Numer. Math. 61, 91-110 (1992)

Numerische Mathematik

Modified block-approximate factorization strategies

Monga-Made Magolu *

Université Libre de Bruxelles, CP 165, Service de Métrologie Nucléaire, 50 av. F.D. Roosevelt, B-1050 Brussels, Belgium

Received September 6, 1990 / Revised version received June 25, 1991

Summary. Two variants of modified incomplete block-matrix factorization with additive correction are proposed for the iterative solution of large linear systems of equations. Both rigorous theoretical support and numerical evidence are given displaying their efficiency when applied to discrete second order partial differential equations (PDEs), even in the case of quasi-singular problems.

Mathematics Subject Classification (1991): 65F10, 65F35, 65B99

1 Introduction

Incomplete block-matrix factorizations have been recognized as useful preconditioning techniques for the iterative solution of large sparse linear systems, particularly efficient when combined with a conjugate gradient like method [4, 5, 9, 10, 13, 26], or when used for the construction of smoothers in multigrid methods [19, 31], or of domain decomposition preconditioners [15].

Among these factorizations, "modified" versions, for which the original matrix and its approximate factorization have equal row sum, are known to be powerful for reducing the spectral condition number (see e.g. [5, 10, 13, 20–22, 25]). Unfortunately, performances of modified methods are sometimes spoiled by round off errors in the presence of *strongly isolated largest eigenvalues* [27, 28]. We intend here to elaborate approximate block factorizations that not only reduce the largest eigenvalues but also (implicitly) cluster the interior eigenvalues and ensure a $O(h^{-1})$ behaviour for the associated spectral condition number, when applied to discrete multidimensional self-adjoint (elliptic) PDEs with (average) mesh size parameter h. This will lead us to investigate a particular family of block approximate factorizations satisfying a perturbed generalized row sum relation, of the form $(A+\Delta)\mathbf{x}=B\mathbf{x}$, A denoting a given matrix, B standing for its approximate factorization while $\mathbf{x}$ and Δ are some appropriate positive vector and nonnegative (pointwise) diagonal matrix. Of course, the stan-

* Research supported by the A.B.O.S. (A.G.C.D.) under project 11, within the co-operation between Belgium and Zaire

dard modified version ($\Delta=0$ and $\mathbf{x}=(1, 1, \ldots, 1)^T$) belongs to that family; on the other hand, the so-called unmodified method will also be found to belong to the same family. We shall discuss two new strategies for choosing the matrix Δ. Both rigorous theoretical support and numerical evidence will be given, displaying the superiority of the new versions over the standard (unmodified and modified) versions.

The paper is organized as follows. General terminology and notation are summarized in Sect. 2. Section 3 describes modified block incomplete factorization with additive correction and reports an existence criterion borrowed from [25]. In Sect. 4, the announced strategies are presented and discussed. We only consider the case of Stieltjes matrices because, as in the case of point approximate factorizations, the conditioning analysis of general positive definite matrices may be reduced to that of Stieltjes matrices (or of a more general class of matrices, called almost Stieltjes matrices, introduced in [11]) by means of spectral equivalence or other similar arguments; we refer to [3], [11] and [16] for practical means to operate this reduction. Results of numerical experiments are presented and discussed in Sect. 5.

2 General terminology and notation

To be short, by SPD we mean symmetric and positive definite. The symbols A^T, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote, respectively, the transpose, the smallest and the largest eigenvalues of the matrix A.

The *order relation* between real matrices and vectors is the usual componentwise order: if $A=(a_{ij})$ and $B=(b_{ij})$ then $A \leqq B$ ($A<B$) if $a_{ij} \leqq b_{ij}$ ($a_{ij}<b_{ij}$) for all i, j; A is called nonnegative (positive) if $A \geqq 0$ ($A>0$). If $A=(a_{ij})$, we denote by $\text{diag}(A)$ the (diagonal) matrix whose entries are $a_{ii}\delta_{ij}$ and we let $\textit{offdiag}(A)=A-\text{diag}(A)$. Similarly, $\textit{tridiag}(A)$ denotes the tridiagonal matrix whose tridiagonal part consists of the three main diagonals of A. By $\mathbf{e}$ we denote the vector with all components equal to unity, by a "(0, 1) matrix", we understand a matrix whose nonzero entries are equal to unity.

A real square matrix is called a *Stieltjes matrix* if it is symmetric positive definite and if its offdiagonal (pointwise) entries are all nonpositive [12, 29].

Hadamard multiplication. We recall that the Hadamard product $A*B$ of the matrices A and B of the same dimensions, with scalar entries a_{ij} and b_{ij}, is the element by element multiplication, i.e. with $(A*B)_{ij}=a_{ij}b_{ij}$, and that the unit matrix with respect to Hadamard multiplication, denoted ε, is the matrix whose entries are all equal to unity.

Standard LU-factorization. By the standard point LU-factorization of a (Stieltjes) matrix A, we understand the factorization $A=LP^{-1}U$ such that U is upper triangular, $P=\text{diag}(U)$ and $L=U^T$.

Partitionings. Any partitioning of an n-vector $\mathbf{x}=(\mathbf{x}_I)$ into block components $\mathbf{x}_I$ of dimensions n_I, $I=1, 2, \ldots, M$ $\left(\text{with } \sum_{I=1}^{M} n_I=n\right)$ is uniquely determined by a partitioning $\pi=(\pi_I)_{1 \leqq I \leqq M}$ of the set $[1, n]$ of the first n integers. We shall assume throughout the paper that all n-vectors are partitioned in blocks accord-

ing to a given such partitioning. The same partitioning π induces also a partitioning of any $n \times n$ matrix A into block components A_{IJ} of dimensions $n_I \times n_J$ and we shall similarly assume that all $n \times n$ matrices are partitioned in this way.

Lower case indices refer to scalar entries and capital indices to block entries. Thus scalar (resp. block) entries of an $n \times n$ π-partitioned matrix A are denoted a_{ij} (resp. A_{IJ}). When needed, scalar entries of block entries of A are denoted $(A_{IJ})_{ij}$, a notation which implies that $i \in \pi_I$ and $j \in \pi_J$. Similar notations are used for vector components, except that we always represent vectors by small letters.

A matrix which is block diagonal (triangular) relative to a π-partitioning will be referred to as π-diagonal (π-triangular). In order to avoid confusion, we sometimes write $\mathrm{diag}_{\mathrm{p}}(A)$ for $\mathrm{diag}(A)$ and $\mathit{offdiag}_{\mathrm{p}}(A)$ for $\mathit{offdiag}(A)$, the subscript "p" stressing that these notions refer to the point partitioning.

Graph notions. We refer to [14] and [17] for the general terminology on matrix graphs with the warning that all graphs considered here are ordered undirected graphs. We recall that the quotient graph of the $n \times n$ (π-partitioned) matrix H with respect to $\pi = (\pi_I)_{1 \leqq I \leqq M}$, denoted $G(H)/\pi$, may be identified with the node set $[1, M]$ together with the edge set E defined by $\{I, J\} \in E$ if and only if $H_{IJ} \neq 0$ or $H_{JI} \neq 0$. The following more specific concepts borrowed from [11] will also be of help.

An *increasing path* in a graph is a path $I_0, I_1, I_2, \ldots, I_l$ such that $I_0 < I_1 < I_2 < \ldots < I_l$ (we have opted as in [11] for denoting a path as a succession of nodes, each of which is a neighbour of the preceding one).

For any node I of a graph G, the *ascent* $As(I)$ of I is defined as

$$As(I) = \{J; \text{ there exists an increasing path from } J \text{ to } I\},$$

(Observe that $I \in As(I)$ because a path of zero length is an increasing path.)

The *maximal increasing length* $l(L)$ of a nonempty subset L of the node set of the graph G is the length of a longest increasing path in the subgraph of G induced by L. We further set $l(\emptyset) = -1$.

The following hypotheses will be made frequently in this work.

(H 1) $A = D - E - F$ is a Stieltjes matrix such that D is π-diagonal, F is strictly π-upper triangular and $E = F^{\mathrm{T}}$.
(H 2) The π-diagonal entries D_{II} of the matrix D are all irreducible.
(H 3) x is a positive vector such that $Ax \geqq 0$.
(H 4) β and γ are symmetric (0, 1) matrices.

3 Modified incomplete block factorizations

For simplicity, we restrict our analysis to factorizations with no fill-in allowed outside the block diagonal part of a given Stieltjes matrix.

We assume (H 1), (H 3) and (H 4). Let P be the π-diagonal matrix whose entries are computed according to the following algorithm

$$P_{11} = D_{11} + \Delta_{11},$$

for $I=2, 3, \ldots, M$

$$P_{II}=D_{II}-\beta_{II}*\left(\sum_{S=1}^{I-1} F_{SI}^{\mathrm{T}}\, K_{SS}\, F_{SI}\right)-\Omega_{II}+\Delta_{II} \tag{3.1}$$

with $K_{SS}=\gamma_{SS}*P_{SS}^{-1}$ and where Ω_{II} and Δ_{II} are nonnegative p-diagonal matrices. The first one is defined by

$$\Omega_{II}\,\mathbf{x}_I=\left(\sum_{S=1}^{I-1} F_{SI}^{\mathrm{T}}\, P_{SS}^{-1}\, F_{SI}-\beta_{II}*\left(\sum_{S=1}^{I-1} F_{SI}^{\mathrm{T}}\, K_{SS}\, F_{SI}\right)\right)\mathbf{x}_I. \tag{3.2}$$

The second one will be specified subsequently.

The matrix $B=(P-E)\,P^{-1}(P-F)$ will be referred to as the modified block incomplete LU-factorization of A associated with $\mathbf{x}$, β, γ and Δ (or the strategy that determines Δ). The matrices $\Omega=(\Omega_{II})$ with $\Omega_{11}=0$ and Δ are called, respectively, the modification and the perturbation matrices.

From the implementation point of view, the executability of the algorithm should be guaranteed. The following result, borrowed from [25: Lemma 2.1, Theorem 2.1], gives conditions under which the matrices P_{ss} are all nonsingular, together with other relevant properties of the algorithm (3.1)–(3.2).

Theorem 3.1. *Adding to* (H 1)–(H 4) *that for each* $I\in\mathscr{L}=\{1, 2, \ldots, M-1\}$ *there exists some* J, $I<J\leqq M$ *such that* $F_{IJ}\neq 0$, *then any modified block incomplete* LU*-factorization* $B=(P-E)\,P^{-1}(P-F)$ *of* A *associated with* $\mathbf{x}$, β, γ *and any nonnegative* p*-diagonal matrix* Δ *is well defined. Further*

(1) P *is a Stieltjes matrix whose* π*-diagonal entries* P_{II} *are all irreducible, with* $(P-F)\,\mathbf{x}\geqq A\,\mathbf{x}+\Delta\,\mathbf{x}$,
(2) B *is positive definite with* $B\,\mathbf{x}=A\,\mathbf{x}+\Delta\,\mathbf{x}$ *and* $\mathit{offdiag}_{\mathrm{p}}(A-B)\leqq 0$,
(3) $\mathit{offdiag}_{\mathrm{p}}(P)\leqq \mathit{offdiag}_{\mathrm{p}}(D)$,
(4) $(I-P^{-1}F)\,\mathbf{x}\geqq 0$.

According as $\Delta=0$ or not, one speaks of unperturbed or perturbed strategies [8]. The first case corresponds to none other than the popular generalized row sum criterion variant of modified methods. Perturbed modified methods fall into two groups. The so-called static methods [8], whose perturbation matrices are given a priori and the dynamic strategies where the matrix Δ is determined, *according to appropriate criteria, during the factorization.* The first group includes the well known generalized SSOR method by Axelsson [1] (see also [7]), the modified (with overcorrection terms) method of Gustafsson [16], the "locally perturbed" method by Beauwens [6], the shifted incomplete Cholesky method discussed, among others, by Kershaw [18] and Manteuffel [23], the ILU_β method of Wittum [30] (and references quoted therein) as well as the relaxed methods of Axelsson and Lindskog [5]; for the latter methods, the corresponding perturbation matrix is given by

$$\Delta=(1-\omega)\,\Omega \tag{3.3}$$

ω being the relaxation parameter, showing that the so-called *unmodified methods* ($\omega=0$, i.e. $\Delta=\Omega$ [5]) may be viewed as statically *perturbed modified methods.* To our knowledge, the first dynamic algorithm is the one introduced by Axelsson

and Barker in [3: Sect. 7.2]. Other dynamic (pointwise) schemes have been proposed in [2], [8] and [24]. Their extension to block methods has been prevented up to now by the poor development of the conditioning analysis of sparse block factorizations, a situation which has recently improved.

4 Approximate block-matrix factorization strategies

Our algorithms rely on the two following results. The first one is a particular case of [25: Theorem 3.1 and Lemma 3.1(3)] whose origin goes back to the work of Beauwens and Ben Bouzid [10]. The second one is essentially Beauwen's analytical approach (see e.g. [8, 10] and references quoted therein).

Theorem 4.1. *We assume* (H 1) *and* (H 3). *Let* P *be a* π*-diagonal Stieltjes matrix such that* $offdiag_p(P) \leqq offdiag_p(D)$. *Set* $E = F^T$ *and* $B = (P-E)\,P^{-1}(P-F)$. *Assume further that* $(P-F)\,\mathbf{x} \geqq 0$ *with* $((P-F)\,\mathbf{x})_I > 0$ *for* $I \in \mathscr{L} = \{1, 2, \ldots, M-1\}$ *and that* $B\mathbf{x} \geqq (1-\tau_b)\,A\mathbf{x}$ *where*

$$\tau_b = \inf\{t > 0;\ tP\mathbf{x} \geqq F\mathbf{x}\} \tag{4.1}$$

Then

$$\lambda_{max}(B^{-1}A) \leqq 1/(1-\tau_b). \tag{4.2}$$

Corollary 4.2. *If, in addition to the assumptions of Theorem* 4.1, *one has that*

$$((F-E)\,\mathbf{x})_I \leqq (k+l_I+1)\,(B\mathbf{x})_I \quad \text{for } I \in \mathscr{L} \tag{4.3}$$

where l_I *denotes the maximal increasing length of* $As(I)$ *in* $G(F)/\pi$ *and* k *stands for some nonnegative parameter, then*

$$\lambda_{max}(B^{-1}A) \leqq k+l+2 \tag{4.4}$$

with $l = \max_{I \in \mathscr{L}} l_I$.

In applications to discrete finite difference or finite element second order (elliptic) PDE's on rectangular meshes with (average) mesh size parameter h and natural ordering for the gridpoints and line or plane partitionings, one has $l = O(h^{-1})$ so that for the unperturbed strategy ($\Delta = 0$ and $\lambda_{min}(B^{-1}A) = 1$, see e.g. [21]), the above corollary leads, when all "goes right", to $O(h^{-1})$ upper bound on the spectral condition number of $B^{-1}A$ *provided that the condition* (4.3) *is satisfied with* k *(smaller than)* $O(h^{-1})$. The upper bound $1/(1-\tau_b)$ of Theorem 4.1 has been found in [21] and [22] to have the broadest field of application, compared to alternate approaches.

In other cases, i.e. when the unperturbed strategy produces a too large condition number, one may force a selected upper eigenvalue bound to hold by keeping control of the inequalities that ensure its existence, artificially adjusting the diagonal entries of the matrix P if necessary. Such a perturbation procedure is known to induce a decrease (in the case of positive corrections) of the smallest eigenvalues of $B^{-1}A$ (see e.g. [3, 8, 24]). In the present case the decrease of

$\lambda_{\min}(B^{-1}A)$ may be estimated by using the following result which generalizes Theorem 4.3 of [10].

Theorem 4.3. *Let A and B be $n\times n$ π-partitioned SPD matrices, Δ be a $n\times n$ nonnegative* p*-diagonal matrix and* $\mathbf{x}$ *denote a positive n-vector such that*

$$\text{(4.5)} \qquad \textit{offdiag}_{\mathrm{p}}(A-B)\leqq 0$$

and

$$\text{(4.6)} \qquad B\mathbf{x}\leqq A\mathbf{x}+\Delta\mathbf{x}$$

Then

$$\text{(4.7)} \qquad \lambda_{\min}(B^{-1}A)\geqq 1/(1+\xi)$$

with

$$\text{(4.8)} \qquad \xi=\max_{\mathbf{z}\neq 0}\frac{(\mathbf{z},\Delta\mathbf{z})}{(\mathbf{z},A\mathbf{z})}$$

If, in addition to the above assumptions, one has that

(1) *There exists two nonnegative* p*-diagonal matrices $\Delta'=(\Delta'_{ii}\,\delta_{ij})$ and $\Delta''=(\Delta''_{ii}\,\delta_{ij})$ such that $\Delta''_{ii}\neq 0$ only for indices $i\in L\subset\{1,2,\ldots,n\}$, and $\Delta=\Delta'+\Delta''$.*
(2) *There exists a family of symmetric nonnegative definite matrices $(A^{(i)})_{i\in L}$ such that*

$$\text{(4.9)} \qquad \forall\mathbf{z}\in\mathbb{C}^n \qquad \sum_{i\in L}(\mathbf{z},A^{(i)}\mathbf{z})\leqq(\mathbf{z},A\mathbf{z}),$$

and

$$\text{(4.10)} \qquad \forall i\in L,\quad \forall\mathbf{z}\in\mathbb{C}^n \qquad \mathbf{z}_i\neq 0\Rightarrow(\mathbf{z},A^{(i)}\mathbf{z})>0$$

Then

$$\text{(4.11)} \qquad \xi\leqq\frac{\lambda_{\max}(\Delta')}{\lambda_{\min}(A)}+\max_{i\in L}(\Delta''_{ii}\,\gamma_i)\leqq\frac{\lambda_{\max}(\Delta')}{\lambda_{\min}(A)}+\gamma\,\lambda_{\max}(\Delta'')$$

where

$$\text{(4.12)} \qquad \gamma_i=\max_{\mathbf{z}_i\neq 0}\frac{|\mathbf{z}_i|^2}{(\mathbf{z},A^{(i)}\mathbf{z})}$$

and

$$\text{(4.13)} \qquad \gamma=\max_{i\in L}\gamma_i$$

Proof. One has by (4.5) and (4.6) that $\textit{offdiag}_{\mathrm{p}}(A+\Delta-B)\leqq 0$ and $(A+\Delta-B)\,\mathbf{x}\geqq 0$, which implies that $A+\Delta-B$ is nonnegative definite [12]; therefore

$(\mathbf{z}, A\mathbf{z})+(\mathbf{z}, \Delta\mathbf{z})\geqq(\mathbf{z}, B\mathbf{z})$ for any complex vector $\mathbf{z}$, which yields (4.7)–(4.8). Next, for any nonzero complex vector $\mathbf{z}$, one has

$$(\mathbf{z}, \Delta''\mathbf{z})=\sum_{i\in L}\Delta''_{ii}|\mathbf{z}_i|^2\leqq\sum_{i\in L}\Delta''_{ii}\gamma_i(\mathbf{z}, A^{(i)}\mathbf{z})\leqq\max_{i\in L}(\Delta''_{ii}\gamma_i)(\mathbf{z}, A\mathbf{z})$$

$$\leqq\gamma\lambda_{\max}(\Delta'')(\mathbf{z}, A\mathbf{z})$$

whence (4.11). □

To get an $O(h^{-1})$ upper bound on the spectral condition number of $B^{-1}A$ when a given upper spectral bound for $B^{-1}A$ is constrained to be $O(h^{-1})$, the matrix of perturbations should be taken so as to have the lower spectral bound (4.7) independent of h or equivalently, so as to have $\xi\leqq O(1)$. In the case where the matrix A arises from discrete second order elliptic PDEs, assuming that the matrix coefficients are normalized so that the p-diagonal entries are $O(1)$, one has $\lambda_{\min}(A)=O(h^2)$. On the other hand, from the procedure elaborated in [6, pp. 109–112] for the construction of a family of symmetric nonnegative definite matrices $(A^{(i)})_{i\in L}$ that satisfy the requirements (4.9) and (4.10), one has for the parameter γ defined by (4.12) and (4.13) that $\gamma\leqq O(h^{-1})$ whenever there exists a positive vector $\mathbf{y}$ such that $A\mathbf{y}\geqq 0$ and

$$O(\mathrm{Card}(L))\leqq O(\mathrm{Card}(L_0)) \tag{4.14}$$

where $\mathrm{Card}(Q)$ stands for the cardinal of Q while

$$L=\{i;\ 1\leqq i\leqq n,\ \Delta''_{ii}\neq 0\} \tag{4.15}$$

and

$$L_0=\{j;\ 1\leqq j\leqq n,\ (A\mathbf{y})_j=O(1)\} \tag{4.16}$$

In the light of (4.11), it emerges from the above considerations that, to get a $O(1)$ lower spectral bound the perturbation matrix Δ should be constructed in such a way that $\Delta\mathbf{x}$ is $O(h^2)$ at most nodes and $O(h)$ at a limited (from the point of view of (4.14)–(4.16)) number of nodes ($i\in L$). How to achieve this goal is the subject of the strategies to come. Observe that unmodified methods do not meet the above requirements since (see (3.2) and (3.3) with $\omega=0$) $(\Delta\mathbf{x})_i=O(1)$ for $i\in\pi_I$ with $I\geqq 2$.

Modified block incomplete factorization strategies

In addition to (H 1)–(H 4), we assume that $(F\mathbf{x})_I>0$ for $I\in\mathscr{L}=\{1, 2, \ldots, M-1\}$. Therefore, by Theorem 3.1 one has that any modified block incomplete LU-factorization $B=(P-E)P^{-1}(P-F)$ of A associated with $\mathbf{x}$, β, γ and any nonnegative p-diagonal matrix Δ may be carried out; further P is a Stieltjes matrix whose π-diagonal entries P_{II} are all irreducible (therefore $P^{-1}>0$ [12]) and $B\mathbf{x}=A\mathbf{x}+\Delta\mathbf{x}$ or equivalently

$$P\mathbf{x}=A\mathbf{x}+\Delta\mathbf{x}+EP^{-1}(P-F)\mathbf{x}+F\mathbf{x} \tag{4.17}$$

which together with Theorem 3.1(1) and the assumption on $F\mathbf{x}$ implies that

$$P\mathbf{x}>0. \tag{4.18}$$

We first consider the unperturbed strategy because it represents the initial step of the other ones.

Strategy 1 At the I^{th} (block) stage, $\text{diag}(P_{II})$ is computed by means of formulas (3.1)–(3.2) with $\Delta_{II}=0$.

The outline of all the other strategies is as follows: at the I^{th} (block) stage

Step 1: P_{II} is computed as in Strategy 1, i.e. without any perturbations for $\text{diag}_{\text{p}}(P_{II})$,

Step 2: For $i\in\pi_I$, test to decide whether the perturbation must be added to the current diagonal entry of P_{II} is performed and, the perturbation is computed and added if necessary.

For exposition purpose, we denote by P_0 the π-diagonal matrix which represents the intermediate aspect of the matrix P in Step 1. Obviously, one has at the end of Step 2 that

$$(P\mathbf{x})_I=(P_0\,\mathbf{x})_I+(\Delta\mathbf{x})_I \tag{4.19}$$

Strategy 2 Our aim here is to impose an upper bound on the upper bound $1/(1-\tau_{\text{b}})$ of Theorem 4.1. Let α be a small positive parameter, say $0<\alpha<1$. By definition of τ_{b} (Eq. (4.1)), if $(P-F)\,\mathbf{x}\geqq\alpha P\mathbf{x}$ then $1/(1-\tau_{\text{b}})\leqq 1/\alpha$. By (4.18) one has $((P-F)\,\mathbf{x})_M=(P\mathbf{x})_M>\alpha(P\mathbf{x})_M$, which means that perturbations are not needed at nodes $i\in\pi_M$, in other words, $(\Delta\mathbf{x})_M=0$.

For $I\in\mathscr{L}=\{1,2,\ldots,M-1\}$, after performing Step 1, compute for $i\in\pi_I$

$$\alpha^{(i)}=\frac{((P_0-F)\,\mathbf{x})_i}{(P_0\,\mathbf{x})_i}. \tag{4.20}$$

If $\alpha^{(i)}\geqq\alpha$, set $(\Delta\mathbf{x})_i=0$; otherwise the current diagonal entry is modified so as to have $((P-F)\,\mathbf{x})_i=\alpha(P\mathbf{x})_i$, which, by (4.19) and (4.20) amounts to taking

$$(\Delta\mathbf{x})_i=\frac{\alpha-\alpha^{(i)}}{1-\alpha}(P_0\,\mathbf{x})_i. \tag{4.21}$$

Clearly, when $\alpha^{(i)}\geqq\alpha$ at all stages, this strategy reduces to the first one, which certainly happens for problems that are successfully covered by Corollary 4.2, provided that α is chosen such that $\alpha\leqq 1/(k+l+2)$ (k being, *in these particular cases*, the smallest possible nonnegative number such that the condition (4.3) is satisfied).

From an implementation point of view, it is of worth noting that (4.20) and (4.21) may be gathered in

$$\Delta_{ii}=\max\{0,(1-\alpha)^{-1}(F\mathbf{x})_i-(P_0\,\mathbf{x})_i\}\frac{1}{\mathbf{x}_i}\quad\text{for } i\in\pi_I, I\in\mathscr{L} \tag{4.22}$$

The result to follow displays the influence of the parameter α on $\Delta\mathbf{x}$ and therefore, through Theorem 4.3, on the smallest eigenvalue of $B^{-1}A$. It takes its inspiration from [24, Theorem 5.1].

Lemma 4.4. *For $i\in\pi_I$ with $I\in\mathscr{L}=\{1, 2, \ldots, M-1\}$, define*

$$\sigma_i=\frac{\alpha^2}{1-\alpha}(F\mathbf{x})_i+\alpha((F-E)\mathbf{x})_i-(A\mathbf{x})_i \tag{4.23}$$

Then

$$0\leqq(\Delta\mathbf{x})_i\leqq\begin{cases}\max\{\sigma_i, 0\} & \text{if } i\notin\pi_M\\ 0 & \text{otherwise.}\end{cases} \tag{4.24}$$

Proof. The assertion is obvious (by construction) if $i\in\pi_M$; otherwise one has, again by construction, that either $(\Delta\mathbf{x})_i=0$ or $((P-F)\mathbf{x})_i=\alpha(P\mathbf{x})_i$. The last relation may be rewritten as $((P-F)\mathbf{x})_i=\alpha(1-\alpha)^{-1}(F\mathbf{x})_i$. On the other hand, at the I^{th} (block) stage, one has, once again by construction, that $((P-F)\mathbf{x})_J\geqq\alpha(P\mathbf{x})_J$ for $J<I$, whence

$$\begin{aligned}0\leqq(\Delta\mathbf{x})_i&=(B\mathbf{x})_i-(A\mathbf{x})_i\\&=((P-F)\mathbf{x})_i-(EP^{-1}(P-F)\mathbf{x})_i-(A\mathbf{x})_i\\&=\frac{\alpha}{1-\alpha}(F\mathbf{x})_i-\left(\sum_{J=1}^{I-1}E_{IJ}P_{JJ}^{-1}((P-F)\mathbf{x})_J\right)_i-(A\mathbf{x})_i\\&\leqq\frac{\alpha}{1-\alpha}(F\mathbf{x})_i-\alpha(F\mathbf{x})_i+\alpha((F-E)\mathbf{x})_i-(A\mathbf{x})_i\end{aligned}$$

which concludes the proof. □

We are now in position to state our main result concerning the condition number associated with the Strategy 2.

Theorem 4.5. *In addition to* (H 1)–(H 4), *we assume that*

(1) $(F\mathbf{x})_I>0$ *for* $I\in\mathscr{L}=\{1, 2, \ldots, M-1\}$,
(2) *the matrix A arises from a discrete approximation of an elliptic second order PDE, with average mesh size parameter h,*
(3) *B is the modified block incomplete LU-factorization of A associated with* $\mathbf{x}$, β, γ *and the Strategy* 2 *with* $\alpha=O(h)$,
(4) *there exists a positive vector* $\mathbf{y}$ *such that* $A\mathbf{y}\geqq 0$ *and* $O(\mathrm{Card}(L))\leqq O(\mathrm{Card}(L_0))$ *where* $L=\{i, i\notin\pi_M, ((F-E)\mathbf{x})_i=O(1)$ *and* $((\alpha(F-E)-A)\mathbf{x})_i>0\}$ *and* $L_0=\{j; (A\mathbf{y})_j=O(1)\}$.
Then

$$\kappa(B^{-1}A)\leqq O(h^{-1}) \tag{4.25}$$

Proof. One has by construction that $\lambda_{\max}(B^{-1}A)\leqq 1/\alpha$. The conclusion then readily follows by Theorem 4.3 and Lemma 4.4, taking into account the comments that follow (the proof of) Theorem 4.3. □

Strategy 3. Here, we target the analytical upper spectral bound $k+l+2$ of Corollary 4.2. We have now to take care of the fulfilment of the condition (4.3). Since $B\mathbf{x}=A\mathbf{x}+\Delta\mathbf{x}$, it is sufficient to set for $i\in\pi_I$

$$(\Delta\mathbf{x})_i=\begin{cases}0 & \text{if } I=M\\ \max\{(k+l_I+1)^{-1}((F-E)\mathbf{x})_i-(A\mathbf{x})_i,0\} & \text{otherwise.}\end{cases} \tag{4.26}$$

It is important to note here that, for discrete PDEs with gridpoints naturally ordered, $((F-E)\mathbf{x})_i$ is proportional to the difference between neighbouring coefficients of the original equation, say $O(h)$ at interior nodes associated with smooth coefficients variations and $O(1)$ along discontinuities or boundaries. *So, to avoid possible $O(1)$ perturbations (and thereby an undesirable decrease of the lower spectral bound)* at the first (block) components where $l_I\leqq O(1)$, one should use $k=O(h^{-1})$, whence the following result on the spectral condition number associated with this strategy.

Theorem 4.6. *In addition to* (H 1)–(H 4), *we assume that*

(1) $(F\mathbf{x})_I>0$ *for* $I\in\mathscr{L}=\{1,2,\ldots,M-1\}$,
(2) *the matrix A arises from a discrete approximation of an elliptic second order PDE, with average mesh size parameter h,*
(3) *B is the modified block incomplete LU-factorization of A associated with* $\mathbf{x}$, β, γ *and the Strategy* 3 *with* $k=O(h^{-1})$,
(4) *there exists a positive vector* $\mathbf{y}$ *such that* $A\mathbf{y}\geqq 0$ *and* $O(\mathrm{Card}(L))\leqq O(\mathrm{Card}(L_0))$ *where* $L=\{i, i\in\pi_I$ *with* $I\neq M$, $((F-E)\mathbf{x})_i=O(1)$ *and* $((k_I^{-1}(F-E)-A)\mathbf{x})_i>0\}$ *with* $k_I=k+l_I+1$, *and* $L_0=\{j;(A\mathbf{y})_j=O(1)\}$.

Then

$$\kappa(B^{-1}A)\leqq O(h^{-1}) \tag{4.27}$$

5 Numerical results

To illustrate the relative merits of the strategies discussed in the previous section, we report here the results of a few numerical experiments performed on two-dimensional test problems. All the problems are discretized by the five-point finite difference approximation on a rectangular uniform grid of mesh size h with lexicographic ordering for the grid nodes. The resulting linear systems, whose matrices are partitioned according to the line partitioning, are solved by the preconditioned conjugate gradient method with the zero vector as initial approximation and the residual error reduction $\|r^{(i)}\|_2/\|r^{(0)}\|_2\leqq 10^{-m}$ ($r^{(i)}$ is the residual in the i-th iteration) as convergence criterion. The preconditioning matrices are the modified block approximate factorizations of the system matrices associated with $\mathbf{x}=\mathbf{e}$, $\beta=tridiag(\varepsilon)$, $\gamma=tridiag(\varepsilon)$ and the perturbation matrices determined by the strategies 1–3. For comparison purposes, we have also included the unmodified method, referred to in what follows as Strategy 0 since as already noticed, it can also be viewed as a member of the same family. Guided by the analytical upper spectral bounds provided by Corollary 4.2, we have run, at a first time, the Strategy 2 with $\alpha=1/(l+2)$ and the Strategy 3 with $k=l+2$, l being defined as in the stated corollary (note that for block

tridiagonal matrices, as occurred in our examples, $l+2$ is none other than the number M of block components associated with the partitioning) and next with $\alpha = 1/[s(l+2)]$ and $k = s(l+2)$ with s variable.

Problem 1 (from [22])

$$-\frac{\partial}{\partial x}\left(a\frac{\partial}{\partial x}u(x,y)\right)-\frac{\partial}{\partial y}\left(a\frac{\partial}{\partial y}u(x,y)\right)=f(x,y) \qquad \text{in } \Omega=(0,1)\times(0,1)$$

$$u(x,y)=0 \qquad \text{on } \Gamma_0=\{(x,y);\,0\leqq x\leqq 1,\,y=1\}$$

$$\frac{\partial}{\partial n}u(x,y)=0 \qquad \text{on } \partial\Omega\backslash\Gamma_0$$

with

$$a=\begin{cases}0.01 & \text{in } \Omega'=(\frac{1}{2},1)\times(\frac{1}{2},1)\\ 1 & \text{elsewhere}\end{cases}$$

and

$$f(x,y)=\begin{cases}1 & \text{in } \Omega'\\ 0 & \text{elsewhere}\end{cases}$$

Further, $h=1/M$. For convenience, M is assumed to be even.

Problem 2 (from [28])

$$-\frac{\partial}{\partial x}\left(a\frac{\partial}{\partial x}u(x,y)\right)-\frac{\partial}{\partial y}\left(a\frac{\partial}{\partial y}u(x,y)\right)=f(x,y) \qquad \text{in } \Omega=(0,1)\times(0,1)$$

$$u(x,y)=0 \qquad \text{on } \Gamma_0=\{(x,y);\,0\leqq x\leqq 1,\,y=0\}$$

$$\frac{\partial}{\partial n}u(x,y)=0 \qquad \text{on } \partial\Omega\backslash\Gamma_0$$

with

$$a=\begin{cases}100 & \text{in } \Omega'=(\frac{1}{4},\frac{3}{4})\times(\frac{1}{4},\frac{3}{4})\\ 1 & \text{elsewhere}\end{cases}$$

and

$$f(x,y)=\begin{cases}100 & \text{in } \Omega'\\ 0 & \text{elsewhere}\end{cases}$$

Here also $h=1/M$ and, again for convenience, M is chosen as a multiple of 4.

Problem 3

$$-\frac{\partial}{\partial x}\left(a_1\frac{\partial}{\partial x}u(x,y)\right)-\frac{\partial}{\partial y}\left(a_2\frac{\partial}{\partial y}u(x,y)\right)+bu(x,y)=f(x,y) \qquad \text{in } \Omega=(0,1)\times(0,1)$$

$$\frac{\partial}{\partial n}u(x,y)=0 \qquad \text{on } \partial\Omega$$

The domain Ω is subdivided as shown in Fig. 1 which displays also the values of the coefficients a_i, $i=1,2$ and b as well as the right-hand side $f(x,y)$. Now $h=1/(M-1)$. For simplicity, we assume $M-1$ to be a multiple of 5.

We shall compare the different strategies both on the basis of the spectral condition number achieved and on the observed number of PCG iterations needed to reach a given error reduction. The first comparison has the quality of independence with respect to the choices made for the RHS and initial approximation but may fail to correctly reflect the actual convergence behaviour. The second one has the advantage of reflecting all auxiliary effects such as convergence improvement due to eigenvalue clustering or convergence reduction due to round off error effects. This double comparison will prove quite useful and lead us to add a third comparison on the basis of an effective spectral condition number (see below).

Of course, all these comparisons neglect the factorization work which represents at any rate only a very small part of the total work to solve the problem, particularly for large problems. Note on the other hand that we do not need to take care of the computational work per iteration since it is the same for all schemes compared here.

For each problem, let Ω_h denote the discretization grid. For $\mathbf{y}=\mathbf{e}$, the sets L and L_0 required in Theorems 4.5 and 4.6 are given by the set of indices associated with

$$L=\Omega_h\cap\begin{cases}\{(x,y);\,0\leqq x\leqq 1,\,y=0\} & \text{for both Problems 1 and 3}\\ \{(x,y);\,\frac{1}{4}\leqq x\leqq\frac{3}{4},\,y=\frac{1}{4}\} & \text{for Problem 2}\end{cases}$$

$$L_0=\Omega_h\cap\begin{cases}\{(x,y);\,0\leqq x\leqq 1,\,y=1-h\} & \text{for Problem 1}\\ \{(x,y);\,0\leqq x\leqq 1,\,y=h\} & \text{for Problem 2}\\ \emptyset & \text{for Problem 3}\end{cases}$$

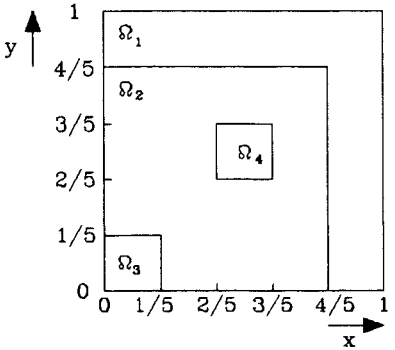
Exception made for Problem 3 where $L_0=\emptyset$ (since $A\mathbf{e}\leqq 0.001h^2\mathbf{e}$), both Card$(L)$ and Card(L_0) are obviously of $O(M)$. We expect therefore $O(1)$ smallest eigenvalues, $O(h^{-1})$ largest eigenvalues and thus $O(h^{-1})$ spectral condition numbers for the preconditioned systems arising from the application of the Strategies 2 and 3 to Problems 1 and 2. This is confirmed by the numerical results gathered in Table 1, but the spectral condition numbers achieved by the Strategies 2 and 3 on Problems 2 and 3 are larger than those obtained with the Strategy 1. On the other hand, the number of PCG iterations required to achieve an error reduction of 10^{-m} are reported in Tables 2–4 for various mesh sizes and various values of m. Also, various values of α (with $1/\alpha=s(l+2)=sM$) for the Strategy 2 and various values of k (with $k=s(l+2)$) for the Strategy 3 have been tested. The results for the largest problems (associated with the smallest mesh sizes) are also illustrated by the convergence histories reported on Figs. 2–4 (with $\alpha=(l+2)/4$ for the Strategy 2 while $k=(l+2)/2$ for the Strategy 3).

From these comparisons, it turns out that the Strategies 2 and 3 behave similarly, require less iterations than Strategy 1 and are not very sensitive to the choice of s.

It is clear that the behaviour of the spectral condition number does not correctly translate the actual behaviour of the preconditioning produced by the Strategies 2 and 3. This comes from the smallest eigenvalue which is strongly

Table 1. Extremal eigenvalues and/or spectral condition number of the preconditioned matrix $B^{-1}A$ associated with the Strategy i ($i=0, 1, 2$ with $\alpha=1/M$, 3 with $k=M$); exponent μ corresponding to the (estimated) asymptotic relationship $\kappa=Ch^{-\mu}$, C denoting a constant; $\upsilon_{min}=\lambda_{min}(B^{-1}A)$, $\upsilon_{max}=\lambda_{max}(B^{-1}A)$ and $\kappa=\kappa(B^{-1}A)$

	Strat. 0	Strat. 1	Strat. 2			Strat. 3		
h^{-1}	κ	κ	υ_{min}	υ_{max}	κ	υ_{min}	υ_{max}	κ
Problem 1								
12	15.9	8.15	0.288	2.152	7.46	0.392	2.590	6.59
24	61.1	33.53	0.292	3.961	13.55	0.383	4.795	12.52
48	242.3	100.8	0.294	8.008	27.25	0.378	10.04	26.59
96	967.2	309.8	0.295	16.20	55.01	0.375	21.05	56.12
192	3866	944.5	0.295	35.23	119.6	0.374	48.25	129.1
μ	2.00	1.61			1.11			1.20
Problem 2								
12	137.9	4.29	0.045	2.356	52.43	0.081	2.777	34.06
24	567.3	13.38	0.045	4.383	97.30	0.081	5.775	71.56
48	2300	51.28	0.045	7.655	169.7	0.080	10.87	135.2
96	9257	150.4	0.045	17.14	379.7	0.080	24.23	301.8
192	37126	456.1	0.045	36.57	810.1	0.080	54.30	676.6
μ	2.00	1.60			1.09			1.16
Problem 3								
10	31198	15.55	$44\ 10^{-6}$	1.950	44018	$58\ 10^{-6}$	2.137	36728
20	$23\ 10^{4}$	111.0	$42\ 10^{-6}$	3.466	82177	$53\ 10^{-6}$	4.141	77635
40	$97\ 10^{4}$	841.2	$41\ 10^{-6}$	7.191	175000	$51\ 10^{-6}$	9.119	179088
80	$40\ 10^{5}$	4133	$41\ 10^{-6}$	14.77	364339	$50\ 10^{-6}$	21.24	427361
160	$16\ 10^{6}$	32235	$40\ 10^{-6}$	31.77	789759	$49\ 10^{-6}$	49.92	10^{6}
μ	2.00	2.96			1.10			1.22



region	a_1	a_2	b	f(x,y)
Ω_1	0.01	0.01	0	0
Ω_2	1	1	0	0
Ω_3	100	100	0	0.1
Ω_4	0.01	1	0.001	0

Fig. 1. Problem 3. Specification of the coefficients and the right hand side of the partial differential equation

Table 2. Number of PCG iterations to achieve $\|r^{(i)}\|_2/\|r^{(0)}\|_2 \leqq 10^{-m}$ for the Strategies 0, 1, 2 with $1/\alpha = sM$ and 3 with $k = M$; $h^{-1} = 48$, 96 for Problems 1 and 2 while $h^{-1} = 40$, 80 for Problem 3

h^{-1}	48								96							
m	3	4	5	6	7	8	9	10	3	4	5	6	7	8	9	10
Problem 1																
Strat. 0	24	27	29	32	34	36	38	40	47	53	58	63	67	70	73	76
Strat. 1	17	20	23	26	30	32	35	38	26	32	37	43	48	53	59	63
Strat. 2																
$S=2$	14	17	19	22	24	26	29	31	21	24	27	31	36	40	43	46
1	14	16	18	20	23	25	27	29	20	23	27	30	33	36	40	43
3/4	14	16	18	20	23	25	27	29	20	23	25	29	33	36	39	43
1/2	14	17	19	21	23	24	26	29	20	23	26	29	32	35	38	41
1/4	15	17	19	21	23	25	27	29	20	24	26	29	31	34	37	40
1/8	18	21	23	25	26	28	30	32	24	27	30	33	35	37	40	42
Strat. 3	14	16	18	21	23	25	28	30	20	23	27	30	34	38	41	45
Problem 2																
Strat. 0	19	21	23	24	26	28	30	31	37	40	44	47	50	53	56	72
Strat. 1	12	15	18	21	23	25	30	31	21	25	29	35	40	44	50	52
Strat. 2																
$S=2$	12	13	16	18	20	23	24	27	18	21	23	27	31	35	38	47
1	11	13	15	17	19	22	24	26	16	18	22	26	29	33	35	39
3/4	11	13	15	17	18	20	23	25	16	19	22	26	28	31	34	44
1/2	12	14	15	17	18	20	22	24	17	19	21	23	27	31	33	36
1/4	12	14	16	18	19	21	22	24	17	19	22	24	27	28	31	34
1/8	14	16	17	19	21	22	24	26	18	20	22	25	26	29	31	40
Strat. 3	11	13	15	18	20	23	24	27	15	20	23	27	30	34	37	48
Problem 3																
h^{-1}	40								80							
m	3	4	5	6	7	8	9	10	3	4	5	6	7	8	9	10
Strat. 0	30	32	34	35	37	45	60	61	57	61	66	68	83	86	111	116
Strat. 1	21	26	30	34	39	43	49	52	41	51	58	67	74	83	94	101
Strat. 2																
$S=2$	20	23	26	28	30	32	46	48	31	35	39	43	47	50	72	75
1	20	22	24	28	30	32	35	48	30	34	37	41	45	47	66	70
3/4	20	22	25	27	28	31	44	46	29	33	37	40	42	61	65	69
1/2	20	22	24	25	29	31	44	46	28	31	35	38	41	61	65	68
1/4	20	22	24	26	28	30	45	46	28	30	33	35	38	42	62	66
1/8	24	26	28	30	32	37	53	54	31	34	36	39	41	66	68	70
Strat. 3	20	23	25	28	30	32	46	48	31	34	38	43	47	66	70	74

Table 3. Number of PCG iterations to achieve $\|r^{(i)}\|_2/\|r^{(0)}\|_2 \leqq 10^{-m}$ for the Strategies 0, 1, 2 with $1/\alpha = sM$ and 3 with $k = M$; $h^{-1} = 192$ for Problems 1 and 2 while $h^{-1} = 160$ for Problem 3

h^{-1}	192							
m	3	4	5	6	7	8	9	10
Problem 1								
Strat. 0	95	109	115	125	131	138	145	168
Strat. 1	44	54	64	76	85	93	104	113
Strat. 2								
$S=2$	32	37	43	49	55	61	69	72
1	30	34	40	44	50	55	60	65
3/4	28	34	39	43	48	53	58	64
1/2	28	33	39	43	47	51	56	60
1/4	29	33	37	40	44	47	51	63
1/8	33	37	41	44	48	51	54	66
Strat. 3	31	36	41	47	53	57	64	68
Problem 2								
Strat. 0	73	78	85	90	97	103	135	146
Strat. 1	37	44	54	62	70	80	87	96
Strat. 2								
$S=2$	28	32	38	44	49	55	60	76
1	23	28	35	40	45	50	54	60
3/4	23	28	34	38	43	47	52	65
1/2	25	28	32	35	40	45	50	63
1/4	25	27	31	35	38	41	44	59
1/8	25	27	31	34	38	41	43	56
Strat. 3	26	29	37	44	47	55	59	75
Problem 3								
h^{-1}	160							
m	3	4	5	6	7	8	9	10
Strat. 0	112	121	128	158	165	217	220	229
Strat. 1	90	109	129	150	170	192	209	228
Strat. 2								
$S=2$	47	53	60	65	100	107	113	119
1	44	49	54	61	67	95	102	106
3/4	42	49	55	58	62	93	97	102
1/2	41	45	50	56	85	89	93	98
1/4	39	43	46	50	55	86	89	92
1/8	43	47	50	54	63	90	93	95
Strat. 3	46	52	59	64	98	105	109	116

Table 4. Number of PCG iterations to achieve $\|r^{(i)}\|_2/\|r^{(0)}\|_2 \leqq 10^{-m}$ for the Strategy 3 with $k=sM$; $h^{-1}=192$ for Problems 1 and 2 while $h^{-1}=160$ for Problem 3

h^{-1}	192							
m	3	4	5	6	7	8	9	10
Problem 1								
$S=3/2$	32	37	43	48	55	61	66	72
1	31	36	41	47	53	57	64	68
3/4	30	35	39	46	52	58	64	68
1/2	31	36	41	46	51	56	61	65
1/4	32	36	40	45	51	56	61	66
1/8	33	39	44	48	52	56	62	69
Problem 2								
$S=3/2$	26	31	38	43	50	54	61	76
1	26	29	37	44	47	55	59	75
3/4	25	29	36	41	47	54	58	73
1/2	24	29	36	41	46	53	67	73
1/4	24	30	36	42	46	51	65	70
1/8	25	31	37	41	45	50	55	72
Problem 3								
h^{-1}	160							
m	3	4	5	6	7	8	9	10
$S=3/2$	48	55	60	67	74	108	115	121
1	46	52	59	64	98	105	109	116
3/4	46	51	58	63	68	101	107	112
1/2	45	51	56	62	68	101	107	113
1/4	46	52	57	64	101	106	110	115
1/8	53	58	63	68	74	111	116	123

isolated from the other ones. In such a situation, the convergence of the PCG process is governed by the so-called effective (or reduced) spectral condition number, i.e. the spectral condition number determined by the spectrum of the preconditioned system deprived of some of its isolated members (see e.g. [3] and references cited therein). By way of illustration, we give in Table 5 the smallest and the largest four eigenvalues as well as the effective spectral condition number (only the smallest eigenvalue is discarded) associated with the Strategy 2 with $s=1$. Table 6 contains the largest four eigenvalues determined by the Strategy 1, say the unperturbed modified method for which the smallest eigenvalues are known to cluster at 1 [5, 13]. A considerable reduction in the effective spectral condition number is observed, even for Problem 3 which is outside the scope of Theorems 4.5 and 4.6.

Finally, the results reported with various values of α and k with $1/\alpha=s(l+2)$ and $k=s(l+2)$ lead to recommend to choose s around 1/4 for the Strategy 2

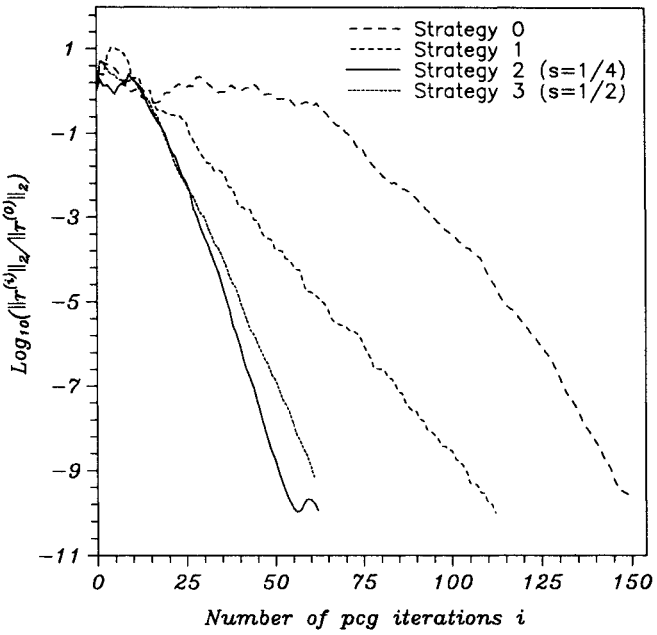


Fig. 2. Error reduction for Problem 1 ($h^{-1} = 192$)

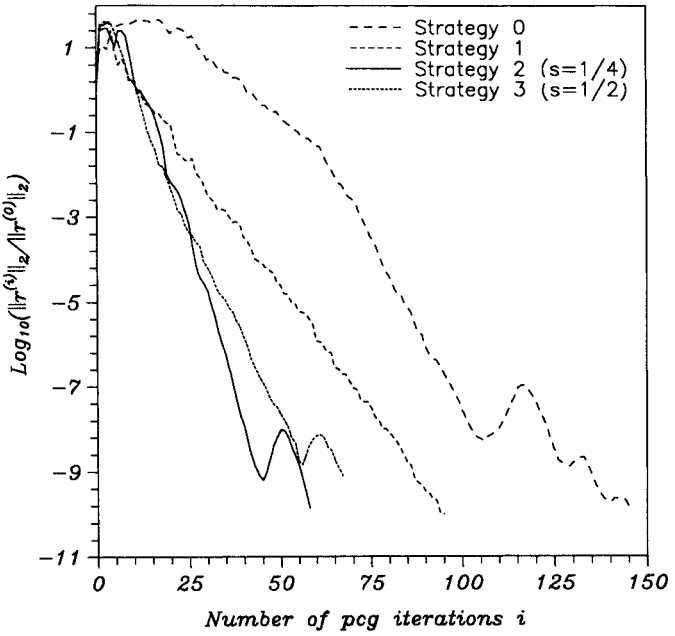


Fig. 3. Error reduction for Problem 2 ($h^{-1} = 192$)

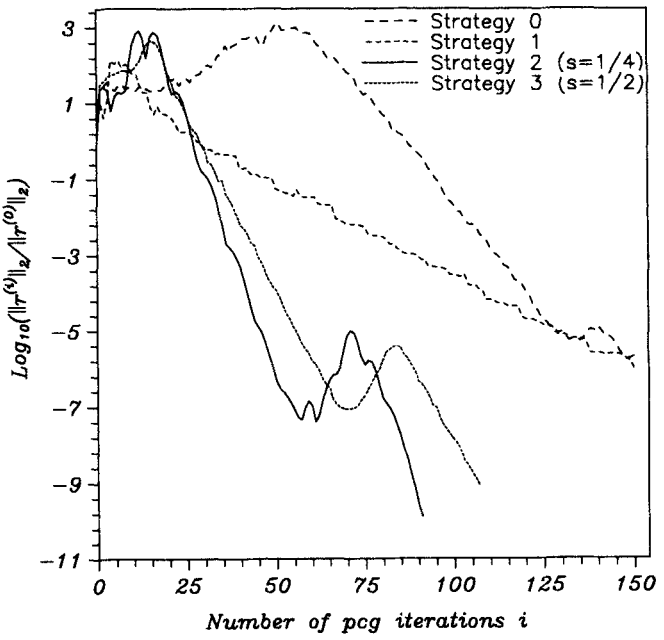


Fig. 4. Error reduction for Problem 3 ($h^{-1} = 160$)

Table 5. Distribution of extremal eigenvalues, spectral condition number (κ) and effective spectral condition number (κ_e) of the matrix $B^{-1}A$ for the Strategy 2 with $\alpha = 1/M$

h^{-1}	Smallest eigenvalues				Largest eigenvalues				κ	κ_e
Problem 1										
12	0.2883	0.8001	0.8870	0.9175	1.4464	1.7357	1.7722	2.1515	7.46	2.69
24	0.2923	0.8435	0.9179	0.9336	2.4150	2.7877	3.3983	3.9611	13.55	4.70
48	0.2939	0.8719	0.9316	0.9411	5.4271	5.6347	6.5622	8.0081	27.25	9.18
96	0.2945	0.8887	0.9297	0.9561	11.190	12.413	14.811	16.203	55.01	18.23
192	0.2947	0.8980	0.9283	0.9649	25.162	29.173	33.586	35.233	119.6	39.24
Problem 2										
12	0.0449	0.8957	0.9363	0.9396	1.5425	1.8278	1.8321	2.3556	52.43	2.63
24	0.0451	0.9251	0.9346	0.9557	3.0915	3.1846	3.2721	4.3833	97.30	4.74
48	0.0451	0.9335	0.9466	0.9701	6.0837	6.1324	7.3066	7.6545	169.7	8.20
96	0.0451	0.9330	0.9602	0.9798	12.642	12.702	14.561	17.138	379.7	18.37
192	0.0451	0.9326	0.9683	0.9860	28.780	30.996	32.494	36.567	810.1	39.21
Problem 3										
10	$44\,10^{-6}$	0.7475	0.9060	0.9113	1.3855	1.5732	1.7036	1.9497	44018	2.61
20	$42\,20^{-6}$	0.7878	0.9270	0.9309	2.5506	2.7338	2.9748	3.4661	82177	4.40
40	$41\,10^{-6}$	0.8173	0.9372	0.9473	4.6558	5.6061	6.0988	7.1905	175000	8.80
80	$41\,10^{-6}$	0.8348	0.9245	0.9619	10.882	12.311	12.690	14.765	364339	17.69
160	$40\,10^{-6}$	0.8430	0.9143	0.9712	23.604	25.743	28.145	31.765	789759	37.68

Table 6. Largest eigenvalues of the preconditioned matrix $B^{-1}A$ associated with the Strategy 1

h^{-1}	Problem 1				Problem 2				h^{-1}	Problem 3			
12	2.16	2.57	2.85	8.15	1.81	2.61	2.72	4.29	10	2.36	3.55	4.12	15.6
24	3.67	5.00	7.84	33.5	4.02	4.80	5.54	13.4	20	4.68	6.95	15.4	111
48	8.39	13.3	44.3	101	8.78	8.83	18.6	51.3	40	17.7	33.2	130	841
96	37.9	102	164	309	21.7	44.3	109	150	80	117	382	1459	4133
192	209	394	757	945	167	246	335	456	160	1581	4290	6580	32235

and s around 1/2 for the Strategy 3 *independently of the problem considered*, stressing that variations from $s=1/8$ to $s=1$ have only a weak influence on the number of PCG iterations.

6 Conclusions

Both theoretical and experimental results display the improvement of the convergence rate of the (block) PCG method brought about by the use of modulated corrections to reduce the upper eigenvalue bounds of (block) preconditioned systems. The theoretical results have shown that a $O(h^{-1})$ spectral condition number could be achieved for a large class of problems. The experimental results have further shown that, if the perturbations required by our strategies do induce a reduction of the smallest eigenvalue, they also have the merit to isolate it, thereby inducing an important auxiliary improvement of the convergence rate and extending the scope of these strategies to a class of quasi-singular problems not covered by the theoretical approach.

Acknowledgements. The author would like to thank Professor R. Beauwens for fruitful suggestions and for his help in correcting the english manuscript, and to express his deep gratitude to the referee whose careful reading resulted in an improvement in the readability of this paper. Thanks are also due to Dr. Y. Notay for useful comments.

References

1. Axelsson, O. (1972): A generalized SSOR method. BIT, **13**, 443–467
2. Axelsson, O. (1989): On the eigenvalue distribution of relaxed incomplete factorization methods and the rate of convergence of preconditioned conjugate gradient method. XV National summer school on Application of Mathematics in Engineering, August 23–31, 1989, Varna, Bulgaria
3. Axelsson, O., Barker, V.A.: Finite Element Solutions of Boundary Value Problems: Theory and Computation. Academic Press, New York, 1984
4. Axelsson, O., Eijkhout, V. (1989): Vectorizable preconditioners for elliptic difference equations in three space dimensions. J. Comput. Appl. Math. **27**, 299–321
5. Axelsson, O., Lindskog, G. (1986): On the eigenvalue distribution of a class of preconditioning methods. Numer. Math. **48**, 479–498
6. Beauwens, R. (1987): Lower eigenvalue bounds for pencils of matrices. Lin. Alg. Appl. **85**, 101–119
7. Beauwens, R. (1985): On Axelssons' perturbations. Lin. Alg. Appl. **68**, 221–242
8. Beauwens, R. (1990): Modified incomplete factorization strategies. In: Axelsson, O., Kolotilina, L. eds., Preconditioned Conjugate Gradient Methods. Lectures Notes in Mathematics No. 1457, Springer, Berlin Heidelberg New York, pp. 1–16

9. Beauwens, R., Ben Bouzid, M. (1987): On sparse block factorization iterative methods. SIAM J. Numer. Anal. **24**, 1066–1076
10. Beauwens, R., Ben Bouzid, M. (1988): Existence and conditioning properties of sparse approximate block factorizations. SIAM J. Numer. Anal. **25**, 941–956
11. Beauwens, R., Wilmet, R. (1989): Conditioning analysis of positive definite matrices by approximate factorizations. J. Comput. Appl. Math. **26**, 257–269
12. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York, 1979
13. Concus, P., Golub, G.H., Meurant, G. (1985): Block preconditioning for the conjugate gradient method. SIAM J. Sci. Statist. Comput. **6**, 220–252
14. George, A., Liu, J.W. (1981): Computer Solution of Large Sparse Positive Definite Systems. Prentice-Hall, Englewood Cliffs, N.J.
15. Goovaerts, D. (1990): Domain Decomposition Methods for Elliptic Partial Differential Equations. PhD Thesis, Katholieke Universiteit Leuven, Department of Computer Science, Leuven, Belgium
16. Gustafsson, I. (1983): Modified incomplete Cholesky (MIC) Methods. In: Evans, D.J. ed., Preconditioning Methods, Theory and Applications. Gordon and Breach, New York London Paris, pp. 265–293
17. Harary, F. (1969): Graph Theory. Addison-Wesley, Reading
18. Kershaw, D.S. (1978): The incomplete Choleski-conjugate gradient method for the iterative solution of systems of linear equations. J. Comput. Phys. **26**, 43–65
19. Kettler, R. (1987): Linear multigrid method for numerical reservoir simulation. Ph. D. Thesis, University of Technology, Dept. of Technical Math. and Informatics, Delft
20. Magolu, M.M. (1991): Conditioning analysis of sparse block approximate factorizations. Appl. Numer. Math. **8**, 25–42
21. Magolu, M.M.: Analytical bounds for block approximate factorization methods. Lin. Alg. Appl. (submitted)
22. Magolu, M.M., Notay, Y. (1991): On the conditioning analysis of block approximate factorization methods. Lin. Alg. Appl. **154–156**, 583–599
23. Manteuffel, T.A. (1980): An incomplete factorization technique for positive definite linear systems. Math. Comput. **34**, 473–497
24. Notay, Y. (1989): Incomplete factorizations of singular linear systems. BIT **29**, 682–702
25. Notay, Y. (1991): Conditioning analysis of modified block incomplete factorizations. Lin. Alg. Appl. **154–156**, 711–722
26. Sonneveld, P. (1989): A fast Lanczos type solver for nonsymmetric linear system. SIAM J. Sci. Statist. Comput. **10**, 36–52
27. Van der Sluis, A., Van der Vorst, H.A. (1986): The rate of convergence of conjugate gradients. Numer. Math. **48**, 543–560
28. Van der Vorst, H.A. (1990): The convergence behaviour of preconditioned CG and CG-S. In: Axelsson, O., Kolotilina, L. eds., Preconditioned Conjugate Gradient Methods. Lectures Notes in Mathematics No. 1457. Springer, Berlin Heidelberg New York, pp. 126–136
29. Varga, R.S. (1962): Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, N.J.
30. Wittum, G. (1989): On the robustness of ILU-smoothing. SIAM J. Sci. Stat. Comput. **10**, 699–717
31. Wittum, G. (1989): Linear iterations as smoothers in Multigrid methods: Theory with applications to incomplete decompositions. Impact Comput. Sci. Engrg. **1**, 180–215