 & 0 \\ 1/2 & 1 & -1 & -1/2 & 1 & -1 & 0 & 0 \\ -1/2 & -1 & 1 & 1/2 & -1 & 1 & 0 & 0 \\ -1 & -1/2 & 1/2 & 1 & -1 & 1 & 0 & 0 \\ --- & --- & --- & --- & --- & --- & --- & --- \\ 1 & 1 & -1 & -1 & 8/5 & -8/5 & 0 & 0 \\ -1 & -1 & 1 & 1 & -8/5 & 8/5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 16/3 & 8/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8/3 & 16/3 \end{array}\right].$$

For the simplest model problem when $a_{1,1} = a_{2,2} = 1$, $h^x_i = h^y_j = h$ we have

$$A_{i,j} = \frac{1}{3}\left[\begin{array}{cccc|cccc} 2 & -1/2 & -1 & -1/2 & 1 & -1 & 1 & -1 \\ -1/2 & 2 & -1/2 & -1 & 1 & -1 & -1 & 1 \\ -1 & -1/2 & 2 & -1/2 & -1 & 1 & -1 & 1 \\ -1/2 & -1 & -1/2 & 2 & -1 & 1 & 1 & -1 \\ -- & --- & --- & --- & --- & --- & --- & --- \\ 1 & 1 & -1 & -1 & 104/15 & 16/15 & 0 & 0 \\ -1 & -1 & 1 & 1 & 16/15 & 104/15 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 104/15 & 16/15 \\ -1 & 1 & 1 & -1 & 0 & 0 & 16/15 & 104/15 \end{array}\right].$$

We use a natural ordering of unknowns: node variables (line by line), horisontal edge variables (row by row) and vertical edge variables (column by column). Then the algebraic system of equations has the form

$$Au^h \equiv \left[\begin{array}{cc} A_n & A_{ne} \\ A_{en} & A_e \end{array}\right]\left[\begin{array}{c} u_n \\ u_e \end{array}\right] = \left[\begin{array}{c} f_n \\ f_e \end{array}\right] \equiv f^h \tag{1.3}$$

where A_n and A_e are square 'nodal' and 'edge' submatrices and u_n, u_e, f_n, f_e are the corresponding subvectors.

If the grid consists of N^2 square elements, $(N-1)^2$ interior nodes, and $2N(N-1)$ interior edge points, the 'nodal' nine-diagonal (block-tridiagonal) matrix A_n has the order $(N-1)^2$, the value 8/3 on the main diagonal and the value $-1/3$ for nonzero off-diagonal entries for the model problem. The 'edge' matrix is block diagonal, i.e.

$$A_e = \text{block diag}\{A^x_e, A^y_e\}$$

and has two diagonal blocks A^x_e, A^y_e. Each of them has the following block-diagonal structure:

$$A^x_e = \text{block diag}\{A^x_{e,i}\}, \quad i = 1, \ldots, N, \quad A^y_e = \text{block diag}\{A^y_{e,j}\}, \quad j = 1, \ldots, N$$

where $A^x_{e,i}$, $A^y_{e,j}$ are tridiagonal nonnegative submatrices of order $N-1$. For the model problem, the diagonal and off-diagonal entries of these submatrices are equal to $208/45$ and $16/45$, respectively. The rectangular submatrices A_{en} and A_{ne} have, in general, six and twelve nonzero entries in each row, respectively.

The grid stencils for the nodes and the mid-edge points are presented in Fig. 2.

For constant coefficients $a_{k,l}$ and a uniform grid the matrix A_n corresponding to the bilinear approximation of the original problem can be defined as

$$\begin{aligned}(A_n u)_{i,j} &= 4(c^x+c^y)u_{ij} - (2c^x-c^y)(u_{i-1,j}+u_{i+1,j}) - (2c^y-c^x)(u_{i,j-1}+u_{i,j+1}) \\ &\quad -(c^x+c^y)(u_{i-1,j-1}+u_{i+1,j+1})/2 \\ &\quad -(c^x+c^y)(u_{i-1,j+1}+u_{i+1,j-1})/2 \qquad (1.4)\\ c^x &= \frac{a_{1,1}h_y}{3h_x},\ c^y = \frac{a_{2,2}h_x}{3h_y}.\end{aligned}$$

It is obvious that A_n is an M-matrix only under strong conditions for the meshsteps and anisotropy coefficients:

$$1/2 \le c^x/c^y \le 2. \qquad (1.5)$$

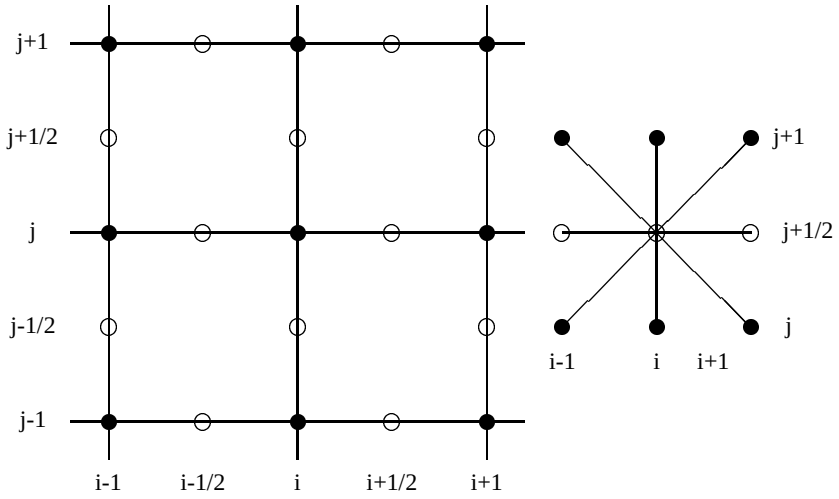

Figure 2. Node stencil (left) and edge stencil (right) with non-zero links.

For the model problem, the matrix A_n has the eigenvalues

$$\lambda_{p,q} = \frac{2}{3}(4-\cos p\pi h - \cos q\pi h - 2\cos p\pi h \cos q\pi h) \ge 2(\pi h)^2, \quad 1 \le p, q \le N-1$$

which yield the spectral condition number

$$\mathrm{cond}(A_n) \le \frac{8}{3(\pi h)^2}.$$

It was shown in [1, 2] that for the model problem the matrix A is spectrally equivalent to the block-diagonal matrix $\bar{A} = \text{block diag}\{A_n, A_e\}$ (see [1]):

$$\mathrm{cond}(A^{-1}\bar{A}) \le \frac{1+\gamma}{1-\gamma} \qquad (1.6)$$

where the constant γ is defined as

$$\gamma = \sup_{u\in V_n, v\in V_e} \frac{a(u,v)}{\|u\|\|v\|} = \sqrt{\frac{5}{11}} \approx 0.67$$

$$\|u\| = \sqrt{a(u,u)}, \quad V_n = \mathrm{Span}\{\varphi_1^{(i,j)}, \varphi_2^{(i,j)}, \varphi_3^{(i,j)}, \varphi_4^{(i,j)}\}$$

$$V_e = \mathrm{Span}\{\varphi_5^{(i,j)}, \varphi_6^{(i,j)}, \varphi_7^{(i,j)}, \varphi_8^{(i,j)}\}.$$

For any vector $v \in V = V_n \oplus V_e$ we can write instead of (1.6) the algebraic inequalities [2]

$$(1-\gamma)(\bar{A}v, v) \le (Av, v) \le (1+\gamma)(\bar{A}v, v) \tag{1.7}$$

which give the estimate of the condition number for the model problem:

$$\mathrm{cond}(A) \le \frac{1+\gamma}{1-\gamma}\mathrm{cond}\,(\bar{A}) \le \frac{1+\gamma}{1-\gamma}\frac{8}{3(\pi h)^2}$$

because of the inequalities

$$\mathrm{cond}(\overline{A}) \le \max\{\mathrm{cond}(A_n), \mathrm{cond}(A_e)\}, \quad \mathrm{cond}(A_e) \le 15/11.$$

Upon eliminating the subvector u_e from system (1.3), we obtain

$$\tilde{A}_n u_n \equiv (A_n - A_{ne}A_e^{-1}A_{en})u_n = \tilde{f}_n \equiv f_n - A_{ne}A_e^{-1}f_e \tag{1.8}$$

$$u_e = A_e^{-1}(f_e - A_{en}u_n). \tag{1.9}$$

Here $\tilde{A}_n$ is the Schur complement matrix which is spectrally equivalent to the matrix A_n and satisfies the following inequalities for any vector $v_n \in V_n$ [2]:

$$\begin{aligned} &(1-\gamma^2)(A_n v_n, v_n) \le (\tilde{A}_n v_n, v_n) \le (A_n v_n, v_n) \\ &\mathrm{cond}(A_n^{-1}\tilde{A}_n) \le \frac{1}{1-\gamma^2} = \frac{11}{6}, \mathrm{cond}(\tilde{A}_n) \le \frac{1}{1-\gamma^2}\mathrm{cond}(A_n) \le \frac{44}{9(\pi h)^2}. \end{aligned} \tag{1.10}$$

For a more general case of anisotropy coefficients and variable meshsteps, the spectral equivalence of the matrices and the estimates of the condition numbers can be obtained as follows.

Suppose we have the spectral equivalence inequalities of type (1.7) for each separate local matrix:

$$\begin{aligned} &(1-\gamma_{i,j}^x)(\overline{A}_{i,j}^x v, v) \le (A_{i,j}^x v, v) \le (1+\gamma_{i,j}^x)(\overline{A}_{i,j}^x v, v),\ 0 \le \gamma_{i,j}^x \le 1 \\ &(1-\gamma_{i,j}^y)(\overline{A}_{i,j}^y v, v) \le (A_{i,j}^y v, v) \le (1+\gamma_{i,j}^y)(\overline{A}_{i,j}^y v, v),\ 0 \le \gamma_{i,j}^y \le 1 \end{aligned} \tag{1.11}$$

where $v \in V^{ij} = V_n^{ij} \oplus V_e^{ij}$ and $\overline{A}_{i,j}^x, \overline{A}_{i,j}^y$ represent block-diagonal parts (each part is of the fourth order) of the matrices $A_{i,j}^x, A_{i,j}^y$, respectively. If we introduce

$$\gamma^x = \max_{i,j}\{\gamma_{i,j}^x\}, \quad \gamma^y = \max_{i,j}\{\gamma_{i,j}^y\}$$

then by summation of (1.11) over all the elements we can obtain (1.6) with

$$\gamma = \max\{\gamma^x, \gamma^y\}.$$

The value of $\gamma_{i,j}^x$ can evidently be found by solving the generalized eigenvalue problem

$$\lambda_p A_{i,j}^x z_p = \overline{A}_{i,j}^x z_p, \quad p = 1, \ldots, 8 \tag{1.12}$$

where $\max_p \lambda_p = 1/(1-\gamma)$.

Direct computations shows that the eigenvalues defined by (1.12) are equal to

$$\lambda_1 = \lambda_2 = \lambda_3 = 0, \lambda_4 = \lambda_5 = \lambda_6 = 1, \lambda_{7,8} = \frac{1}{1 \pm \sqrt{5/6}} \tag{1.13}$$

which are the same for all the elements $\Omega_{i,j}^h$.

Remark 1.1. These computations were done by an indirect application of MATLAB, The direct use of the latter for the generalized eigenvalue problem (1.12) failed due to the singularity of the matrices $A_{i,j}^x, \overline{A}_{i,j}^x$. But a close examination of their structure immediately reveals three zero and two unit eigenvalues with the five eigenvectors:

$$z_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad z_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad z_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad z_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad z_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

As for the remaining eigenvectors z_q, from the orthogonality conditions $(z_p, z_q) = 0$ for $p = 1, \ldots, 5$ and $q = 6, 7, 8$ which have the form

$$z_{p,1} = -z_{p,2}, \ z_{p,3} = -z_{p,4}, \ z_{p,5} = z_{p,6} = 0, \ z_{p,7} = -z_{p,8}$$

we get the generalized eigenvalue problem for the nonsingular matrices of the third order

$$\lambda_p \begin{bmatrix} 2 & -1 & 2 \\ -1 & 2 & -2 \\ 2 & -2 & 16/5 \end{bmatrix} z_p' = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 16/5 \end{bmatrix} z_p', \quad p = 6, 7, 8.$$

The solution of this problem gives the rest eigenvalues in (1.13).

Thus, we can define

$$\gamma^x = \gamma^y = \sqrt{5/6} \approx 0.913$$

because the matrix $A^y_{i,j}$ has the entries and the matrix structure similar to $A^x_{i,j}$.

Theorem 1.1. *Let A be defined in accordance with* (1.2), (1.3) *and* $\overline{A} =$ block diag $\{A_n, A_e\}$. *Then for the piecewise constant coefficients* $a_{k,k}$ *with nonnegative* $a_{1,2} = a_{2,1}$ *and for a nonuniform grid inequality (*1.7) holds with $\gamma = \sqrt{5/6}$ independent of $a_{k,k}$ and h^x_i, h^y_j.

In each element $\Omega^h_{i,j}$ the approximate solution is given as

$$\tilde{u}^h(x,y) = \sum_{k=1}^{8} c_k^{(i,j)} \varphi_k(x,y)$$

where the coefficients $c_k^{(i,j)}$ take the values of the components of u_n and u_e from (1.3) at the nodes and mid-edge points, respectively.

In general, for a non-uniform grid the error of the approximate solution is estimated in $L_2(\Omega)$-norm by the inequality (see, e.g. [2])

$$\|u - \tilde{u}^h\|_{L_2} \le Ch^3$$

where the constant C does not depend on h.

Because of the property of the basis functions (1.1), in the local notation of Fig. 1 we can find

$$c_k^{(i,j)} = \tilde{u}^h_k = \tilde{u}^h(x_k, y_k), \quad k = 1,2,3,4$$

and, for example,

$$c_5^{(i,j)} = \tilde{u}^h_5 - (\tilde{u}^h_1 + \tilde{u}^h_2)/2, \quad c_7^{(i,j)} = \tilde{u}^h_7 - (\tilde{u}^h_1 + \tilde{u}^h_4)/2. \tag{1.14}$$

Therefore the numerical solution of the algebraic system (1.3) provides at the mid-edge points the second finite differences of the finite element solution $\tilde{u}^h$, which approximate the values $-\dfrac{h_x^2}{8}\dfrac{\partial^2 u}{\partial x^2}$ and $-\dfrac{h_y^2}{8}\dfrac{\partial^2 u}{\partial y^2}$.

In Subsection 2.5, we shall prove for the isotropic case ($a_{1,1} = a_{2,2} = 1$) that the approximate solution of the serendipity FEM on the uniform rectangular grid $x_{i+1} = x_i + h_x$, $y_{j+1} = y_j + h_y$, under additional assumptions of smoothness, has the fourth-order error estimate

$$\|u - \tilde{u}\|_{L_2} = O(h^4), \quad h = \max\{h_x, h_y\}.$$

2. ITERATIVE SOLVERS

Due to the large size of the system in (1.3) we use iterative methods with different preconditioners for its solution. Here we consider both the methods based on the original matrix in (1.3) and the methods based on the reduced system (1.8).

2.1. Explicit incomplete factorization method

First we consider the application of the explicit incomplete factorization method EXIF [9] to solve the whole system (1.3). The above method is a formal generalization of the symmetric successive overrelaxation (SSOR) method and the Incomplete Cholesky algorithm ICCG(0), see [2, 9]. The method is given by the following preconditioned conjugate gradient (CG) iterative process:

$$\begin{aligned} &r^0 = f - Au^0, \quad p^0 = B^{-1}u^0 \\ &u^k = u^{k-1} + \alpha_{k-1}p^{k-1}, \quad \alpha_k = \frac{(B^{-1}r^k, r^k)}{(Ap^k, p^k)} \\ &r^k = r^{k-1} - \alpha_{k-1}Ap^{k-1} \\ &p^k = B^{-1}r^k + \beta_k p^{k-1}, \quad \beta_k = \frac{(B^{-1}r^k, r^k)}{(B^{-1}r^{k-1}, r^{k-1})}. \end{aligned} \tag{2.1}$$

Here for the original matrix

$$A = D_A - L_A - U_A, \quad D_A = \operatorname{diag}(A), \quad L_A = U_A^t \tag{2.2}$$

the preconditioning matrix B is written as

$$\begin{aligned} B &= (G_A - L_A)G_A^{-1}(G_A - U_A) \\ G_A &= \frac{1}{\omega}D_A - \operatorname{diag}(L_A G_A^{-1} U_A) - \vartheta S, \quad 0 < \omega < 2, \quad 0 \le \vartheta \le 1 \end{aligned} \tag{2.3}$$

and the diagonal matrix S is defined by the row sum criterion (the compensation principle)

$$Se = \left[L_A G_A^{-1} U_A - \operatorname{diag}(L_A \bar{G}_A U_A) - \frac{\omega - 1}{\omega} D_A \right] e$$

where e is the vector with unit entries and ω, ϑ are the relaxation and compensation parameters, respectively. For $\vartheta = 0$ the formulae (2.2), (2.3) reduce to the SSOR-CG method.

The number of iterations $k(\varepsilon)$ needed to satisfy the inequality

$$(A^{-1}r^k, r^k)/(A^{-1}r^0, r^0) \le \varepsilon^2 \ll 1$$

is estimated as

$$k(\varepsilon) \le \frac{1}{2}|\ln\varepsilon|\sqrt{\varkappa} + 1$$

where $\varkappa$ is the condition number of the matrix product $B^{-1}A$.

Because of the complicated structure of the original 'serendipity' matrix A, in the next section the estimation of the number of iterations for the method (2.1)–(2.3) is shown only experimentally. It should be noted that the convergence of the EXIF method for $0 < \vartheta \leq 1$ is proved for M-matrices only; for the general symmetric positive definite matrix the strict proof exists only for $\vartheta = 0$, which corresponds to the SSOR method. We emphasize that in the matrix A each 'nodal' and 'edge' row has 21 and 9 non-zero entries, respectively (see Fig. 2). Thus, at any iteration one vector-matrix operation Ap^k requires asymptotically, for big N, $M_A = 39N^2 + O(N)$ multiplications since there are $(N-1)^2$ 'nodal' and $2N(N-1)$ 'edge' rows in total. About the same number of arithmetical operations is needed to compute the vector $B^{-1}r^k$ if r^k is given. And, in addition, since the total number of unknowns equals $(N-1)(3N-1)$, the implementation of the CG formulae (2.1) requires $M_{CG} = 15N^2 + O(N)$ multiplications at each iteration, because one iteration of the CG requires five multiplications per unknown. Thus, the preconditioned explicit incomplete factorization method (2.1)–(2.3) requires the total number of multiplications to solve the serendipity equation (1.3):

$$M_{\mathrm{EXIF}} \approx 93N^2 k(\varepsilon) \tag{2.4}$$

plus some relatively small number of operations before iterations to compute the entries of the preconditioning matrix factors and the initial vectors r^0 and p^0.

Our further considerations are based on the special block matrix structure of the serendipity FEM, namely first we compute the nodal vector u_n from the reduced system (1.8), in which case the mid-point edge vector u_e can easily be found from (1.9).

2.2. Two-level Jacobi method

Formally, the simplest Jacobi method for (1.8) is written in the form

$$u_n^k = A_n^{-1}A_{ne}A_e^{-1}A_{en}u_n^{k-1} + A_n^{-1}\tilde{f}_n \tag{2.5}$$

and can be symmetrized by multiplication of (2.5) by the matrix $L_e^{-1}A_{en}$ from the left (here L_e is the left Cholesky decomposition factor, i.e. $A_e = L_eU_e$, $L_e = U_e^t$) and by the definition of a new variable $v_e^k = L_e^{-1}A_{en}v_n^k$:

$$\begin{gathered} v_e^k = Tv_e^{k-1} + g_e, \quad T = L_e^{-1}A_{en}A_n^{-1}A_{ne}U_e^{-1} \\ g_e = L_e^{-1}A_{en}A_n^{-1}(f_n - A_{ne}A_e^{-1}f_e). \end{gathered} \tag{2.6}$$

The improved version of (2.6) is the CG iterative process which is presented formally by (2.1) with the changes of the notations: $f \to g_e$, $A \to I - T$, $u^k \to v_e^k$, $B = I$. Since this method includes the computationally expensive multiplication by the matrix A_n^{-1}, a reasonable modification of the algorithm consists of the two-level iterative process

$$\tilde{v}_e^k = \tilde{T}_k\tilde{v}_e^{k-1} + g_e \quad (+\text{CG acceleration}), \quad \tilde{T}_k = L_e^{-1}A_{en}\hat{A}_n^{-1}A_{ne}U_e \tag{2.7}$$

where $\hat{A}_n^{-1}$ denotes some approximation of the inverse matrix. In fact, the intermediate matrix-vector multiplication $w_k = \hat{A}_n^{-1} A_{ne} U_e^{-1} \tilde{v}_e^{k-1}$ in (2.7) is replaced by the iterative (CG) solution of the system

$$A_n w_k = A_{ne} U_e^{-1} \tilde{v}_e^{k-1}.$$

Thus, we have two-level iterations with, in general, dynamic preconditioning, because the matrices $\tilde{T}_k$ in (2.7) are different at different steps.

In terms of the original iterated vectors u_n^k, the considered algorithm (2.5) is equivalent to the preconditioned solution of equation (1.8):

$$A_n(u_n^k - u_n^{k-1}) = r_n^{k-1} \equiv \tilde{f}_n - \tilde{A}_n u_n^{k-1} \tag{2.8}$$

with the number of iterations defined by the value cond $(A_n^{-1}\tilde{A}_n)$.

2.3. One-level solution of the reduced system

We also consider a one-level iterative process to solve (1.8), which is based on the application of the preconditioning matrix of the nine-diagonal matrix A_n. More precisely, we use the preconditioner of the implicit incomplete factorization method IMIF93 [9]:

$$\begin{aligned} B_n &= (G_n - L_n) G_n^{-1} (G_n - U_n) \\ G_n &= D_n - L_n (G_n^{-1})^{(3)} U_n - \vartheta S_n \\ S_n e &= [L_n G_n^{-1} U_n - L_n (G_n^{-1})^{(3)} U_n] e \end{aligned} \tag{2.9}$$

where $(G_n^{-1})^{(3)}$ represents the three-diagonal portrait of the matrix G_n^{-1}, D_n and $L_n = U_n^t$ are the block-diagonal and the block-low triangular parts of the matrix $A_n = D_n - L_n - U_n$, respectively. This means that B_n is approximate factorization of the matrix A_n but not of the matrix $\tilde{A}_n$ of the system (1.8) to be solved. The implementation of this preconditioned CG method can be done by the formulae (2.1) with the changes of the notations

$$B \to B_n, \quad A \to \tilde{A}_n, \quad u^k \to u_n^k, \quad f \to \tilde{f}_n. \tag{2.10}$$

The computational cost of one iteration for the method (2.1), (2.9), (2.10) can be estimated as follows. If we save the entries of matrix factorizations for A_e and B_n, then the numbers of multiplications in matrix-vector operations $\tilde{A}_n p^k$ and $B_n^{-1} r^k$ are equal to

$$M_{\tilde{A}} = 39N^2 + O(N), \quad M_B = 12N^2 + O(N)$$

respectively. Thus, the implementation of one iteration requires

$$M = 5N^2 + M_{\tilde{A}} + M_B \approx 56N^2 + O(N) \tag{2.11}$$

arithmetic operations, because the vector dimension in this case equals $(N-1)^2$.

Lemma 2.1. *The following estimate is valid for a uniform grid under conditions* (1.4), (1.5):

$$\mathrm{cond}\,(B_n^{-1}A_n) \le \frac{N+3}{4}.$$

This inequality can be proved by direct application of the results from [11].

Since the arithmetic costs of the decompositions of A_e and B_n are estimated by the value $O(N^2)$, we can formulate the following result.

Theorem 2.1. *The number of arithmetic operations in the method* (2.1), (2.9), (2.10), *under the conditions of Lemma* 2.1, *is estimated by the value*

$$M_{\mathrm{IMIF}} \le |\ln\varepsilon|\ 14N^2\sqrt{\frac{11N}{6}}.$$

This result is a direct consequence of (1.10), Lemma 2.1, and the evident inequality

$$\mathrm{cond}(B_n^{-1}\tilde{A}_n) \le \mathrm{cond}(B_n^{-1}A_n)\ \mathrm{cond}(A_n^{-1}\tilde{A}_n).$$

For the iterative solution of the serendipity algebraic system (1.3), the reasonable choice of the initial values u_n^0, u_e^0 is presented by the solution u_{nn} of the nodal algebraic system

$$A_n u_{nn} = f_{nn} \tag{2.12}$$

where the right- hand vector f_{nn} is defined by the bilinear basis functions only. Since $u_n^0 = u_{nn}$ has the error $O(h^2)$, in accordance with (1.9), (1.14), the acceptable choice can be given by

$$u_e^0 = A_e^{-1}(f_e + U u_n^0)$$

or obtained computing the second-order differences of the vector u_n^0.

It should be noted that the implementation of each iteration (2.1), (2.9), (2.12) for nodal unknowns requires $\approx 26N^2$ multiplications and the number of iterations is $\sqrt{11/6}$ times less. Thus, for the model problem the total cost of solving (2.12) by the IMIF method is

$$M_{nn} \le \frac{1}{2}|\ln\varepsilon|\ 13N^2\sqrt{N}.$$

It should be noted that the convergence of the IMIF method is proved for M-matrices as it was done at the end of Subsection 2.1 for the EXIF method. Therefore the general case is investigated experimentally in the following section.

2.4. Defect correction method

It is also possible to use the efficient defect correction algorithm (see [2, 10]) for the solution of the algebraic system (1.8). In this classical approach a solution of an auxiliary equation with the same right-hand side but with a more easily invertible operator A_n is used to obtain the solution of a given equation. Thus, we first compute the initial solution and the residual

$$u_n^0 = A_n^{-1}\tilde{f}_n, \quad r_n^0 = \tilde{f}_n - \tilde{A}_n u_n^0. \tag{2.13}$$

At the second step we solve the equation

$$A_n u_n^1 = \tilde{f}_n + r_n^0 \tag{2.14}$$

with u_n^0 as the initial value for the iterative solution of (2.14). Here one can check the norm of the residual $r_n^1 = \tilde{f}_n - \tilde{A}_n u_n^1$ and compute the next approximations $u_n^2, \ldots$, if necessary.

The properties of such an approach are the following. Because of (2.13), (2.14), we can write

$$u_n^1 = P\tilde{f}_n, \quad P = 2A_n^{-1} - A_n^{-1}\tilde{A}_n A_n^{-1}.$$

Thus, for the error $u_n - u_n^1$ of this approximation to the solution of equation (1.3) we have

$$u_n - u_n^1 = (\tilde{A}_n^{-1} - P)\tilde{f}_n = (I - A_n^{-1}\tilde{A}_n)^2 \tilde{A}_n^{-1}\tilde{f}_n = A_n^{-1}(A_n - \tilde{A}_n)(I - A_n^{-1}\tilde{A}_n)u_n.$$

In a similar way it is easy to show that if we define the corrections as

$$r^{k-1} = \tilde{f}_n - \tilde{A}_n u^{k-1}, \quad A_n u^k = \tilde{f}_n + r^0 + r^1 + \ldots + r^{k-1} \tag{2.15}$$

then the error (defect) satisfies the relation

$$u_n - u_n^k = (I - A_n^{-1}\tilde{A}_n)^{k+1} u_n.$$

Let us denote by $\hat{u}_n$, $\hat{f}_n$ the nodal restrictions of the exact solution and the right-hand side function of the original boundary value problem and let

$$z_{nn} = \hat{u}_n - u_{nn} = h^{\alpha} w_h, \quad z_n = \hat{u}_n - u_n = h^{\beta}\tilde{w}_h \tag{2.16}$$

be the errors of the bilinear solution (2.12) and the serendipity solution, respectively ($0 < \alpha < \beta$, w_h and $\tilde{w}_h$ are grid restrictions of smooth functions). Then from the above it follows that

$$u_n - u_n^1 = A_n^{-1}(A_n - \tilde{A}_n)[z_{nn} - z_n + A_n^{-1}(f_{nn} - \tilde{f}_n)].$$

If $\tilde{f}_n$ is a smooth approximation of f_n, $\tilde{f}_n = \hat{f}_n + O(h^{\alpha})$, then from (2.16) we have $u_n - u_n^1 = O(h^{2\alpha})$. But because the smoothness property of the difference $f_n - \tilde{f}_n$

is not evident, we investigate the efficiency of the defect correction method experimentally in the next subsection.

The above approximation principle of constructing the defect correction approach can be reformulated in algebraic terms and iterations can be accelerated by the CG method. Indeed we can rewrite (2.15) in the form (2.8)

$$A_n(u^k - u^{k-1}) = r^{k-1} \tag{2.17}$$

and use the formulae (2.1) with the changes

$$u^k \to u_n^k, \quad B \to A_n, \quad f \to \tilde{f}_n, \quad A \to \tilde{A}_n. \tag{2.18}$$

Lemma 2.2. *The number of iterations in the method* (2.1), (2.18) *is estimated by the value*

$$k(\varepsilon) \le \frac{1}{2}|\ln \varepsilon| \sqrt{\frac{1}{1-\gamma^2}} + 1.$$

This result for the preconditioned CG method follows directly from (1.10). But we must keep in mind that the computation of the vector $A_n^{-1} r^k$ is in fact done iteratively by the method IMIF, for example. Thus, in this case we obtain a two-level iterative process with a variable-step preconditioning.

It is easy to show that, based on the corresponding estimate, algorithm (2.1), (2.18) has the same convergence rate and the same number of outer iterations $k(\varepsilon)$ as the method (2.7). And because of the necessity to solve the auxiliary nodal subsystems in (2.6), (2.7), the computational complexity of (2.1), (2.18) seems to be more preferable.

Let m_k denote the number of inner iterations at the k-th outer iteration (m_k is proportional to $|\ln \varepsilon_k|$, where ε_k is variable, in general, intrinsic accuracy), $M_k = (12 + 9 + 5)N^2 = 26N^2$ is the number of multiplications at each iteration of the IMIF method.

Theorem 2.2. *The number of arithmetic operations in the two-level iterative process* (2.1), (2.18) *is estimated as*

$$M_{2L} \le 44K + 26(m_1 + \ldots + m_k) \tag{2.19}$$

where $K = k(\varepsilon)$ *is defined in Lemma* 2.2.

Obviously, minimization of the computational complexity for this algorithm requires the optimization of the values $m_k(\varepsilon_k)$ for given ε.

2.5. Modified defect correction method

We now consider a modification of the defect correction method to solve iteratively system (1.8), where we aim to obtain an $O(h^4)$ accuracy. In the following we show

that this accuracy can be obtained with the two modified steps (2.13), (2.14) of the defect correction method. For this purpose, we should consider in more detail the approximation property of the serendipity FEM for the uniform rectangular grid.

Let us introduce the notations (see Fig. 2):

$$u_n = \{u_{i,j}\}, \quad u_e = \{\hat{u}_e, \check{u}_e\}, \quad \hat{u}_e = \{v_{i\pm 1/2,j}\}, \quad \check{u}_e = \{v_{i,j\pm 1/2}\}$$

for the nodal, horizontal, and vertical mid-edge components of the unknown solution u^h in (1.3).

Upon rescaling equation (1.3) to the conventional finite difference form and if $u_{i,j}$, $v_{i\pm 1/2,j}$, $v_{i,j\pm 1/2}$ are the restrictions of the smooth enough functions, for $a_{1,1} = a_{2,2} = 1$, we can write

$$\begin{aligned}(A_n u_n)_{i,j} \;=\; & \frac{2}{3}\Big(\frac{1}{h_x^2}+\frac{1}{h_y^2}\Big)u_{i,j} - \frac{1}{3}\Big(\frac{1}{h_y^2}-\frac{1}{2h_x^2}\Big)(u_{i,j-1}+u_{i,j+1}) \\ & -\frac{1}{3}\Big(\frac{1}{h_x^2}-\frac{1}{2h_y^2}\Big)(u_{i-1,j}+u_{i+1,j}) \\ & -\frac{1}{12}\Big(\frac{1}{h_x^2}+\frac{1}{h_y^2}\Big)(u_{i+1,j+1}+u_{i-1,j-1}+u_{i-1,j+1}+u_{i+1,j-1})\end{aligned}$$

$$\begin{aligned}(A_{n,e} u_e)_{i,j} \;=\; & \frac{1}{3}\Big[\frac{1}{h_y^2}(v_{i+1/2,j}+v_{i-1/2,j}) + \frac{1}{h_x^2}(v_{i,j+1/2}+v_{i,j-1/2})\Big] \\ & -\frac{1}{6h_x^2}(v_{i+1,j-1/2}+v_{i+1,j+1/2}+v_{i-1,j-1/2}+v_{i-1,j+1/2}) \\ & -\frac{1}{6h_y^2}(v_{i+1/2,j+1}+v_{i+1/2,j-1}+v_{i-1/2,j+1}+v_{i-1/2,j-1})\end{aligned}$$

$$\begin{aligned}(A_e \hat{u}_e + A_{e,n} u_n)_{i+1/2,j} \;=\; & 16\Big(\frac{2}{3h_x^2}+\frac{1}{5h_y^2}\Big)v_{i+1/2,j} \\ & +8\Big(\frac{1}{3h_x^2}-\frac{1}{5h_y^2}\Big)(v_{i+1/2,j+1}+v_{i+1/2,j-1}) \\ & +\frac{1}{h_y^2}[2(u_{i,j}+u_{i+1,j}) - u_{i,j+1} - u_{i+1,j+1} - u_{i,j-1} - u_{i+1,j-1}]\end{aligned}$$

$$
\begin{aligned}
(A_e\check{u}_e + A_{e,n}u_n)_{i,j+1/2} \;=\; & 16\Big(\frac{2}{3h_y^2}+\frac{1}{5h_x^2}\Big)v_{i,j+1/2} \\
& +8\Big(\frac{1}{3h_y^2}-\frac{1}{5h_x^2}\Big)(v_{i+1,j+1/2}+v_{i-1,j+1/2}) \\
& +\frac{1}{h_x^2}[2(u_{i,j}+u_{i,j+1})-u_{i+1,j} \\
& -u_{i+1,j+1}-u_{i-1,j}-u_{i-1,j+1}]. \qquad (2.20)
\end{aligned}
$$

By means of the direct Taylor expansion we can write for the errors z_{nn} and z_n of the bilinear and the serendipity solutions, see (1.8), (2.12), (2.16):

$$A_n z_{nn} = \psi_{nn} = h^2 w_h, \quad \tilde{A}_n z_n = \psi_n = h^4 \tilde{w}_h. \qquad (2.21)$$

Lemma 2.3. *Let A_n and $\tilde{A}_n = A_n - A_{n,e}A_e^{-1}A_{e,n}$ be the matrices defined in* (1.3), (1.8), (2.20). *Then for a sufficiently smooth solution u of the original differential operator problem we have*

$$z_{nn} = h^2 e_h, \quad z_n = h^4 \tilde{e}_h \qquad (2.22)$$

for each nodal point $(x_i, y_j) \in \Omega_h$, where e_n and $\tilde{e}_n$ are smooth functions that do not depend on h.

Proof. We first note that from (2.21) it follows that

$$z_{nn} = h^2 A_n^{-1} w_h, \quad z_n = h^4 \tilde{A}_n^{-1} \tilde{w}_h$$

because A_n, $\tilde{A}_n$ are the the diagonal block and the Schur complement of the positive definite matrix A. Thus, they have a bounded inverse and $e_h, \tilde{e}_h$ in (2.22) are the restrictions of the smooth functions.

The modified defect correction method takes the following form. To solve the Schur complement system

$$\tilde{A}_n u_n \equiv (A_n - A_{ne}A_e^{-1}A_{en})u_n = f_n - A_{ne}A_e^{-1} f_e \equiv \tilde{f}_n$$

we first solve the bilinear approximation equation

$$A_n u_n^0 = f_{nn}. \qquad (2.23)$$

Then we compute u_n^1, which is the next approximation to u_n, by solving

$$A_n u_n^1 = f_{nn} + r_n^0, \quad r_n^0 = \tilde{f}_n - \tilde{A}_n u_n^0. \qquad (2.24)$$

Theorem 2.3. *Consider a uniform rectangular mesh Ω_h with the bilinear approximation u_n^0 and the modified defect correction solution u_n^1 [see (2.20), (2.21)] for the model problem. Then for a sufficiently smooth solution u we have an h^4 error expansion*

$$\hat{u}_n - u_n^1 = O(h^4).$$

Proof. Since u_n^1 can be represented in the form

$$u_n^1 = (2A_n^{-1} - A_n^{-1}\tilde{A}_n A_n^{-1})\,\tilde{f}_n - A_n^{-1}(\tilde{A}_n A_n^{-1} - I)(f_{nn} - \tilde{f}_n)$$

for the error $z_n^1 = u_n - u_n^1$ the following equalities hold:

$$\begin{aligned} u_n - u_n^1 &= (I - A_n^{-1}\tilde{A}_n)^2 \tilde{A}_n^{-1}\tilde{f}_n + A_n^{-1}(\tilde{A}_n A_n^{-1} - I)(f_{nn} - \tilde{f}_n) \\ &= A_n^{-1}(A_n - \tilde{A}_n)(u_n - A_n^{-1}\tilde{f}_n) + A_n^{-1}(\tilde{A}_n A_n^{-1} - I)(f_{nn} - \tilde{f}_n) \\ &= A_n^{-1}(A_n - \tilde{A}_n)(u_n - u_n^0) = A_n^{-1}(A_n - \tilde{A}_n)(z_{nn} - z_n). \end{aligned}$$

Thus, if z_{nn} and z_n are the nodal restrictions of the smooth functions of at least the second order and A_n, $\tilde{A}_n$ approximate the differential operator with the same order, then the considered iteration u_n^1 approximates the serendipity solution with the error $O(h^4)$ on the uniform mesh, i.e. with the optimal order of the discretization error.

For a non-uniform mesh, one can still get the improvement from $O(h^2)$ for the bilinear approximation up to $O(h^3)$ for the serendipity approximation.

The cost of computing the defect correction approximated solution is only two solutions with the bilinear matrix. This solution is normally obtained by iteration, and the stopping criterion in computing the solution and the correction should be absolute, i.e. of order $O(h^4)\|f\|$.

The number of multiplications in the modified defect correction method (MDC) is given by the formula

$$M_{\mathrm{MDC}} = (22k + 39)N^2$$

where k is the total number of IMIF iterations at the first and the second steps of the method and $39N^2$ is the cost of the computation of the residual r_n^0.

3. RESULTS OF NUMERICAL EXPERIMENTS

In this section, we present the results of the numerical solution of the serendipity algebraic system for the model boundary value problem

$$-\frac{\partial^2 u}{\partial x^2} - \sigma\frac{\partial^2 u}{\partial y^2} = f(x,y), \quad x,y \in \Omega = (0,2)^2, \quad u|_\Gamma = g(x,y)$$

where $f = 0$ and the function g corresponds to the exact solution

$$u(x,y) = x^4 - 6x^2y^2/\sigma + y^4/\sigma^2. \tag{3.1}$$

The computations were made by the modified code FVSDE2 [10] on square grids with the numbers of meshsteps $N \times N$, $N = 16,\ 32,\ 64,\ 128$. In all the experiments the stopping criterion (tolerance) of iterations was

$$\|r^k\|_2 / \|r^0\|_2 \le \varepsilon \tag{3.2}$$

with the same initial values $u^0 = 0$.

First we present in the Table 1 the number of iterations k and the resulting error in maximum vector norm $\delta = \|u - u^h\|_\infty$ for the solution of system (1.3) by the EXIF method with $\varepsilon = 10^{-9}$, $\omega = \vartheta = \sigma = 1$ (the isotropic case). We can see that the error of the serendipity solution has the fourth order, $\delta = O(h^4)$, and that the number of iterations N is almost proportional to $\sqrt{N}$. The exceptional case here is $N = 128$, where the iterative tolerance $\varepsilon = 10^{-9}$ is close to the error. For this dimension and, for example, for $\varepsilon = 10^{-11}$ we get $k = 72$ and $\delta = 1.56 \times 10^{-8}$.

The numbers of iterations depend on the values of relaxation and compensation parameters ω, ϑ, which is demonstrated for the isotropic case in Tables 2, 3 for grids 16×16 and 64×64, respectively. The results given in Tables 2, 3 show the existence of optimal parameters $\omega_0 > 1$ (or $\omega_0 < 1$), $\vartheta_0 < 1$. But it is advisable to choose $\omega = \vartheta = 1$ and to skip this optimization problem since the difference in the numbers of iterations is about 25% only.

It should be noted that the situation changes for $\sigma \ne 1$ when the EXIF solver diverges for some values of the compensation parameter $\vartheta > 0$. It can be explained by the fact that the convergence of this algorithm is proved for $\vartheta = 0$, which corresponds to the SSOR-CG method. Table 4 gives such examples for $\sigma = 3 \times 64$. Here '$*$' represents the singular behaviour of the iterative process when either G_A is nonpositive or the process diverges.

Similar results on the accuracy and the convergence rate of iterations for different σ are given in Table 5 for the iterative solution of the reduced system of equations (1.8) for nodal unknowns by the preconditioned CG method (2.1), (2.9), i.e. the IMIF procedure with the compensation parameter $\vartheta = 1$. In each square of the table the number of iterations k and the error δ are presented for $\varepsilon = 10^{-11}$ by the top and the bottom values, respectively. One can see that the IMIF method has a good enough convergence rate even for the non-M-matrix.

Table 1.
EXIF method, $\varepsilon = 10^{-9}$, $\omega = \vartheta = 1$.

N	16	32	64	128
k	19	27	40	58
δ	6.37×10^{-5}	3.98×10^{-6}	2.51×10^{-7}	2.33×10^{-8}

Table 2.
EXIF method, the numbers of iterations for grid 16×16, $\varepsilon = 10^{-9}$.

$\omega \setminus \vartheta$	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.2	40	39	39	38	37	36	34	32	28	23	19
.4	34	33	32	31	30	28	26	24	21	17	19
.6	28	27	26	25	24	23	21	20	18	15	19
.8	23	23	22	21	20	19	18	17	16	15	19
1.0	20	19	18	18	17	16	16	15	15	15	19
1.2	17	16	16	16	15	15	15	15	15	16	19
1.4	17	16	16	16	15	15	15	16	16	16	19
1.6	19	19	18	18	17	17	17	17	17	18	19
1.8	26	24	23	22	22	21	20	20	19	19	19
2.0	34	32	31	28	27	25	24	22	21	20	19

Table 3.
EXIF method, the numbers of iterations for grid 64×64, $\varepsilon = 10^{-9}$.

$\omega \setminus \vartheta$	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.2	149	147	143	140	135	130	122	112	99	80	40
.4	122	118	114	110	105	99	93	85	76	62	40
.6	100	97	94	90	86	81	76	71	63	51	40
.8	84	81	78	75	72	69	65	60	53	45	40
1.0	69	67	65	63	61	58	55	51	47	41	40
1.2	58	56	54	52	51	49	47	45	42	37	40
1.4	48	47	46	44	43	42	41	39	37	33	40
1.6	39	38	37	36	36	35	34	33	32	30	40
1.8	33	32	31	31	31	30	30	30	30	31	40
2.0	78	74	69	65	61	59	57	55	53	50	40

Table 4.
EXIF method, the numbers of iterations for grid 64×64, $\varepsilon = 10^{-9}$, $\sigma = 3 \times 64$.

$\omega \setminus \vartheta$	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.2	319	308	296	283	267	249	228	204	175	184	*
.4	226	217	208	198	187	174	161	155	183	*	*
.6	174	168	161	154	146	143	150	184	*	*	*
.8	140	135	131	127	130	142	178	*	*	*	*
1.0	114	112	110	114	129	170	*	*	*	*	*
1.2	94	93	95	109	147	*	*	*	*	*	*
1.4	76	76	83	113	*	*	*	*	*	*	*
1.6	57	60	71	166	*	*	*	*	*	*	*
1.8	63	74	105	*	*	*	*	*	*	*	*
2.0	228	198	*	*	*	*	*	*	*	*	*

Table 5.
Reduced system (1.8), IMIF method (2.1), (2.9), $\vartheta = 1, \varepsilon = 10^{-11}$.

$\sigma \setminus N$	16	32	64	128
3/64	31	39	52	71
	1.87×10^{-3}	1.25×10^{-4}	8.10×10^{-6}	5.16×10^{-7}
3/8	24	32	44	62
	1.69×10^{-4}	1.06×10^{-5}	6.61×10^{-7}	4.13×10^{-8}
1	19	26	37	53
	6.37×10^{-5}	3.98×10^{-6}	2.48×10^{-7}	1.56×10^{-8}
3	23	30	43	61
	2.11×10^{-5}	1.32×10^{-6}	8.26×10^{-7}	5.16×10^{-8}
3×8	28	32	42	61
	3.89×10^{-6}	2.63×10^{-7}	1.71×10^{-8}	1.09×10^{-9}
3×64	29	30	32	41
	2.20×10^{-6}	1.49×10^{-7}	1.09×10^{-8}	7.41×10^{-10}

Table 6.
Bilinear FEM, IMIF method (2.1), (2.9), $\varepsilon = 10^{-9}, 10^{-11}, \sigma = \vartheta = 1$.

N	16	32	64	128
	11	16	22	31
k	13	18	26	37
	1.85×10^{-2}	4.61×10^{-3}	1.15×10^{-3}	2.88×10^{-4}
δ	1.85×10^{-2}	4.61×10^{-3}	1.15×10^{-3}	2.88×10^{-4}

Table 7.
The numbers of outer and inner iterations, two-level method (2.1), (2.18).

	$\varepsilon_k \setminus N$	16	32	64	128
	10^{-7}	13(49)	15(69)	18(98)	20(136)
$\sigma = 1$	10^{-9}	12(68)	12(101)	12(144)	12(206)
	10^{-11}	11(96)	11(137)	11(196)	11(280)
	10^{-7}	22(41)	24(50)	25(71)	28(106)
$\sigma = 3 \times 64$	10^{-9}	24(58)	24(67)	24(97)	25(153)
	10^{-11}	24(75)	24(92)	24(137)	24(214)
	10^{-7}	23(69)	25(91)	28(121)	32(166)
$\sigma = 3/64$	10^{-9}	22(99)	23(138)	24(190)	23(259)
	10^{-11}	22(138)	22(196)	22(274)	22(377)

Table 8.
Error at two steps of the MDC method (2.23), (2.24), $\varepsilon = 10^{-11}$, exact solution (3.1).

	N	16	32	64	128
	k_1	13	18	26	37
$\sigma = 1$	k_2	9	12	15	19
	δ_1	1.85×10^{-2}	4.61×10^{-3}	1.15×10^{-3}	2.88×10^{-4}
	δ_2	2.84×10^{-5}	2.18×10^{-6}	1.50×10^{-7}	9.80×10^{-9}
	k_1	4	5	7	11
$\sigma = 3 \times 64$	k_2	3	3	3	4
	δ_1	8.41×10^{-5}	2.05×10^{-5}	5.11×10^{-6}	1.28×10^{-6}
	δ_2	2.69×10^{-5}	2.29×10^{-6}	1.51×10^{-7}	1.04×10^{-8}
	k_1	10	14	20	27
$\sigma = 3/64$	k_2	6	8	9	10
	δ_1	0.349	8.71×10^{-2}	2.18×10^{-2}	5.44×10^{-3}
	δ_2	8.75×10^{-3}	1.09×10^{-3}	6.84×10^{-5}	4.32×10^{-6}

Table 9.
Error at two steps of the MDC method (2.23), (2.24),
$\varepsilon = 10^{-11}$, exact solution (3.3).

N	16	32	64	128
	13	19	27	38
k	10	14	18	23
	1.85×10^{-1}	4.6×10^{-2}	1.2×10^{-2}	2.9×10^{-3}
δ	1.5×10^{-3}	1.2×10^{-4}	8.2×10^{-6}	5.5×10^{-7}

The efficiency of the approximation and convergence rates can be compared with the data in Table 6 for the nodal system of equations (2.12) obtained by the bilinear FEM for the isotropic case ($\sigma = 1$), by the same preconditioned IMIF-CG algorithm (2.1), (2.9). In each square the number of iterations k and the error δ are presented for $\varepsilon = 10^{-9},\ 10^{-11}$ by the top and the bottom values, respectively. We see that $\varepsilon = 10^{-9}$ would suffice for the iterations with the given convergence rate which is less than $O(h^4)$.

In Table 7, for $\sigma = 1,\ 3 \times 64, 3/64$ we present the numbers of outer iterations k and the total numbers (in brackets) of inner iterations $\sum_{k'=1}^{k} m_{k'}$ for the two-level CG method (2.1), (2.18), i.e. the accelerated 'classical' defect correction method, for the outer tolerance $\varepsilon = 10^{-9}$ and different inner tolerances $\varepsilon_k = 10^{-7}, 10^{-9}, 10^{-11}$. The resulting accuracy δ here is approximately the same as in Table 5.

As for the defect correction method without CG in outer iterations (2.8), which corresponds to (2.1) with $\alpha_k = 1,\ \beta_k = 0$, it gives the worst result: even on the grid 16×16, for example, it gives 22 outer iterations with 100 inner ones for $\varepsilon = \varepsilon_k = 10^{-9}$.

Tables 8, 9 give the numbers k_1, k_2 of iterations by the IMIF method both at the first and the second steps, respectively, and accuracy δ_1, δ_2 in the maximum norm for the modified defect correction method (2.23), (2.24). Table 8 corresponds to the exact solution (3.1) for three different values $\sigma = 1, 3/64, 3 \times 64$. Note that the stopping criterion at the second step, i.e. the computation of u_n^1 with u_n^0 as an initial guess had the form (3.2) but the initial residual r^0 was taken from the first step of the method. The data in Table 8 shows that this method is good enough for different anisotropy coefficients.

Similar results for the solution of the Laplace equation with the exact solution

$$u(x,y) = x^6 - 15x^4y^2 + 15x^2y^4 - y^6 \tag{3.3}$$

are given in Table 9.

An examination of the last rows of Tables 8, 9 shows that the convergence rate is $O(h^4)$ since the error decrease rate is approximately 16 when the grid dimension doubles. In the last experiments it turned out that $\varepsilon = 10^{-9}$ was not sufficient to yield the $O(h^4)$ convergence when passing from $N = 64$ to $N = 128$ and therefore $\varepsilon = 10^{-11}$ was used.

4. CONCLUSION

The comparative analysis of the numerical experiments allows us to make the following conclusions about the computational complexity of the considered algorithms.

(a) The resulting accuracy of the serendipity FEM for the considered test problem has order $O(h^4)$ for the uniform grid, which gives a great advantage over the bilinear FEM of the second order. For example, for the test problems the serendipity approximation has higher accuracy for the grid 16×16 than the bilinear FEM for the grid 128×128.

(b) The direct application of the explicit and implicit incomplete factorization methods is sufficiently efficient in the sense that the number of iterations for the 'serendipity' system is only slightly larger than that for the 'bilinear' system. In both cases the number of iterations is proportional to $\sqrt{N}$ in accordance with the theoretical estimate for the EXIF and IMIF methods applied to the 'bilinear' nodal system. In the anisotropic case the IMIF method reveals a good convergence even for the non-M-matrix. But the divergence of the EXIF method for some compensation parameter $\vartheta > 0$ (EXIF convergences are reliable in its SSOR variant only) shows the necessity of further investigations along this line.

(c) Since the accuracy of the serendipity system is high [$O(h^3)$ or even $O(h^4)$] for a uniform grid, in practice it is sufficient to use a very coarse mesh. Then there is a little need for a multilevel solver (AMLI or multigrid) because the incomplete factorization method is about as fast as or even faster than multilevel methods when solving sufficiently coarse systems. Also, it is simpler to implement when there is no need for several meshes.

(d) The experiments for the MDC method for anisotropic problems show a good convergence in terms of both the number of iterations and accuracy for a wide class of problems when the anisotropy coefficient σ is varied.

(e) The final comparison of the four considered algorithms for the model problem is presented in Table 10. Here the values of the coefficient c_m from the formula $M_m = c_m N^2$ for the total numbers of multiplications are given for different grids and methods, all with the same iterative stopping criterion. The coefficients c_m are calculated from the formulae (2.4), (2.11), (2.19) and from the numbers of iterations presented in Tables 1, 4, 6–8 with $\varepsilon = \varepsilon_k = 10^{-9}$.

From Table 10 it is evident that in the considered examples the modified defect correction method has a significant advantage over the other algorithms under investigation.

An alternative to the modified defect correction is to use a Richardson h^2-extrapolation of bilinear element approximations for meshes Ω_h and $\Omega_{h/2}$. However, this method is efficient only for regular meshes and smooth solutions. On the other hand, the serendipity element approximation also gives accurate solutions for less regular solutions.

This investigation on the serendipity FEM gives useful information even for the model BVP on square mesh. Of course, the extension of these problems to more general PDEs, 3D geometry, and nonuniform grids is of special practical interest. For such problems the matrix A_e does not take such a simple form as for the case considered here. However, it turns out that systems with A_e can still be solved efficiently, using an inner iteration method.

Table 10.
The numbers of multiplications per node for different grids and methods for $\varepsilon = 10^{-9}$, exact solution (3.1), σ=1.

	16×16	32×32	64×64	128×128
EXIF	1767	2511	3720	5394
IMIF	896	1232	1680	2552
2-level	2530	3154	4272	5884
MDC	435	589	765	1007

REFERENCES

1. O. Axelsson, On multigrid methods of two-level type. In: *Multigrid Methods* (Eds. W. Hackbush and U. Trottenberg). LNLM, Springer-Verlag, Berlin–New York, 1982, Vol. 960, pp. 352–367.
2. O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems. Theory and Computations.* Academic Press, New York, 1984.
3. O. Axelsson and P. S. Vassilevski, Algebraic multilevel preconditioning methods, I. *Numer Math.* (1989) **56**, 157–177.

4. O. Axelsson and P. S. Vassilevski, Variable-step multilevel preconditioning methods, I. Self-adjoint and positive definite elliptic problems. *Numer. Linear. Alg. Appl*. (1994), No. 1, 75–101.
5. O. Axelsson, Yu. R. Hakopian, and Yu. A. Kuznetsov, Multilevel preconditioning for perturbed finite element matrices. IMA *J. Numer. Anal*. (1997) **17**, No. 1, 125–149.
6. O. Axelsson and M. Larin, An algebraic multilevel iteration method for finite element matrices. *J. Comp. Math.* (1997) **89**, 135–153.
7. O. Axelsson, *Iterative Solution Methods.* Cambridge University Press, New York, 1994.
8. Y. L. Gurieva and V. P. Il'in, Finite volume approaches for 2-D BVP's: algorithms, data structures, software and experiments. *Report No. 9715*, Dep. of Math., University of Nijmegen, The Netherlands, 1997.
9. V. P. Il'in, *Iterative Incomplete Factorization Methods*. World. Sci. Publ., Singapore, 1993.
10. V. P. Il'in, On approach to improvement of difference solutions. *Sib. J. Comp. Math.* (1992) **1**, No. 3, 205–214.
11. V. P. Il'in, On estimates of the convergence rate of iterative incomplete factorization methods. *Russ. J. Numer. Anal. Math. Modelling* (1999) **14**, No. 2, 125–136.