Lower eigenvalue bounds for singular pencils of matrices

Magolu monga-Made *

Service de Métrologie Nucléaire, C.P. 165, Université Libre de Bruxelles, 50 avenue F.D. Roosevelt,
B-1050 Bruxelles, Belgium

Received 1 October 1990
Revised 15 March 1991

Abstract
Magolu, M.-M., Lower eigenvalue bounds for singular pencils of matrices, Journal of Computational and

Beauwens' procedure for obtaining lower eigenvalue bounds for (regular) pencils of matrices $A - \gamma B$ is
simplified and extended to the singular case. The theory is then compared, through a particular perturbed
modified incomplete factorization, with Notay's generalization of another approach, initiated by Gustafsson,
and developed by Axelsson and Barker and by Wilmet.

Keywords: Polynomiably accelerated iterative methods, preconditioning, treediagonal matrices, incomplete

1. Introduction

The present work is concerned with the a priori analysis of the convergence behaviour of
currently accelerated iterative methods for solving large and sparse preconditioned (by
incomplete factorization) positive semi-definite linear systems of algebraic equations. Relevant
features of incomplete factorizations under consideration will be recalled in Section 3 below.
We refer to [2–11,15,16,19,25,30–32,34] for more detailed information in the regular case and
to [18,26–29] for the singular case.

Such analyses rest primarily on the determination of a priori lower and upper bounds for the
positive eigenvalues of the matrix $B^+A$ where $A$ is a (Stieltjes) matrix (the finite-difference or
the finite-element matrix of the system under consideration or some spectrally equivalent [4]
one), while $B^+$ denotes the Moore–Penrose inverse [12] of its approximate factorization $B$.

We deal here only with lower bounds for which two techniques have been developed. The
first one, introduced by Gustafsson [15,16], and successively investigated by Axelsson and Barker [4], Wilmet [33] and Notay [27] covers, under its most recent version [27, Theorem 4.1],

* Research supported by the A.B.O.S. (A.G.C.D) under project 11, within the cooperation between Belgium and
Zaire.
both the regular and the singular cases. The other technique introduced by Beauwens [8] was found more accurate by Wilmet [33] but it covers only the regular case. Our main purpose is to reformulate it under a form suitable for both the regular and the singular case and to compare its accuracy with Notay's version of the first technique. The incomplete factorization we shall consider to perform our comparison may be seen as the point version of a particular modified block approximate factorization (with additive modulated corrections) technique investigated in [23, Section 4: Strategy 3]. Recent developments of upper bound theories may be found in [3,9-11,27-29].

The outline of the paper is as follows. Terminology and notation are defined in Section 2. Relevant features of modified incomplete factorization algorithms are summarized in Section 3. Our extension of Beauwens' technique is described in Section 4 and applied to specific examples in Section 5. In Section 6, numerical results are presented and commented on.

2. General terminology and notation

We write $A^t$, $A^+$, $N(A)$, $\sigma(A)$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote, respectively, the transpose, the Moore–Penrose inverse [12], the null space, the spectrum, the smallest and the largest eigenvalues of the matrix $A$.

The order relation between real matrices and vectors (of the same dimension) is the usual componentwise order: if $A = (a_{ij})$ and $B = (b_{ij})$, then $A < B$ ($A < B$) if $a_{ij} < b_{ij}$ ($a_{ij} < b_{ij}$) for all $i, j$; $A$ is called nonnegative (positive) if $A \geq 0$ ($A > 0$). If $A = (a_{ij})$, we denote by $\text{diag}(A)$ the diagonal matrix whose diagonal entries coincide with those of $A$ and we let $\text{offdiag}(A) = A - \text{diag}(A)$. By $e$ we denote the vector with all components equal to unity.

Graph notions

All graph concepts used in this work refer to ordered undirected graphs [14,17]. For the sake of easy reference we recall from [1,11] the following uncommonly used notations.

An increasing path in a graph $G$ is a path $i_0, i_1, i_2, \ldots, i_l$ such that $i_0 < i_1 < i_2 < \cdots < i_l$.

For any node $i$ of a graph $G$, the ascent $\text{As}(i)$ and the descent $\text{Ds}(i)$ of $i$ are defined by

$$\text{As}(i) = \{j; \text{ there exists an increasing path from } j \text{ to } i\}.$$  
$\text{Ds}(i) = \{j; \text{ there exists an increasing path from } i \text{ to } j\}.$

Observe that $i \in \text{As}(i)$ and $i \in \text{Ds}(i)$ because a path of zero length is an increasing path.

For any nonempty subset $M$ of the node set of a graph $G$, the set of the ascents (descents) of the nodes of $M$ is denoted by $\text{As}(M)$ ($\text{Ds}(M)$). We further set $\text{As}(\emptyset) = \text{Ds}(\emptyset) = \emptyset$.

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

A node $j$ of a graph $G$ is called precursor (successor) of another node $i$ if $(i, j)$ is an edge of $G$ with $j < i$ ($i < j$). The set of precursors (successors) of $i$ is denoted by $P(i)$ ($S(i)$). If $M$ is a nonempty subset of the node set of $G$, the set of precursors (successors) of the nodes of $M$ is denoted by $P(M)$ ($S(M)$). We further set $P(\emptyset) = S(\emptyset) = \emptyset$.

In a graph $G$, any node $i$ such that $P(i) = \emptyset$ is called an initial node or a root of $G$. 
For any pair of nodes $i$ and $j$ ($i \neq j$) of a graph $G$, we denote by
\[ \text{Pc}(i, j) = P(i) \cap P(j) \]
their set of common precursors; we further define
\[ \text{Pc}(G) = \bigcup_{i \neq j} \text{Pc}(i, j) , \]
if $G$ is the graph of a matrix $A$; we also write $\text{Pc}(A)$ for $\text{Pc}(G)$.

A graph $G$ is called a tree if it is connected and acyclic (i.e., it has no cycles [17]) or equivalently if every two nodes of $G$ are joined by a unique path [1].

A rooted tree is a tree in which one node is distinguished as the unique root. In such a tree, precursors (successors) are also called fathers (sons).

A spanning tree in a graph $G$ is a tree (subgraph of $G$) whose node set coincides with that of $G$.

A treediegonal matrix is a matrix whose graph is a tree. Tridiagonal matrices are a particular case of treediegonal matrices [20].

3. Modified incomplete factorizations with additive corrections

We shall restrict our attention to the case of (singular) Stieltjes matrices. The reason is that, in practice, a discretization by finite differences or by finite elements of self-adjoint PDEs gives rise, in general, to matrices that are either Stieltjes or spectrally equivalent to Stieltjes matrices [4,11,16,27,29]. As specified in the title of the section, our analysis is only concerned with the so-called (perturbed) modified incomplete factorization. We briefly recall (for more details we refer to the above-mentioned papers) that the offdiagonal entries of the upper triangular factor $U$ of a modified incomplete factorization
\[ B = U^t P^{-1} U^t , \quad \text{with} \quad P = \text{diag}(U) , \] (3.1)
of a (singular) Stieltjes matrix $A$, are computed as in “unmodified” incomplete factorization (see, e.g., [5,18,32]), while its diagonal entries are determined so as to satisfy a generalized row sum relation of the form
\[ Bx = Ax + \Delta x , \] (3.2)
where $x$ denotes a positive vector such that $Ax \geq 0$, while $\Delta$ stands for a nonnegative diagonal matrix often referred to as the perturbation (or correction) matrix [4,10,16,29].

Letting $\Delta = 0$ results in the standard (unperturbed) modified incomplete factorization. Guided by various motivations, different ways for choosing (explicitly or implicitly) $\Delta$ have been followed in the literature ([2-5,7-11,15,16,19,25,27-29,34] and references cited therein). In the case one intends to use the matrix $B$ as preconditioner in a polynomially accelerated iterative process, the choice of $\Delta$ should be done so as to minimize, within given constraints, the spectral condition number (i.e., the ratio of the largest to the smallest nonzero eigenvalues) of the pencil of matrices $A - \gamma B$.

As concerns the unperturbed methods, 1 is known to be the smallest nonzero eigenvalue of $A - \gamma B$ (see, e.g., [6,27]), while the existence of useful upper spectral bounds is subject to certain conditions: the upper triangular factor $U$ must be “sufficiently” strictly diagonally dominant or both diagonally dominant and “$S/P$ consistently ordered” (see [9,10,28,29]). The
quality of such upper bounds depends on the extent to which the matrix $U$ complies with the required conditions. In less favourable situations, perturbations should be called in to reduce the upper bound to a given order of magnitude, often $O(\tau^{-1})$ for discrete PDEs with (average) mesh parameter $\tau$; the resulting spectral condition number is $O(h^{-1})$ provided that the nonzero smallest eigenvalue remains $O(1)$. In the regular case, two general techniques for estimating the smallest eigenvalue have been worked out. The first one was initiated by Gustafsson [15], extensively developed by Axelsson and Barker [4] and by Wilmet [33], and improved and extended to the singular case by Notay [27]. The other one was elaborated by Beauwens [8]. In the next section, we aim at generalizing Beauwens’ procedure in order to also cover the singular case.

4. Lower eigenvalue bounds

4.1. Theoretical framework

The following results describe the theoretical framework of our analysis. The first one essentially defines the common kernel of Notay’s and Beauwens’ approaches. The next one departs from the latter approach which it actually simplifies and generalizes.

Theorem 4.1. Let $A$ and $B$ be $n \times n$ real symmetric nonnegative definite matrices, $x$ a positive vector and $\Delta$ a nonnegative diagonal matrix such that

1. $\text{offdiag}(A - B) \leq 0$,
2. $Bx \leq Ax + \Delta x$.

Then for every complex $n$-vector $z$,

$$ (z, Bz) \leq (z, Az) + (z, \Delta z). \quad (4.1) $$

If moreover,

$$ N(B) \subset N(A) \quad (4.2) $$

and

$$ \max_{z \neq 0} \frac{(z, \Delta z)}{(z, Az)} \leq \xi, \quad (4.3) $$

then

$$ \lambda_{\min}(B^+A) = \min_{v \in \sigma(B^+A), v \neq 0} v \geq \frac{1}{1 + \xi}. \quad (4.4) $$

Proof. One has by the assumptions (1) and (2) that $(A + \Delta - B)x \geq 0$ and $\text{offdiag}(A + \Delta - B) \leq 0$, which implies that $A + \Delta - B$ is a Stieltjes matrix (see [13] or [27, Theorem 2.1 (3)]), hence nonnegative definite; (4.1) then readily follows. Next, for any $z \neq 0$, $z \in N(A)^\perp$, one has

$$ (z, Bz) \leq \left(1 + \frac{(z, \Delta z)}{(z, Az)}\right)(z, Az) < (1 + \xi)(z, Az), $$

which, together with (4.2) and [26, Theorem 3.1 (5)] yields (4.4). □
Theorem 4.2. Let $A$ and $\Delta$ be $n \times n$ real matrices such that $A$ is symmetric and nonnegative definite with positive diagonal entries, while $\Delta$ is diagonal and nonnegative. Let $L$ be a subset of $\{1, 2, \ldots, n\}$. Set $D = \text{diag}(A)$. Assume further that

1. there exist two nonnegative diagonal matrices $\Delta' = (\Delta'_{ii})$ and $\Delta'' = (\Delta''_{ii})$ such that $\Delta''_{ii} = 0$ for all indices $i \notin L$ and $\Delta = \Delta' + \Delta''$;
2. there exists a family of symmetric nonnegative definite matrices $(A_i)_{i \in L}$ such that

\[ N(A_i) = N(A) \quad (4.5) \]

and

\[ \forall z \in \mathbb{C}^n: \sum_{i \in L} (z, A_i z) \leq (z, A z). \quad (4.6) \]

Then

\[ \max_{z \in N(A)^\perp} \frac{(z, \Delta z)}{(z, A z)} \leq \frac{\lambda_{\text{max}}(D')}{\lambda_{\text{min}}(D^{-1}A)} + \max_{i \in L} (\Delta''_{ii} \gamma_i), \quad (4.7) \]

where

\[ \gamma_i = \max_{z \in N(A)^\perp} \frac{|z_i|^2}{(z, A_i z)} \quad (4.8) \]

and

\[ \lambda_{\text{min}}(D^{-1}A) = \min_{v \in \sigma(D^{-1}A)} v, \quad (4.9) \]

while $D'$ is the diagonal matrix whose diagonal entries $a'_{ii}$ are defined by

\[ a'_{ii} = \frac{\Delta'_{ii}}{a_{ii}}, \quad \text{for } i = 1, 2, \ldots, n. \quad (4.10) \]

Proof. For any nonzero complex $n$-vector $z \in N(A)^\perp$ one has

\[ (z, \Delta z) \leq \lambda_{\text{max}}(D')(z, Dz) + \sum_{i \in L} \Delta''_{ii} |z_i|^2 \leq \lambda_{\text{max}}(D')(z, Dz) + \sum_{i \in L} \Delta''_{ii} \gamma_i (z, A_i z) \leq \lambda_{\text{max}}(D')(z, Dz) + \max_{i \in L} (\Delta''_{ii} \gamma_i)(z, A z). \]

On the other hand, one has obviously that $N(D) \subset N(A)$. The conclusion readily follows by applying [26, Theorem 3.1 (5)] to the pencil of matrices $A - \gamma D$. \(\square\)

The combined use of the above two theorems leads for $\sigma(B^+ A) \setminus \{0\}$ to the following lower bound:

\[ \lambda_{\text{min}}(B^+ A) \geq \left(1 + \frac{\lambda_{\text{max}}(D')}{\lambda_{\text{min}}(D^{-1}A)} + \max_{i \in L} (\Delta''_{ii} \gamma_i)\right)^{-1}, \quad (4.11) \]
which generalizes some known results [6-8] and provides an alternative approach to tackle the so-called “dynamic” methods where the perturbation matrix $A$ is implicitly determined during the factorization process according to appropriate criteria [4, Section 7.2], [10,22-24,27,29]. Since for discrete I'LDEs the order of magnitude of $\lambda_{\min}(D^{-1}A)$ is in general known, the remaining major problem for the lower bound (4.11) is to determine a family of symmetric nonnegative definite matrices that satisfy the conditions (4.5) and (4.6) and for which the parameters defined in (4.8) are easy to compute.

There is no loss of generality to assume at this stage that $A$ is irreducible since a reducible symmetric matrix may be permuted to block diagonal form with irreducible or zero diagonal blocks. On the other hand, we are primarily concerned with the case where $A$ is a Stieltjes matrix and, as shown in the Appendix, the involved parameters can easily be computed when the matrices $A_i$ are (suitable) treediagonal Stieltjes matrices. A practical construction procedure based on these remarks will now be described.

4.2. Practical determination of the matrices $A_i$

As in [8], we first proceed with a step-by-step description and check afterwards that any family $\{A_i\}_{i \in L}$ obtained in this way fulfils the required conditions. We assume that $A$ is an $n \times n$ irreducible Stieltjes matrix, $x$ is a positive vector such that $Ax \geq 0$ and $L$ is a subset of $\{1, 2, ..., n\}$.

\textbf{Step 1.} Set $J = \{j; (Ax)_j > 0\}$. Note that $J \neq \emptyset$ if and only if $A$ is nonsingular [13].

\textbf{Step 2.} For each $i \in L$, choose a graph $G(A_i)$ subject to the restriction that it must be a spanning tree for $G(A)$. If $J \neq \emptyset$, define a mapping $g$ from $L$ into $J$. Set $n_j = \text{Card}(g^{-1}(j))$ for $j \in J$, where Card($H$) denotes the cardinal number of $H$.

\textbf{Step 3.} For each $i \in L$, choose arbitrarily the nonzero offdiagonal entries $a_{kl}^{(i)}$ of $A_i$, subject only to the following constraints:

\begin{equation}
    a_{kl}^{(i)} = a_{kl}^{(i)} < 0, \tag{4.12}
\end{equation}

\begin{equation}
    \sum_{i \in L} a_{kl}^{(i)} \geq a_{kl}, \tag{4.13}
\end{equation}

\textbf{Step 4.} For each $i \in L$, determine the diagonal entries $a_{kk}^{(i)}$ of $A_i$ according to the formulas

\begin{equation}
    a_{kk}^{(i)} = \begin{cases} 
    - \sum_{l \neq k} a_{kl}^{(i)} \frac{x_l}{x_k}, & \text{if } J = \emptyset \text{ or if } J \neq \emptyset \text{ and } k \neq g(i), \\
    - \sum_{l \neq k} a_{kl}^{(i)} \frac{x_l}{x_k} + \frac{1}{n_k x_k} (Ax)_k, & \text{if } J \neq \emptyset \text{ and } k = g(i). 
\end{cases} \tag{4.14}
\end{equation}

\textbf{Lemma 4.3.} Let $A$ be an $n \times n$ irreducible Stieltjes matrix, $L$ be a subset of $\{1, 2, ..., n\}$, $x$ be a positive vector such that $Ax \geq 0$ and $(A_i)_{i \in L}$ be a family of matrices determined according to the above construction procedure. Then

1. for all $i \in L$, $A_i$ is an irreducible Stieltjes matrix;
2. for all $i \in L$, $N(A_i) = N(A);
3. for every complex $n$-vector $z$, $\sum_{i \in L} (z, A_i z) \leq (z, Az)$. 

Proof. From Step 4 one easily establishes that, for \( i \in L \),

\[
(A_i, x)_k = \begin{cases} 
0, & \text{if } J = \emptyset \text{ or if } J \neq \emptyset \text{ and } k \neq g(i), \\
\frac{1}{n_k} (Ax)_k, & \text{if } J \neq \emptyset \text{ and } k = g(i),
\end{cases}
\]

whence \( A_i x \geq 0 \), which combined with (4.12) and the fact that \( A_i \) is a treediagonal matrix (by construction, see Step 2) implies that \( A_i \) is an irreducible Stieltjes matrix. Next, by (4.15) and [27, Theorem 2.1 (a)-(c)] (or [13]) one has for fixed \( i \in L \) that \( A \) and \( A_i \) are simultaneously either nonsingular or singular with \( N(A) = N(A_i) = \text{Span}(x) \). Finally, with \( A_0 = A - \sum_{i \in L} A_i \) one readily checks that offdiag\( (A_0) \leq 0 \) (by (4.13)) and that

\[
(A_0, x)_k = \begin{cases} 
(Ax)_k, & \text{if } J = \emptyset \text{ or if } J \neq \emptyset \text{ but for all } i \in L, g(i) \neq k, \\
0, & \text{otherwise},
\end{cases}
\]

whence it follows that \( A_0 \) is nonnegative definite, which yields (3). \( \square \)

5. Examples of application

To illustrate the results that can be obtained in practice by the technique developed in Section 4, we now examine two typical problems we shall also analyze by Notay's approach [27]. We consider the system matrices deduced from the five-point central difference approximation of the two-dimensional PDEs (we use a uniform grid of mesh size \( h \) and the lexicographic ordering in the \((x, y)\)-plane for the gridpoints):

\[
\begin{align*}
-a \Delta u(x, y) &= f(x, y), & \text{in } \Omega = (0, 1) \times (0, 1), \\
u(x, y) &= 0, & \text{on } \Gamma_0, \\
\frac{\partial}{\partial n} u(x, y) &= 0, & \text{on } \partial \Omega \setminus \Gamma_0,
\end{align*}
\]

with

\[
a = \begin{cases} 
1, & \text{in } \Omega' = (0, \frac{1}{2}) \times (0, \frac{1}{2}), \\
1, & \text{elsewhere}
\end{cases}
\]

We consider the following two problems.

Problem 1. \( \Gamma_0 = \{(x, y); 0 \leq x \leq 1 \text{ and } y = 1 \text{ or, } x = 1 \text{ and } 0 \leq y \leq 1\} \), and

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

Problem 2. \( \Gamma_0 = \emptyset \) while \( f(x, y) \) satisfies the compatibility condition

\[
\int_{\Omega} f(x, y) \, dx \, dy = 0.
\]
We give in Fig. 1 the graphs of the corresponding matrices $A$ by using the notation of [8]. With this notation, each diagonal entry $a_{ii}$ of $A$ is written inside a circle which represents the node $i$, while each nonzero offdiagonal entry $a_{ij}$ is written along the edge $(i, j)$. In both problems, the order of the matrix $A$ is $n = N^2$ where $N = h^{-1}$ for Problem 1 and $N = h^{-1} + 1$ for Problem 2; moreover, $G(A)$ is connected, $\text{offdiag}(A) < 0$ and $Ae \geq 0$, showing that $A$ is an irreducible Stieltjes matrix, nonsingular in the case of Problem 1 and singular in the other one where $Ae = 0$. For convenience we have assumed that $N$ is a multiple of 4 (plus one) in the first (second) example. With respect to the lexicographic ordering of the $(q, r)$-plane of Fig. 1, the $i$th node corresponds to $i = (r-1)N + q$.

The approximate factorization $B$ of $A$ we shall investigate is defined as follows:

$$B = U'P^+U,$$

where the offdiagonal entries of $U$ are determined by the relation

$$\text{offdiag}(U' + U) = \text{offdiag}(A)$$
(i.e., no fill-in is permitted during the factorization process), while $P$ is computed so as to have

$$Be = Ae + \Delta e,$$

the diagonal entries of the perturbation matrix $\Delta$ being given by

$$\Delta_{ii} = \begin{cases} \max \left\{ \frac{1}{t + l_i + 1}((U' - U)e)_i - (Ae)_i, 0 \right\}, & \text{if } i \in \text{As}(\text{Pc}(U)), \\ 0, & \text{otherwise}, \end{cases}$$

where $t$ denotes a $O(h^{-1})$ parameter and $l_i = l(\text{As}(i))$, say the maximal increasing length of $\text{As}(i)$ in $G(U)$. Observe at this stage that $G(U)$ and $G(A)$ are equal, so that by inspection on the graphs in Fig. 1 one readily checks that for both problems

$$\text{As}(\text{Pc}(U)) = \text{Pc}(U) = \{i = (r-1)N + q; q < N, r < N\},$$

$$l_i = r + q - 2, \quad \text{for } i = (r-1)N + q,$$

$$\Delta_{ii} = \begin{cases} O(h), & \text{if } i \in L = \{1, 2, \ldots, N-1\} \cup \{SN + 1; 1 < s < N - 2\}, \\ 0, & \text{otherwise}. \end{cases}$$

Hence by [27, Theorem 2.2], $U$ is an M-matrix with $Ue > 0$ and $B$ is positive definite (with $N(B) = \{0\} \subset N(A)$). Consequently, $P^+ = P^{-1}$.

The justification of our choice of the perturbation matrix $\Delta$ goes back to a paper by Beauwens and Wilmet [11, Theorem 3.2] whose result has been extended to the singular case by Notay [27, Theorem 3.1]; but as mentioned in Section 1, the method may also be considered as the point version of a particular modified block approximate factorization, first investigated in [22, Section 5]. Set $l = l(\text{As}(\text{Pc}(U)))$, i.e., the maximal increasing length of $\text{As}(\text{Pc}(U))$ in $G(U)$. For both problems one has $l = 2N - 4$, so that the above-mentioned two theorems lead to

$$\lambda_{\max}(B^{-1}A) \leq t + l + 2 = O(h^{-1}).$$

Since $0 \leq \Delta e = O(h)e$ and $\lambda(D^{-1}A) = O(h^2)$, it is necessary to impose $D' \leq O(h^2)$ in (4.11) (or equivalently $\Delta' \leq O(h^2)$ in Theorem 4.2) in order to get $O(1)$ lower bounds on $\lambda_{\min}(B^{-1}A)$.
Problem 1.

\((2 < \rho_0 < N-1)\)
\((2 < \rho_0 < N-1)\)
\((\rho_0 = \rho_0)\)

\(q_0 = \rho_0 = \frac{3N+1}{4}\)

Problem 2.

\((2 < \rho_0 < N-1)\)
\((2 < \rho_0 < N-1)\)
\((q_0 = \rho_0 = \frac{3N+1}{4})\)

Fig. 1. Matrices expressed by graph notation. The \(i\)th node is \(i = (r - 1)N + q\).
and therefore $O(h^{-1})$ upper bounds on the spectral condition numbers. Relations (5.5) and (5.8) enable us to take $\Delta' = 0$. In so doing, (4.11) becomes

$$
\lambda_{\min}(B^{-1}A) \geq \left( 1 + \max_{i \in L} (\Delta_{ii} \gamma_i) \right)^{-1},
$$

(5.10)

where the diagonal matrix $\Delta = (\Delta_{ii} \delta_i)$ and the set $L$ are defined by (5.5) and (5.8), respectively, and the parameters $\gamma_i$ are to be computed through Lemma A.1 or A.2 of the Appendix.

For the same method, Notay's theory [27, Theorem 4.1], which requires the existence of a family of Stieltjes matrices $\{A_i = (a_{ij})_{i \in L}, i \in L\}$ such that $G(A_i)$ is a path or a cycle including node $i$, $A_i e \geq 0$ and the condition (4.6) is satisfied, provides the following lower bound:

$$
\lambda_{\min}(B^{-1}A) \geq \left( 1 + \rho(\lambda_{\min}(D^{-1}A))^{-1} + \delta(\lambda_{\min}(D^{-1}A))^{-1/2} \right)^{-1},
$$

(5.11)

with

$$
\rho = \max_{i \in L} \rho_i,
$$

(5.12)

$$
\delta = \max_{i \in L} \delta_i,
$$

(5.13)

where

$$
\rho_i = \frac{\Delta_{ii}}{(e, D_i e)}
$$

(5.14)

and

$$
\delta_i = \frac{\xi_i \Delta_{ii}}{\alpha_i},
$$

(5.15)

with

$$
D_i = \text{diag}(A_i),
$$

(5.16)

$$
\xi_i = \begin{cases} 
2^{1/2}, & \text{if } G(A_i) \text{ is a path,} \\
2^{-1/2}, & \text{if } G(A_i) \text{ is a cycle,}
\end{cases}
$$

(5.17)

$$
\alpha_i = \min_{(k, l) \in E_i} (-a_{kl}^{(i)}),
$$

(5.18)

$E_i$ denoting the edge set of $G(A_i)$. It is interesting to note that our step-by-step technique of Section 4 may also serve to construct Notay's family of matrices $(A_i, i \in L)$ provided that in Step 2, one imposes $G(A_i)$ to be a path or a cycle instead of a spanning tree for $G(A)$ (see [27, p.691]).

We shall now examine each problem in detail. For exposition purposes we split the set $L$ as follows:

$$
L = L_1 \cup L_2,
$$

(5.19)

with

$$
L_1 = \{i; \text{mod}(i, N) = 1, i \neq (N - 1)N + 1\}, \quad L_2 = \{2, 3, \ldots, N - 1\}.
$$

**Problem 1. Beuwens' approach.**

**Step 1.** $J = \{j = (r - 1)N + q; r = N \text{ or } q = N\}.
Fig. 2. Trees associated with \((A_i)_{i \in L} \ (L = L_1 \cup L_2)\); nodes \(i\) are represented by empty circles; empty squares define nodes \(g(i)\) in the regular case.

\textbf{Step 2.} The trees we have chosen for the matrices \(A_i, i \in L\), as well as the subset \(\{g(i); i \in L\}\) of \(J\) which determines the application \(g\), are displayed in Fig. 2. Clearly

\[
n_j = \begin{cases} 
0, & \text{if } j \in J' = \{(N - 1)N + 1, N^2\}, \\
1, & \text{if } j \in J \setminus J'. 
\end{cases}
\]

\textbf{Step 3.} For each \(i \in L\), let \(P_{g(i)}\) denote the unique path from \(i\) to \(g(i)\) with respect to \(G(A_i)\). Let \(E_i\) stand for the edge set of \(G(A_i)\). Let further \(\epsilon\) represent an arbitrary — as small as we need — positive number, e.g., \(\epsilon < \epsilon_0\) where \(\epsilon_0\) denotes the relative machine accuracy of floating-point arithmetic. For \((k, l) \in E_i\) we set

\[
a_{k l}^{(i)}(\epsilon) = \begin{cases} 
\frac{a_{k l} + \epsilon}{m_{k l}}, & \text{if } (k, l) \in P_{g(i)}, \\
\frac{a_{k l}}{n_{k l}}, & \text{if } \{r; r \in L, (k, l) \in P_{r(r)}\} = \emptyset, \\
-\frac{\epsilon}{n_{k l} - m_{k l}}, & \text{if } \{r; r \in L, r \neq i, (k, l) \in P_{r(r)}\} \neq \emptyset,
\end{cases}
\]

with

\[
m_{k l} = \text{Card}(\{r; r \in L, (k, l) \in P_{r(r)}\}),
\]

\[
n_{k l} = \text{Card}(\{r; r \in L, (k, l) \in E_r\}).
\]

\textbf{Step 4.} We apply the stated formulas.

Hence, with the help of Lemma A.1 of the Appendix, one gets

\[
\max_{i \in L} (\Delta_i, y_i) = 2^{-(d + 3)} \frac{N}{t + 1},
\]
which combined with (5.10) yields

$$\lambda_{\min}(B^{-1}A) \geq \left(1 + 2^{-1}(d + 3)\frac{N}{t + 1}\right)^{-1}. \quad (5.20)$$

Since $t = O(h^{-1}) = O(N)$, the latter lower bound is obviously $O(1)$.

Before going further, let us point out that for each tree $G(A_i)$, only the path from $i$ to $g(i)$ participates to the calculation of $\gamma_i$ and that this path is the one one would normally choose as $G(A_i)$ when applying the theory developed in [8, pp.107–112], resulting in essentially the same bound.

**Problem 1. Notay's approach.**

*Step 1.* $J = \{j = (r - 1)N + q; q = N \text{ or } r = N\}$.

*Step 2.* Figure 3 depicts the graphs (simple paths) $G(A_i)$, $i \in L$, we have selected. It also specifies the mapping $g$. The parameters $n_j$, $j \in J$, are obviously as in the previous approach.

*Step 3.* For $i \in L$ and for $(k, l) \in E_i$ we take $a^{(i)}_{kl} = a_{kl}$.

*Step 4.* The needed rules are applied.

Hence, by (5.11)–(5.18),

$$\lambda_{\min}(B^{-1}A) \geq \left(1 + \frac{2d}{t + 1} \left(\frac{2}{\lambda_{\min}(D^{-1}A)((3d + 1)N - 2)} + \left(\frac{2}{\lambda_{\min}(D^{-1}A)}\right)^{1/2}\right)^{1/2}\right)^{-1}, \quad (5.21)$$

which is $O(1)$ since $t = O(N)$ and $\lambda_{\min}(D^{-1}A) = O(h^2)$. However, on account of the simultaneous influence of the parameters $d$ and $(\lambda_{\min}(D^{-1}A))^{1/2}$, the occurrence $d \gg 1$ (which corresponds to a quasi singular problem) causes the latter lower bound to strongly degenerate in comparison with (5.20) (see numerical results to come). A way to avoid this degeneration consists in getting rid of the parameter $d$ by a judicious choice of paths associated with $(A_i)_{i \in L}$. It is obvious from a careful look over formulas (5.11)–(5.18) and Fig. 1 that one should try to avoid paths crossing over two "regions" where the matrix coefficients are "strongly" different. In practice, such a procedure would reveal rather intricate; suffice it to imagine a problem where several "regions" with different values of the parameter $d$ are adjacent and where one would have to specify which coefficients are "strongly" different. In our case, two regions may easily be delimited according as the matrix coefficients depend on the parameter $d$. 

---

Fig. 3. Paths associated with $(A_i)_{i \in L}$ ($L = L_1 \cup L_2$) for Notay's approach; empty circles represent nodes $i$; empty squares define nodes $g(i)$ in the regular case.
or not, which leads us to the paths displayed in Fig. 4. Note that the mapping $g$, which is not essential in Notay's approach, is not defined here. Therefore, keeping Step 3 above unaltered one readily finds that

$$\lambda_{\min}(B^{-1}A) \geq \left(1 + \frac{2}{t+1} \left(\frac{2}{3N\lambda_{\min}(D^{-1}A)} + \left(\frac{2}{\lambda_{\min}(D^{-1}A)}\right)^{1/2}\right)\right)^{-1},$$

a far better bound than (5.21) whenever $d \gg 1$. For $d = 1$ (in fact for values of $d$ around 1) one easily checks that the lower bound (5.22) is (slightly) less accurate than (5.21).

**Problem 2.** As explained at the beginning of the present section, the matrix $A$ is here singular with $Ae = 0$, so that $J = \emptyset$. We shall therefore specify only the graphs associated with $(A_i)_{i \in L}$ as well as how to compute the nonzero offdiagonal entries of $A_i$; as soon as they are known, the determination of the diagonal entries is straightforward by formulas (4.14).

**Beauwens' approach.** We keep the same trees as in the previous problem (see Fig. 2). For ease of presentation, we shall write $P_i$ for fixed $i \in L$ to denote the path delimited in Fig. 2 by an empty circle and an empty square which we shall call the main path (or the trunk) of the tree $G(A_i)$; the remaining paths will be termed secondary paths (or branches). The parameters $\gamma_i$ are now to be computed through Lemma A.2 of the Appendix (see (A.5)-(A.8)), applied to the matrices $A_i = (a_{ij}^{(i)})$, $i \in L$. For fixed $i \in L$, let $(k, l)$ be an edge of $G(A_i)$; one of the nodes $k$ and $l$, denoted here by $S(k, l)$, is necessary successor (or son) of the other one, with respect to the (node set) ordering one obtains by considering (as required in Lemma A.2) the node $i$ as the root of $G(A_i)$ and as the first node, and by numbering sons after fathers. We shall denote the number of descendants of $S(k, l)$ with respect to the latter ordering by $d(a_{kl}^{(i)})$. Then one has by (A.7), (A.8),

$$\gamma_i = \frac{1}{N^4} \sum_{(k, l) \in E_i} |a_{kl}^{(i)}|^{-1}(d(a_{kl}^{(i)}))^2,$$

where $E_i$ stands for the edge set of $G(A_i)$.
To optimize the lower bound (5.10), the ideal would be to determine the nonzero offdiagonal entries of $A_i$, $i \in L$, so as to have all the quantities $\Delta_{ii} \gamma_i$ equal to each other. Unfortunately, from a practical point of view, such a procedure would be costly and tedious because of the variations of $\Delta_{ii}$, $i \in L$. That is why, besides conditions (4.12), (4.13) ((4.13) being satisfied with an equal sign), we shall impose only to the contribution of each edge of secondary paths to be the same in each $\gamma_i$ for all $i \in L_1$ (respectively, $i \in L_2$). Before doing so we set for $1 \leq k \leq n$ and $1 \leq l \leq n$,

$$m_{kl} = \text{Card}([i; i \in L, (k, l) \in P_i]),$$

$$p_k = \text{int} \left( \frac{k - 1}{N} \right) + 1$$

(int($x$) indicates the integer part of $x$) and

$$n_k = \text{mod}(k, N).$$

- **Along main paths:** we set

$$a_{kl}^{(i)} = \frac{a_{kl}}{m_{kl} + 1}.$$  

Obviously

$$d(a_{kl}^{(i)}) = \begin{cases} (N - n_k)N, & \text{if } i \in L_1, \\ (N - p_k)N, & \text{if } i \in L_2. \end{cases}$$

- **Along secondary paths:**

  **Case 1:** $i \in L_1$. Nonzero offdiagonal entries of $A_i$ are of the form $a_{kk+N}^{(i)}$. Further,

$$d(a_{kk+N}^{(i)}) = \begin{cases} N - p_k, & \text{if } i \in L_{pk}^{(1)} = \{1, \, N + 1, \, 2N + 1, \ldots, (p_k - 1)N + 1\}, \\ p_k, & \text{if } i \in L_1 \setminus L_{pk}^{(1)}. \end{cases}$$

Then, after translation of the above considerations into mathematical language,

$$a_{kk+N}^{(i)} = \begin{cases} v_{pk}, & \text{if } i \in L_{pk}^{(1)}, \\ w_{pk}, & \text{if } i \in L_1 \setminus L_{pk}^{(1)}. \end{cases}$$

where $v_{pk}$ and $w_{pk}$ are the solutions of the system

$$\begin{align*}
p_kv_{pk} + (N - 1 - p_k)w_{pk} &= \frac{a_{kk+N}}{m_{kk+N} + 1}, \\
\left(\frac{N - p_k}{v_{pk}}\right)^2 &= \frac{p_k^2}{w_{pk}}.
\end{align*}$$

**Case 2:** $i \in L_2$. Nonzero offdiagonal entries of $A_i$ are of the form $a_{kk+1}^{(i)}$. Further,

$$d(a_{kk+1}^{(i)}) = \begin{cases} n_k, & \text{if } i \in L_{n_k}^{(2)} = \{n_k + 1, \, n_k + 2, \ldots, N - 1\}, \\ N - n_k, & \text{if } i \in L_2 \setminus L_{n_k}^{(2)}. \end{cases}$$
Now
\[ a_{kk+1}^{(i)} = \begin{cases} y_{nk}, & \text{if } i \in L_{nk}^{(2)}, \\ z_{nk}, & \text{if } i \in L_{2} \setminus L_{nk}^{(2)}, \end{cases} \tag{5.33} \]
where \( y_{nk} \) and \( z_{nk} \) are the solutions of the system
\[ \begin{aligned}
(n_k - 1)z_{nk} + (N - 1 - n_k)y_{nk} &= \frac{a_{kk+1}}{m_{kk+1} + 1}, \\
\frac{n_k^2}{y_{nk}} &= (N - n_k)^2.
\end{aligned} \tag{5.34} \]

Thus, by (5.5) and (5.23), elementary algebraic calculations lead to
\[ \max_{i \in L} \Delta_{ii} \gamma_i = \Delta_{11} \gamma_1 = \frac{d}{t + 1} \gamma, \]
where
\[ \gamma = \frac{1}{384} \left( 55 + \frac{585}{d} \right) N + a_0 + \frac{a_1}{N} + \frac{a_2}{N^2} + \frac{a_3}{N^3} + \frac{a_4}{N^4}, \tag{5.35} \]
with
\[ a_0 = \frac{216}{d + 1} - \frac{981}{d}, \quad a_1 = \frac{248 + 408}{d} - \frac{288}{d + 1}, \quad a_2 = \frac{96}{d + 1} - \frac{164}{d}, \]
\[ a_3 = 9 - \frac{9}{d}, \quad a_4 = 3 + \frac{9}{d} - \frac{24}{d + 1}, \]
whence through (5.10),
\[ \lambda_{\min}(B^{-1}A) \geq \left( 1 + \frac{d}{t + 1} \gamma \right)^{-1}. \tag{5.36} \]

Observe that our treatment is such that all the parameters \( \gamma_i \) are of the same order of magnitude, which guarantees a good lower bound in the case where the quantities \( \{(U^i - U)e\}_i, \ i \in L \) (see (5.5)), vary smoothly since, remember, the closer the \( \Delta_{ii} \gamma_i, \ i \in L \), the better the lower bound. For \( d \gg 1 \), the required condition is not satisfied, whence the risk of obtaining less accurate bounds (which nevertheless exhibit the correct order of magnitude of \( \lambda_{\min}(B^{-1}A) \)). This is indeed the case for, e.g., \( d = 100 \) (see numerical results to come). A complete, and costly, optimization requires to also take into account the variation of the parameter \( l_i, \ i \in L \), defined by (5.7). However, an improvement may cheaply be achieved by imposing only the quantities \( \{(U^i - U)e\}_i, \ i \in L \), to be equal to each other. Applied to the case of \( d = 100 \), such a particular partial optimization produced lower bounds only three times smaller than those ones computed from Notay's approach with "optimized" paths and the actual value of \( \lambda_{\min}(D^{-1}A) \).

It is worth mentioning that the above considerations concern only singular problems with nondegenerate nonzero smallest eigenvalues (i.e., independent of possible discontinuities in
material coefficients of the boundary value problem) which do no hold for, e.g., problem (5.1) with \( \Gamma_0 = \emptyset \), \( d \ll 1 \) and \( \Omega' = \{(x, y) ; 0 \leq x \leq 1, \frac{1}{3} \leq y \leq \frac{2}{3}\} \), in which case the arguments developed in (5.24)–(5.34) give rise to lower bounds that compete with those derived from Notay's approach with "optimized paths" and the exact value of \( \lambda_{\min}(D^{-1}A) \).

**Problem 2. Notay's approach.** For comparison purposes, we use successively the "straightforward" paths of Fig. 3 and the "improved" paths of Fig. 4 as graphs associated with \( A_i, i \in L \), so that, setting \( a_{kl}^{(i)} = a_{kl} \) for \((k, l) \in E_i\), one readily finds through (5.11)–(5.18) that

\[
\lambda_{\min}(B^{-1}A) \geq \left( 1 + \frac{2d}{t+1} \left( \frac{2}{(3d+1)(N-1)\lambda_{\min}(D^{-1}A)} + \sqrt{\frac{2}{\lambda_{\min}(D^{-1}A)}} \right) \right)^{-1},
\]

(5.37)

\[
\lambda_{\min}(B^{-1}A) \leq \left( 1 + \frac{2}{t+1} \left( \frac{2}{3(N-1)\lambda_{\min}(D^{-1}A)} + \sqrt{\frac{2}{\lambda_{\min}(D^{-1}A)}} \right) \right)^{-1}.
\]

(5.38)

Note that

\[
\lambda_{\min}(D^{-1}A) = \frac{1}{2} \left( 1 - \cos \frac{\pi}{N-1} \right), \quad \text{for } d = 1.
\]

### 6. Numerical results and concluding remarks

We present in Tables 1 and 2 the results of numerical experiments realized on both test problems. Table 1 contains the actual smallest nonzero eigenvalue of \( B^{-1}A \), the lower bounds computed from the analytical expressions obtained above, the spectral condition number of \( B^{-1}A \) as well as the actual smallest nonzero eigenvalue of \( D^{-1}A \) which is needed in Notay's bounds. We have also included the effective (or reduced) spectral condition number where extremal (isolated) eigenvalues (here only the smallest one) are not taken into consideration (see, e.g., [4,5,30]). The "perturbed modified incomplete factorization (without fill-in)" (5.1)–(5.4), termed here PMIF(0), has been carried out with \( t = 0.5 N \). From a comparison point of view, it is clear from the involved lower bounds that the parameter \( t \) does not play a crucial role. As observed during the derivation of the formulas used in Table 1, both approaches give rise to the correct order of magnitude of \( \lambda_{\min}(B^{-1}A) \), with the warning that for Notay's theory one must always choose the required graphs (paths) with care, taking into account possible discontinuities in the system matrix coefficients. Note that, besides the extension to the singular case, the essential merit of Notay's approach compared with all its precursors [4,16,33] is that restrictive conditions on paths are considerably reduced, which enables one to avoid discontinuities (responsible for degenerate lower bounds in [33]) when necessary. However, this relative freedom in choosing paths turns out to be a serious drawback in the perspective of an automatic computation of lower bounds.
Table 1
Numerically computed nonzero smallest eigenvalues $v_0$ and $v_{\min}$ of respectively $D^{-1}A$ and $B^{-1}A$, lower spectral bounds provided by (5.20), (5.21), (5.22), (5.36), (5.37) and (5.38), spectral ($\kappa$) and effective spectral ($\kappa_e$) condition numbers for the matrix $B^{-1}A$ associated with PMIF(0) ($t = 0.5 N$)

<table>
<thead>
<tr>
<th>Problem 5.1 ($d = 1$)</th>
<th>$N$</th>
<th>$v_0$</th>
<th>$v_{\min}$</th>
<th>(5.20)</th>
<th>(5.21)</th>
<th>(5.22)</th>
<th>$\kappa$</th>
<th>$\kappa_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>86 10^{-4}</td>
<td>0.449</td>
<td>0.226</td>
<td>0.147</td>
<td>0.138</td>
<td>10.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>21 10^{-4}</td>
<td>0.439</td>
<td>0.213</td>
<td>0.138</td>
<td>0.130</td>
<td>21.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
<td>54 10^{-5}</td>
<td>0.433</td>
<td>0.207</td>
<td>0.134</td>
<td>0.126</td>
<td>45.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96</td>
<td>13 10^{-5}</td>
<td>0.430</td>
<td>0.203</td>
<td>0.132</td>
<td>0.123</td>
<td>93.30</td>
</tr>
<tr>
<td>Problem 5.1 ($d = 100$)</td>
<td></td>
<td>12</td>
<td>20 10^{-5}</td>
<td>0.039</td>
<td>0.011</td>
<td>$34 \cdot 10^{-5}$</td>
<td>0.009</td>
<td>166.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>50 10^{-6}</td>
<td>0.036</td>
<td>0.010</td>
<td>$32 \cdot 10^{-5}$</td>
<td>0.009</td>
<td>361.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
<td>13 10^{-6}</td>
<td>0.035</td>
<td>0.010</td>
<td>$30 \cdot 10^{-5}$</td>
<td>0.008</td>
<td>767.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96</td>
<td>31 10^{-7}</td>
<td>0.034</td>
<td>0.010</td>
<td>$30 \cdot 10^{-5}$</td>
<td>0.008</td>
<td>1617</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem 5.2 ($d = 1$)</th>
<th>$N$</th>
<th>$v_0$</th>
<th>$v_{\min}$</th>
<th>(5.36)</th>
<th>(5.37)</th>
<th>(5.38)</th>
<th>$\kappa$</th>
<th>$\kappa_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>17 10^{-3}</td>
<td>0.736</td>
<td>0.282</td>
<td>0.220</td>
<td>0.210</td>
<td>11.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>43 10^{-4}</td>
<td>0.723</td>
<td>0.257</td>
<td>0.203</td>
<td>0.194</td>
<td>23.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>49</td>
<td>11 10^{-4}</td>
<td>0.715</td>
<td>0.244</td>
<td>0.194</td>
<td>0.185</td>
<td>47.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97</td>
<td>27 10^{-5}</td>
<td>0.711</td>
<td>0.237</td>
<td>0.189</td>
<td>0.180</td>
<td>97.04</td>
</tr>
<tr>
<td>Problem 5.2 ($d = 100$)</td>
<td></td>
<td>13</td>
<td>29 10^{-3}</td>
<td>0.778</td>
<td>0.042</td>
<td>$45 \cdot 10^{-4}$</td>
<td>0.270</td>
<td>8.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>73 10^{-4}</td>
<td>0.766</td>
<td>0.036</td>
<td>$41 \cdot 10^{-4}$</td>
<td>0.251</td>
<td>17.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>49</td>
<td>19 10^{-4}</td>
<td>0.758</td>
<td>0.033</td>
<td>$39 \cdot 10^{-4}$</td>
<td>0.240</td>
<td>35.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97</td>
<td>46 10^{-5}</td>
<td>0.755</td>
<td>0.032</td>
<td>$37 \cdot 10^{-4}$</td>
<td>0.235</td>
<td>73.98</td>
</tr>
</tbody>
</table>

On some parallel computer architectures, operations like inner products and vector norms (which require global communication) are excessively time-consuming. Because of the possibility of avoiding the above-mentioned operations, the Chebyshev iterative relaxation method is preferred in such machines over the popular conjugate gradient acceleration provided that (accurate) bounds for extreme eigenvalues are available [3,21], whereas the high interest of automatic computation of lower bounds. In this respect, particularly attractive is Beauwens' approach where "useful" graphs do not depend critically on the coefficients of the original matrix, and which, this is its major (and undeniable) advantage to our mind, does not require (at least for the factorization strategy we discuss here) the knowledge of some expensive to estimate parameter like $\lambda_{\min}(D^{-1}A)$, to the price (for some discontinuities) of an optimization effort. Even for methods where $\lambda_{\min}(D^{-1}A)$ is unavoidable (as in [27, Strategy 2]), good lower bounds can be obtained (see, e.g., [24]). We observe in passing that in the case where one disposes a "good" estimate for $\lambda_{\min}(D^{-1}A)$, the heuristic formula proposed by Beauwens [10] is more appropriate (see, e.g., [24,29] for numerical evidence). We refer to [28] for an automatic calculation of upper eigenvalue bounds.

In Table 2 are reported (for the sake of comparison) the number of preconditioned conjugate (PCG) iterations to achieve the relative residual error $\|r^{(i)}\|_2/\|r^{(0)}\|_2 \leq 10^{-m}$ ($r^{(i)}$ is the residual in the $i$th iteration), with zero as initial guess, and PMIF(0) with $t = cN$ (various values of $c$), the unmodified incomplete factorization IC(0) (see [18,29] for the singular case) and the unperturbed modified incomplete factorization MIC(0) as preconditioners. Concerning the practical use and the conditioning analysis of the MIC(0) method in the case of singular
Table 2
Number of PCG iterations to achieve $\|r^{(i)}\|_2 / \|r^{(0)}\|_2 \leq 10^{-m}$ for the unmodified method IC(0), the unperturbed modified method MIC(0) and the perturbed modified method PMIF(0) with $t = cN$

<table>
<thead>
<tr>
<th>$h^{-1}$</th>
<th>96</th>
<th>192</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem 5.1 ($d = 100$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC(0)</td>
<td>70</td>
<td>82</td>
</tr>
<tr>
<td>MIC(0)</td>
<td>57</td>
<td>68</td>
</tr>
<tr>
<td>PMIF(0)</td>
<td>57</td>
<td>68</td>
</tr>
<tr>
<td>$c = 0.0$</td>
<td>57</td>
<td>64</td>
</tr>
<tr>
<td>0.2</td>
<td>46</td>
<td>52</td>
</tr>
<tr>
<td>0.5</td>
<td>44</td>
<td>50</td>
</tr>
<tr>
<td>0.7</td>
<td>43</td>
<td>49</td>
</tr>
<tr>
<td>1.0</td>
<td>42</td>
<td>49</td>
</tr>
<tr>
<td>2.0</td>
<td>43</td>
<td>50</td>
</tr>
</tbody>
</table>

| Problem 5.2 ($d = 1$) |      |      |
| IC(0)    | 72   | 82   |
| MIC(0)   | 42   | 55   |
| PMIF(0)  | 44   | 53   |
| $c = 0.0$| 44   | 53   |
| 0.2      | 34   | 41   |
| 0.5      | 32   | 41   |
| 0.7      | 32   | 41   |
| 1.0      | 33   | 41   |
| 2.0      | 35   | 43   |

| Problem 5.2 ($d = 100$) |      |      |
| IC(0)    | 63   | 87   |
| MIC(0)   | 69   | 88   |
| PMIF(0)  | 69   | 88   |
| $c = 0.0$| 38   | 46   |
| 0.2      | 30   | 36   |
| 0.5      | 27   | 35   |
| 0.7      | 27   | 35   |
| 1.0      | 29   | 36   |
| 2.0      | 31   | 38   |
problems, detailed information may be found in [27–29]; we just mention that during the factorization process, zero diagonal entries are to be exchanged for arbitrary positive numbers (often 1). In the case of Problem 2, the right-hand side of the linear system has been chosen such that the solution to the system coincides with the discrete values of the function \( u_0(x, y) = (1 + x)^2(1 + y)(2 - y)e^{xy} \) (from [27]) on the grid. Note however that, in practical singular problems, due to rounding errors, the compatibility condition is generally lost after discretization; techniques for avoiding divergence of PCG methods that may occur in such a case are discussed at great length in [18,29].

For a wide range of (two-dimensional) PDEs including Problem 1 with \( d \gg 1 \) and Problem 2, PMIF(0) gives better results than IC(0) and MIC(0); in particular for large systems (small values of \( h \)), considerable reductions of the number of iterations have been observed. Nevertheless for regular problems with weak variations of coefficients, MIC(0) and PMIF(0) realized essentially the same performances. The rate of convergence of the PCG-PMIF(0) method has also been observed to be remarkably insensitive to the choice of the constant \( c \); however, too small values (close to 0) are not advisable. The choice \( c = 0.5 \) is recommended. Observe on the other hand that in the case of \( c = 0 \), all the lower spectral bounds derived herein (as well as the actual smallest nonzero eigenvalue(s)) become \( O(h) \), causing PMIF(0) to behave like a relaxed method (RIC(0) [5]) where the relaxation parameter is chosen away from 1.

Appendix

**Lemma A.1.** Let \( A = (a_{kl}) \) be a treediagonal \( n \times n \) Stieltjes matrix and \( x > 0 \) such that \( (Ax)_r = 0 \) for \( r \neq j \) while \( (Ax)_j > 0 \); then, for any \( i, 1 \leq i \leq n \), one has that

\[
\gamma_i = \max_{z \neq 0} \frac{|z_i|^2}{(z, Az)} = \left[ \sum_{(k, l) \in P_{ij}} \left( \frac{1}{x_k |a_{kl}| x_l} \right) + \frac{1}{x_j (Ax)_j} \right] x_i^2,
\]

(A.1)

where \( P_{ij} \) denotes the (unique) path from \( i \) to \( j \).

**Proof.** Observe first that there is no loss of generality in assuming \( z_i \neq 0 \) in the definition of \( \gamma_i \) and that

\[
\gamma_i = \max_{z \in \mathbb{C}^n} \frac{|z_i|^2}{(z, Az)} = \max_{z \in \mathbb{C}^n} \frac{x_i^2}{(z, Az)}.
\]

Let now \( D = (d_{kl}) \) be a diagonal matrix and for any \( z \in \mathbb{C}^n \), we set \( w = D^{-1}z \); by [6, Lemma 3.1] one has that \((z, Az) = (w, DADw), \) whence

\[
\frac{(z, Az)}{x_i^2} = \sum_{k=1}^{n} \frac{x_k}{x_i} (Ax)_k \left| \frac{z_k}{x_k} \right|^2 - \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \frac{x_k}{x_i} a_{kl} \frac{x_l}{x_i} \frac{z_k}{x_k} \frac{z_l}{x_l} = \frac{x_j}{x_i} (Ax)_j \left| \frac{z_j}{x_j} \right|^2 + \sum_{e \in E} c_e |a_e|^2,
\]

where \( E \) represents the edge set of \( G(A) \), i.e., \( E = \{ e; e = (k, l), a_{kl} \neq 0, k < l \} \) and for any
\[ e = (k, l) \in E, \quad \alpha_c = \frac{z_k}{x_k} - \frac{z_l}{x_l}, \quad \text{while} \quad c_c = -(x_k/x_l)a_{kl}(x_l/x_k). \]

Since \( G(A) \) is a tree, the \( n - 1 \) independent variables \( z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \) \((z_i = x_i \text{ is fixed})\) may be exchanged for the \( n - 1 \) independent variables \((\alpha_c)_c \in E\) or for the \( n \) variables \((\alpha_c)_c \in E\) and \( z_i \) with the constraint \( z_i = x_i \). The latter constraint may be written

\[
\sum_{e \in P_i} \alpha_c + \frac{z_j}{x_j} = \frac{z_i}{x_i} = 1. \tag{A.2}
\]

Introducing

\[
\beta_c = c_{c^{1/2}} \alpha_c, \quad \text{for} \ e \in E,
\]

and

\[
\beta_0 = c_0^{1/2} \frac{z_j}{x_j}, \quad \text{with} \ c_0 = \frac{x_j(A_0)_j}{x_i^2},
\]

we have

\[
\gamma_i^{-1} = \min \left( |\beta_0|^2 + \sum_{c \in E} |\beta_c|^2 \right),
\]

where the minimum is to be taken under the constraint (A.2), i.e.,

\[
\sum_{c \in P_i} \left( \frac{\beta_c}{c_c^{1/2}} \right) + \frac{\beta_0}{c_0^{1/2}} = 1. \tag{A.3}
\]

In other words, \( 1/\gamma_i \) is the square of the distance \( d \) from the origin to the hyperplane of equation (A.1) in the \( n \)-dimensional space \( \beta_0, (\beta_c)_c \in E \). Since

\[
d^2 = \left( \sum_{e \in P_i} \left( \frac{1}{c_c} \right) + \frac{1}{c_0} \right)^{-1},
\]

we have

\[
\gamma_i = \sum_{e \in P_i} \left( \frac{1}{c_c} \right) + \frac{1}{c_0}, \tag{A.4}
\]

which concludes the proof. \( \square \)

**Lemma A.2.** Let \( A = (a_{kl}) \) be a treedagonal \( n \times n \) singular Stieltjes matrix and \( x > 0 \) such that \( Ax = 0; \) then

\[
\gamma_i = \max_{z \in \mathbb{C}^n \setminus \{0\}} \frac{|z_i|^2}{(z, Az)} = \left( \sum_{k=1}^{n-1} \sum_{l \in S(k)} \frac{w_i^2}{a_{kl}|x_l|} \right) x_i^2, \tag{A.5}
\]

where

\[
w_i = \frac{\sum_{r \in D_1(l)} x_r^2}{\sum_{r=1}^{n} x_r^2}, \tag{A.6}
\]
provided that the node $i$ is distinguished as the root of the tree associated with $A$ and the node set of this tree is ordered by taking the root as first node and numbering sons after fathers.

Proof. We have obviously

$$\gamma_i = \max_{z \perp x} \frac{|z_i|^2}{(z, Az)} = \max_{z \perp x} \frac{x_i^2}{(z, Az)},$$

and, by [6, Lemma 3.11] and the assumption $Ax = 0$,

$$\frac{(z, Az)}{x_i^2} = - \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \frac{x_k}{x_i} a_{kl} \frac{x_l}{x_i} \left| \frac{z_k}{x_k} - \frac{z_l}{x_l} \right|^2 = \sum_{e \in E} c_e |\alpha_e|^2,$$

where $E$ represents the edge set of $G(A)$, i.e., $E = \{e; e = (k, l), a_{kl} \neq 0, k < l\}$ and for any $e = (k, l) \in \bar{E}$, $\alpha_e = z_k/x_k - z_l/x_l$ while $c_e = -(x_k/x_l)a_{kl}(x_l/x_k)$. Since $G(A)$ is a tree, the $n - 1$ independent variables $z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$ may be exchanged for the $n - 1$ independent variables $(\alpha_e)_{e \in E}$; the relations expressing $z_j$ in terms of $\alpha_e$ are

$$\sum_{e \in P_{ij}} \alpha_e + \frac{z_j}{x_j} = \frac{z_i}{x_i} = 1,$$

where $P_{ij}$ denotes the (unique) path from the root $i$ to $j$. The constraint $z \perp x$ may then be written, successively,

$$\sum_{r=1}^{n} x_r z_r = 0,$$

$$\sum_{r=1}^{n} \left( \sum_{e \in P_{ir}} \alpha_e \right) x_r^2 = \sum_{r=1}^{n} x_r^2,$$

or equivalently

$$\sum_{e \in E} \sum_{r \in Ds(e)} x_r^2 = \sum_{r=1}^{n} x_r^2,$$

where for each $e = (k, l) \in E$, $Ds(e)$ is defined as $Ds(e) = Ds(l)$. Hence

$$\sum_{e \in E} w_e \alpha_e = 1,$$

where for each $e = (k, l) \in E$,

$$w_e = \frac{\sum_{r \in Ds(l)} x_r^2}{n}.$$

Introducing $\beta_e = \frac{c_e^{1/2}}{2} \alpha_e$, we now have to minimize

$$\frac{(z, Az)}{x_i^2} = \sum_{e \in E} |\beta_e|^2,$$
under the constraint
\[ \sum_{e \in E} \frac{w_e}{c_e^{1/2}} \beta_e = 1, \]
in the \( n - 1 \) dimensional space \((\beta_e)_{e \in E}\). This leads to
\[ \gamma_i = \sum_{e \in E} \frac{w_e^2}{c_e}, \]
which ends the proof. \( \Box \)

**Remark A.3.** For \( x = (1, 1, \ldots, 1)' \) the formulas (A.5) and (A.6) become
\[ \gamma_i = \frac{1}{n} \sum_{k = 1}^{n-1} \sum_{l \in S(k)} \frac{w_l^2}{|a_{kl}|}, \]
\[ w_l = -\text{Card}(D_s(l)). \]

**Acknowledgements**

The author wishes to thank Professor R. Beauwens for suggesting this work and giving valuable ideas, and Mr. Y. Notay for useful comments. Numerous improvements and suggestions from the referee are also gratefully acknowledged.

**References**


