Preconditioned iterative methods for the nine-point approximation to the convection–diffusion equation

M. Bhuruth*, M.K. Jain, A. Gopaul 1

Department of Mathematics, Faculty of Science, University of Mauritius, Reduit, Mauritius

Received 3 April 2000

Abstract

Iterative methods preconditioned by incomplete factorizations and sparse approximate inverses are considered for solving linear systems arising from fourth-order finite difference schemes for convection–diffusion problems. Simple recurrences for implementing the ILU(0) factorization of the nine-point scheme are derived. Different sparsity patterns are considered for computing approximate inverses for the coefficient matrix and the quality of the preconditioner is studied in terms of plots of the field of values of the preconditioned matrices. In terms of algebraic properties of the preconditioned matrices, our experimental results show that incomplete factorizations give a preconditioner of better quality than approximate inverses. Comparison of the convergence rates of GMRES applied to the preconditioned linear systems is done with respect to the field of values, Ritz and harmonic Ritz values of the preconditioned matrices. Numerical results show that the GMRES residual norm decreases rapidly when the difference between the Ritz and harmonic Ritz values becomes small. We also describe the results of experiments when some well-known Krylov subspace methods are used to solve the linear system arising from the compact fourth-order discretizations. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Convection–diffusion equation; Fourth-order scheme; Sparse approximate inverse; Field of values; Harmonic Ritz values

1. Introduction

We consider the two-dimensional convection–diffusion equation

\[-\Delta u + p(x, y)u_x + q(x, y)u_y + s(x, y)u = w\]

(1)
on the unit square \(\Omega = \{(x, y): 0 \leq x, y \leq 1\}\) with Dirichlet boundary conditions \(u(x, y) = v(x, y)\) on \(\partial\Omega\). Large sparse linear systems arising from the central difference discretizations of (1) have

* Corresponding author.
E-mail address: mbhuruth@uom.ac.mu (M. Bhuruth).

1 The work of the third author was supported by a postgraduate research scholarship from the University of Mauritius.
often been used as test problems for studying the efficiency of iterative schemes. Let $h$ denote the mesh size of a uniform grid on $\Omega$ and let $\mathcal{R} = \mathcal{P}h/2$ denote the cell Reynolds number where

$$\mathcal{P} = \max \left( \sup_{(x,y) \in \Omega} |p(x,y)|, \sup_{(x,y) \in \Omega} |q(x,y)| \right).$$

In many practical problems, the cell Reynolds number is high, i.e., the convective terms dominate the diffusion term. In such cases, difficulties are encountered when iterative solution methods are used for solving the linear system arising from central difference approximations of (1). For example, relaxation methods do not converge when the value of $\mathcal{R}$ exceeds a certain constant.

Recently, there has been much interest in studying iterative solution methods for linear systems arising from higher order compact finite difference schemes [2,22,28]. Fourth-order schemes for elliptic equations with nonlinear first derivative terms of the form

$$-\Delta u = F(x,y,u,\partial_x u, \partial_y u)$$

in a bounded two-dimensional region with Dirichlet boundary conditions have been described in [16]. These schemes at a given mesh point involve the nearest eight neighbouring mesh points. For the linear problem (1), the fourth-order scheme derived using this approach is mathematically equivalent to that given in [17]. For high cell Reynolds numbers, the discrete solution given by the central difference scheme oscillates whereas the fourth-order scheme produces oscillation-free solutions [16]. The accuracy of the numerical solution of convection–diffusion problems using central differences, upwind scheme and fourth-order scheme is compared in [28] and the results show that for small to moderate cell Reynolds numbers, the discrete solution obtained by the fourth-order scheme is of much higher accuracy. In this paper, we are mainly interested in the quality of sparse approximate inverse and ILU(0) preconditioners that can be obtained from the fourth-order approximations of convection–diffusion equations of the form (1). We also study the convergence of preconditioned iterations in relation to the field of values, Ritz values and harmonic Ritz values of the preconditioned matrices.

Let $n$ be the number of uniformly spaced gridpoints in each direction and let $h = 1/(n+1)$ be the grid size. For the constant coefficient convection–diffusion equation with $s \equiv 0$, $p(x,y) \equiv p$ and $q(x,y) \equiv q$, where $p$ and $q$ are positive constants, the fourth-order difference scheme has the computational stencil given by

$$
\begin{align*}
-(1 + \gamma)(1 - \delta) & \quad -(4 - 4\delta + 2\delta^2) & \quad -(1 - \gamma)(1 - \delta), \\
-(4 + 4\gamma + 2\gamma^2) & \quad 20 + 4\gamma^2 + 4\delta^2 & \quad -(4 - 4\gamma + 2\gamma^2), \\
-(1 + \gamma)(1 + \delta) & \quad -(4 + 4\delta + 2\delta^2) & \quad -(1 - \gamma)(1 + \delta),
\end{align*}
$$

(2)

where $\gamma = ph/2$ and $\delta = qh/2$ are the cell Reynolds numbers. Let

$$Au_h = w_h$$

(3)

be the resulting linear system from the stencil (2). Zhang [27] has shown that the coefficient matrix $A$ is irreducibly diagonally dominant if $|\gamma| \leq 1$ and $|\delta| \leq 1$. Block iterative methods for solving the linear system arising from the fourth-order discretization of the constant coefficient problem have been considered in [2]. Numerical experiments have shown that block relaxation schemes are convergent for large values of the cell Reynolds number. It has also been shown in [25] that the
point Jacobi method for solving (3) converges for any initial guess. This unconditional convergence property avoids the use of defect-correction techniques for achieving high accuracy with multigrid algorithms [26].

Preconditioning large linear systems using sparse approximate inverses $M \approx A^{-1}$ are currently of much interest. In this paper, we consider preconditioning the matrices arising from fourth-order discretizations using incomplete LU factorizations [18] and sparse approximate inverses [6]. An outline of this paper is as follows. In Section 2, we consider incomplete factorizations and for the matrix $A$ corresponding to the nine-point difference operator, we show how the ILU(0)-preconditioner can be implemented using simple recurrences. We consider the computation of sparse approximate inverses for the coefficient matrix in Section 3 and in Section 4, we consider the convergence of GMRES in relation to the field of values, Ritz and harmonic Ritz values of the preconditioned matrices. In Section 5 we describe some numerical experiments and compare the performance of the different preconditioned Krylov subspace algorithms for solving the linear systems arising from the fourth-order discretizations.

2. Incomplete LU preconditioning

The existence of incomplete LU factorization preconditioners for $M$-matrices was first proved by Meijerink and Van der Vorst [18]. For matrices with indefinite symmetric parts, problems such as inaccuracy due to very small pivots and unstable triangular solves may occur [5]. More accurate variants of ILU to improve the efficiency of the ILU preconditioner with zero level of fill-in (ILU(0)) have been proposed by Saad [21].

2.1. Fourth-order scheme and ILU(0) preconditioning

On a square mesh with meshsize $h = 1/(n + 1)$, we denote by $u_{i,j}$ the approximate value of a function $u$ at the grid point $(x_i, y_j)$. Let

$$A_{ij} = \frac{p(x_i, y_j)h}{2}, \quad \Psi_{ij} = \frac{q(x_i, y_j)h}{2} \quad \text{and} \quad \Pi_{ij} = \frac{s(x_i, y_j)h^2}{2}.$$ 

Then, the compact fourth-order scheme for (1) can be written in the form

$$\begin{bmatrix}
\chi_{i-1,j+1} - 1 & \chi_{i,j+1} - 4 & \chi_{i+1,j+1} - 1 \\
\chi_{i-1,j} - 4 & (20 + \chi_{i,j}) & \chi_{i+1,j} - 4 \\
\chi_{i-1,j-1} - 1 & \chi_{i,j-1} - 4 & \chi_{i+1,j-1} - 1
\end{bmatrix} u_{i,j} = \frac{h^2}{2} \begin{bmatrix}
1 - \Psi_{i,j} \\
1 + A_{i,j} & 8 & 1 - A_{i,j} \\
1 + \Psi_{i,j}
\end{bmatrix} w_{i,j},$$

where

$$\chi_{i+1,j} = 4A_{i,j} + \frac{1}{2}(3A_{i+1,j} - 2A_{i,j} - A_{i-1,j}) + \Pi_{i+1,j}(1 - hA_{i,j}) - \frac{1}{2}A_{i,j}(3A_{i+1,j} + A_{i-1,j}),$$

$$\chi_{i-1,j} = -4A_{i,j} + \frac{1}{2}(A_{i+1,j} + 2A_{i,j} - 3A_{i-1,j}) + \Pi_{i-1,j}(1 + hA_{i,j}) - \frac{1}{2}A_{i,j}(A_{i+1,j} + 3A_{i-1,j}),$$

$$\chi_{i,j+1} = 4\Psi_{i,j} + \frac{1}{2}(3\Psi_{i,j+1} - 2\Psi_{i,j} - \Psi_{i,j-1}) + \Pi_{i,j+1}(1 - h\Psi_{i,j}) - \frac{1}{2}\Psi_{i,j}(3\Psi_{i,j+1} + \Psi_{i,j-1}),$$

$$\chi_{i,j-1} = -4\Psi_{i,j} + \frac{1}{2}(\Psi_{i,j+1} + 2\Psi_{i,j} - 3\Psi_{i,j-1}) + \Pi_{i,j-1}(1 + h\Psi_{i,j}) - \frac{1}{2}\Psi_{i,j}(\Psi_{i,j+1} + 3\Psi_{i,j-1}),$$
\[
\chi_{i,j} = 8\Pi_{i,j} - 2(A_{i+1,j} - A_{i-1,j}) - 2(\Psi_{i,j+1} - \Psi_{i,j-1})
+ 2\Psi_{i,j}(\Psi_{i,j+1} + \Psi_{i,j-1}) + 2A_{i,j}(A_{i+1,j} + A_{i-1,j}),
\]
\[
\chi_{i+1,j+1} = \frac{1}{2}(A_{i,j} + A_{i,j+1}) + \frac{1}{2}(\Psi_{i,j} + \Psi_{i+1,j}) - \frac{1}{2}(A_{i,j}\Psi_{i+1,j} + \Psi_{i,j}A_{i,j+1}),
\]
\[
\chi_{i+1,j-1} = \frac{1}{2}(A_{i,j} + A_{i,j-1}) - \frac{1}{2}(\Psi_{i,j} + \Psi_{i+1,j}) + \frac{1}{2}(A_{i,j}\Psi_{i+1,j} + \Psi_{i,j}A_{i,j-1}),
\]
\[
\chi_{i-1,j+1} = -\frac{1}{2}(A_{i,j} + A_{i,j+1}) + \frac{1}{2}(\Psi_{i,j} + \Psi_{i-1,j}) + \frac{1}{2}(A_{i,j}\Psi_{i-1,j} + \Psi_{i,j}A_{i,j+1}),
\]
\[
\chi_{i-1,j-1} = -\frac{1}{2}(A_{i,j} + A_{i,j-1}) - \frac{1}{2}(\Psi_{i,j} + \Psi_{i-1,j}) - \frac{1}{2}(A_{i,j}\Psi_{i-1,j} + \Psi_{i,j}A_{i,j-1}).
\]

We write the nine-point difference stencil for (1) in the form
\[
\begin{pmatrix}
g_{ij} & h_{ij} & l_{ij} \\
a_{ij} & b_{ij} & c_{ij} \\
d_{ij} & e_{ij} & f_{ij}
\end{pmatrix}.
\]

The entries in the stencil are the non-zero elements on the row of the matrix \( A \) corresponding to the \((i,j)\)th grid point. We consider an incomplete \( LD^{-1}U \) factorization of \( A \) [4], where \( L \) is lower triangular, \( D \) is diagonal and \( U \) is upper triangular. Since the ILU(0) factorization allows no fill-in in the nonzero positions of \( A \), the difference stencils for the \( L \), \( D^{-1} \) and \( U \) factors are given by
\[
\begin{pmatrix}
r_{ij} & \beta_{ij} \\
\gamma_{ij} & \eta_{ij} & \xi_{ij}
\end{pmatrix}, \quad \begin{pmatrix}
\beta^{-1}_{ij} \\
\delta_{ij}
\end{pmatrix}, \quad \begin{pmatrix}
v_{ij} & \theta_{ij} & \psi_{ij}
\end{pmatrix}.
\]

The stencil corresponding to the matrix product \( LD^{-1}U \) is given by
\[
\text{stencil}_{ij}(LD^{-1}U) = \gamma_{ij}\beta^{-1}_{i-1,j-1} \times \text{stencil}_{i-1,j-1}(U) + \eta_{ij}\beta^{-1}_{i,j-1} \times \text{stencil}_{i,j-1}(U)
+ \xi_{ij}\beta^{-1}_{i+1,j-1} \times \text{stencil}_{i+1,j-1}(U) + r_{ij}\beta^{-1}_{i-1,j} \times \text{stencil}_{i-1,j}(U)
+ \text{stencil}_{ij}(U).
\]

The computational molecule corresponding to the matrix product \( LD^{-1}U \) with no fill-in is given by
\[
\begin{pmatrix}
g_{ij} & h_{ij} & l_{ij} \\
a_{ij} & b_{ij} & c_{ij} \\
d_{ij} & e_{ij} & f_{ij}
\end{pmatrix},
\]

where
\[
g_{ij} = v_{ij} + r_{ij}\theta_{i-1,j}\beta^{-1}_{i-1,j},
\]
\[
h_{ij} = \theta_{ij} + r_{ij}\psi_{i-1,j}\beta^{-1}_{i-1,j},
\]
\[
\tilde{l}_{ij} = \psi_{ij},
\]
\[
\tilde{d}_{ij} = \gamma_{ij},
\]
\[ \tilde{e}_{ij} = \eta_{ij} + \gamma_{ij} \delta_{i-1,j-1} \beta_{i-1,j-1}^{-1}, \]
\[ \tilde{f}_{ij} = \zeta_{ij} + \eta_{ij} \delta_{i,j-1} \beta_{i,j-1}^{-1}, \]
\[ \tilde{a}_{ij} = \rho_{ij} + \eta_{ij} v_{i,j-1} \beta_{i,j-1}^{-1} + \gamma_{ij} \theta_{i-1,j-1} \beta_{i-1,j-1}^{-1}, \]
\[ \tilde{b}_{ij} = \beta_{ij} + \gamma_{ij} \psi_{i-1,j-1} \beta_{i-1,j-1}^{-1} + \eta_{ij} \theta_{i,j-1} \beta_{i,j-1}^{-1} + \zeta_{ij} v_{i+1,j-1} \beta_{i+1,j-1}^{-1} + \rho_{ij} \delta_{i-1,j} \beta_{i-1,j}^{-1}, \]
\[ \tilde{c}_{ij} = \delta_{ij} + \eta_{ij} \psi_{i-1,j} \beta_{i,j-1}^{-1} + \zeta_{ij} \theta_{i+1,j-1} \beta_{i+1,j-1}^{-1}. \]

Comparing the stencils (5) and (7), we find that \( \psi_{ij} = l_{ij} \) and \( \gamma_{ij} = d_{ij} \). The remaining entries of the factors \( L, D^{-1}, U \) can be computed as follows:
\[
\rho_{ij} = a_{ij} - \eta_{ij} v_{i,j-1} \beta_{i,j-1}^{-1} - \gamma_{ij} \theta_{i-1,j-1} \beta_{i-1,j-1}^{-1},
\]
\[
v_{ij} = g_{ij} - \rho_{ij} \theta_{i-1,j} \beta_{i-1,j}^{-1},
\]
\[
\theta_{ij} = h_{ij} - \rho_{ij} \psi_{i-1,j} \beta_{i-1,j}^{-1},
\]
\[
\eta_{ij} = e_{ij} - \gamma_{ij} \delta_{i-1,j-1} \beta_{i-1,j-1}^{-1},
\]
\[
\zeta_{ij} = f_{ij} - \eta_{ij} \delta_{i,j-1} \beta_{i,j-1}^{-1},
\]
\[
\delta_{ij} = c_{ij} - \eta_{ij} \psi_{i-1,j} \beta_{i-1,j}^{-1} - \zeta_{ij} \theta_{i+1,j-1} \beta_{i+1,j-1}^{-1}
\]
and finally, the entry \( \beta_{ij} \) is given by
\[
\beta_{ij} = b_{ij} - \gamma_{ij} \psi_{i-1,j-1} \beta_{i-1,j-1}^{-1} - \eta_{ij} \theta_{i,j-1} \beta_{i,j-1}^{-1} - \zeta_{ij} v_{i+1,j-1} \beta_{i+1,j-1}^{-1} - \rho_{ij} \delta_{i-1,j} \beta_{i-1,j}^{-1}. \tag{8}
\]

The cost of computing the ILU(0) factorization is \( 16n^2 \). For the constant coefficient problem, the coefficient matrix is an \( M \)-matrix when \( |\gamma| \leq 1 \) and \( |\delta| \leq 1 \) and thus the incomplete factorization exists in this case. For more general problems, numerical experiments indicate that no particular problem occurs during the computation of the incomplete factorization.

3. Sparse approximate inverse preconditioning

Preconditioning using sparse approximate inverses has attracted much interest since the preconditioning operation can be achieved by a matrix–vector multiplication. Currently, there exists several algorithms for constructing sparse approximate inverse preconditioners. One often used approach for construction of an approximate inverse is based on minimization of the Frobenius norm. A sparsity pattern \( S \) of the approximate inverse is prescribed before performing the minimization and adaptive strategies are used to update the sparsity pattern as the algorithm proceeds to compute an approximate inverse.
3.1. Overview of Frobenius norm minimization techniques

Let \( A \) be an \( N \times N \) large, sparse, nonsymmetic and nonsingular matrix. Methods based on the Frobenius norm to construct an approximate inverse \( M \) of \( A \) minimize the quantity

\[
||AM - I||_F^2 = \sum_{k=1}^{N} ||(AM - I)e_k||_2^2,
\]

(9)

where \( e_k \) is the \( k \)th unit vector of \( \mathbb{R}^N \). The minimization is carried out among all such matrices \( M \) that satisfy a given sparsity pattern. Denoting by \( m_k \) the \( k \)th column of the matrix \( M \), (9) decouples into \( N \) independent least-squares problems

\[
\min_{m_k} ||Am_k - e_k||_2, \quad k = 1, 2, \ldots, N.
\]

(10)

Solution of the least-squares problem can be easily done. The main problem is the choice beforehand of a good sparsity pattern for entries of the matrix \( A^{-1} \). Techniques to augment the sparsity pattern were first given by Cosgrove et al. [6]. They showed that by filling in some location on a column \( m_k \), the residual \( ||Am_k - e_k||_2 \) may be reduced. Other ways for augmenting the sparsity pattern have been proposed by Grote and Huckle [10] and Gould and Scott [8]. In the following we briefly outline the SPAI algorithm of Grote and Huckle [10].

Consider the minimization of \( ||Am_k - e_k||_2 \). We let \( J \) denote the set of indices in \( m_k \) with nonzero entries and let \( I \) denote the shadow of \( J \) in \( A \), that is, the set of the indices of the nonzero rows in the submatrix \( A(:,J) \). Let \( \hat{m}_k = m_k(J) \), \( \hat{A} = A(I,J) \) and \( \hat{e}_k = e_k(I) \). Then,

\[
\min_{\hat{m}_k} ||\hat{A}\hat{m}_k - \hat{e}_k||_2 = \min_{m_k} ||A m_k - e_k||_2.
\]

(11)

The least squares problem in (11) is solved by finding the \( QR \) decomposition of \( \hat{A} \) in the form

\[
\hat{A} = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = (Y Z) \begin{pmatrix} R \\ 0 \end{pmatrix}.
\]

(12)

The solution to the least squares problem is then given by [8]

\[
\hat{m}_k = R^{-1} Y^T \hat{e}_k
\]

(13)

and the residual norm is

\[
||r_k||_2 = ||A(:,J)\hat{m}_k - \hat{e}_k||_2.
\]

(14)

The next step in the SPAI algorithm is to improve on the current solution \( \hat{m}_k \). This is done by dynamically choosing new entry positions in \( m_k \) and solving the problem for the enlarged set \( J \cup \hat{J} \), where \( \hat{J} \) is the set of profitable indices which will further decrease the residual norm. For one possible new index \( j \), Grote and Huckle suggest the use of the one-dimensional minimization problem

\[
\min_{j} ||A(m_k + \lambda e_j) - e_k||_2
\]

(15)

to determine the optimal \( j \). The solution to the minimization problem, \( \rho_j \), is such that

\[
\rho_j^2 = ||r_k||_2 - \frac{(r_k^T A e_j)^2}{||A e_j||_2^2}.
\]

(16)
For the SPAI algorithm, the index set \( \tilde{J} \) may contain one or more new indices. These indices are chosen if \( \rho_j \) is less than the mean of all the \( \rho_j \)'s. Another criteria for introducing a new index \( j \) makes use of the minimization problem

\[
\min_{m_k(J \cup \{j\})} \|A(:,J \cup \{j\})m_k(J \cup \{j\}) - \hat{e}_k\|_2. \tag{17}
\]

The solution to the minimization problem (17) is given by [8]

\[
\sigma_j = \frac{\|r_k\|_2^2}{\|Y^T \hat{A} \hat{e}_j\|_2^2} - \frac{(r_k^T A e_j)^2}{\|A e_j\|_2^2 - \|Y^T \hat{A} \hat{e}_j\|_2^2}. \tag{18}
\]

New indices are added to \( J \) as long as \( \|r_k\|_2 > \varepsilon \) or when the maximum allowed fill-in in a column has been reached.

### 3.2. Approximate sparsity patterns

Banded sparsity patterns in the approximate inverse have been considered by Grote and Simon [11]. Such patterns are reasonable approximations for the sparsity since the entries in the inverse of diagonally dominant matrices decay rapidly away from the diagonal [19]. Other a priori approximate sparsity patterns for \( M \), have been described by Huckle [13]. Considering the characteristic polynomial for the matrix \( A \), we can find coefficients \( \{z_j\}_0^N \) such that

\[
A^{-1} = -(z_N A^{N-1} + \cdots + z_1 I)/z_0.
\]

It then follows that the pattern of \( A^{-1} \) denoted by \( S(A^{-1}) \) is such that

\[
S(A^{-1}) \subseteq S((I + A)^{N-1}).
\]

Huckle also suggests that the pattern of \( A^T \) be included in the sparsity pattern of \( A^{-1} \). For the SPAI algorithm, he gives an upperbound for the sparsity pattern with \( \mu \) steps of adding new entries in \( M \). This upper bound can be taken as the pattern of \( (A^T A)^{\mu - 1} A^T \).

### 3.3. Computation of approximate inverses

We describe the computation of the approximate inverses for matrices arising from the fourth-order discretizations of convection–diffusion problems. We have used our MATLAB programs for the SPAI algorithm. Each least-squares problem is solved by computing a \( QR \)-decomposition of \( \hat{A} \) using the Householder orthogonalization method [14]. We choose enlarged index sets using the criterion given by (16). For each column, the computation is stopped when \( \|r_k\|_2 \leq \varepsilon \) or when the number of nonzero entries in \( m_k \) has reached a maximum fill-in of \( \text{lif} \). We always start with a diagonal structure for the sparse approximate inverse \( M \).

**Problem 1.** We consider the constant coefficient problem

\[
-\Delta u + \tau u = 0, \quad (x,y) \in (0,1) \times (0,1),
\]

with Dirichlet boundary conditions determined from the exact solution

\[
u(x,y) = e^{\tau/2} \sin \frac{\pi y}{\sinh \xi (2e^{-\tau/2} \sinh \xi x + \sinh \xi (1-x))}, \tag{20}
\]

where \( \xi^2 = \pi^2 + \tau^2/4 \).
We choose a mesh size of $h = \frac{1}{25}$, i.e., $n = 24$. This gives a matrix with dimension $N = 576$ and the number of nonzero entries (nnz) is 4900. We consider three cases: $\tau = 50$, $\tau = 100$ and $\tau = 200$. The spectral condition numbers of $A$ are 89.5, 72.4 and 81.7, respectively.

Table 1 shows the Frobenius norms of $AM - I$, the number of flops and the number of nonzero entries for two different values of $\varepsilon$ for the constant coefficient problem. In our numerical experiments, we have varied the l1l parameter so that the bound $||AM - I||_F^6 \sqrt{N} \leq 10$, Theorem 3.1 is satisfied. In our case this bound is 9.6 when $\varepsilon = 0.4$ and 4.8 when $\varepsilon = 0.2$. We observe that for $\varepsilon = 0.2$, the number of nonzero entries in the matrix $M$ is more than four times the number of nonzero entries for the case $\varepsilon = 0.4$.

We have also computed sparse approximate inverses for $\tau = 50$ using upper bounds for the sparsity pattern of $M$. The first pattern $\mathcal{S}_1$ is SPAI with upper bound $(A^T A)^T$. For this pattern, we augment the sparsity pattern if $||r_k|| > \varepsilon$ or the amount of fill-in in $m_k$ is less than the number of nonzero entries in the $k$th column of $(A^T A)^T$. In the second pattern $\mathcal{S}_2$, we choose the most profitable fill-in positions from the sparsity pattern of $(A^T A)^T$. The results are shown in Table 2.

Comparing the two sparsity patterns, we find that good approximations are obtained with the pattern $\mathcal{S}_1$ in the sense that the condition number of the preconditioned matrices are much less than the condition number of $A$. However, more work and a greater number of fill-in are necessary. Comparing the entries in Tables 1 and 2 for $\tau = 50$ and $\varepsilon = 0.2$, we find that unrestricted SPAI (SPAI with criterion (16) for determining enlarged sets) is slightly superior to SPAI with sparsity pattern $\mathcal{S}_1$ as a lesser amount of work and fill-in are required for unrestricted SPAI. The condition number of the preconditioned matrix $AM$ is 36.5 for unrestricted SPAI and 38.4 for SPAI with sparsity pattern $\mathcal{S}_1$. However, the amount of fill-in for $\mathcal{S}_1$ is double the amount for the sparsity pattern obtained with unrestricted SPAI.
Table 3
SPAI for Problem 2: $\varepsilon = 0.4$, $\text{lfil} = 15$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Flops</th>
<th>$|AM - I|_F$</th>
<th>Cond($AM$)</th>
<th>nnz($M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1.2E6</td>
<td>5.6548</td>
<td>3.8413E3</td>
<td>1236</td>
</tr>
<tr>
<td>24</td>
<td>3.6E6</td>
<td>7.4519</td>
<td>5.7003E3</td>
<td>3545</td>
</tr>
<tr>
<td>32</td>
<td>2.3E5</td>
<td>12.3860</td>
<td>1.7724E4</td>
<td>1024</td>
</tr>
<tr>
<td>64</td>
<td>9.4E5</td>
<td>24.4061</td>
<td>6.7309E4</td>
<td>4096</td>
</tr>
</tbody>
</table>

Table 4
SPAI for Problem 2: $\varepsilon = 0.2$

<table>
<thead>
<tr>
<th>$n$</th>
<th>lfil</th>
<th>Flops</th>
<th>$|AM - I|_F$</th>
<th>Cond($AM$)</th>
<th>nnz($M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>80</td>
<td>1.4E8</td>
<td>3.1874</td>
<td>1.5081E4</td>
<td>13036</td>
</tr>
<tr>
<td>24</td>
<td>40</td>
<td>5.5E7</td>
<td>4.7361</td>
<td>2.5148E3</td>
<td>14037</td>
</tr>
<tr>
<td>32</td>
<td>72</td>
<td>8.9E7</td>
<td>5.9088</td>
<td>4.1358E3</td>
<td>25063</td>
</tr>
<tr>
<td>64</td>
<td>96</td>
<td>3.9E8</td>
<td>10.6670</td>
<td>1.3296E4</td>
<td>102329</td>
</tr>
</tbody>
</table>

**Problem 2.** We consider the separable convection–diffusion equation

$$-\Delta u + 10xu_x + 10yu_y - 100u = f, \quad (x, y) \in \Omega = (0, 1) \times (0, 1).$$

The function $f$ is chosen such that the exact solution is $u(x, y) = xe^{xy} \sin \pi x \sin \pi y$. We consider matrices $A$ arising from fourth-order discretizations. We consider different mesh sizes with $n = 16, 24, 32$ and 64. The matrices $A$ are highly ill-conditioned, for example, the condition number of $A$ for $n = 24$ is 10386 and for $n = 64$ the condition number is 67413. Tables 3 and 4 show the results for Problem 2. We have observed that the bound $\|AM - I\|_F \leq \sqrt{N \varepsilon}$ holds when the maximum fill-in parameter lfil = 15 in the case $\varepsilon = 0.4$. For $\varepsilon = 0.2$, larger lfil values are required to satisfy the bound.

**4. Convergence of nonsymmetric Krylov iterations**

For a normal matrix or for a nonnormal diagonalizable matrix with a well-conditioned eigenvector matrix, the convergence of GMRES is essentially determined by the eigenvalues of the matrix $A$ [20, p. 783, Theorems 2 and 3]. For nonnormal matrices whose eigenvalues are highly sensitive to small perturbations in the entries of the matrix, useful alternatives to study the convergence are plots of the pseudospectrum [24], the field of values [9, p. 56], Ritz and harmonic Ritz values [7]. Since the calculation of the field of values is simpler than the calculation of the pseudospectrum in most cases, we study the field of values of the preconditioned matrices. Ritz and Harmonic Ritz values can be computed during the course of a GMRES iteration. We study plots of these quantities in order to relate the convergence of GMRES in terms of these quantities.
The field of values of an $N \times N$ matrix $A$ with complex entries is the set of all complex numbers
\[ \mathcal{F}(A) \equiv \left\{ \frac{z^* A z}{z^* z}, z \in \mathbb{C}^N, z \neq 0 \right\}. \]

The field of values of a matrix $A$ can be computed via the Lanczos method with continuation [3]. If the Hermitian part, $A_H = \frac{1}{2}(A + A^*)$ of the matrix $A$ is positive definite, then the GMRES method converges. In terms of the field of values, the statement that $A_H$ is positive definite is equivalent to the half-plane condition [20], that is, $\mathcal{F}(A)$ lies in the open right half plane.

Bounds based on the field of values of the coefficient matrix and of its inverse have been used by Starke to predict the convergence behaviour of preconditioned nonsymmetric Krylov iterations when hierarchical basis and additive multilevel preconditioners are used as preconditioning strategies [23]. It is also known [9, p. 56] that if the field of values does not contain zero and if $\mathcal{F}(A) \subset D = \{ z \in \mathbb{C} : |z - c| \leq s \}$, then the relative residual norm at the $k$th GMRES iteration satisfies
\[ \frac{||r_k||_2}{||r_0||_2} \leq 2 \left( \frac{s}{|c|} \right)^k. \]

We have computed the field of values of the different matrices arising from fourth-order discretizations of convection-diffusion equations. We have used the routine (`fv.m`) in the Test Matrix Toolbox for MATLAB [12]. Consider Problem 1 given in Section 3.3 with $\tau = 50$. We consider a grid size with $h = \frac{1}{25}$. In Fig. 1 we show the field of values of the preconditioned matrices. The eigenvalues are denoted by crosses ($\times$) on the plots. We find that in all cases the half plane condition is satisfied and thus GMRES will converge. We also remark that the field of values of the ILU(0) preconditioned system is further away from zero than in the other cases. In the case of a single Jordan block, Ipsen [15] has shown that GMRES converges fast if the field of values is far away from zero. In this more general case, our numerical experiments also indicate that the GMRES applied to the ILU(0)-preconditioned system converges faster than for the SPAI preconditioned systems.

Next we consider the coefficient matrix $A$ arising from fourth-order discretizations of Problem 2. In Fig. 2 we show the field of values of the matrix $A$ and the different preconditioned matrices. We remark that $0 \in \mathcal{F}(A)$ in all cases and the field of values are flatter in this case than in the case of Problem 1.

To further illustrate the convergence of GMRES applied to the linear system arising from fourth-order discretizations, we consider the Ritz and harmonic Ritz values which can be computed during a GMRES iteration. For a given $N \times N$ matrix $A$ and an initial vector $r_0 \in \mathbb{R}^N$, after $m$ steps of Arnoldi’s algorithm [1], we obtain $m + 1$ orthogonal vectors $v_1, v_2, \ldots, v_m, v_{m+1}$ which satisfy the relation
\[ AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T, \]
where $V_m = [v_1, v_2, \ldots, v_m]$ and $e_m$ is the $m$th unit vector in $\mathbb{R}^m$. The columns of $V_m$ form an orthonormal basis of the Krylov space
\[ \mathcal{K}(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}. \]

The entries $h_{ij}$ of the upper Hessenberg matrix $H_m$ are given by $(Av_j, v_i)$.

The eigenvalues of the matrix $H_m$ are called the Ritz values whereas harmonic Ritz values are the zeros of the GMRES residual polynomial [7].
To illustrate the convergence of GMRES in terms of plots of Ritz and harmonic Ritz values, we again consider Problems 1 and 2 described in Section 3.3. In Figs. 3–5, we show plots of the maximum and minimum norms of the Ritz and harmonic Ritz values, the convergence history of GMRES as a function of $m$, the dimension of the Krylov subspace and the computed Ritz and harmonic Ritz values for the ILU(0) and SPAI-preconditioned linear systems, respectively.

We first consider Problem 1 with $h = \frac{1}{25}$. For the ILU(0)-preconditioned system we find that when the difference between the Ritz and harmonic Ritz values is large, the norm of the GMRES residual decreases very slowly and as from $m = 8$, this difference becomes small and we remark that the residual norm decreases rapidly from $2.586 \times 10^{-1}$ when $m = 8$ to $4.2676 \times 10^{-6}$ when $m = 13$, that is,
in five more iterations. For the SPAI-preconditioned system with $\varepsilon = 0.4$, we find that the difference in the Ritz and harmonic Ritz values becomes small as from $m = 20$ and we observe that the residual norm starts to decrease rapidly from $1.8654 \times 10^0$ when $m = 20$ to $1.5618 \times 10^{-6}$ in 12 further iterations, whereas for $\varepsilon = 0.2$, the difference becomes small as from $m = 10$. The residual norm is $9.3617 \times 10^0$ when $m = 10$ and decreases to $2.2594 \times 10^{-6}$ in 11 more iterations. The above numerical results show that the difference in the Ritz and harmonic Ritz values is a good indication of the rate of decrease of the GMRES residual. If this difference is small, then the residual norm decreases rapidly.
Fig. 3. Minimum and maximum norm of Ritz values and residual norm of GMRES for Problem 1: ILU(0) preconditioned system—\( \tau = 50 \). (a) Minimum norms, (b) maximum norms, (c) GMRES, (d) Ritz and harmonic Ritz values.

In Fig. 6, we show the results obtained when we use GMRES to solve the linear system arising from discretization of Problem 2 described in Section 3.3. We choose a mesh size \( h = \frac{1}{25} \). The matrix \( A \) is indefinite with the two leftmost eigenvalues being equal to \(-0.7427 \) and \(-0.4413 \) and the two rightmost eigenvalues are 31.5793 and 31.6959. We observe similar results as for Problem 1. We find that the GMRES residual norm decreases very slowly when \( m < 40 \). The difference between the Ritz and harmonic Ritz values becomes very small when the dimension of the Krylov subspace reaches a value which is about 60 and we observe that the residual norm starts to decrease at a faster rate reaching the value \( 8.5419 \times 10^{-6} \) when \( m = 85 \).
Fig. 4. Minimum and maximum norm of Ritz values and residual norm of GMRES for Problem 1: $\tau = 50$, SPAI ($\varepsilon = 0.4$).
(a) Minimum norms, (b) maximum norms, (c) GMRES, (d) Ritz and harmonic Ritz values.

5. Numerical experiments

In this section we describe our experimental results when different Krylov subspace methods are used for solving the linear systems arising from the fourth-order discretizations. For each experiment, we give the iteration and flop counts required by the iterative method to satisfy the stopping criteria

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq 10^{-7}.$$  

We first consider the constant coefficient problem (Problem 1) with $\tau = 50$. Tables 5–7 show the iteration and flop counts when GMRES, BICGSTAB and QMR are used as iterative methods.

We observe that QMR and BICGSTAB require approximately the same amount of work for the unpreconditioned linear system. Both these methods require a lesser amount of work than GMRES to solve the unpreconditioned system. For the case $n = 24$, we observe that GMRES-ILU(0) converges
This faster convergence can be explained by our earlier observation that the field of values of the ILU-preconditioned system lies further away from zero than in the other cases. Unrestricted SPAI and SPAI with the sparsity \( S_2 = (A^T A) A^T \) as upper bound give approximately the same number of iterations and require approximately the same amount of work to achieve convergence. Comparing the performance of the three iterative methods when used to solve the preconditioned systems, we observe that BICGSTAB perform better than GMRES and QMR in terms of the number of flop counts required.

We consider the variable coefficient problem (Problem 2 in Section 3.3). Tables 8–10 show the numerical results obtained (div indicates that the iterative method fails to produce convergent iterations).

For the case \( n = 64 \), we observe that BICGSTAB fails to solve the unpreconditioned and the SPAI(\( \varepsilon = 0.4 \))-preconditioned systems. We also observe that for \( \varepsilon = 0.4 \), SPAI does not give a good
Fig. 6. Minimum and maximum norm of Ritz values and residual norm of GMRES for Problem 2. (a) Minimum norms, (b) maximum norms, (c) GMRES, (d) Ritz and harmonic Ritz values.

Table 5

<p>| Iteration and flop counts required by GMRES for Problem 1: $\tau = 50$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>$\tau$</th>
<th>GMRES</th>
<th>ILU(0)</th>
<th>SPAI 0.4</th>
<th>SPAI 0.2</th>
<th>SPAI $\mathcal{S}_2(0.4)$</th>
<th>SPAI $\mathcal{S}_2(0.2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>17(4.2E5)</td>
<td>10(2.5E5)</td>
<td>22(6.7E5)</td>
<td>16(5.3E5)</td>
<td>25(8.5E5)</td>
<td>17(6.5E5)</td>
</tr>
<tr>
<td>20</td>
<td>25(1.7E6)</td>
<td>13(7.2E5)</td>
<td>30(2.4E6)</td>
<td>19(1.4E6)</td>
<td>28(2.2E6)</td>
<td>19(1.6E6)</td>
</tr>
<tr>
<td>60</td>
<td>33(4.9E6)</td>
<td>15(1.7E6)</td>
<td>39(7.0E6)</td>
<td>22(3.7E6)</td>
<td>41(7.8E6)</td>
<td>22(3.8E6)</td>
</tr>
<tr>
<td>64</td>
<td>73(8.0E7)</td>
<td>26(1.6E7)</td>
<td>86(1.1E8)</td>
<td>37(3.2E7)</td>
<td>86(1.1E8)</td>
<td>40(3.4E7)</td>
</tr>
</tbody>
</table>
preconditioner in the sense that there is not much improvement in the amount of work and number of iterations required when solving the unpreconditioned and preconditioned systems. We also find that the ILU-preconditioner performs better than SPAI with $\varepsilon = 0.2$.

**Problem 3.** We consider the convection–diffusion equation

$$-\Delta u + \frac{\sigma}{2}(1 + x^2)u_x + \tau u_y = f, \quad (x, y) \in \Omega = (0, 1) \times (0, 1).$$

The function $f$ is chosen such that the exact solution is $u(x, y) = xe^{xy} \sin \pi x \sin \pi y$. We let $\sigma = \tau = 10$. The results are given in Tables 11–13.

The above results show that ILU gives a better preconditioner than SPAI and that BICGSTAB converges faster than QMR and GMRES in the case of Problem 3.
Table 9
Iteration and flop counts required by QMR for Problem 2

<table>
<thead>
<tr>
<th>n</th>
<th>QMR</th>
<th>ILU(0)</th>
<th>SPAI $\varepsilon = 0.4$</th>
<th>SPAI $\varepsilon = 0.2$</th>
<th>SPAI $S_2(\varepsilon = 0.4)$</th>
<th>SPAI $S_2(\varepsilon = 0.2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>78(1.8E6)</td>
<td>30(1.0E6)</td>
<td>68(1.9E6)</td>
<td>34(2.6E6)</td>
<td>64(1.9E6)</td>
<td>37(2.1E6)</td>
</tr>
<tr>
<td>24</td>
<td>110(5.7E6)</td>
<td>36(2.7E6)</td>
<td>90(5.9E6)</td>
<td>56(6.1E6)</td>
<td>85(5.6E6)</td>
<td>43(5.7E6)</td>
</tr>
<tr>
<td>32</td>
<td>146(1.3E7)</td>
<td>44(6.0E6)</td>
<td>145(1.4E7)</td>
<td>68(1.3E7)</td>
<td>121(1.2E7)</td>
<td>65(1.3E7)</td>
</tr>
<tr>
<td>64</td>
<td>250(9.3E7)</td>
<td>79(4.4E7)</td>
<td>256(1.0E8)</td>
<td>127(1.0E8)</td>
<td>223(8.7E7)</td>
<td>127(1.0E8)</td>
</tr>
</tbody>
</table>

Table 10
Iteration and flop counts required by BICGSTAB for Problem 2

<table>
<thead>
<tr>
<th>n</th>
<th>BICGSTAB</th>
<th>ILU(0)</th>
<th>SPAI $\varepsilon = 0.4$</th>
<th>SPAI $\varepsilon = 0.2$</th>
<th>SPAI $S_2(\varepsilon = 0.4)$</th>
<th>SPAI $S_2(\varepsilon = 0.2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>117(2.9E6)</td>
<td>28(1.0E6)</td>
<td>87(2.6E6)</td>
<td>34(2.6E6)</td>
<td>66(2.1E6)</td>
<td>38(2.2E6)</td>
</tr>
<tr>
<td>24</td>
<td>237(1.4E7)</td>
<td>37(3.0E6)</td>
<td>133(9.6E6)</td>
<td>56(6.3E6)</td>
<td>130(9.1E6)</td>
<td>45(6.2E6)</td>
</tr>
<tr>
<td>32</td>
<td>181.5(1.9E7)</td>
<td>41(6.0E6)</td>
<td>222(2.4E7)</td>
<td>78(1.6E7)</td>
<td>140(1.1E7)</td>
<td>71(1.4E7)</td>
</tr>
<tr>
<td>64</td>
<td>div</td>
<td>111(6.6E7)</td>
<td>div</td>
<td>176(1.5E8)</td>
<td>div</td>
<td>190(1.6E8)</td>
</tr>
</tbody>
</table>

Table 11
Iteration and flop counts required by GMRES for Problem 3

<table>
<thead>
<tr>
<th>n</th>
<th>GMRES</th>
<th>ILU(0)</th>
<th>SPAI $\varepsilon = 0.4$</th>
<th>SPAI $\varepsilon = 0.2$</th>
<th>SPAI $S_2(\varepsilon = 0.4)$</th>
<th>SPAI $S_2(\varepsilon = 0.2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>36(1.4E6)</td>
<td>11(2.8E5)</td>
<td>36(1.5E6)</td>
<td>18(6.1E5)</td>
<td>38(1.6E6)</td>
<td>18(6.2E5)</td>
</tr>
<tr>
<td>24</td>
<td>53(6.4E6)</td>
<td>15(9.5E5)</td>
<td>53(6.4E6)</td>
<td>25(2.3E6)</td>
<td>55(6.9E6)</td>
<td>27(2.5E6)</td>
</tr>
<tr>
<td>32</td>
<td>69(1.8E7)</td>
<td>20(2.6E6)</td>
<td>69(1.8E7)</td>
<td>34(6.5E6)</td>
<td>73(2.0E7)</td>
<td>32(5.9E6)</td>
</tr>
<tr>
<td>64</td>
<td>135(2.5E8)</td>
<td>34(2.4E7)</td>
<td>135(2.5E8)</td>
<td>57(6.3E7)</td>
<td>141(2.7E8)</td>
<td>60(6.3E7)</td>
</tr>
</tbody>
</table>

Table 12
Iteration and flop counts required by QMR for Problem 3

<table>
<thead>
<tr>
<th>n</th>
<th>QMR</th>
<th>ILU(0)</th>
<th>SPAI $\varepsilon = 0.4$</th>
<th>SPAI $\varepsilon = 0.2$</th>
<th>SPAI $S_2(\varepsilon = 0.4)$</th>
<th>SPAI $S_2(\varepsilon = 0.2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>41(9.3E5)</td>
<td>10(3.4E5)</td>
<td>42(9.9E5)</td>
<td>18(7.0E5)</td>
<td>41(9.7E5)</td>
<td>18(7.3E5)</td>
</tr>
<tr>
<td>24</td>
<td>64(3.3E6)</td>
<td>15(1.2E6)</td>
<td>63(3.4E6)</td>
<td>27(2.5E6)</td>
<td>61(3.3E6)</td>
<td>28(2.6E6)</td>
</tr>
<tr>
<td>32</td>
<td>82(7.6E6)</td>
<td>19(2.6E6)</td>
<td>82(7.9E6)</td>
<td>35(5.9E6)</td>
<td>84(8.1E6)</td>
<td>35(5.9E6)</td>
</tr>
<tr>
<td>64</td>
<td>82(6.4E7)</td>
<td>19(2.2E7)</td>
<td>82(6.6E7)</td>
<td>35(5.3E7)</td>
<td>162(6.3E7)</td>
<td>70(5.5E7)</td>
</tr>
</tbody>
</table>

6. Conclusion

We have described preconditioned Krylov subspace methods for solving nonsymmetric linear systems arising from fourth-order discretizations of convection diffusion problems. Simple recurrences for implementing the ILU(0) factorization of the nine-point scheme have been derived. The convergence of GMRES has been studied by examining plots of the field of values of the preconditioned matrices and by studying the difference between the Ritz and harmonic Ritz values which can be
cheaply computed during a GMRES iteration. Our numerical results show that if the difference between the Ritz and harmonic Ritz values is small, then the GMRES residual norm decreases at a fast rate. Finally, we have studied the performance of ILU(0) and SPAI preconditioning strategies in terms of the amount of work required by three popular iterative methods. Our numerical results show that the ILU preconditioning performs well in most cases than the preconditioning using an approximate inverse and of the three iterative methods studied, BICGSTAB performs better than QMR and GMRES in most of the cases.

References


