STAIR MATRICES AND THEIR GENERALIZATIONS WITH APPLICATIONS TO ITERATIVE METHODS II: ITERATION ARITHMETIC AND PRECONDITIONINGS *

HAO LU †

Abstract. Iteration arithmetic is formally introduced based on iteration multiplication and α-addition which is a special multisplitting. This part focuses on construction of convergent splittings and approximate inverses for Hermitian positive definite matrices by applying stair matrices, their generalizations and iteration arithmetic. Analysis of the splittings and the approximate inverses is also presented. Application of some of the results extends the classical convergence result of the SSOR method. In particular, multiplication symmetrization and addition symmetrization are introduced, which produce Hermitian positive definite approximations for the inverse of an Hermitian positive definite matrix. Furthermore, preconditioning average is introduced to improve some preconditioning methods. Numerical results show a significant improvement of preconditioning average to the approximate inverse preconditionings if an anisotropic elliptic equation is solved.

Key words. stair matrices and their generalization, iteration method, convergence rate, iteration arithmetic, multiplication symmetrization, addition symmetrization, preconditioning average, parallel computation, anisotropic elliptic equation

AMS subject classifications. 65F10, 65F15, 65F50

1. Introduction. Stair matrices and their generalizations are introduced in the first part [7]. This class of matrices provides bases of matrix splittings. Iterative methods based on the matrices are easily performed on a parallel computing platform. By applying stair matrices and their generalizations, a generalization of the SOR method is also introduced in [7]. The SOR theory on determination of the optimal parameter is extended to the generalization. The asymptotic rate of convergence of the new method is derived for Hermitian positive definite matrices. These extend some elegant results of the SOR method in Varga [9], [10] and Young [11], [12].

This paper continues the study of application of stair matrices and their generalizations to iterative methods focusing on construction of convergent splittings and preconditionings for Hermitian positive definite matrices. First, some basic techniques of iterative methods are summarized, including multiplication and α-addition which is a special multisplitting [8]. Then based on these two basic operators, iteration arithmetic is introduced. Let A be a Hermitian positive definite matrix. It is shown that iteration arithmetic indeed provides efficient ways in construction of convergent splittings of A and approximation of the inverse of A. In particular, multiplication symmetrization and addition symmetrization are formally introduced. The trace of multiplication symmetrization is easily found in the literature. For example, the SSOR method is the multiplication symmetrization of the SOR method [10] and [12]. Symmetrization techniques result Hermitian positive definite approximations of the inverse of A, thus yielding efficient preconditionings for preconditioned conjugate gradient methods. Analysis of the splittings and the approximate inverses by using iteration arithmetic and symmetrization is also presented. A result on convergence of the SOR method and the generalization of the SOR method [7] is presented in term of A-norm, which generalizes some results of the SOR method in [12]. Applying this result and a result on iteration arithmetic, we immediately extend the fundamental

*Version October 12, 1999

†Institute for Scientific Computation, Texas A&M University, College Station, Texas 77843-3404, USA (na.hlu@na-net.ornl.gov)

result on convergence of the SSOR method due to Habetler and Wachspress [5], and Ehrlich [3], and Young [12]. Furthermore, preconditioning average is introduced to improve the approximate inverse preconditionings. However, the issue is addressed in a general framework, which can be applied to improve any preconditioning method under certain conditions. Finally, numerical examples are presented to illustrate the preconditioning techniques. If an anisotropic elliptic equation is solved, preconditioning average significantly improves the performance of the approximate inverse preconditionings presented in the paper, showing independence of anisotropy somehow.

2. Preliminaries. In this section we briefly mention stair matrices, their generalizations and some preliminary techniques in iterative methods including multiplication and a special multisplitting, called α-addition in the present paper. We denote by $A = (a_{ij})_{n\times n}$ an $n \times n$ matrix. The entries a_{ij} can be $n_i \times n_j$ blocks. In the case a_{ij} are blocks we still treat them as basic entries. If we emphasize that entries of a matrix are blocks, A_{ij} is adopted to represent the (i, j)th entry instead of a_{ij}.

2.1. Stair matrices and their generalizations. We now recall stair matrices and their generalizations introduced in the first part [7]. All notation is the same as that in [7].

DEFINITION 2.1. *A tridiagonal matrix* $A = \text{tridiag}(a_{i,i-1}, a_{ii}, a_{i,i+1})$ *is called a stair matrix if one of the following conditions is satisfied*

I. $a_{i,i-1} = 0$, $a_{i,i+1} = 0$, $i = 1, 3, \ldots, 2\lfloor\frac{n-1}{2}\rfloor + 1$;

II. $a_{i,i-1} = 0$, $a_{i,i+1} = 0$, $i = 2, 4, \ldots, 2\lfloor\frac{n}{2}\rfloor$.

A stair matrix is of type I if condition I is satisfied and is of type II if condition II holds.

A stair matrix is denoted by $A = \text{stair}(a_{i,i-1}, a_{ii}, a_{i,i+1})$. In particular, $A = \text{stair1}(a_{i,i-1}, a_{ii}, a_{i,i+1})$ and $A = \text{stair2}(a_{i,i-1}, a_{ii}, a_{i,i+1})$ represent a stair matrix of type I and a stair matrix of type II, respectively.

LEMMA 2.2. *An* $n \times n$ *stair matrix* $A = \text{stair}(a_{i,i-1}, a_{ii}, a_{i,i+1})$ *is nonsingular if and only if* a_{ii}, $i = 1, 2, \ldots, n$ *are nonsingular. Furthermore, if* A *is nonsingular then*

$$A^{-1} = D^{-1}(2D - A)D^{-1}, \tag{2.1}$$

where $D = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})$.

A stair linear system $A\mathbf{x} = \mathbf{b}$ is solved by the following algorithm.

ALGORITHM I. This algorithm solves the stair linear system $A\mathbf{x} = \mathbf{b}$. The solution overwrites $\mathbf{b}$. In the algorithm $b_i = 0$ if $i < 1$ or $i > n$.

```
if (A is of type I)
    for i = 1 : 2 : 2⌊(n−1)/2⌋ + 1
        b_i = a_ii^{-1} b_i
    endfor i
    for i = 2 : 2 : 2⌊n/2⌋
        b_i = a_ii^{-1} (b_i − a_{i,i−1} b_{i−1} − a_{i,i+1} b_{i+1})
    endfor i
endif
if (A is of type II)
    for i = 2 : 2 : 2⌊n/2⌋
        b_i = a_ii^{-1} b_i
    endfor i
    for i = 1 : 2 : 2⌊(n−1)/2⌋ + 1
```

$b_i = a_{ii}^{-1}(b_i - a_{i,i-1}b_{i-1} - a_{i,i+1}b_{i+1})$

endfor i

endif.

The generalizations of stair matrices are recursively defined by

- $\mathcal{L}_n^1 = \{A : A \text{ is an } n \times n \text{ matrix and } A = \text{stair}(a_{i,i-1}, a_{ii}, a_{i,i+1})\}$,
- $\mathcal{L}_n^k = \{A : A \text{ is an } n \times n \text{ matrix and } A = \text{stair}(A_{i,i-1}, A_{ii}, A_{i,i+1})$, where each diagonal block A_{ii} is an $n_i \times n_i$ matrix and $A_{ii} \in \mathcal{L}_{n_i}^r$ with $r < k\}$.

As shown in [7] $\mathcal{L}_n^k \subset \mathcal{L}_n^{k+1}$, $k = 1, 2, \ldots$ and $\mathcal{L}_n^k = \mathcal{L}_n^n$ if $k \geq n$. We denote $\mathcal{L}_n \equiv \mathcal{L}_n^n$.

The matrices in $\mathcal{L}_n$ have much in common with triangular matrices. If $S \in \mathcal{L}_n$, the linear system $S\mathbf{x} = \mathbf{b}$ is easily solved by recursively performing Algorithm I. In particular, the solution process is easily parallelized for sparse matrices [7].

2.2. Multiplication and α-addition. Split a nonsingular matrix $A = M - N$ with a nonsingular matrix M. A basic iterative method for the linear system $A\mathbf{x} = \mathbf{b}$ is given by

$$O: \quad \mathbf{x}^n = M^{-1}(N\mathbf{x}^{n-1} + \mathbf{b}), \tag{2.2}$$

which is, in particular, a linear operator from $\mathbb{C}^n$ to $\mathbb{C}^n$. We call O an iterator corresponding to the splitting $A = M - N$ and $M^{-1}N$ the iteration matrix of O. One of basic requirements to a splitting $A = M - N$ is that the linear system with the coefficient matrix M must be easily solved. The traditional way is to choose a triangular matrix M [10], [12]. Matrices in $\mathcal{L}_n$ provide us a lot of new choices. An iterator O is convergent if and only if the spectral radius $\rho(M^{-1}N) < 1$. Since (2.2) is equivalent to

$$\mathbf{x}^n = x^{n-1} + M^{-1}(\mathbf{b} - A\mathbf{x}^{n-1}), \tag{2.3}$$

knowing how to solve the linear system with the coefficient matrix M suffices to fulfill (2.3). Sometime we even don't need to know M and N explicitly.

Based on some splittings, the most common way to construct a new convergent splitting without explicitly knowing M and N is multiplication of iterators. For example, the SSOR and the ADI methods are typically the results of iteration multiplication. Another way is multisplitting. See [8] for details.

Let $A = M_1 - N_1$ and $A = M_2 - N_2$ be two splittings with nonsingular matrices M_1 and M_2. They yield two basic iterative methods

$$O_1: \quad \mathbf{x}^n = M_1^{-1}(N_1\mathbf{x}^{n-1} + \mathbf{b}), \tag{2.4}$$

$$O_2: \quad \mathbf{x}^n = M_2^{-1}(N_2\mathbf{x}^{n-1} + \mathbf{b}). \tag{2.5}$$

Performing O_1 first and then performing O_2 yield the following new iteration:

$$\mathbf{x}^{n-1/2} = M_1^{-1}(N_1\mathbf{x}^{n-1} + \mathbf{b}),$$
$$\mathbf{x}^n = M_2^{-1}(N_2\mathbf{x}^{n-1/2} + \mathbf{b}).$$

This defines the multiplication of O_1 and O_2 by

$$O: \qquad \mathbf{x}^n = M_1^{-1}N_1M_2^{-1}N_2\mathbf{x}^{n-1} + (M_2^{-1}N_2M_1^{-1} + M_2^{-1})\mathbf{b}. \tag{2.6}$$

We denote $O = O_2O_1$. If $M_2^{-1}N_2M_1^{-1} + M_1^{-1}$ is nonsingular, which is satisfied if O_2 is convergent, the multiplication of O_1 and O_2 is actually a basic iteration corresponding to the splitting $A = M - N$, where M is the matrix whose inverse is given by

$$M^{-1} = M_2^{-1}N_2M_1^{-1} + M_2^{-1} = M_1^{-1} + M_2^{-1} - M_2^{-1}AM_1^{-1}. \tag{2.7}$$

The iteration matrix is given by

$$M^{-1}N = M_1^{-1}N_1M_2^{-1}N_2. \tag{2.8}$$

Based on (2.7) the linear system $M^{-1}\mathbf{c}$ is solved in two steps. First we solve $\mathbf{d} = M_1^{-1}\mathbf{c}$ and then compute

$$M^{-1}\mathbf{c} = \mathbf{d} + M_2^{-1}(\mathbf{c} - A\mathbf{d}).$$

Let α be a nonnegative constant satisfying $0 \leq \alpha \leq 1$. The α-addition of O_1 and O_2, denoted by $O = O_1(\alpha+)O_2$, is a weighted average of O_1 and O_2 defined by

$$\mathbf{x}^n = (\alpha M_1^{-1}N_1 + (1-\alpha)M_2^{-1}N_2)\mathbf{x}^{n-1} + (\alpha M^{-1} + (1-\alpha)M_2^{-1})\mathbf{b}, \tag{2.9}$$

which is a special multisplitting [8]. If $\alpha M_2 + (1-\alpha)M_1$ is nonsingular, so is $\alpha M_1^{-1} + (1-\alpha)M_2^{-1}$ because $(\alpha M_1^{-1} + (1-\alpha)M_2^{-1}) = M_1^{-1}(\alpha M_2 + (1-\alpha)M_1)M_2^{-1}$. Furthermore, if $\alpha M_1^{-1} + (1-\alpha)M_2^{-1}$ is nonsingular, then the α-addition of O_1 and O_2 is a basic iteration corresponding to the splitting $A = M - N$ with

$$M^{-1} = \alpha M_1^{-1} + (1-\alpha)M_2^{-1} \tag{2.10}$$

and the iteration matrix is given by

$$M^{-1}N = \alpha M_1^{-1}N_1 + (1-\alpha)M_2^{-1}N_2. \tag{2.11}$$

By iteration arithmetic we mean the restriction of arithmetic of iterators that involves only multiplication and addition. The addition refers to α-addition. By an arithmetic iterator we mean an operator of iteration arithmetic. Given iterators $O_1, \ldots, O_k$, notation $p(O_1, \ldots, O_k)$ represents an arithmetic iterator of $O_1, \ldots, O_k$. For convenience, for k numbers $r_1, \ldots, r_k$ we also use $p(r_1, \ldots, r_k)$ to represent the same arithmetic operator on $r_1, \ldots, r_k$. For example, if $p(O_1, O_2)$ is the multiplication of O_1 and O_2, then $p(r_1, r_2) = r_1r_2$ is the product of r_1 and r_2.

3. Convergent splittings and approximate inverses. In this section we show how to construct convergent splittings and approximate inverses for Hermitian positive definite matrices based on iteration arithmetic. Throughout the section A stands for a Hermitian positive definite matrix unless specialized.

For a matrix B denote by B^* the conjugate transpose of B and define A-norm by $\|B\|_A = \|A^{1/2}BA^{-1/2}\|_2$. We now show the following basic result. The first part is essentially the same as the result in [12] (Theorem 5.3, page 79).

THEOREM 3.1. *Let A be a Hermitian positive definite matrix and $A = M - N$. Then*

a) M is nonsingular and $\|M^{-1}N\|_A < 1$ if and only if $M + M^* > A$, and

b) any eigenvalue of $M^{-1}N$ satisfies $|\lambda(M^{-1}N)| \geq 1$ if M is nonsingular and $M + M^* \leq A$.

Proof. Note that if a matrix Q satisfies $Q + Q^* \geq A$, then Q is nonsingular.

Denote $C = A^{1/2}M^{-1}NA^{-1/2} = I - A^{1/2}M^{-1}A^{1/2}$. We find that

$$\begin{aligned} CC^* &= (I - A^{1/2}M^{-1}A^{1/2})(I - A^{1/2}(M^*)^{-1}A^{1/2}) \\ &= I - A^{1/2}M^{-1}A^{1/2} - A^{1/2}(M^*)^{-1}A^{1/2} + A^{1/2}M^{-1}A(M^*)^{-1}A^{1/2} \\ &= I - A^{1/2}M^{-1}(M + M^* - A)(M^*)^{-1}A^{1/2}, \end{aligned} \tag{3.1}$$

which implies the conclusion of a).

If $M + M^* \le A$, then $(-N) + (-N)^* = 2A - (M + M^*) \ge A$, which implies that N is nonsingular. Applying a) to the splitting $A = (-N) - (-M)$ shows that $\|N^{-1}M\|_A \le 1$. Therefore, any eigenvalue of $N^{-1}M$ satisfies $|\lambda(N^{-1}M)| \le 1$, showing the conclusion of b). □

For a Hermitian positive definite matrix A denote

$$S_A = \{O : \ O \text{ is an iterator corresponding to a splitting } A = M - N \text{ satisfying } M + M^* > A\}.$$

Theorem 3.1 shows that O is convergent if $O \in S_A$. Let O be an iterator corresponding to a splitting $A = M - N$ with a nonsingular matrix M. Define $\|O\|_A = \|M^{-1}N\|_A$ and $\rho(O) = \rho(M^{-1}N)$. We show the following result on iteration arithmetic.

THEOREM 3.2. *Let A be a Hermitian positive definite matrix and $O_i \in S_A$, $i = 1, \ldots, k$. Then any arithmetic iterator $O = p(O_1, \ldots, O_k)$ belongs to S_A and $\|O\|_A \le p(\|O_1\|_A, \ldots, \|O_k\|_A)$.*

Proof. Let $O_1, O_2 \in S_A$ be two iterators corresponding to splittings $A = M_1 - N_1$ and $A = M_2 - N_2$, respectively. Then the multiplication of O_1 and O_2 is a basic iteration corresponding to the splitting $A = M - N$. The inverse of M is given by (2.7). It follows from (2.8) and a) of Theorem 3.1 that

$$\|M^{-1}N\|_A \le \|M_1^{-1}N_1\|_A \|M_2^{-1}N_2\|_A < 1.$$

Applying a) of Theorem 3.1 again shows that the multiplication of O_1 and O_2 belongs to S_A. Similarly, the α-addition of O_1 and O_2 belongs to S_A. Hence, $O \in S_A$ follows immediately from induction. Following (2.8) and (2.11) we find that $\|O\|_A \le p(\|O_1\|_A, \|O_2\|_A)$ if O is the multiplication or the α-addition of O_1 and O_2. The inequality $\|O\|_A \le p(\|O_1\|_A, \ldots, \|O_k\|_A)$ follows from induction too. □

Assume that we know a number of iterators $O_1, \ldots, O_k \in S_A$. According to Theorem 3.2 any arithmetic iterator of $O_1, \ldots, O_k$ is a convergent iterator. For example, let $O \in S_A$ be an iterator corresponding to a splitting $A = M - N$. Define the power of O by $O^2 = OO$ and $O^k = O^{k-1}O$ for $k > 2$. Theorem 3.2 shows that $O^k \in S_A$ corresponding to the splitting $A = P_k - Q_k$. Applying (2.7) and (2.8) shows that

$$P_k^{-1} = M^{-1} \sum_{i=0}^{k-1} (NM^{-1})^i \tag{3.2}$$

and the iteration matrix $P_k^{-1}Q_k = (M^{-1}N)^k$.

For any $O \in S_A$ corresponding to a splitting $A = M - N$, we have

$$\|M^{-1}A - I\|_A = \|M^{-1}N\|_A < 1. \tag{3.3}$$

Therefore, M^{-1} is a fair approximate inverse of A. This approximation can be improved by iteration arithmetic. The kth power of O yields an approximate inverse M_k^{-1} given by (3.2), which is the truncation of Neumann series and was first studied as preconditioning in [2]. However, M is usually not Hermitian. This brings some difficulty when M is applied as a preconditioner for a Hermitian positive definite matrix, in particular, if a preconditioned conjugate gradient method is involved. By using iteration arithmetic the difficulty can be overcome by multiplication symmetrization or addition symmetrization defined as follows.

DEFINITION 3.3. *Let A be a Hermitian matrix and O be an iterator corresponding to a splitting $A = M - N$. Denote by O_* the iterator corresponding to the splitting*

$A = M^* - N^*$. *The multiplication symmetrization of* O *is defined by* $m(O) = OO_*$ *and the addition symmetrization of* O *is defined by* $a(O) = \frac{1}{2}(O + O_*)$.

If $O \in S_A$ then $O_* \in S_A$ because $M + M^* > A$. Theorem 3.2 shows $m(O), a(O) \in S_A$. The trace of multiplication symmetrization is easily found in the literature. The SSOR is the multiplication symmetrization of the SOR method. Let $A = M_m - N_m$ and $A = M_a - N_a$ be splittings of $m(O)$ and $a(O)$, respectively. Then M_m and M_a are Hermitian positive definite matrices due to the following lemma.

LEMMA 3.4. *Let* A *be a Hermitian positive definite matrix and* $O \in S_A$ *be an iterator corresponding to a splitting* $A = M - N$. *If* M *is Hermitian then* M *is positive definite,* $\rho(M^{-1}N) = \|M^{-1}N\|_A$ *and*

$$\kappa(M^{-1}A) \le \frac{1 + \rho(M^{-1}N)}{1 - \rho(M^{-1}N)}. \tag{3.4}$$

Proof. A straightforward computation shows that

$$\begin{aligned} \rho(M^{-1}N) &= \rho(A^{1/2}M^{-1}NA^{-1/2}) = \rho(I - A^{1/2}M^{-1}A^{1/2}) \\ &= \|I - A^{1/2}M^{-1}A^{1/2}\|_2 = \|M^{-1}N\|_A, \end{aligned} \tag{3.5}$$

which also implies that

$$1 - \rho(M^{-1}N) \le \lambda(M^{-1}A) \le 1 + \rho(M^{-1}N).$$

Therefore, M is positive definite because $\lambda(A^{1/2}M^{-1}A^{1/2}) = \lambda(M^{-1}A) > 0$ and (3.4) follows immediately. □

COROLLARY 3.5. *Let* A *be a Hermitian positive definite matrix and* $O, J \in S_A$. *Then* $(OJ)_* = J_*O_*$ *and* $(O(\alpha+)J)_* = O_*(\alpha+)J_*$.

Proof. The proof is trivial. □

Theorem 3.2 shows how to construct convergent splittings and approximate inverses based on iteration arithmetic. To do this, we need to know some basic iterators in S_A. This is easily fulfilled by applying Theorem 3.1. Split $A = D - E - E^*$, where D is a Hermitian positive definite matrix. It follows from Lemma 5.6 in [7] that the eigenvalues of the Jocabi matrix $D^{-1}(E + E^*)$ are real and strictly less than one. Let $M = D_1 - E$ and $N = -D_2 + E^*$, where $D = D_1 + D_2$. Applying Theorem 3.1 shows that $\|M^{-1}N\|_A < 1$ if and only if $D_1 > D_2$, which provides a lot of convergent splittings for A. For example, if D is the diagonal or the block diagonal of A and $E \in \mathcal{L}_n$, we have plenty of choices of diagonal or block diagonal matrices for D_1 and D_2 such that $D_1 > D_2$. A special case is the SOR splitting with $D_1 = D/\omega$ and $D_2 = (1 - 1/\omega)D$. In the following theorem a bound of $\|M^{-1}N\|_A$ is presented for the SOR method and the generalization of the SOR method [7] with $0 < \omega < 2$.

THEOREM 3.6. *Let* A *be a Hermitian positive definite matrix and split* $A = M - N$ *with* $M = D/\omega - E$ *and* $N = (1/\omega - 1)D + E^*$, *where* D *is a Hermitian positive definite matrix and* ω *is a real parameter. Then* $\|M^{-1}N\|_A < 1$ *if and only if* $0 < \omega < 2$, *and if* $0 < \omega < 2$ *then*

$$\|M^{-1}N\|_A \le \begin{cases} \sqrt{1 - \dfrac{\omega(2-\omega)(1-\beta)}{1 - \beta\omega + r^2\omega^2}} & \text{if } r^2\omega^2 \ge \omega - 1, \\ \sqrt{1 - \dfrac{\omega(2-\omega)(1-\alpha)}{1 - \alpha\omega + r^2\omega^2}} & \text{if } r^2\omega^2 < \omega - 1, \end{cases} \tag{3.6}$$

where $r \geq \|D^{-1/2}ED^{-1/2}\|_2$, α and $\beta < 1$ are a lower bound and an upper bound of the Jacobi matrix $D^{-1}(E+E^*)$, respectively. Furthermore,

$$\min_{0<\omega<2} \|M^{-1}N\|_A \leq \begin{cases} \dfrac{\sqrt{4r^2-\beta^2}}{\sqrt{1-2(\beta-2r^2)}+1-\beta} & \text{if } 4r^2 > \beta, \\ \dfrac{2r}{1+\sqrt{1-4r^2}} & \text{if } \alpha \leq 4r^2 \leq \beta, \\ \dfrac{\sqrt{4r^2-\alpha^2}}{\sqrt{1-2(\alpha-2r^2)}+1-\alpha} & \text{if } 4r^2 < \alpha. \end{cases} \tag{3.7}$$

Proof. Since $A = D - E - E^*$ and $M + M^* = 2D/\omega - E - E^*$, we find that $M + M^* > A$ if and only if $0 < \omega < 2$. Applying Theorem 3.1 shows the first part of the theorem.

If $0 < \omega < 2$, it follows from (3.1) that

$$\begin{aligned} \|M^{-1}N\|_A^2 &= \lambda_{\max}(I - \frac{2-\omega}{\omega}A^{1/2}M^{-1}D(M^*)^{-1}A^{1/2})) \\ &= \lambda_{\max}(I - \frac{2-\omega}{\omega}D^{1/2}(M^*)^{-1}AM^{-1}D^{1/2})) \\ &= 1 - \omega(2-\omega)\min_{\mathbf{y}\in\mathbb{C}^n} \frac{\mathbf{y}^*(I - D^{-1/2}(E+E^*)D^{-1/2})\mathbf{y}}{\mathbf{y}^*(I - D^{-1/2}E^*D^{-1/2})(I - D^{-1/2}E^*D^{-1/2})\mathbf{y}} \\ &\leq 1 - \omega(2-\omega)\min_{\mathbf{y}\in\mathbb{C}^n} \frac{1 - \mathbf{y}^*(D^{-1/2}(E+E^*)D^{-1/2})\mathbf{y}/\mathbf{y}^*\mathbf{y}}{1 - \mathbf{y}^*(D^{-1/2}(E+E^*)D^{-1/2})\mathbf{y}/\mathbf{y}^*\mathbf{y} + \omega^2 r^2} \\ &\leq 1 - \omega(2-\omega)\frac{1-x}{1-\omega x + \omega^2 r^2}, \end{aligned}$$

where $x = \mathbf{y}^*(D^{-1/2}(E+E^*)D^{-1/2}\mathbf{y}$. The rest of the proof is essentially the same as that of Theorem 5.7 in [7]. □

Note that if D is the diagonal of A, then the diagonal of $E + E^*$ is zero, which implies that $D^{-1/2}(E+E^*)D^{-1/2}$ is neither positive definite nor negative definite. Thus, $\alpha \leq 0$ and $4r^2 < \alpha$ never occurs. This is the case for the SOR method and the generalization of the SOR method in [7].

By comparison of Theorem 3.6 with Theorem 5.7 in [7], the bounds given by (3.6) and (3.7) are very similar to the bounds of $\rho(M^{-1}N)$ given by Theorem 5.7 in [7]. The slight difference is the requirements of r in two results, say, $r \geq \|D^{-1/2}ED^{-1/2}\|_2$ in Theorem 3.6 while $r \geq r(D^{-1/2}ED^{-1/2})$ in the other one, where $r(D^{-1/2}ED^{-1/2})$ is the numerical radius of $D^{-1/2}ED^{-1/2}$. However, they are two independent results because $\rho(M^{-1}N) \leq \|M^{-1}N\|_A$ and $r(D^{-1/2}ED^{-1/2}) \leq \|D^{-1/2}ED^{-1/2}\|_2$ [4].

Let O be an iterator corresponding to a splitting $A = M - N$, where M satisfies the conditions of Theorem 3.6 with $0 < \omega < 2$. By applying Theorem 3.2 and Theorem 3.6, it is straightforward to obtain bounds of $\|m(O)\|_A$, $\|a(O)\|_A$, $\|O^k\|_A$ and so on. For example, assume that the diagonal of $E + E^*$ is zero and let $\sigma = \|D^{-1/2}ED^{-1/2}\|_2$. Applying Theorem 3.2 and Theorem 3.6 shows that if $4\sigma^2 > \beta$, then

$$\begin{aligned} \|m(O)\|_A \leq \|M^{-1}N\|_A^2 &\leq \frac{4\sigma^2 - \beta^2}{((1-2\beta+4\sigma^2)^{1/2}+1-\beta)^2} \\ &= \Big(1 - \frac{1-\beta}{(1-2\beta+4\sigma^2)^{1/2}}\Big)\Big/\Big(1 + \frac{1-\beta}{(1-2\beta+4\sigma^2)^{1/2}}\Big) \end{aligned} \tag{3.8}$$

with $\omega_0 = \frac{2}{1+(1-2\beta+4\sigma^2)^{1/2}}$ and if $4\sigma^2 \le \beta$, then

$$\|m(O)\|_A \le \|M^{-1}N\|_A^2 = \frac{4\sigma^2}{(1+(1-4\sigma^2)^{1/2})^2} \le \frac{1-(1-4\sigma^2)^{1/2}}{1+(1-4\sigma^2)^{1/2}} \tag{3.9}$$

with $\omega_0 = \frac{2}{1+(1-4\sigma^2)^{1/2}}$. Let $\gamma = \max(\sigma^2, 1/4)$. Then $4\gamma \ge 1 > \beta$ because $\beta < 1$. Applying (3.8) shows

$$\|m(O)\|_A \le \Big(1 - \frac{1-\beta}{(1-2\beta+4\gamma)^{1/2}}\Big)\Big/\Big(1 + \frac{1-\beta}{(1-2\beta+4\gamma)^{1/2}}\Big) \tag{3.10}$$

with $\omega_0 = \frac{2}{1+(1-2\beta+4\gamma)^{1/2}}$. In particular, if $\sigma^2 \le 1/4$ then $4\gamma = 1$, inequality (3.10) becomes

$$\|m(O)\|_A \le \Big(1 - \Big(\frac{1-\beta}{2}\Big)^{1/2}\Big)\Big/\Big(1 + \Big(\frac{1-\beta}{2}\Big)^{1/2}\Big). \tag{3.11}$$

with $\omega_0 = \frac{2}{1+(2(1-\beta))^{1/2}}$. Applying (3.10) and (3.11) to the SSOR method we immediately obtain the fundamental result on convergence of the SSOR method due to Habetler and Wachspress [5], and Ehrlich [3], and Young [12] summarized in Young's book [12] (Theorem 3.1, page 464). However, straightforwardly applying (3.8) and (3.9) further improves the fundamental result on the SSOR method.

Now we proceed to estimate $\rho(m(O))$ and $\rho(a(O))$ for $O \in S_A$. Let B be an $n \times n$ matrix and $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of B. We denote

$$\tau(B) = \max_{1 \le i \le n} |\mathrm{Re}(\lambda_i)| \tag{3.12}$$

and define $\tau(O) = \tau(M^{-1}N)$.

THEOREM 3.7. *Let A be a Hermitian positive definite matrix and $O \in S_A$. Then $\rho(m(O)) = \|O\|_A^2 \ge \rho(O)^2$ and $\tau(O) \le \rho(a(O)) \le \|O\|_A$.*

Proof. Let O be an iterator corresponding to a splitting $A = M - N$ and denote $C = A^{1/2}M^{-1}NA^{-1/2} = I - A^{1/2}M^{-1}A^{1/2}$. We find

$$C^* = I - A^{1/2}(M^*)^{-1}A^{1/2} = A^{1/2}(M^*)^{-1}N^*A^{-1/2},$$

which implies that $\rho(O_*) = \rho(O)$ and $\|O_*\|_A = \|O\|_A$. A straightforward calculation shows that

$$\begin{aligned}\rho(m(O)) &= \rho(M^{-1}N(M^*)^{-1}N^*) = \rho(A^{1/2}M^{-1}NA^{-1/2}A^{1/2}(M^*)^{-1}N^*A^{-1/2}) \\ &= \rho(CC^*) = \|C\|_2^2 = \|M^{-1}N\|_A^2 \ge \rho(O)^2.\end{aligned}$$

Let λ be an arbitrary eigenvalue of $M^{-1}N$. Then λ is an eigenvalue of C. Assume that $\mathbf{x}$ is the corresponding eigenvector, i.e., $C\mathbf{x} = \lambda\mathbf{x}$. Computing $\rho(a(O))$ we find that

$$\begin{aligned}\rho(a(O)) &= \frac{1}{2}\rho(M^{-1}N + (M^*)^{-1}N^*) \\ &= \frac{1}{2}\rho(A^{1/2}(M^{-1}N + (M^*)^{-1}N^*)A^{-1/2}) \\ &= \frac{1}{2}\rho(C + C^*) = \frac{1}{2}\|C + C^*\| \\ &= \frac{1}{2}\max_{\mathbf{y}\in\mathbb{C}^n, \mathbf{y}\neq 0} \frac{|\mathbf{y}^*(C+C^*)\mathbf{y}|}{\mathbf{y}^*\mathbf{y}} \ge \frac{|\mathbf{x}^*(C+C^*)\mathbf{x}|}{\mathbf{x}^*\mathbf{x}} \\ &= |\mathrm{Re}(\lambda)|,\end{aligned}$$

which implies that $\rho(a(O)) \geq \tau(O)$. The inequality $\rho(a(O)) = \|a(O)\|_A \leq \|O\|_A$ follows from Theorem 3.2. □

As basic methods $m(O)$ and O^2 need same computational cost at each iteration. However, $m(O)$ cannot be faster than O^2 because

$$\rho(m(O)) = \|O\|_A^2 \geq \|O^2\|_A \geq \rho(O^2).$$

The advantage of $m(O)$ is that it produces a Hermitian positive definite preconditioner. Since

$$\begin{aligned} A^{1/2}M_m^{-1}A^{1/2} &= A^{1/2}M_m^{-1}AA^{-1/2} = I - A^{1/2}M_m^{-1}N_mA^{1/2} \\ &= I - A^{1/2}M^{-1}NA^{-1/2}A^{1/2}(M^*)^{-1}N^*A^{-1/2} = I - CC^*, \end{aligned}$$

where $C = I - A^{1/2}M^{-1}A^{1/2} = A^{1/2}M^{-1}NA^{-1/2}$, applying Theorem 3.7 shows

$$\begin{aligned} \lambda_{\max}(M_m^{-1}A) &= 1 - \lambda_{\min}(CC^*), \\ \lambda_{\min}(M_m^{-1}A) &= 1 - \lambda_{\max}(CC^*) = 1 - \|O\|_A^2. \end{aligned}$$

Therefore, the condition number of $M_m^{-1}A$ is given by

$$\kappa(M_m^{-1}A) = \frac{1 - \lambda_{\min}(CC^*)}{1 - \|O\|_A^2}. \tag{3.13}$$

Addition symmetrization also yields a Hermitian positive definite preconditioner. Applying Lemma 3.4 and Theorem 3.7 shows

$$\kappa(M_a^{-1}A) \leq \frac{1 + \|O\|_A}{1 - \|O\|_A} = \frac{(1 + \|O\|_A)^2}{1 - \lambda_{\min}(CC^*)}\kappa(M_m^{-1}A). \tag{3.14}$$

In practice, $\lambda_{\min}(CC^*)$ is very close to zero and $\kappa(M_a^{-1}A) \lessapprox 4\kappa(M_m^{-1}A)$. Hence, if multiplication symmetrization yields a good preconditioner, addition symmetrization can yield a reasonably good preconditioner too. For example, consider the matrix arising from the Dirichlet problem on the unit square discretized by a central difference scheme

$$A = \text{blocktridiag}(A_{i,i-1}, A_{ii}, A_{i,i+1}),$$

where $A_{i,i-1} = A_{i,i+1} = -I$ and $A_{ii} = \text{tridiag}(-1, 4, -1)$. Split $A = D - L - L^T$, where D is the diagonal of A and L is the strictly lower triangular part of A. Let O be the iterator corresponding to the SOR splitting $A = M - N$ with $M = D/\omega - L$, where $0 < \omega < 2$. It is well known that $\rho(D^{-1}(L+L^T)) = \cos\pi h$ and is easily checked that $\|D^{-1/2}LD^{-1/2}\|_2 \leq 1/2$. With $\omega = 2/(1 + (2(1-\beta))^{1/2})$ it follows from (3.6) that

$$\|O\|_A \leq \frac{1 - \sin(\pi h/2)}{1 + \sin(\pi h/2)} \approx 1 - \pi h.$$

Therefore, applying (3.13) and (3.14) shows

$$\begin{aligned} \kappa(M_m^{-1}A) &\leq \frac{1}{1 - \|O\|_A^2} \approx \frac{1}{2\pi}h^{-1}, \\ \kappa(M_a^{-1}A) &\leq \frac{2}{1 - \|O\|_A} \approx \frac{2}{\pi}h^{-1}. \end{aligned}$$

An obvious advantage of addition symmetrization preconditioning over multiplication symmetrization preconditioning is that the first one is more easily performed on a parallel computing platform. Since $m(O)$ cannot be faster than O^2, the approximate inverse generated by $m(O^k)$ with a proper positive integer k is recommended if multiplication symmetrization is applied for preconditioning.

4. Preconditioning average. Let A be a Hermitian positive definite matrix. Straightforward application of approximate inverses and symmetrization provides preconditioners to solve the linear system $A\mathbf{x} = \mathbf{b}$ as shown in §3. In this section, we improve those approximate inverse preconditionings by introducing preconditioning average. However, the issue is addressed in a general framework, which can be used to improve any preconditioning method.

Assume that there are a matrix B and a unitary matrix U satisfying

$$A = U^*BU. \tag{4.1}$$

Let C_1 be a preconditioner of A and C_2 be a preconditioner of B. Then U^*C_2U is another preconditioner of A and $(U^*C_2U)^{-1} = U^*C_2^{-1}U$. Following the idea of α-addition of iterators , we define a preconditioner C of A whose inverse is given by

$$C^{-1} = \alpha C_1^{-1} + \beta U^*C_2^{-1}U, \tag{4.2}$$

where α and β are nonnegative number satisfying $\alpha + \beta > 0$. This approach is called preconditioning average. In practice, U is often a permutation matrix. Usually, we assume that C_1 and C_2 are Hermitian positive definite matrices. Therefore, C is a Hermitian positive definite matrix. Since

$$C^{-1}\mathbf{d} = \alpha C_1^{-1}\mathbf{d} + \beta U^*C_2^{-1}U\mathbf{d}$$

for a vector $\mathbf{d}$, solving the linear system $C\mathbf{z} = \mathbf{d}$ is straightforward.

To understand the behavior of the preconditioner defined by (4.2) we first state some results on convergence of a preconditioned conjugate gradient method. Let D be an $n \times n$ matrix with positive eigenvalues $\lambda_1, \ldots, \lambda_n$ and denote

$$\mu(D) = \Big(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\Big)^n \Big/ \prod_{i=1}^{n}\lambda_n. \tag{4.3}$$

It is readily seen that $\mu(D) = \big(\frac{1}{n}\mathrm{tr}(D)\big)^n/\det(D)$, where $\mathrm{tr}(D)$ is the trace of D and $\det(D)$ is the determinant of D. Following Kaporin [6] we illustrate the following results. The results are also found in [1].

a) Let A and B be Hermitian positive matrices then

$$\mu(\alpha A + \beta B) \le \max(\mu(A), \mu(B)),$$

where α and β are nonnegative constants.

b) Let A be a Hermitian positive definite matrix and C be a Hermitian positive definite preconditioner of the linear system $A\mathbf{x} = \mathbf{b}$. Then the smaller the value of $\mu(C^{-1}A)$ the faster the convergence of the preconditioned conjugate gradient method.

Note that in [6] and [1] the results are stated for symmetric positive definite matrices. Following their proofs we find that the results are true for Hermitian positive definite matrices.

Let C be the preconditioner defined by (4.2). Because

$$\begin{aligned}\mu(C^{-1}A) &= \mu(A^{1/2}C^{-1}A^{1/2}) = \mu(\alpha A^{1/2}C_1^{-1}A^{1/2} + \beta A^{1/2}U^*C_2^{-1}UA^{1/2}) \\ &= \mu(\alpha A^{1/2}C_1^{-1}A^{1/2} + \beta U^*B^{1/2}C_2B^{1/2}U),\end{aligned}$$

applying a) shows that

$$\mu(C^{-1}A) \le \max(\mu(C_1^{-1}A), \mu(C_2^{-1}B)) \tag{4.4}$$

For condition number we show a similar inequality. Note that the assumption $A = U^*BU$ implies $A^{1/2} = U^*B^{1/2}U$. A straightforward computation shows that

$$\begin{aligned}\lambda_{\min}(C^{-1}A) &= \lambda_{\min}(A^{1/2}C^{-1}A^{1/2}) \\ &= \lambda_{\min}(\alpha A^{1/2}C_1^{-1}A^{1/2} + \beta A^{1/2}U^*C_2^{-1}UA^{1/2}) \\ &= \lambda_{\min}(\alpha A^{1/2}C_1^{-1}A^{1/2} + \beta U^*UA^{1/2}U^*C_2^{-1}UA^{1/2}U^*U)) \\ &= \lambda_{\min}(\alpha A^{1/2}C_1^{-1}A^{1/2} + \beta U^*B^{1/2}C_2^{-1}B^{1/2}U) \\ &\ge \alpha\lambda_{\min}(A^{1/2}C_1^{-1}A^{1/2}) + \beta\lambda_{\min}(B^{1/2}C_2^{-1}B^{1/2}) \\ &= \alpha\lambda_{\min}(C_1^{-1}A) + \beta\lambda_{\min}(C_2^{-1}B).\end{aligned}$$

Similarly, we find that

$$\lambda_{\max}(C^{-1}A) \le \alpha\lambda_{\max}(C_1^{-1}A) + \beta\lambda_{\max}(C_2^{-1}B).$$

Therefore, the condition number of $C^{-1}A$ is bounded by

$$\kappa(C^{-1}A) \le \frac{\alpha\lambda_{\max}(C_1^{-1}A) + \beta\lambda_{\max}(C_2^{-1}B)}{\alpha\lambda_{\min}(C_1^{-1}A) + \beta\lambda_{\min}(C_2^{-1}B)} \le \max(\kappa(C_1^{-1}A), \kappa(C_2^{-1}B)). \tag{4.5}$$

In particular, if $B = A$ and $C_2 = C_1$, (4.4) and (4.5) show

$$\kappa(C^{-1}A) \le \kappa(C_1^{-1}A), \qquad \mu(C^{-1}A) \le \mu(C_1^{-1}A) \tag{4.6}$$

Inequalities (4.4), (4.5) and (4.6) are only rough estimates. We proceed to provide a concrete example to show that preconditioning average indeed improves some preconditioning methods.

LEMMA 4.1. *Let A and B be Hermitian positive definite matrices. If $A - B$ is Hermitian positive semidefinite and $A - B \neq 0$, then* $\det(A) > \det(B)$.

Proof. Denote $D = A - B$. Then $A^{-1/2}BA^{-1/2} = I - A^{-1/2}DA^{-1/2}$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $A^{-1/2}BA^{-1/2}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the corresponding eigenvectors. Since $D \neq 0$ is Hermitian positive semidefinite, there is at least one $\mathbf{x}_k$, $1 \le k \le n$ such that $\mathbf{x}_k^*A^{-1/2}DA^{-1/2}\mathbf{x}_k > 0$. On the other hand,

$$\lambda_i = \frac{\mathbf{x}_i^*A^{-1/2}BA^{-1/2}\mathbf{x}_i}{\mathbf{x}_i^*\mathbf{x}_i} = 1 - \frac{\mathbf{x}_i^*A^{-1/2}DA^{-1/2}\mathbf{x}_i}{\mathbf{x}_i^*\mathbf{x}_i}.$$

This shows $\lambda_i \le 1$ for $i = 1, \ldots, n$ and $\lambda_k < 1$. Finally, computing the rate of $\det(B)/\det(A)$ we find that

$$\frac{\det(B)}{\det(A)} = \det(A^{-1})\det(B) = \det(A^{-1/2}BA^{-1/2}) = \prod_{i=1}^{n}\lambda_i < 1,$$

which concludes the proof of the lemma. □

Assume that $B = A$ and U is a permutation matrix such that $U^2 = I$. The conditions are satisfied for some problems in practice. For example, matrices arising from an elliptic equation

$$-\frac{\partial}{\partial x}\left(a_1\frac{\partial}{\partial x}u\right) - \frac{\partial}{\partial y}\left(a_2\frac{\partial}{\partial y}u\right) = f \quad \text{on } \Omega = (0,1)\times(0,1), \tag{4.7}$$

$$u|_{\partial\Omega} = g$$

discretized by a central difference scheme or certain finite element methods satisfy our assumptions if $a_1(x,y) = a_2(x,y)$. Details will be given in the following section. Let C_1 be a preconditioner of A. Choosing $C_2 = C_1$ and $\alpha = \beta = 1$, we now show that the second inequality in (4.6) is strict.

Due to $U^2 = I$, an eigenvalue of U is either 1 or -1. Since U is a permutation matrix, thus an orthogonal matrix, we find $U^* = U$, i.e., U is a symmetric matrix. Equation (4.1) implies $AU = UA$. Because A and U are Hermitian matrices, it follows from the well-known result that there exist a unitary matrix P such that

$$P^*AP = \mathrm{diag}(\lambda_1,\ldots,\lambda_n), \qquad P^*UP = \begin{pmatrix} I_m & 0 \\ 0 & -I_k \end{pmatrix}, \tag{4.8}$$

where $\lambda_1,\ldots,\lambda_n$ are the eigenvalues of A, and m and k are the numbers of the eigenvalues 1 and -1 of U, respectively. Let $A^{1/2} = P\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_n})P^*$ and partition

$$G_1 \equiv P^*A^{1/2}C_1^{-1}A^{1/2}P = \begin{pmatrix} A_{11} & A_{12} \\ A_{22} & A_{22} \end{pmatrix}, \tag{4.9}$$

where A_{11} is an $m\times m$ matrix, and A_{22} is a $k\times k$ matrix, and $A_{21}^* = A_{12}$. Applying (4.8) we find that $A^{1/2}UC_1^{-1}UA^{1/2} = UA^{1/2}C_1^{-1}A^{1/2}U$ and

$$\begin{aligned} G_2 &\equiv P^*UA^{1/2}C_1^{-1}A^{1/2}UP \\ &= P^*UPP^*A^{1/2}C_1^{-1}A^{1/2}PP^*UP \\ &= \begin{pmatrix} I_m & 0 \\ 0 & -I_k \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} I_m & 0 \\ 0 & -I_k \end{pmatrix} \\ &= \begin{pmatrix} A_{11} & -A_{12} \\ -A_{21} & A_{22} \end{pmatrix}. \end{aligned} \tag{4.10}$$

Let $G = G_1 + G_2$. Then $\mu(C_1^{-1}A) = \mu(G_1)$, $\mu(C^{-1}A) = \mu(G)$ and

$$G = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}.$$

It is obvious that $\mathrm{tr}(G_1) = \mathrm{tr}(A_{11}) + \mathrm{tr}(A_{22}) = \mathrm{tr}(G)$. The decomposition of

$$G_1 = \begin{pmatrix} I_m & 0 \\ A_{21}A_{11}^{-1} & I_k \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}$$

shows that $\det(G_1) = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12})$. If A_{12} is a non-zero matrix, which is often the case if C_1 is generated by a block preconditioning, Lemma 4.1 shows

that $\det(A_{22} - A_{21}A_{11}^{-1}A_{12}) < \det(A_{22})$. Therefore

$$\begin{aligned}\mu(C_1^{-1}A) = \mu(G_1) &= \big(\frac{1}{n}\mathrm{tr}(G_1)\big)^n/\det(G_1)\\ &= \big(\frac{1}{n}\mathrm{tr}(G)\big)^n/\big(\det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12})\big)\\ &> (\frac{1}{n}\mathrm{tr}(G))^n/(\det(A_{11})\det(A_{22})) = \mu(G)\\ &= \mu(C^{-1}A).\end{aligned}$$

Although we only show that the preconditioner given by (4.2) provides faster convergence when applied to a preconditioned conjugate gradient method for the isotropic case $a_1(x,y) = a_2(x,y)$, as we will see in the numerical section of the paper preconditioning average significantly improves the performance of the approximate inverse preconditionings proposed in the previous section.

5. Numerical examples. In this section we present some numerical examples using the approximate inverse preconditionings discussed in §3 and preconditioning average to solve (4.7).

The discretization of (4.7) by a central difference scheme with a uniform meshsize h and the lexicographic order of the mesh points yields the following linear system

$$A\mathbf{x} = \mathbf{b}, \tag{5.1}$$

where A is a block tridiagonal matrix given by

$$A = \mathrm{blocktridiag}(-A_{i,i-1}, A_{ii}, -A_{i,i+1})$$

with tridiagonal matrices A_{ii} and diagonal matrices $A_{i,i-1}$ and $A_{i,i+1}$.

Let B be the difference matrix of (4.7) discretized with the uniform meshsize h and the columnwise order of the mesh points. It is readily verified that $A = UBU$ and $U^2 = I$, where U is the permutation matrix corresponding to the permutation

$$\begin{pmatrix} 1 & 2 & \cdots & m & m+1 & \cdots & 2m & \cdots m^2 \\ 1 & n+1 & \cdots & (m-1)m+1 & 2 & \cdots & (m-1)m+2 & \cdots m^2 \end{pmatrix}.$$

In particular, if $a_1(x,y) = a_2(x,y)$, then $A = B$.

Let $D = \mathrm{blockdiag}(A_{11}, \ldots, A_{mm})$ and

$$P = \mathrm{stair1}(A_{i,i-1}, 0, A_{i,i+1}), \qquad Q = \mathrm{stair2}(A_{i,i-1}, 0, A_{i,i+1}).$$

We split $A = M - N$ by defining

$$M = D/\omega - P, \qquad N = (1/\omega - 1)D - Q,$$

where $0 < \omega < 2$ is a parameter. The matrix B is of the same form as A. We split $B = M_1 - N_1$ in the same way.

Let $O \in S_A$ be the iterator corresponding to the splitting $A = M - N$ and $O_1 \in S_B$ be the iterator corresponding to the splitting $B = M_1 - N_1$. Linear system (5.1) is solved by preconditioned conjugate gradient methods. The right-hand side of the linear system is chosen such that the function $u(x,y) = x(1-x)y(1-y)e^{xy}$ generates the solution on the grid. Let k be a positive integer. The preconditioners adopted are

- A_k, the approximation of A^{-1} generated by $a(O^k)$;
- M_k, the approximation of A^{-1} generated by $m(O^k)$;
- $C_a = A_k + U\widetilde{A}_k U$, where $\widetilde{A}_k$ is the approximation of B^{-1} generated by $a(O_1^k)$;
- $C_m = M_k + U\widetilde{M}_k U$, where $\widetilde{M}_k$ is the approximation of B^{-1} generated by $m(O_1^k)$.

We consider six examples. The meshsize is chosen to be $h = 1/128$ for every one. The stopping criterion is

$$\|\mathbf{r}_i\|_2/\|\mathbf{r}_0\|_2 < 10^{-7}, \tag{5.2}$$

where $\mathbf{r}_i = \mathbf{b} - A\mathbf{x}^{(i)}$ is the ith residual and the initial guess is $\mathbf{x}^{(0)} = (1, 1, \ldots, 1)^T$. We run with two parameters used frequently in practice. One is the optimal parameter $\omega = 1.9329$ of the block SOR method for the model problem $a_1(x, y) = a_2(x, y) = 1$. The other one is $\omega = 1$. The results are presented by iteration numbers of the preconditioned conjugate gradient methods with different preconditioners mentioned above. Notation N_c represents the iteration number of the conjugate gradient method.

Example 1: The model problem $a_1(x, y) = a_2(x, y) = 1$.

Example 2: A discontinuous coefficients given by

$$a_1(x, y) = a_2(x, y) = \begin{cases} 10^4 & \text{if } (x - 0.5)^2 + (y - 0.5)^2 \le 0.125, \\ 1 & \text{otherwise.} \end{cases}$$

Example 3: Anisotropic and discontinuous coefficients given by $a_2(x, y) = 1$ and

$$a_1(x, y) = \begin{cases} 10^3 & \text{if } (x, y) \in [0.25, 0.75] \times [0.25, 0.75], \\ 10^{-3} & \text{otherwise.} \end{cases}$$

Example 4: Again anisotropic and discontinuous coefficients given by $a_1(x, y) = 1$ and

$$a_2(x, y) = \begin{cases} 10^3 & \text{if } (x, y) \in [0.25, 0.75] \times [0.25, 0.75], \\ 10^{-3} & \text{otherwise.} \end{cases}$$

This example is used to test the different ordering of mesh points to the methods. Linear system (5.1) is the same as that of Example 3 if equation (4.7) of Example 3 is discretized with the columnwise ordering of the mesh points.

Example 5: Anisotropic coefficients in some parts of the domain given by

$$a_1(x, y) = \begin{cases} 10^{-5} & \text{if } (x, y) \in [0, 0.7] \times [0, 0.7], \\ 1 & \text{otherwise.} \end{cases}$$

$$a_2(x, y) = \begin{cases} 10^{-5} & \text{if } (x, y) \in [0.3, 1] \times [0.3, 1], \\ 1 & \text{otherwise.} \end{cases}$$

Example 6: Again anisotropic coefficients in some parts of the domain given by

$$a_1(x, y) = \begin{cases} 10^6 & \text{if } (x, y) \in [0.2, 0.3] \times [0.2, 0.3], \\ 1 & \text{otherwise.} \end{cases}$$

$$a_2(x, y) = \begin{cases} 10^6 & \text{if } (x, y) \in [0.7, 0.8] \times [0.7, 0.8], \\ 1 & \text{otherwise.} \end{cases}$$

TABLE 5.1
Iteration numbers, Example 1, $N_c = 294$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_k	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	113	213	106	119	137	112	127	99
2	61	90	58	57	87	65	78	58
3	43	56	40	36	69	50	62	45
4	33	40	32	27	58	42	53	38
5	28	31	27	21	52	37	47	34
6	23	25	23	18	47	34	42	30

TABLE 5.2
Iteration numbers, Example 2, $N_c = 9582$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_p	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	183	342	149	168	221	182	181	139
2	97	146	81	79	140	105	113	83
3	68	81	57	51	110	81	89	65
4	53	64	46	38	94	68	76	55
5	44	70	39	30	83	60	68	49
6	38	40	34	25	76	54	61	44

TABLE 5.3
Iteration numbers, Example 3, $N_c = 13499$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_k	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	259	466	78	104	294	248	77	66
2	140	203	48	52	180	136	53	43
3	95	125	46	38	143	105	44	36
4	71	92	31	31	122	88	39	32
5	58	70	33	27	109	79	36	30
6	49	57	25	23	99	70	34	28

TABLE 5.4
Iteration numbers, Example 4, $N_c = 13931$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_k	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	1145	2023	78	104	1320	1073	77	66
2	606	879	48	52	845	631	53	43
3	411	553	46	38	670	493	44	36
4	312	339	31	31	574	419	39	32
5	254	308	33	27	513	371	36	30
6	214	249	25	23	466	388	34	28

TABLE 5.5
Iteration numbers, Example 5, $N_c = 10428$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_k	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	196	374	85	88	238	196	102	73
2	103	159	57	43	150	112	65	44
3	71	99	38	29	119	87	54	35
4	54	71	31	22	101	69	46	30
5	46	54	27	18	84	61	41	27
6	39	44	25	16	77	55	38	24

TABLE 5.6
The iteration numbers, Example 6, $N_c = 2569$

	$\omega = 1.9329$				$\omega = 1.0$			
k	A_k	M_k	C_a	C_m	A_k	M_k	C_a	C_m
1	853	1544	125	144	1032	823	143	101
2	450	662	68	68	657	488	96	72
3	317	412	46	45	526	382	77	56
4	246	295	35	32	446	321	67	48
5	207	226	30	26	394	282	60	42
6	127	183	27	22	363	260	54	38

As we see from Tables 5.1–5.6, even with $k = 1$ the approximate inverse preconditioners A_k and M_k substantially reduce the iteration number of the conjugate gradient method for each example. For isotropic problems the preconditioners C_a and C_m improve A_k and M_k consistently with our analysis in §4. For anisotropic problems, C_a and C_m significantly improve A_k and M_k, showing some independence of anisotropy.

Since A_k, M_k, C_a and C_m are constructed by using block stair matrices, they are easily performed on a parallel computing platform. Among them C_a is certainly the best choice for parallel computation.

Based on the splittings $A = M - N$ and $B = M_1 - N_1$, there are a number of ways to construct preconditioners for (5.1) by using arithmetic iterators and symmetrization techniques. We briefly mention a few of them. Since $A = UBU$, the splitting $B = M_1 - N_1$ yields a splitting of A by $A = \widetilde{M} - \widetilde{N}$, where $\widetilde{M} = UM_1U$ and $\widetilde{N} = UN_1U$. Let $\widetilde{O}$ be the iterator corresponding to the splitting $A = \widetilde{M} - \widetilde{N}$. Due to $\widetilde{M} + \widetilde{M}^* = U(M_1 + M_1^*)U > UBU = A$, applying Theorem 3.1 shows $\widetilde{O} \in S_A$. Denote $E = O\widetilde{O}$ and $J = 0.5(O+\widetilde{O})$. Then for a positive integer k, the approximate inverses generated by $a(E^k)$, $a(J^k)$, $m(E^k)$ and $a(J^k)$ provides us other four preconditioners.

REFERENCES

[1] O. AXELSSON, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[2] P. F. DUBOIS, A. GREENBAUM, AND G. H. RODRIGUE, *Approximating the inverse of a matrix for use in iterative algorithms on vector processors*, Computing, 22 (1979), pp. 257–268.

[3] L. W. EHRLICH, *The block symmetric successive overrelaxation method*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 807–826.

[4] M. Goldberg and E. Tadmor, *On numerical radius and its applications*, Linear Algebra Appl., 42 (1982), pp. 263–284.
[5] G. J. Habetler and E. L. Wachspress, *Symmetric successive overrelaxation in solving difusion difference equations*, Math. Comp., 15 (1961), pp. 356–362.
[6] I. E. Kaporin, *An alternative approach to estimation of the conjugate gradient iteration number*, in Numerical Methods and Software, Y. A. Kuznetsov, ed., Acad. Sci. USSR., Moscow, 1990, p. (in Russian).
[7] H. Lu, *Stair matrices and their generalizations with applications to iterative methods I: A generalization of the successive overrelaxation method*, SIAM J. Numer. Anal., (to appear).
[8] D. P. O'Leary and R. E. White, *Multisplittings of matrices and parallel solution of linear systems*, SIAM J. Algebraic Discrete Methods, 6 (1985), pp. 630–640.
[9] R. S. Varga, *p-cyclic matrices: a generalization of the Young-Frankel successive overrelaxation scheme*, Pacific J. Math, 9 (1959), pp. 617–628.
[10] ———, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
[11] D. M. Young, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc, 76 (1954), pp. 92–111.
[12] ———, *Iterative Solution for Large Systems*, Academic Press, New York, 1971.