Journal of Mathematical Sciences, Vol. 101, No. 4, 2000

FACTORIZED SPARSE APPROXIMATE INVERSE PRECONDITIONINGS. III. ITERATIVE CONSTRUCTION OF PRECONDITIONERS

A. Yu. Yeremin, L. Yu. Kolotilina, and A. A. Nikishin

UDC 519.612.2

This paper presents new results of the theoretical study of factorized sparse approximate inverse (FSAI) preconditionings. In particular, the effect of the a posteriori Jacobi scaling and the possibility of constructing FSAI preconditioners iteratively are analyzed. A simple stopping criterion for the termination of local iterations in constructing approximate FSAI preconditioners using the PCG method is proposed. The results of numerical experiments with 3D finite-element problems from linear elasticity are presented. Bibliography 21 *titles.*

1. Introduction

This paper considers the so-called factorized sparse approximate inverse (FSAI) preconditionings for linear algebraic systems with symmetric positive-definite (SPD) coefficient matrices. FSAI preconditionings were introduced and theoretically studied in [19]. The application of FSAI preconditionings to the solution of 3-dimensional finite-element problems on massively parallel computers was considered in [20]. An alternative approach to the construction of sparse approximate inverses in factored form, based on an algorithm for constructing two sets of A-biconjugate vectors, was suggested in [2, 3]. In comparison with the nonfactorized sparse approximate inverse (SAI) preconditionings, introduced much earlier [1] and studied rather intensivity in the last decade (see, e.g., [18, 9, 13, 4, 5, 8]), the FSAI preconditionings have the obvious advantage of preserving the symmetry and positive definiteness of the original matrix.

In this paper, the lower triangular sparsity pattern S of an FSAI preconditioner for an SPD matrix A is assumed to have been fixed beforehand, and the problem of the optimal selection of S is not considered. This is in contrast with the approaches used in [2, 3] and in [13, 4, 5, 8], where the sparsity pattern is selected during the computation of the preconditioner. The present paper mainly addresses the following two issues. First, in Sec. 2, we give additional evidence to support the necessity of incorporating the Jacobi scaling in constructing FSAI preconditioners. To this end, we theoretically compare two types of FSAI preconditioners (that differ by a diagonal scaling matrix), the first of which corresponds to the unconstrained minimization of the Frobenius norm of the corresponding residual matrix over all lower triangular preconditioning matrices of a fixed sparsity pattern S, whereas the second one corresponds to the minimization of the same functional but under the additional constraint that the preconditioned matrix has all of its diagonal entries equal to 1.

The results of this comparison show that the unconstrained minimization of the Frobenius norm yields preconditioned matrices with smaller eigenvalues, which frequently leads to spectral condition numbers significantly larger than when the a posteriori Jacobi scaling is used. This exhibits the potential danger of basing the construction of sparse approximate inverse preconditioners on the unconstrained minimization of the Frobenius norm.

The second issue addressed in this paper is the possibility of constructing FSAI preconditioners iteratively. The necessity of the iterative construction is due to two main reasons. First, this provides the possibility of reducing the costs of constructing FSAI preconditioner, which can be quite large, especially on a sequential computer. Second, in some cases (e.g., when factorized sparse approximate inverses are used to approximate the inverses to Schur complements during the construction of an incomplete block factorization preconditioner), it is unreasonable to form the matrix for which we need an approximate inverse explicitly, for instance, because of the memory considerations. In such situations, direct solution methods simply cannot be applied.

Translated from *Zapiski Nauchnykh Seminarov POMI*, Vol. 248, 1998, pp. 17–48. Original article submitted November 3, 1997.

Concerning the iterative construction of FSAI preconditioners, considered theoretically in Sec. 3 and numerically in Sec. 4, our main conclusion is that FSAI preconditioners do not need to be computed with high accuracy, and a simple stopping criterion based on the relative variation of the diagonal entries of the preconditioner under construction can be used.

2. Optimal and quasioptimal FSAI preconditionings

In this section, we recall (see [19, 20]) and compare two types of FSAI preconditionings for an SPD matrix A. Preconditionings of the first type are introduced for purely theoretical reasons. They are referred to as optimal because, by definition, they are required to minimize the Frobenius norm of the corresponding residual matrix over the set of all lower triangular matrices of a prescribed sparsity pattern. We also consider another type of FSAI preconditionings, which actually possess the same minimization property but under the additional constraint that the resulting preconditioned matrix has unit diagonal entries. Thus, preconditionings of the second type are quasioptimal from the standpoint of minimizing the corresponding Frobenius norm. On the other hand, it turns out that from the viewpoint of optimizing the spectrum distribution of preconditioned matrices, the quasioptimal preconditioners are superior to the optimal ones.

Let A be an $n\times n$ SPD matrix and let S,

$$\{(i,j):\ i<j\} \subseteq S \subseteq \{(i,j):\ i\neq j\}, \tag{2.1}$$

be a fixed lower triangular sparsity pattern. Further, let $A = L_A\, L_A^T$ be the Cholesky decomposition of A. The lower triangular matrix $G^{(0)} = (g_{ij}^{(0)})_{i,j=1}^n$ is defined as the minimizer of the functional $\|I - HL_A\|_F^2 = \operatorname{tr}\,[(I-HL_A)(I-HL_A)^T]$ over all matrices $H = (h_{ij})_{i,j=1}^n$ such that $h_{ij} = 0$ whenever $(i,j) \in S$, i.e., over all matrices H of sparsity pattern S. Thus, the matrix $G^{(0)}$ can be regarded as the optimal sparse approximate inverse of sparsity pattern S to the Cholesky factor L_A, and it is natural to use $G^{(0)}$ as a preconditioner for A, where the preconditioned matrix is of the form $G^{(0)}AG^{(0)^T}$.

As is not difficult to see [19], the matrix $G^{(0)}$ is determined by the equations

$$\begin{aligned}
&g_{ij}^{(0)} = 0, && (i,j) \in S;\\
&(G^{0)}A)_{ij} = 0, && i\neq j\,\&\,(i,j)\notin S;\\
&(G^{(0)}A)_{ii} = \ell_{ii}, && i = 1,\dots,n,
\end{aligned} \tag{2.2}$$

where $L_A = (\ell_{ij})_{i,j=1}^n$. Since all of the principal submatrices of the SPD matrix A are nonsingular, system (2.2) is uniquely solvable, and thus $G^{(0)}$ is uniquely determined. However, $G^{(0)}$ cannot be computed unless the diagonal entries of L_A are available.

We also consider another lower triangular preconditioning matrix $G = (g_{ij})_{i,j=1}^n$ of sparsity pattern S, which is defined below. First one constructs the auxiliary matrix $\widetilde{G} = (\widetilde{g}_{ij})_{i,j=1}^n$ defined by the following relations:

$$\begin{aligned}
&\widetilde{g}_{ij} = 1, && i = j;\\
&\widetilde{g}_{ij} = 0, && (i,j) \in S;\\
&(\widetilde{G}A)_{ij} = 0, && (i,j)\notin S\,\&\,i\neq j.
\end{aligned} \tag{2.3}$$

Then one computes the diagonal matrix $D = \operatorname{diag}(d_1,\dots,d_n)$ defined by the equality

$$D^2 = \operatorname{diag}\left(\widetilde{G}\,A\,\widetilde{G}^T\right) \tag{2.4}$$

and, finally, one sets

$$G = D^{-1}\,\widetilde{G}. \tag{2.5}$$

As is trivial to see, the corresponding preconditioned matrix GAG^T has unit diagonal entries:

$$(GAG^T)_{ii} = 1, \quad i = 1, \dots, n, \tag{2.6}$$

i.e., GAG^T is Jacobi scaled.

As far as we know, the matrices defined by (2.3)–(2.5) first appeared in [8], where they were introduced in a different (recursive) way in the context of approximating banded SPD matrices. In [14] (see also [15–17]), these matrices (described slightly differently) arose as the minimizers of Kaporin's functional

$$\beta(HAH^T) = \frac{\operatorname{tr} HAH^T}{n} \left(\det HAH^T\right)^{-1/n} \tag{2.7}$$

over the set of all lower triangular matrices of a fixed sparsity pattern. Finally, in [19] the same matrices were suggested as a practically computable replacement for the optimal sparse approximate inverses (2.2).

Our first result shows that the preconditioner G defined by (2.3)–(2.5), which is optimal w.r.t. minimizing Kaporin's β, actually possesses a more general optimality property, which can be formulated in terms of the diagonal entries of G.

Lemma 2.1. *Let A be an SPD matrix and let S be a fixed lower triangular sparsity pattern. If the matrix G is defined by* (2.3)–(2.5) *and the matrix $H = (h_{ij})_{i,j=1}^n$ satisfies the conditions*

$$\forall (i,j) \in S \quad h_{ij} = 0 \tag{2.8}$$

and

$$\left(HAH^T\right)_{ii} \le 1, \quad i = 1, \dots, n, \tag{2.9}$$

then

$$g_{ii} \ge h_{ii}, \quad i = 1, \dots, n. \tag{2.10}$$

Proof. Let i, $1 \le i \le n$, be fixed. By (2.6) and (2.9), we have

$$0 \le \left[(G-H)A(G-H)^T\right]_{ii} \le 2\left[1 - (GAH^T)_{ii}\right],$$

whence

$$\left(GAH^T\right)_{ii} \le 1. \tag{2.11}$$

Taking into account that, by (2.3)–(2.5) and (2.8)

$$\left(GAH^T\right)_{ii} = \sum_{j=1}^n (GA)_{ij}\, h_{ij} = (GA)_{ii}\, h_{ii},$$

and, similarly,

$$1 = \left(GAG^T\right)_{ii} = (GA)_{ii}\, g_{ii}, \tag{2.12}$$

from (2.11) we obtain that

$$(GA)_{ii}\, h_{ii} \le (GA)_{ii}\, g_{ii},$$

which implies (2.10). □

Using this simple result, one can easily derive some important implications and, in particular, the quasioptimality of matrices (2.3)–(2.5) regarded as sparse approximate inverses to the Cholesky factor L_A.

Theorem 2.1. *Let, for an $n \times n$ SPD matrix A and a fixed lower triangular sparsity pattern S, the matrix $G = (g_{ij})_{i,j=1}^n$ be defined by (2.3)-(2.5) and let the matrix $H = (h_{ij})_{i,j=1}^n$ of sparsity pattern S be such that*

$$\left(HAH^T\right)_{ii} = 1, \quad i = 1, \dots, n. \tag{2.13}$$

Then

$$\|I - GL_A\|_F \le \|I - HL_A\|_F.$$

Proof. Obviously, it is sufficient to show that

$$\left[(I - GL_A)(I - GL_A)^T\right]_{ii} \le \left[(I - HL_A)(I - HL_A)^T\right]_{ii}, \quad i = 1, \dots, n. \tag{2.14}$$

Since, in view of (2.6) and (2.13), we have

$$\left[(I - GL_A)(I - GL_A)^T\right]_{ii} = 2\left(1 - g_{ii}\,\ell_{ii}\right), \quad i = 1, \dots, n, \tag{2.15}$$

$$\left[(I - HL_A)(I - HL_A)^T\right]_{ii} = 2\left(1 - h_{ii}\,\ell_{ii}\right), \quad i = 1, \dots, n, \tag{2.16}$$

relations (2.14) trivially follow from inequalities (2.10), established in Lemma 2.1. □

Theorem 2.1 states that the matrix G defined by (2.3)–(2.5) is the minimizer of $\|I - HL_A\|_F$ over all matrices H of the same sparsity pattern S that satisfy the addition constraint $\operatorname{diag}(HAH^T) = I_n$. Thus, G proves to be a quasioptimal sparse approximate inverse to L_A.

Based on Lemma 2.1, it is also possible to reestablish the above-mentioned minimization property of the matrix G defined by (2.3)–(2.5) w.r.t. the functional (2.7) in a very simple way.

Theorem 2.2. *For any $n \times n$ SPD matrix A and a fixed lower triangular sparsity pattern S, the matrix G defined by* (2.3)–(2.5) *minimizes the functional* (2.7) *over all nonsingular matrices H of the same sparsity pattern S.*

Proof. Let

$$\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_n), \quad \text{where } \gamma_i = \left[(HAH^T)_{ii}\right]^{1/2}, \quad i = 1, \dots, n.$$

Then

$$\beta(HAH^T) = \frac{\Sigma\gamma_i^2}{n}\left(\det HAH^T\right)^{-1/n} \ge (\Pi\gamma_i^2)^{1/n}\left(\det HAH^T\right)^{-1/n} = \det\left(\overline{H}A\overline{H}^T\right)^{-1/n},$$

where we set $\overline{H} = \Gamma^{-1}H$. Since $(HAH^T)_{ii} = 1$, $i = 1, \dots, n$, and $\overline{H} = (\overline{h}_{ij})_{i,j=1}^n$ has the same sparsity pattern S. Lemma 2.1 ensures that

$$\overline{h}_{ii} = h_{ii}/\gamma_i \le g_{ii}, \quad i = 1, \dots, n,$$

whence

$$\det\left(\overline{H}A\overline{H}^T\right)^{-1/n} \ge \left(\det GAG^T\right)^{-1/n} = \beta\left(GAG^T\right).$$

□.

Now we will establish the explicit relation between G and $G^{(0)}$.

Lemma 2.2. *Let, for an $n \times n$ SPD matrix $A = L_A L_A^T$, where $L_A = (\ell_{ij})_{i,j=1}^n$ is the Cholesky factor of A, and a fixed lower triangular sparsity pattern S, the matrices $G^{(0)}$ and G be defined by* (2.2) *and* (2.3)-(2.5), *respectively. Then*

$$G^{(0)} = \Delta G, \tag{2.17}$$

where

$$\Delta = \operatorname{diag}(\delta_1, \dots, \delta_n), \quad \delta_i = \ell_{ii}\,g_{ii} \le 1, \quad i = 1, \dots, n. \tag{2.18}$$

Proof. First we note that, in view of (2.3),

$$0 < \left(\widetilde{G}A\widetilde{G}^T\right)_{ii} = \sum_{j=1}^{n}(\widetilde{G}A)_{ij}\,\widetilde{g}_{ij} = (\widetilde{G}A)_{ii}, \quad i = 1,\dots,n, \tag{2.19}$$

and thus, by (2.3)–(2.5), the matrix G satisfies the following relations:

$$\begin{aligned} (GA)_{ii} &= d_i^{-1}\,(\widetilde{G}A)_{ii} > 0, && i = 1,\dots,n;\\ (GA)_{ij} &= d_i^{-1}\,(\widetilde{G}A)_{ij} = 0, && i \neq j \,\&\, (i,j) \notin S;\\ g_{ij} &= d_i^{-1}\,\widetilde{g}_{ij} = 0, && (i,j) \in S. \end{aligned}$$

Comparing these relations with (2.2) and making use of the unique solvability of (2.2), we arrive at the conclusion that there exists a diagonal matrix $\Delta = \operatorname{diag}(\delta_1,\dots,\delta_n)$ with positive diagonal entries such that (2.17) holds true. Obviously, $\delta_i = g_{ii}^{(0)}/g_{ii}$, $i = 1,\dots,n$, and thus, by (2.17) and (2.6),

$$\left(G^{(0)}AG^{(0)^T}\right)_{ii} = \left(g_{ii}^{(0)}/g_{ii}\right)^2 (GAG^T)_{ii} = \left(g_{ii}^{(0)}/g_{ii}\right)^2, \quad i = 1,\dots,n. \tag{2.20}$$

On the other hand, in view of (2.2) we have

$$\left(G^{(0)}AG^{(0)^T}\right)_{ii} = \sum_{j=1}^{n}(G^{(0)}A)_{ij}\,g_{ij}^{(0)} = \ell_{ii}\,g_{ii}^{(0)}, \quad i = 1,\dots,n. \tag{2.21}$$

Comparing (2.20) with (2.21), we see that

$$g_{ii}^{(0)} = g_{ii}^2\,\ell_{ii}, \quad i = 1,\dots,n, \tag{2.22}$$

which implies that

$$\delta_i = g_{ii}^{(0)}/g_{ii} = g_{ii}\,\ell_{ii}, \quad i = 1,\dots,n.$$

Thus, to complete the proof of Lemma 2.2, it remains to ascertain that $g_{ii}\,\ell_{ii} \leq 1$, $i = 1,\dots,n$. Indeed, for any i, $1 \leq i \leq n$, we have

$$1 = (GAG^T)_{ii} = \left[GL_A(GL_A)^T\right]_{ii} = (GL_A)_{ii}^2 + \sum_{j\neq i}(GL_A)_{ij}^2 \geq (g_{ii}\,\ell_{ii})^2.$$

□

Based on Lemma 2.2, we will first show that, independently of A and S,

$$1 \leq \frac{\|I - GL_A\|_F}{\|I - G^{(0)}\,L_A\|_F} \leq \sqrt{2}$$

and, furthermore, $\|I - GL_A\|_F$ approaches $\|I - G^{(0)}\,L_A\|_F$ as $G^{(0)}$ approaches L_A^{-1}. Thus, though the matrix G defined by (2.3)–(2.5) is worse as an approximate inverse to L_A than $G^{(0)}$ defined by (2.2), the values of $\|I - GL_A\|_F$ and $\|I - G^{(0)}\,L_A\|_F$ cannot differ by a factor larger than $\sqrt{2}$, and the difference between the two norms decreases as the optimal approximate inverse $G^{(0)}$ approaches L_A^{-1}. Note also that, by (2.22), $g_{ii}^{(0)}\,\ell_{ii} = 1$ if and only if $g_{ii}\,\ell_{ii} = 1$.

Theorem 2.3. *Let, for an $n \times n$ SPD matrix A and a fixed lower triangular sparsity pattern S, the sparse approximate inverses $G^{(0)}$ and G to L_A be defined by (2.2) and (2.3)–(2.5), respectively. Then*

$$\|I - G^{(0)}L_A\|_F \le \|I - G\,L_A\|_F \le \left[\frac{2}{1 + \min\limits_{1\le i\le n}\{g_{ii}^{(0)}\,\ell_{ii}\}^{1/2}}\right]^{1/2} \|I - G^{(0)}\,L_A\|_F. \tag{2.23}$$

Proof. The left-hand-side inequality in (2.23) follows from the optimality of the matrix $G^{(0)}$ with respect to $\|I - G^{(0)}\,L_A\|_F$. The right-hand-side inequality in (2.23) is a consequence of the following relations:

$$\begin{aligned}
\left[(I - G^{(0)}\,L_A)(I - G^{(0)}\,L_A)^T\right]_{ii} &= 1 - 2g_{ii}^{(0)}\,\ell_{ii} + \left(G^{(0)}AG^{(0)^T}\right)_{ii}\\
&\overset{(2.21)}{=} 1 - g_{ii}^{(0)}\,\ell_{ii} \overset{(2.22)}{=} 1 - g_{ii}^2\,\ell_{ii}^2 = (1 - g_{ii}\,\ell_{ii})(1 + g_{ii}\,\ell_{ii}) \ge \left(1 + \min_{1\le i\le n}\{g_{ii}\,\ell_{ii}\}\right)(1 - g_{ii}\,\ell_{ii})\\
&\overset{(2.22),(2.15)}{=} \left(1 + \min_{1\le i\le n}\{g_{ii}^{(0)}\,\ell_{ii}\}^{1/2}\right)\frac{[(I - GL_A)(I - GL_A)^T]_{ii}}{2}, \quad i = 1, \dots, n.
\end{aligned}$$

□

The result below, concerning the eigenvalues of GAG^T and $G^{(0)}AG^{(0)^T}$, is also a consequence of Lemma 2.2.

Theorem 2.4. *Under the hypotheses of Theorem 2.3,*

$$\lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right) \le \lambda_i(GAG^T) \le \left[\min_{1\le i\le n}\{g_{ii}^{(0)}\,\ell_{ii}\}\right]^{-1} \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right), \quad i = 1, \dots, n, \tag{2.24}$$

where the eigenvalues of both matrices are numbered in the same monotone order, say, nonincreasingly.

Proof. Using (2.17) and (2.18) we derive

$$\begin{aligned}
\lambda_i(GAG^T) &= \lambda_i\left(\Delta^{-1}\,G^{(0)}\,A\,G^{(0)^T}\,\Delta^{-1}\right) \ge \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\lambda_{\min}(\Delta^{-2})\\
&= \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\min_{1\le i\le n}\{\delta_i^{-2}\} \ge \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right), \quad i = 1, \dots, n,
\end{aligned}$$

which proves the left-hand-side inequalities in (2.24). The remaining inequalities can be established in a similar way:

$$\begin{aligned}
\lambda_i(GAG^T) &\le \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\lambda_{\max}(\Delta^{-2}) = \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\max_{1\le i\le n}\{\delta_i^{-2}\}\\
&= \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\max_{1\le i\le n}\{(g_{ii}\,\ell_{ii})^{-2}\} = \lambda_i\left(G^{(0)}\,A\,G^{(0)^T}\right)\max_{1\le i\le n}\left\{(g_{ii}^{(0)}\,\ell_{ii})^{-1}\right\}, \quad i = 1, \dots, n.
\end{aligned}$$

□

As Theorem 2.4 shows, no eigenvalue of the matrix GAG^T can be smaller than the corresponding eigenvalue of the matrix $G^{(0)}AG^{(0)^T}$. In particular, we may expect that $\lambda_{\min}(GAG^T) > \lambda_{\min}(G^{(0)}AG^{(0)^T})$, which demonstrates the advantage of using G as a preconditioner for A if one knows that $\lambda_{\max}(GAG^T)$ is bounded from above by a reasonable constant. For instance, if A is an H-matrix, then $\lambda_{\max}(GAG^T) < 2$ [19, Theorem 4.2]. On the other hand, from Theorem 2.4 it follows that the corresponding eigenvalues of the matrices GAG^T and $G^{(0)}AG^{(0)^T}$ approach each other as the diagonal entries $g_{ii}^{(0)}$ (or, equivalently, g_{ii}) approach ℓ_{ii}^{-1}, $i = 1, \dots, n$. Thus, the closer to I_n any of the preconditioned matrices GAG^T and $G^{(0)}AG^{(0)^T}$, the less the difference between them.

Theorem 2.4 implies the following relations for the smallest eigenvalues of $G^{(0)}AG^{(0)^T}$ and GAG^T.

Corollary 2.1. *Under the hypotheses of Theorem 2.3,*

$$\lambda_{\min}\left(G^{(0)}AG^{(0)^T}\right) \le \lambda_{\min}(GAG^T) \le \lambda_{\min}^{1/2}\left(G^{(0)}AG^{(0)^T}\right). \tag{2.25}$$

Proof. In view of the left-hand-side inequality in (2.24), it is sufficient to establish the relation

$$\lambda_{\min}\left(G^{(0)}AG^{(0)^T}\right) \ge \lambda_{\min}^2(GAG^T),$$

which stems from the right-hand-side inequality in (2.24), equality (2.22), and the relation

$$\lambda_{\min}(GAG^T) \le \min_{1\le i\le n}\{g_{ii}^2\,\ell_{ii}^2\}.$$

The latter relation immediately follows from the well-known inequality (see, e.g., [12, p. 191])

$$\sigma_{\min}(B) \le \min_{1\le i\le n} \mid \lambda_i(B) \mid, \tag{2.26}$$

where $\sigma_{\min}(B)$ is the smallest singular value of B, applied to the matrix $B = GL_A$. □

Two-sided bounds for the smallest eigenvalues of $G^{(0)}AG^{(0)^T}$ and GAG^T in terms of $g_{ii}^{(0)}\,\ell_{ii} = (G^{(0)}AG^{(0)^T})_{ii}$ are provided in the next theorem.

Theorem 2.5. *Under the previous notation,*

$$\min_{1\le i\le n}\left\{g_{ii}^{(0)}\,\ell_{ii}\right\}\left(\frac{n-1}{n}\right)^{n-1}\prod_{i=1}^{n} g_{ii}^{(0)}\,\ell_{ii} \le \lambda_{\min}\left(G^{(0)}AG^{(0)^T}\right) \le \left[\min_{1\le i\le n}\left\{g_{ii}^{(0)}\,\ell_{ii}\right\}\right]^2, \tag{2.27}$$

$$\left(\frac{n-1}{n}\right)^{n-1}\prod_{i=1}^{n} g_{ii}^{(0)}\,\ell_{ii} \le \lambda_{\min}(GAG^T) \le \min_{1\le i\le n}\left\{g_{ii}^{(0)}\,\ell_{ii}\right\}. \tag{2.28}$$

Proof. The upper bounds in (2.27) and (2.28) follow from the general inequality (2.26) and the relation $g_{ii}^2\,\ell_{ii} = g_{ii}^{(0)}$ (see (2.22)). Both lower bounds in (2.27) and (2.28) readily follow from the bound [11, Theorem 1]

$$\sigma_{\min}(A) \ge \left(\frac{n-1}{n}\right)^{\frac{n-1}{2}} \mid \det A \mid \frac{r_{\min}(A)}{\prod\limits_{i=1}^{n} r_i(A)}, \tag{2.29}$$

where $A \in \mathbb{C}^{n\times n}$, $r_i(A) = (\sum\limits_{j=1}^{n} |a_{ij}|^2)^{1/2}$, and $r_{\min}(A) = \min\limits_{1\le i\le n} r_i(A)$, if one applies it to the matrices $G^{(0)}L_A$ and GL_A and uses the relations

$$r_i^2(G^{(0)}\,L_A) = \left(G^{(0)}\,A\,G^{(0)^T}\right)_{ii} \overset{(2.21)}{=} g_{ii}^{(0)}\,\ell_{ii},\ r_i^2(GL_A) = (GAG^T)_{ii} \overset{(2.6)}{=} 1,$$

$$\det(G^{(0)}\,L_A) = \prod_{i=1}^{n} g_{ii}^{(0)}\,\ell_{ii},\ \ det(GL_A) = \prod_{i=1}^{n} g_{ii}\,\ell_{ii} \overset{(2.22)}{=} \left[\prod_{i=1}^{n} g_{ii}^{(0)}\,\ell_{ii}\right]^{1/2}.$$

□

Remark 2.1. As has been shown in [11], the lower bound (2.29) in its general form can be easily deduced (by applying an appropriate matrix scaling) from the bound

$$\sigma_{\min}(A) \ge \left(\frac{n-1}{n}\right)^{\frac{n-1}{2}} \mid \det A \mid \ \left(\ge e^{-1/2} \mid \det A \mid\right), \tag{2.30}$$

which is valid for any matrix $A \in \mathbb{C}^{n\mathrm{x}n}$ such that $(AA^*)_{ii} = r_i^2(A) = 1$, $i = 1, \dots, n$. We note that the proof of (2.30) presented in [11] is unnecessarily complicated, and this bound is actually almost trivial. Indeed, for any $A \in \mathbb{C}^{n\mathrm{x}n}$, we have

$$\sigma_{\min}^2(A) = \lambda_n(AA^*) = \frac{\det AA^*}{\prod\limits_{i=1}^{n-1} \lambda_i(AA^*)} \geq \frac{|\det A|^2}{\left[\frac{\sum_{i=1}^{n-1} \lambda_i(AA^*)}{n-1}\right]^{n-1}} \geq |\det A|^2 \left(\frac{n-1}{\operatorname{tr} AA^*}\right)^{n-1},$$

whence (2.30) immediately follows, provided that $\operatorname{tr} AA^* \leq n$.

Note that the interval (2.27) for $\lambda_{\min}(G^{(0)} A G^{(0)^T})$ is obtained by multiplying the interval (2.28) for $\lambda_{\min}(GAG^T)$ by the quantity $\min\limits_{1 \leq i \leq n} g_{ii}^{(0)} \ell_{ii}$, which is strictly less than 1 unless $G = G^{(0)} = L_A^{-1}$.

We conclude our comparison of the two preconditioners $G^{(0)}$ and G by considering a simple example, which shows that the relation

$$\lambda_{\min}(GAG^T) = O\left(\lambda_{\min}^{1/2}\left(G^{(0)} A G^{(0)^T}\right)\right)$$

may actually occur in the case where the Cholesky factor L_A of A has a large off-diagonal entry. In view of (2.25), the latter relation corresponds to the case where the ratio $\lambda_{\min}(GAG^T)/\lambda_{\min}\left(G^{(0)} A G^{(0)^T}\right)$ is of the greatest possible order. Note that since G and $G^{(0)}$ differ by only a diagonal scaling matrix, it is sufficient to consider the case of the diagonal preconditioning matrices $G = D$ and $G^{(0)} = D^{(0)}$.

Example 2.1. Consider the 2×2 parametric matrix

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1+\alpha^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}, \quad \alpha > 0.$$

Applying the symmetric Jacobi scaling to A, we arrive at the matrix

$$B = \begin{bmatrix} 1 & \frac{\alpha}{\sqrt{1+\alpha^2}} \\ \frac{\alpha}{\sqrt{1+\alpha^2}} & 1 \end{bmatrix}$$

with the eigenvalues

$$\lambda_1(B) = 1 + \frac{\alpha}{\sqrt{1+\alpha^2}}, \quad \lambda_2(B) = 1 - \frac{\alpha}{\sqrt{1+\alpha^2}}.$$

Since, obviously,

$$\frac{\alpha}{1+\alpha} \leq \frac{\alpha}{\sqrt{1+\alpha^2}} \leq 1,$$

we see that

$$2 - \frac{1}{1+\alpha} \leq \lambda_1(B) \leq 2,$$

whence

$$\lambda_1(B) = O(1), \quad \text{as } \alpha \to +\infty.$$

Therefore,,

$$\lambda_2(B) = \frac{\det B}{\lambda_1(B)} = \frac{1}{(1+\alpha^2)\lambda_1(B)} = O(\alpha^{-2}), \quad \text{as } \alpha \to +\infty.$$

On the other hand, by minimizing $\|DL_A - I\|_F$ we obtain the matrix

$$B^{(0)} = D^{(0)} A D^{(0)} = \begin{bmatrix} 1 & \frac{\alpha}{1+\alpha^2} \\ \frac{\alpha}{1+\alpha^2} & \frac{1}{1+\alpha^2} \end{bmatrix}.$$

Clearly,

$$1 \le \lambda_1(B^{(0)}) \le \operatorname{tr} B^{(0)} = 1 + \frac{1}{1+\alpha^2} = O(1), \quad \alpha \to +\infty.$$

Hence,

$$\lambda_2(B^{(0)}) = \frac{\det B^{(0)}}{\lambda_1(B^{(0)})} = \frac{1}{(1+\alpha^2)^2 \lambda_1(B^{(0)})} = O(\alpha^{-4}), \quad \alpha \to +\infty.$$

Thus, we see that as $\alpha \to +\infty$,

$$\lambda^2_{\min}(B)/\lambda_{\min}(B^{(0)}) \to 1,$$

whereas

$$\lambda_{\max}(B) = O(1), \quad \lambda_{\max}(B^{(0)}) = O(1).$$

Therefore, for this example, in which the presence in L_A of the large off-diagonal entry α is responsible for the occurrence of the small diagonal entry $(B^{(0)})_{22} = \frac{1}{1+\alpha^2} = O(\alpha^{-2})$, from the standpoint of minimizing the spectral condition number the Jacobi scaling indeed proves to be superior to that minimizing $\|DL_A - I\|_F$ and, moreover,

$$\left(\operatorname{cond} B\right)^2 / \operatorname{cond} B^{(0)} \to 1, \quad \alpha \to +\infty.$$

3. Approximate FSAI preconditioners

In order to construct an FSAI preconditioner (2.3)–(2.5) for an SPD matrix A, one needs to solve linear equations (2.3), which, as is easy to realize, decouple into n independent linear systems (referred to as local), each of which determines the nonzero entries of the corresponding row of the matrix $\widetilde{G}$. As was mentioned in Sec. 1, in some cases the application of direct methods to the solution of local linear systems can prove to be either impossible or too expensive. For this reason, in this section we consider the possibility of constructing the FSAI preconditioners (2.3)–(2.5) approximately by applying an iterative solution technique.

The general theoretical basis for replacing an explicit SPD preconditioner $K \approx A^{-1}$ for an SPD matrix A by an SPD approximation $\widetilde{K}$ to K is as follows. Let $\widetilde{K} = K + E$ and assume that $\|E\|_2$ is small. Then

$$\widetilde{K}A = KA + EA,$$

and thus the perturbation of KA induced by perturbing K is EA, and for the spectral norm of the latter matrix we have

$$\|EA\|_2 \le \lambda_{\max}(A)\, \|E\|_2.$$

Therefore, if $\lambda_{\max}(A)$ is not considerably greater than 1, then the norm of the matrix $\widetilde{K}A - KA$ will be almost as small as that of the matrix $E = \widetilde{K} - K$.

This simple argument leads us to the conclusion that the less the norm of the original matrix A the less harmful the replacement of the initial explicit preconditioner for A by an approximation to it.

In the case of a symmetric preconditioning of the form $A \to GAG^T$, we have the following result.

Theorem 3.1. *Let A be an $n \times n$ SPD matrix, and let $n \times n$ matrices G and H be invertible and satisfy the assumption*

$$\|I - HG^{-1}\|_2 \le \varepsilon < 1. \tag{3.1}$$

Then

$$(1-\varepsilon)^2 \lambda_i(GAG^T) \le \lambda_i(HAH^T) \le (1+\varepsilon)^2 \lambda_i(GAG^T), \quad i = 1, \dots, n, \tag{3.2}$$

where the eigenvalues of both GAG^T and HAH^T are nonincreasingly ordered.

Proof. Denote $F = I - HG^{-1}$. Since, by assumption, $\|F\|_2 \le \varepsilon < 1$, we have

$$\|GH^{-1}\|_2 = \|(I-F)^{-1}\|_2 \le (1 - \|F\|_2)^{-1} \le (1-\varepsilon)^{-1}. \tag{3.3}$$

Therefore, using the standard singular-values inequality (see, e.g., [12, Theorem 3.3.16(d)]) and (3.3), we derive

$$\lambda_i(GAG^T) = \sigma_i^2(GL_A) = \sigma_i^2(GH^{-1}HL_A) \le \sigma_1^2(GH^{-1})\sigma_i^2(HL_A)$$
$$= \|GH^{-1}\|_2^2\lambda_i(HAH^T) \le (1-\varepsilon)^{-2}\lambda_i(HAH^T), \quad i = 1,\dots,n,$$

where $A = L_A L_A^T$ as in Sec. 2 and by σ_i we denote the nonincreasingly ordered singular values. This establishes the left-hand-side inequality in (3.2). Similarly,

$$\lambda_i(HAH^T) = \sigma_i^2(HG^{-1}GL_A) \le \lambda_i(GAG^T)\|HG^{-1}\|_2^2, \quad i = 1,\dots,n,$$

and the proof is completed by taking into account that, by virtue of (3.1),

$$\|HG^{-1}\|_2 = \|I - F\|_2 \le 1 + \varepsilon.$$

□

Now, after providing the above general considerations, we come back to the FSAI preconditioners. Since our approach to the construction of FSAI preconditioners is based on minimizing the Frobenius norm, it is of interest to consider the growth of the value of the target functional when the exact FSAI preconditioner G is replaced by an approximation H to G that is of the same sparsity pattern.

Let an SPD matrix $A = L_A L_A^T$ and a lower triangular sparsity pattern S satisfying (2.1) be fixed, and let the matrices $\widetilde{G}$ and G be defined in accordance with (2.3)–(2.5). Consider an arbitrary matrix $\widetilde{H} = (\widetilde{h}_{ij})_{i,j=1}^n$ that satisfies the relations

$$\widetilde{h}_{ii} = 1, \quad i = 1,\dots,n,$$
$$\widetilde{h}_{ij} = 0, \quad (i,j) \in S, \tag{3.4}$$

and define

$$H = \widetilde{D}^{-1}\widetilde{H}, \tag{3.5}$$

where $\widetilde{D} = \operatorname{diag}(\widetilde{d}_1,\dots,\widetilde{d}_n)$,

$$\tilde{d}_i^2 = (\widetilde{H}A\widetilde{H}^T)_{ii}, \quad i = 1,\dots,n. \tag{3.6}$$

Then for the matrix $H = (h_{ij})_{i,j=1}^n$ the relations

$$h_{ij} = 0, \qquad (i,j) \in S,$$
$$(HAH^T)_{ii} = 1, \quad i = 1,\dots,n, \tag{3.7}$$

obviously hold. Therefore, by Lemma 2.1,

$$h_{ii} \le g_{ii}, \quad i = 1,\dots,n, \tag{3.8}$$

and, by Theorem 2.1,

$$\|I - HL_A\|_F \ge \|I - GL_A\|_F. \tag{3.9}$$

Actually, in the case considered, we can derive explicit expressions for $\|I - HL_A\|_F^2$ and for the difference $\|I - HL_A\|_F^2 - \|I - GL_A\|_F^2$.

Lemma 3.1. *Under the previous notation, we have*

$$\|I - HL_A\|_F^2 = 2\sum_{i=1}^{n} (1 - h_{ii}\,\ell_{ii}) \tag{3.10}$$

and

$$\|I - HL_A\|_F^2 - \|I - GL_A\|_F^2 = \sum_{i=1}^{n} (EAE^T)_{ii}\, g_{ii}\,\ell_{ii}, \tag{3.11}$$

where we denote $L_A = (\ell_{ij})_{i,j=1}^n$ and $E = G - H$.

Proof. Equality (3.10) immediately follows from the relations

$$\left[(I - HL_A)(I - HL_A)^T\right]_{ii} = 2\,(1 - h_{ii}\,\ell_{ii})\,, \quad i = 1, \dots, n.$$

On the other hand, we have

$$2\,(1 - h_{ii}\,\ell_{ii}) = 2\,(1 - g_{ii}\,\ell_{ii}) + 2\,(g_{ii} - h_{ii})\;\ell_{ii} = 2\,(1 - g_{ii}\,\ell_{ii}) + 2\left(1 - \frac{h_{ii}}{g_{ii}}\right) g_{ii}\,\ell_{ii}, \quad i = 1, \dots, n,$$

and since

$$\left[(I - GL_A)(I - GL_A)^T\right]_{ii} = 2(1 - g_{ii}\,\ell_{ii}), \quad i = 1, \dots, n, \tag{3.12}$$

we conclude that

$$\|I - HL_A\|_F^2 = \|I - GL_A\|_F^2 + 2\sum_{i=1}^{n} \left(1 - \frac{h_{ii}}{g_{ii}}\right) g_{ii}\,\ell_{ii}. \tag{3.13}$$

Thus, to complete the proof, it only remains to show that

$$(EAE^T)_{ii} = 2\left(1 - \frac{h_{ii}}{g_{ii}}\right), \quad i = 1, \dots, n. \tag{3.14}$$

Indeed,

$$\begin{aligned} (EAE^T)_{ii} &= \left[(G - H)A(G - H)^T\right]_{ii} = (GAG^T)_{ii} + (HAH^T)_{ii} - 2(GAH^T)_{ii} \\ &= 2\Big[1 - \sum_{j=1}^{n} (GA)_{ij}\, h_{ij}\Big] = 2\left[1 - (GA)_{ii}\, h_{ii}\right] \\ &\overset{(2.12)}{=} 2\left[1 - (GAG^T)_{ii} \frac{h_{ii}}{g_{ii}}\right] = 2\left(1 - \frac{h_{ii}}{g_{ii}}\right), \quad i = 1, \dots, n. \end{aligned} \tag{3.15}$$

Lemma 3.1 is proved. □

Here, three comments are in order. First, for a fixed FSAI preconditioner G, the difference $\|I - HL_A\|_F^2 - \|I - GL_A\|_F^2$ depends only on the diagonal entries of H (see (3.13)). Second, (3.11) provides the representation of $\|I - HL_A\|_F^2$ in the form of two summands, the first of which depends only on the sparsity pattern S, whereas the second one accounts for the inexactness in computing G. Third, the closer $g_{ii}\,\ell_{ii}$ to 0 (i.e., see (3.12), the worse the quality of the ith row of the exact FSAI preconditioner G), the smaller the coefficient with which $(EAE^T)_{ii}$ (i.e., the squared A-norm of the ith row of the error matrix $E = G - H$) occurs in (3.11). Thus, from the standpoint of minimizing $\|I - HL_A\|_F^2$, it is most essential to find high-quality approximations to those rows of G for which the corresponding $g_{ii}\,\ell_{ii}$ are close to 1, whereas the rows for which $g_{ii}\,\ell_{ii}$ are small do not need to be approximated very accurately.

Using Lemma 3.1, we can easily obtain a lower bound in terms of the quantities related to the Frobenius-norm minimization for the spectral norm of the matrix $I - HG^{-1}$, occurring in Theorem 3.1.

Corollary 3.1. *Under the previous assumptions,*

$$\left\|I - HG^{-1}\right\|_2 \geq \left[\frac{\|I - HL_A\|_F^2 - \|I - GL_A\|_F^2}{n}\right]^{1/2}.$$

Proof. Taking into account that, by (2.18), $\ell_{ii}\, g_{ii} \leq 1$, $i = 1, \dots, n$, we derive

$$\sum_{i=1}^{n} (EAE^T)_{ii}\, g_{ii}\, \ell_{ii} \leq \operatorname{tr} EAE^T = \operatorname{tr} L_A^T\, E^T\, EL_A = \operatorname{tr} L_A^T\, G^T \left(G^{-T}\, E^T\, EG^{-1}\right) GL_A$$

$$\leq \lambda_{\max}\left(G^{-T}\, E^T\, EG^{-1}\right) \operatorname{tr} L_A^T\, G^T\, GL_A = \left\|EG^{-1}\right\|_2^2 \operatorname{tr} GAG^T = n \left\|I - HG^{-1}\right\|_2^2,$$

and the result follows from (3.11). □

Note that in the case $S = \{(i,j) : i < j\}$, the FSAI preconditioner G coincides with the inverse L_A^{-1}, and the above lower bound reduces to the trivial bound

$$\|I - HL_A\|_2 \geq \frac{\|I - HL_A\|_F}{\sqrt{n}}.$$

In fact, it would be very helpful to have a nontrivial upper bound for the spectral norm $\left\|I - HG^{-1}\right\|_2$ in terms of the quantities related to the target functional. Unfortunately, we are able to derive only an almost trivial bound

$$\left\|I - HG^{-1}\right\|_2^2 = \left\|EG^{-1}\right\|_2^2 \leq \|EL\|_2^2 \left\|L^{-1}\, G^{-1}\right\|_2^2 \leq \frac{\operatorname{tr} EAE^T}{\lambda_{\min}(GAG^T)} \overset{(3.15)}{=} 2\frac{\sum_{i=1}^{n}\left(1 - \frac{h_{ii}}{g_{ii}}\right)}{\lambda_{\min}(GAG^T)}. \tag{3.16}$$

In particular, this bound shows that whenever

$$\frac{g_{ii} - h_{ii}}{g_{ii}} < \frac{\lambda_{\min}(GAG^T)}{2n}, \quad i = 1, \dots, n, \tag{3.17}$$

we certainly have

$$\left\|I - HG^{-1}\right\|_2^2 < 1,$$

and thus Theorem 3.1 is applicable.

Our next purpose is to provide two-sided bounds for the quantities $(EAE^T)_{ii}$ occurring in (3.11). To this end, we need the following result.

Lemma 3.2. *Under the previous notation, we have*

$$h_{ii}^{-2} - g_{ii}^{-2} = (\widetilde{E}\, A\, \widetilde{E}^T)_{ii}, \quad i = 1, \dots, n, \tag{3.18}$$

where $\widetilde{E} = \widetilde{G} - \widetilde{H}$ *is the intermediate error matrix.*

Proof. We have

$$(\widetilde{E}\, A\, \widetilde{E}^T)_{ii} = \left[(\widetilde{G} - \widetilde{H})A(\widetilde{G} - \widetilde{H})^T\right]_{ii} = (\widetilde{G}\, A\, \widetilde{G}^T)_{ii} + (\widetilde{H}\, A\, \widetilde{H}^T)_{ii} - 2(\widetilde{G}\, A\, \widetilde{H}^T)_{ii}, \quad i = 1, \dots, n. \tag{3.19}$$

But since, in view of (2.3) and (3.4),

$$(\widetilde{G}\, A\, \widetilde{H}^T)_{ii} = \sum_{j=1}^{n} (\widetilde{G}\, A)_{ij}\, \widetilde{h}_{ij} = (\widetilde{G}\, A)_{ii}, \quad i = 1, \dots, n,$$

and, similarly

$$(\widetilde{G}\, A\, \widetilde{G}^T)_{ii} = (\widetilde{G}\, A)_{ii}, \quad i = 1, \dots, n.$$

from (3.19) it follows that

$$(\widetilde{E}\, A\, \widetilde{E}^T)_{ii} = (\widetilde{H}\, A\, \widetilde{H}^T)_{ii} - (\widetilde{G}\, A\, \widetilde{G}^T)_{ii}, \quad i = 1, \dots, n,$$

and (3.18) stems from (2.3)–(2.5) and (3.4)–(3.6). □

Remark 3.1. Relation (3.18) shows, in particular, that the decrease of $(\widetilde{E}\, A\, \widetilde{E}^T)_{ii}$, i.e., of the squared A-norm of the ith row of the matrix $\widetilde{E}$, is equivalent to the increase of the corresponding diagonal entry h_{ii} of the approximate FSAI preconditioner H and, in view of (3.10), to the decrease of $\|I - HL_A\|_F$.

Lemma 3.3. *Under the previous notation,*

$$h_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii} \le (EAE^T)_{ii} \le 2h_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}, \;\; i = 1, \dots, n; \tag{3.20}$$

$$(EAE^T)_{ii} \le g_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}, \;\; i = 1, \dots, n. \tag{3.21}$$

Proof. In order to prove the right-hand-side inequality in (3.20), we make use of (3.14) and Lemma 3.2. In this way we obtain

$$(EAE^T)_{ii} = 2\Big(1 - \frac{h_{ii}}{g_{ii}}\Big) \le 2\Big(1 - \frac{h_{ii}^2}{g_{ii}^2}\Big) = 2h_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}, \;\; i = 1, \dots, n.$$

The left-hand-side inequality in (3.20) is also derived by using (3.14) and Lemma 3.2:

$$(EAE^T)_{ii} = 2\Big(1 - \frac{h_{ii}}{g_{ii}}\Big) = 1 + \Big(1 - \frac{h_{ii}}{g_{ii}}\Big)^2 - \frac{h_{ii}^2}{g_{ii}^2} \ge 1 - \frac{h_{ii}^2}{g_{ii}^2} = h_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}, \;\; i = 1, \dots, n.$$

Finally, by using Lemma 3.2 and (3.8), inequality (3.21) can be deduced as follows:

$$\begin{aligned}(EAE^T)_{ii} &= 2\Big(1 - \frac{h_{ii}}{g_{ii}}\Big) = 2h_{ii}\Big(\frac{1}{h_{ii}} - \frac{1}{g_{ii}}\Big) = h_{ii}\frac{(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}}{\big(h_{ii}^{-1} + g_{ii}^{-1}\big)/2} \\ &\le h_{ii}(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}\, h_{ii}^{1/2}\, g_{ii}^{1/2} \le g_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}, \;\; i = 1, \dots, n.\end{aligned}$$

□

Now, instead of two matrices $\widetilde{H}$ and H, we consider two sequences of matrices $\{\widetilde{H}^{(k)}\}_{k\ge1}$ and $\{H^{(k)}\}_{k\ge1}$, where, for each k, the matrices $\widetilde{H}^{(k)}$ and $H^{(k)}$ satisfy the corresponding counterparts of relations (3.4)-(3.6). We assume that the matrices $\widetilde{H}^{(k)}$, $k \ge 1$, are obtained as successive (P)CG approximations to the matrix $\widetilde{G}$ defined in (2.3). More precisely, this can be described as follows. As already mentioned, the basic equation of system (2.3) decouples into independent "local" linear systems of the form

$$\tilde{g}_i\, A_i = b_i, \quad i \ge 2, \quad n_i \ge 1, \tag{3.22}$$

where

$$\begin{aligned} &A_i = A[J_i, J_i], \;\; \tilde{g}_i = \widetilde{G}[\{i\}, J_i], \;\; b_i = -A[\{i\}, J_i], \\ &J_i = \{j : j \ne i \,\&\, (i,j) \notin S\}, \;\; |\, J_i \,| = n_i, \end{aligned} \tag{3.23}$$

and $B[I, J]$ denotes the submatrix B that corresponds to the row index set I and the column index set J.

We assume that the row vectors $\tilde{h}_i^{(k)} = \widetilde{H}^{(k)}[\{i\}, J_i]$, $i \ge 2$, $n_i \ge 1$, $k \ge 1$, which completely determine $\widetilde{H}^{(k)}$, are the approximate solutions of the corresponding local systems (3.22) obtained after k iterations of the PCG method [7]. Then we obviously have

$$(\widetilde{E}^{(k)}\,A\,\widetilde{E}^{(k)^T})_{ii} = \begin{cases} 0, & n_i = 0, \\ \tilde{e}_i^{(k)}\,A_i\,\tilde{e}_i^{(k)^T}, & n_i \ge 1, \end{cases} \quad i = 1, \dots, n, \tag{3.24}$$

where $\widetilde{E}^{(k)} = \widetilde{G} - \widetilde{H}^{(k)}$ and $\tilde{e}_i^{(k)} = \tilde{g}_i - \tilde{h}_i^{(k)}$, $i = 1, \dots, n$, $k \ge 0$. Now the application of the well-known bound [7] for the PCG method yield the estimate

$$(\widetilde{E}^{(k)}\,A\,\widetilde{E}^{(k)^T})_{ii} \le 4\Big(\frac{\sqrt{\varkappa_i} - 1}{\sqrt{\varkappa_i} + 1}\Big)^{2k}(\widetilde{E}^{(0)}\,A\,\widetilde{E}^{(0)^T})_{ii}, \;\; n_i \ge 1, \;\; k \ge 1, \tag{3.25}$$

where $\varkappa_i = \lambda_{\max}(M_i^{-1}A_i)/\lambda_{\min}(M_i^{-1}A_i)$ is the spectral condition number of the preconditioned matrix $M_i^{-1}A_i$ and $A_i = M_i - N_i$.

Taking into account Lemmas 3.1 and 3.3 and (3.25), we arrive at the following upper bound for $\|I - H^{(k)}L_A\|_F^2$ as a function of k.

Theorem 3.2. *Let, for a fixed $n\times n$ SPD matrix A and a fixed sparsity pattern S of the form* (2.1), *$G = (g_{ij})_{i,j=1}^{n}$ be the exact FSAI preconditioner defined by* (2.3)–(2.5) *and let the sequences $\{H^{(k)}\}$ and $\{\widetilde{H}^{(k)}\}$ of $n\times n$ matrices of the sparsity pattern S, satisfying the conditions*

$$(\widetilde{H}^{(k)})_{ii} = 1, \quad H^{(k)} = \widetilde{D}^{(k)^{-1}}\,\widetilde{H}^{(k)}, \quad (H^{(k)}\,A\,H^{(k)^T})_{ii} = 1, \quad i = 1,\dots,n, \quad k \ge 1,$$

be constructed by applying the PCG method to the local systems (3.22)–(3.23). *Then*

$$\begin{aligned} \|I - GL_A\|_F^2 &\le \|I - H^{(k)}\,L_A\|_F^2 \\ &\le \|I - GL_A\|_F^2 + 4\sum_{i:n_i\ge 1} g_{ii}^2\left(\frac{\sqrt{\varkappa_i}-1}{\sqrt{\varkappa_i}+1}\right)^{2k}(\widetilde{E}^{(0)}\,A\,\widetilde{E}^{(0)^T})_{ii}\,g_{ii}\,\ell_{ii}, \quad k\ge 1, \end{aligned} \tag{3.26}$$

where, for $1\le i\le n$ such that $n_i \ge 1$, $\varkappa_i = \lambda_{\max}(M_i^{-1}\,A)/\lambda_{\min}(M_i^{-1}\,A)$, $A_i = M_i - N_i$, and $\widetilde{E}^{(0)} = \widetilde{G} - \widetilde{H}^{(0)}$ is the initial intermediate error matrix.

Remark 3.2. Clearly, in constructing an approximate FSAI preconditioner H using the PCG method, one can perform a different number of iterations for each of the local systems. In this case, in the right-hand side of (3.26) one must replace k by the corresponding k_i's.

We conclude our theoretical analysis of iteratively constructed approximate FSAI preconditioners by considering the behavior of Kaporin's function $\beta(H^{(k)}\,A\,H^{(k)^T}) = \beta(H^{(k)^T}\,H^{(k)}\,A)$ as a function of k. The role of this functional in the theory of preconditioning is clarifed by the following theorem.

Theorem 3.3 [15]. *Let A be an SPD matrix and $M \approx A^{-1}$ be an SPD preconditioner for A. Then the kth PCG iterate y_k for the system $Ay = f$ satisfies the error bound*

$$\|r_k\|_M \le \left[\beta(MA)^{n/k} - 1\right]^{k/2}\|r_0\|_M, \quad 1\le k\le n-1,$$

where $r_k = f - Ay_k$.

Thus, using β it is possible (provided that $\beta < 2$ is close enough to 1) to bound the convergence rate of the PCG method in terms of the corresponding norm of the residuals. Therefore, the deterioration of the preconditioning quality when passing from an exact FSAI preconditioner G to an approximate FSAI preconditioner H can be evaluated in terms of the ratio

$$\frac{\beta(HAH^T)}{\beta(GAG^T)} = \left(\prod_{i=1}^{n}\frac{g_{ii}}{h_{ii}}\right)^{2/n} = \left[\prod_{i=1}^{n}\left(1 + g_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}\right)\right]^{1/n} \le 1 + \max_{i:n_i\ge 1}\left\{g_{ii}^2(\widetilde{E}\,A\,\widetilde{E}^T)_{ii}\right\}, \tag{3.27}$$

where we have used Lemma 3.2. Note that, by Theorem 2.2, $\beta(HAH^T)/\beta(GAG^T) \ge 1$.

Relations (3.27) and (3.25) immediately imply the following estimate for the ratio $\beta(H^{(k)}\,A\,H^{(k)^T})/\beta(GAG^T)$, where the matrices $H^{(k)}$ are the PCG approximations to G of the same sparsity pattern.

Theorem 3.4. *Under the hypotheses of Theorem 3.2,*

$$1 \le \frac{\beta(H^{(k)}\,A\,H^{(k)^T})}{\beta(GAG^T)} \le 1 + 4\max_{i:n_i\ge 1}\left\{g_{ii}^2\left(\frac{\sqrt{\varkappa_i}-1}{\sqrt{\varkappa_i}+1}\right)^{2k}(\widetilde{E}^{(0)}\,A\,\widetilde{E}^{(0)^T})_{ii}\right\}, \quad k\ge 1.$$

4. Numerical results

In this section, we present the results of numerical experiments on using approximate FSAI preconditioners instead of the exact ones when solving linear systems of equations with SPD coefficient matrices resulting from finite-element approximations of 3D problems of linear elasticity.

For our experiments, we have chosen two linear-elasticity problems for orthotropic materials considered previously in [10]. The first one corresponds to a cubic domain with zero-displacement conditions on three pairwise-adjacent faces. The second problem is the standard channel problem of structural analysis. Three test systems of linear algebraic equations of the form $Ax = b$ with SPD coefficient matrices* have been obtained by applying the hierarchical p-version of the conforming FEM with $p = 3$ for the cubic domain and with $p = 1$ and $p = 4$ for the channel problem to the corresponding variational formulations (see [10] for more details) and by symmetrically scaling the resulting coefficient matrices by the Jacobi method. The first linear system, which corresponds to the cubic domain, is relatively well conditioned ($\mathrm{cond} A = 1.1 \cdot 10^7$), whereas the other two are quite ill conditioned ($\mathrm{cond} A > 3.0 \cdot 10^9$).

For each of the test systems, by applying the Cholesky method, we have constructed the exact FSAI preconditioners corresponding to two sparsity patterns, referred to as "original" and "enlarged." The "original" sparsity pattern coincides with the sparsity pattern of the lower triangular part of the coefficient matrix in question, whereas the enlarged one is the sparsity pattern with one level of fill-in.

Approximate FSAI preconditioners have been constructed by using the (unsymmetrically) preconditioned block conjugate gradient method (PBCG) [21]. The choice of the block version of the CG method is based on the fact that, for the test problems considered, local linear systems with the same coefficient matrices and several (≥ 3) right-hand sides naturally arise. Further, as implied by Theorem 2 in [21], the PBCG method possesses the property of minimizing the A-norm of individual error vectors, which guarantees that the diagonal entries of successively constructed approximate FSAI preconditioners converge monotonically (see Sec. 3). Finally, it should be mentioned that the local linear systems have also been preconditioned with FSAI preconditioners the sparsity pattern of which reproduce those of the lower triangular parts of the "local" coefficient matrices.

The main practical implication of the considerations in Sec. 3 is that the sequences $\{h_{ii}^{(k)}\}$ monotonically converge to the corresponding g_{ii}'s, and the magnitudes of the diagonal entries $h_{ii}^{(k)}$, which determine the values of $\|I - H^{(k)} L_A\|_F$ and $\beta(H^{(k)} A H^{(k)^T})$, can be accepted as measures of the quality of the corresponding rows of the preconditioner $H^{(k)}$. Based on this, for the termination of local iterations we have used the following stopping criterion:

$$\max_{1 \leq i \leq n} \frac{h_{ii}^{(k+1)} - h_{ii}^{(k)}}{h_{ii}^{(k)}} \leq \mathrm{tol} \quad (\mathrm{tol} > 0). \tag{4.1}$$

Note that at each iteration for the ith local linear system, which determines the ith row of the matrix $\widetilde{H}^{(k)}$, the quantity

$$h_{ii}^{(k)^{-1}} = (\widetilde{H}^{(k)} A \widetilde{H}^{(k)^T})_{ii}^{1/2}$$

must be computed, whereas the off-diagonal entries of the final preconditioning matrix H are computed only once upon termination of all local iteration processes.

The numerical results for three test linear systems, each preconditioned with several exact and approxi-

*The test linear systems are available by request from the authors through *ftp* in the ASCII format.

mate FSAI preconditioners, are presented in Table 1. This table adopts the following notation:

N is the order of the linear system;

NZ is the number of nonzero entries in the coefficient matrix A;

tol is the threshold parameter used in the local stopping criterion (4.1), and tol = 0 means that the preconditioner has been computed by the Cholesky method;

$\mathrm{IT}_{\mathrm{loc}}$ is the number of local iterations needed to satisfy (4.1) with the corresponding value of tol;

IT is the number of global FSAI-CG iterations required to satisfy the global stopping criterion

$$\log_{10}(\|r_k\|_2/\|r_0\|_2) < -10,$$

where r_k is the residual at the kth iteration of the PCG method;

$\lambda_{\max}(\lambda_{\min})$ is the largest (smallest) eigenvalue of the tridiagonal matrix made up of the coefficients of the PCG method;

cond = $\lambda_{\max}/\lambda_{\min}$

Table 1

Test system 1: N = 5040, NZ = 688108

S	tol	$\mathrm{IT}_{\mathrm{loc}}$	IT	$\lambda_{\min}$	$\lambda_{\max}$	cond
Original	0.0	–	218	8.9145e-03	3.9629	444.55
Original	1.0-e1	1	218	8.9145e-03	3.9629	444.55
Enlarged	0.0	–	73	3.4061e-02	2.0882	61.307
Enlarged	1.0e-1	3	82	2.7588e-02	2.2096	80.095
Enlarged	1.0e-2	4	76	3.1572e-02	2.1416	67.830
Enlarged	1.0e-3	8	73	3.3242e-02	2.0910	62.901
Enlarged	1.0e-4	12	73	3.3905e-02	2.0876	61.571

Test system 2: N = 3088, NZ = 733872

S	tol	$\mathrm{IT}_{\mathrm{loc}}$	IT	$\lambda_{\min}$	$\lambda_{\max}$	cond
Original	0.0	–	2855	3.8945e-07	4.0550	1.0412e+07
Original	1.0e-1	2	2823	3.8945e-07	4.0550	1.0412e+07
Enlarged	0.0	–	418	3.1321e-06	2.0657	6.5952e+05
Enlarged	1.0e-1	9	440	3.0968e-06	2.2011	7.1075e+05
Enlarged	1.0e-2	11	425	3.1796e-06	2.0703	6.5112e+05
Enlarged	1.0e-3	21	417	3.1602e-06	2.0638	6.5306e+05
Enlarged	1.0e-4	28	411	3.1403e-06	2.0635	6.5712e+05

Test system 3: N = 6183, NZ = 790755

S	tol	$\mathrm{IT}_{\mathrm{loc}}$	IT	$\lambda_{\min}$	$\lambda_{\max}$	cond
Original	0.0	–	> 3000	5.8359e-07	4.0447	6.9307e+06
Original	0.0	–	1065	2.1491e-06	2.8593	1.3304e+06
Enlarged	1.0e-1	9	1385	1.3032e-06	3.1800	2.4401e+06
Enlarged	1.0e-2	33	1172	1.8281e-06	2.9511	1.6143e+06
Enlarged	1.0e-3	43	1050	2.1332e-06	2.8609	1.3411e+06
Enlarged	1.0e-4	60	1025	2.0805e-06	2.8604	1.3748e+06

Analyzing the data in Table 1, we can draw two main conclusions. First, the local stopping criterion (4.1) works quite satisfactorily, and the number of global iterations required monotonically decreases if we decrease the value of tol. It is of interest to note that the number of global iterations required to satisfy the global stopping criterion for an iteratively constructed FSAI preconditioner can be even less than that for the corresponding "exact" FSAI preconditioner (see data for test systems 2 and 3), provided that the value of tol used in the local stopping criterion (4.1) is small enough. So far we have no explanation of this phenomenon. Second, if the coefficient matrix is relatively well conditioned and/or the exact FSAI preconditioner is relatively poor, then tol $= 10^{-1}$ or tol $= 10^{-2}$ is a reasonable choice, whereas for ill-conditioned systems and for FSAI preconditioners of a high quality the value of tol must be decreased up to 10^{-3} or 10^{-4}.

5. Conclusions and final remarks

In this paper, new results of the theoretical and experimental study of factorized sparse approximate inverse preconditionings have been presented. The importance of incorporating the Jacobi scaling into the construction of FSAI preconditioners has been explained. It has also been demonstrated that exact FSAI preconditioners can be replaced by approximate FSAI preconditioners almost without deteriorating the resulting preconditioning quality. This is achieved by using a simple criterion for the termination of local iterations that is based on the monotone convergence of the diagonal entries of the sequence of approximate preconditioning matrices under construction. It is important to emphasize that approximate FSAI preconditioners do not need to be computed very accurately, and, as our experience suggests, the value tol $= 10^{-3}$ in (4.1) is suitable in practice.

Another conclusion implied by the numerical results presented is that at least for the class of problems considered, a reasonable sparsity pattern of the FSAI preconditioner can be prescribed a priori, and the sparsity pattern referred to as "enlarged" in Sec. 4 is a possible choice.

Finally, it should be noted that in this paper we have deliberately avoided all issues concerning the comparison of exact and approximate FSAI preconditioners (assuming that the former can actually be constructed) from the viewpoint of their time/cost efficiency, because without specifying a sufficiently narrow class of problems and a computing platform, such a comparison can be only speculative. Accordingly, the main purpose of this contribution is to demonstrate the possibility of replacing exact FSAI preconditioners by approximate ones without a significant convergence deterioration.

This work was supported in part by the Netherlands Organization for Scientific Research (NWO) under grant 047 003 017, by INTAS under grant INTAS-93-377 EXT, and by INTAS and RFBR under grant INTAS-RFBR 95-0098.

Translated by L. Yu. Kolotilina.

REFERENCES

1. M. W. Benson and P. O. Frederickson, "Iterative solution of large sparse linear systems arising in certain multidimensional approximation problems," *Utilitas Math.*, **22**, 127–140 (1982).
2. M. Benzi, C. D. Meyer, and M. Tůma, "A sparse approximate inverse preconditioner for the conjugate gradient method," *SIAM J. Sci. Comput.*, **17**, 1135–1149 (1996).
3. M. Benzi and M. Tůma, "A sparse approximate inverse preconditioner for nonsymmetric linear systems," Tech. Rep. No. 653, Oct. 1995, Inst. Computer Science, Academy of Sciences of the Czech Republic (1995).
4. E. Chow and Y. Saad, "Approximate inverse preconditioners for general sparse matrices," Res. Rep. UMSI 94/101, Univ. of Minnesota, Supercomputer Inst., Minneapolis, Minnesota (1994).
5. E. Chow and Y. Saad, "Approximate inverse techniques for block-partitioned matrices," Res. Rep. UMSI 95/13, Univ. of Minnesota, Supercomputer Inst., Minneapolis, Minnesota (1995).
6. P. Concus, G. H. Golub, and D. P. O'Leary, "A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations," in: *Sparse Matrix Computations*, J. R. Bunch and D. J. Rose, Academic Press, New York (1976), pp. 309–332.

7. P. Dewilde and E. F. Deprettere, "Approximate inversion of positive matrices with applications to modelling," in: *Modelling, Robustness, and Sensitivity Reduction in Control Systems*, R. F. Curtain, ed., NATO ASI Series, **F 34**, 211-238 (1987).
8. N. I. M. Gould and J. A. Scott, "On approximate-inverse preconditioners," Tech. Rep. RAL 95-026, Rutherford Appleton Lab., Chilton, England (1995).
9. M. Grote and H. Simon, "Parallel preconditioning and approximate inverses on the Connection Machine," in: *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing*, R. Sincovec et al., eds., SIAM, Philadelphia (1993), pp. 519–523.
10. F. A. Gruzinov, L. Yu. Kolotilina, and A. Yu. Yeremin, "Block SSOR preconditionings for high order FE systems, III: Incomplete BSSOR preconditionings," *Numer. Linear Algebra Appl.*, **4**, 393–423 (1997).
11. Y. P. Hong and C. T. Pan, "A lower bound for the smallest singular value," *Linear Algebra Appl.*, **172**, 27–32 (1992).
12. R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press (1991).
13. T. Huckle and M. Grote, "A new approach to parallel preconditioning with sparse approximate inverses," Manuscript SCCM 94-03, Scientific Computing and Computational Mathematics Program, Stanford Univ., Stanford, California (1994).
14. I. E. Kaporin, "An alternative approach to estimating the convergence rate of the CG method," in: *Numerical Methods and Soltware* [in Russian], Yu. A. Kuznetsov, ed., Dept. of Numerical Mathematics, USSR Academy of Sciences, Moscow (1990), pp. 55-72.
15. I. E. Kaporin, "Explicitly preconditioned conjugate gradient method for the solution of unsymmetric linear systems," *Int. J. Comput. Math.*, **40**, 169–187 (1992).
16. I. E. Kaporin, "New convergence results and preconditioning strategies for the conjugate gradient method," *Numer. Linear Algebra Appl.*, **1**, 179–210 (1994).
17. I. E. Kaporin, "The trace/determinant condition number: several properties and application to polynomial preconditionings," *Numer. Linear Algebra Appl.* (submitted).
18. L. Yu. Kolotilina and A. Yu. Yeremin, "On a family of two-level preconditionings of the incomplete block factorization type," *Sov. J. Numer. Math. Model.*, **1**, 293–320 (1986).
19. L. Yu. Kolotilina and A. Yu. Yeremin, "Factorized sparse approximate inverse preconditionings. I: Theory," *SIAM J. Matrix Anal. Appl.*, **14**, 45–58 (1993).
20. L. Yu. Kolotilina and A. Yu. Yeremin, "Factorized sparse approximate inverse preconditioning. II: Solution of 3D FE systems on massively parallel computers," *Int. J. High Speed Computing*, **7**, 191–215 (1995).
21. D. P. O'Leary, "The block conjugate gradient algorithm and related methods," *Linear Algebra Appl.*, **29**, 293–322 (1980).