

A Nearly Optimal Preconditioning based on Recursive Red–black Orderings

Yvan Notay* and Zakaria Ould Amar

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

Considering matrices obtained by the application of a five-point stencil on a 2D rectangular grid, we analyse a preconditioning method introduced by Axelsson and Eijkhout, and by Brand and Heinemann. In this method, one performs a (modified) incomplete factorization with respect to a so-called ‘repeated’ or ‘recursive’ red–black ordering of the unknowns while fill-in is accepted provided that the red unknowns in a same level remain uncoupled.

Considering discrete second order elliptic PDEs with isotropic coefficients, we show that the condition number is bounded by $\mathcal{O}(n^{\frac{1}{2} \log_2(\sqrt{5}-1)})$ where n is the total number of unknowns ($\frac{1}{2} \log_2(\sqrt{5}-1) = 0.153$), and thus, that the total arithmetic work for the solution is bounded by $\mathcal{O}(n^{1.077})$. Our condition number estimate, which turns out to be better than standard $\mathcal{O}(\log^2 n)$ estimates for any realistic problem size, is purely algebraic and holds in the presence of Neumann boundary conditions and/or discontinuities in the PDE coefficients.

Numerical tests are reported, displaying the efficiency of the method and the relevance of our analysis. © 1997 by John Wiley & Sons, Ltd.

Numer. Linear Algebra Appl., Vol. 4, 369–391 (1997)
(No. of Figures: 2 No. of Tables: 4 No. of Refs: 25)

KEY WORDS iterative methods for linear systems; acceleration of convergence; preconditioning

* Correspondence to Y. Notay, Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium.

Contract grant sponsor: Fonds National de la Recherche Scientifique; contract grant sponsor: Belgian State—Prime Ministers Service—Federal Office for Scientific, Technical and Cultural Affairs; contract grant sponsor: ULB-IBM; contract grant number: IT-IF-14.

1. Introduction

We consider here the iterative solution of large sparse symmetric positive definite linear system

$$Ax = b. \quad (1.1)$$

For such systems, the conjugate gradient acceleration is known as a powerful technique provided efficient preconditioning can be used (see e.g., [4]).

The preconditioner B is a symmetric positive definite matrix such that solving a system with B , which has to be done at each iteration, is much easier than solving (1.1), while the (spectral) condition number of the preconditioned system

$$\kappa(B^{-1}A) = \frac{\nu_{\max}(B^{-1}A)}{\nu_{\min}(B^{-1}A)} \quad (1.2)$$

(i.e. the ratio of the largest and the smallest eigenvalue), on which the convergence rate depends, has to be as small as possible and anyway much less than the original condition number $\kappa(A)$. Indeed, the number of iteration k_ϵ is bounded above by

$$\kappa_\epsilon \leq \frac{1}{2} \sqrt{\kappa(B^{-1}A)} \ln \frac{2}{\epsilon} + 1 \quad (1.3)$$

where ϵ is the relative error in norm $(\cdot, A)^{1/2}$. If the eigenvalue distribution between ν_{\min} and ν_{\max} is dense, this bound is known sharp, see e.g., [3,21].

For systems arising from the discretization of second order discrete elliptic PDEs, it is now common to use preconditioners obtained by computing an approximate factorization of the system matrix, or, more precisely, of some Stieltjes approximation of the system matrix (in the general case where it has some positive off-diagonal entries), see e.g. [3,4]. Mostly, these factorizations are carried out with respect to a natural ordering of the unknowns, or some other ordering presenting similar properties. In the best cases, the condition number for 2D isotropic problems is of order \sqrt{n} , where n is the number of unknowns. This is an order of magnitude better than the original conditioning $\kappa(A) = \mathcal{O}(n)$.

Here we consider a special technique which works when (the Stieltjes approximation of) the system matrix corresponds to a five point stencil applied to a regular rectangular grid. It originates in the late eighties with the use, in some multigrid algorithms, of intermediate skew meshes, see [1] and the references therein. Indeed, when such meshes are added to the standard multigrid framework, it becomes possible to define the matrix on each successive grid according to an incomplete factorization process.

This was considered in Axelsson and Eijkhout [5,6,7]. They proved in particular that some W cycle version is of optimal order, on the basis of the theory developed by Axelsson and Vassilevski [10] (see also [2,8] for more recent results). However, in this case, the preconditioner is only implicitly defined and not so easy to implement, especially on a parallel platform (see [9,17], for instance). Now, in the V cycle case, the preconditioner reduces to a *modified* incomplete factorization (A and B have same row-sum) computed with respect to a recursive red–black ordering of the unknowns with fill-in allowed provided the red unknowns in a same level remain uncoupled.

For this method, Axelsson and Eijkhout [5,6,7] proved, considering the model Poisson

problem,

$$\kappa(B^{-1}A) \leq \frac{1}{2} \left(1 + \frac{\sqrt{5}}{5}\right) \left(\frac{1 + \sqrt{5}}{2}\right)^\ell \approx 0.7 (1.6)^\ell \tag{1.4}$$

where ℓ is the number of levels; if $\ell = \log_2 h^{-1}$ (see Section 2), this becomes

$$\kappa(B^{-1}A) \leq 0.7 h^{-0.69}. \tag{1.5}$$

The same method was also investigated by Brand and Heinemann [14] and Brand [13]. In the latter paper, it is proved, for the same model problem,

$$\kappa(B^{-1}A) \leq (\sqrt{2})^\ell \tag{1.6}$$

i.e., with $\ell = \log_2 h^{-1}$,

$$\kappa(B^{-1}A) \leq h^{-0.5}. \tag{1.7}$$

Here we show that

$$\kappa(B^{-1}A) \leq \frac{\sqrt{5}(\sqrt{5} - 1)^{\ell-1}}{1 + (-1)^\ell \left(\frac{3-\sqrt{5}}{2}\right)^{\ell-1}} \approx 1.8 (1.23)^\ell, \tag{1.8}$$

i.e.,

$$\kappa(B^{-1}A) \leq 1.8 h^{-0.306}. \tag{1.9}$$

This improves the above results by nearly an order of magnitude. Moreover, it will be seen that it is better than standard $\log^2 h^{-1}$ estimates for any realistic grid size, showing that the method is (as efficient as methods) of nearly optimal order ¹.

The result (1.8) is based on an algebraic computable upper bound, which can be evaluated analytically for some model examples. Resorting to numerical computation, we address its behaviour in the presence of Neumann boundary conditions and/or discontinuities in the PDE coefficients, and show that the bound derived for the model problem still holds, i.e., that the method is robust, provided that the actual grid results from successive refinements applied to a coarse grid, and that the ordering procedure has been carefully applied to leave in the last level all coarse grid nodes. The practical meaning of this restriction is described in the next section.

Finally, we compare our bound with the actual condition number. In the case where the bound (1.8) holds, it overestimates the condition number by a factor of about two only. On the other hand, we obtain effectively better condition numbers and more robustness when leaving all coarse grid nodes in the last level, in comparison with, for instance, the straightforward application of the ordering procedure given in [13].

In the remainder of the paper, we shall assume that the system matrix is a Stieltjes matrix, i.e., that it is positive definite with non-positive off-diagonal entries.

Actually, only the matrix to factorize has to be such, and it may be any relevant (i.e., ‘spectrally equivalent’) approximation of the system matrix. Techniques to derive such approximations are given in [3,4,11]. We shall also assume for convenience that A is (not strictly) diagonally dominant, but the general Stieltjes case is readily included by transform-

¹ Strictly speaking, ‘nearly optimal’ is reserved in the literature for $\mathcal{O}(\log^\sigma h^{-1})$, $\sigma > 0$ condition numbers.

ing (1.1) into $A'x' = b'$, where $A' = \Delta A \Delta$, $x' = \Delta^{-1}x$ and $b' = \Delta b$, with $\Delta = (v_i \delta_{ij})$ such that $v > 0$, $Av \geq 0$ (such a vector always exists, see [12]).

We shall also assume that the system matrix (the matrix to factorize) has a non-zero pattern corresponding to the application of a five point stencil on a regular rectangular grid (note that only the grid has to be regular, not the boundary). Deriving such approximations is less common, but it should be noted that the matrix resulting from the five point finite difference or linear finite element discretization is spectrally equivalent to the matrices resulting from higher order finite difference or finite element schemes [4,6].

With these assumptions, we exclude the cases where the system (1.1) is singular, as it may arise in the discrete PDE context when the solution is only defined up to a constant. Actually, we could include such cases on the basis of the results in [18,19,20]. Basically, the kernel of A is then spanned by the constant vector, and, as B has same row-sum, the preconditioner is also singular. However, it is clear from the results in [19] that, in the context considered here, it will have the same kernel as A . Hence, if the system (1.1) is consistent, the system $Bg_k = b - Ax_k$ to solve at each iteration is also consistent and one may perform conjugate gradient iterations. The bound (1.3) still holds, where $\kappa(B^{-1}A)$ is to be understood as the ratio of the extreme non trivial eigenvalues of the pencil $A - \nu B$ [18]. Hence, the spectral bounds derived in this paper, which are based on inequalities of the type $\underline{\nu}(z, Bz) \leq (z, Az) \leq \bar{\nu}(z, Bz) \quad \forall z \in C^n$, are readily extended to the case of A, B singular. We do not make explicitly this extension here, only in order to avoid the use of a more involved formalism, with generalized inverses and so on. Interested readers will find all relevant details in the above quoted works. See also [20] for considerations about the stability.

The remainder of this paper is organized as follows: recursive red–black orderings are described in Section 2; the associated factorization algorithm is given in Section 3; in Section 4, we prove our algebraic upper bound and evaluate it analytically for the model problem; results of numerical experiments are reported in Section 5.

1.1. General terminology and notation

All vectors belong to \mathcal{C}^n ; all matrices are $n \times n$ real matrices. The symbol A^t denotes the transpose 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 $A \leq B$ ($A < B$) if $a_{ij} \leq b_{ij}$ ($a_{ij} < b_{ij}$) for all i, j ; A is called non-negative (positive) if $A \geq 0$ ($A > 0$). If $A = (a_{ij})$, we denote by $\text{diag}(A)$ the (diagonal) matrix whose entries are $a_{ii}\delta_{ij}$ and we let $\text{offdiag}(A) = A - \text{diag}(A)$; $\text{upp}(A)$ is the strictly upper triangular part of A , i.e. $(\text{upp}(A))_{ij} = a_{ij}$ if $j > i$ and 0 otherwise. The Hadamard product $A * B$ of the matrices $A = (a_{ij})$ and $B = (b_{ij})$ of the same dimensions is defined using element by element multiplication : $(A * B)_{ij} = a_{ij}b_{ij}$. $e = (1 \dots 1)^t$ is the vector with all components equal to unity.

2. Recursive red–black orderings

Consider a regular rectangular grid on which the nodes are connected according to a five-point scheme. The grid may be ordered red–black, and it is known that the elimination of the red nodes results in a reduced system in which the remaining nodes are coupled according to a (skew) nine point scheme.

Within the framework of an incomplete factorization method, one may neglect those fill entries which do not correspond to the (skew) five-point connection scheme. If the row-sum is preserved, one then approximates a nine-point stencil by a (spectrally equivalent) five point one [5,6,7,13]. The reduced system may then also be ordered red-black, which allows us to repeat the process until the number of remaining nodes is sufficiently small to allow an exact factorization (acceptance of all new fills).

This is the basic idea behind the method. However, in practice, the ordering has to be set up prior to the factorization.

In view of this, we label the nodes in the grid G (whose boundary may be arbitrary) by the couple (i, j) of their row index j and column index i . We then define L_1 as the red nodes of a conventional red-black ordering, i.e., L_1 is the set of nodes such that $i + j$ is either odd or even, depending on the case. If one decides that (\hat{i}, \hat{j}) has to be a black node, this may be formally written:

$$L_1 = \{(i, j) \in G \mid i + j = \hat{i} + \hat{j} + 1 \pmod{2}\} \tag{2.1}$$

Next, we consider a (skew) five-point connection scheme between the remaining nodes, i.e., we assume connections between (i, j) and $(i \pm 1, j \pm 1)$, $(i, j) \notin L_1$. The red nodes of a red-black ordering for such a connectivity pattern are the nodes (i, j) for which i is either odd or even, depending on the case. Formally:

$$L_2 = \{(i, j) \in G \setminus L_1 \mid i = \hat{i} + 1 \pmod{2}\} \tag{2.2}$$

where, here again (\hat{i}, \hat{j}) is left in the remaining set $G \setminus (L_1 \cup L_2)$. Note that, equivalently,

$$L_2 = \{(i, j) \in G \setminus L_1 \mid j = \hat{j} + 1 \pmod{2}\} \tag{2.3}$$

while

$$G \setminus (L_1 \cup L_2) = \{(i, j) \in G \mid i = \hat{i} \pmod{2} \text{ and } j = \hat{j} \pmod{2}\} \tag{2.4}$$

Hence, the remaining grid is the original grid from which one has deleted one row and one column out of two. Repeating both preceding steps on this reduced grid leads to

$$L_3 = \{(i, j) \in G \setminus (L_1 \cup L_2) \mid i + j = \hat{i} + \hat{j} + 2 \pmod{4}\} \tag{2.5}$$

$$L_4 = \{(i, j) \in G \setminus (L_1 \cup L_2 \cup L_3) \mid i = \hat{i} + 2 \pmod{4}\} \tag{2.6}$$

Here again, (\hat{i}, \hat{j}) has been left in the remaining set $G \setminus (\cup_{i=1}^4 L_i)$.

In an ℓ step recursive red black (or RRB) ordering, this process is repeated until L_ℓ is set. One may check that a general definition for L_k , $1 \leq k \leq \ell$, is

$$L_k = \{(i, j) \in G \setminus (\cup_{i=1}^{k-1} L_i) \mid i + j = \hat{i} + \hat{j} + 2^{\frac{k-1}{2}} \pmod{2^{\frac{k+1}{2}}}\} \tag{2.7}$$

for k odd, and

$$L_k = \{(i, j) \in G \setminus (\cup_{i=1}^{k-1} L_i) \mid i = \hat{i} + 2^{\frac{k}{2}-1} \pmod{2^{\frac{k}{2}}}\} \tag{2.8}$$

for k even. $L_{\ell+1}$ is then defined by

$$L_{\ell+1} = G \setminus \left(\bigcup_{i=1}^{\ell} L_i \right) \tag{2.9}$$

so that $\mathcal{L} = (L_k)_{k=1, \dots, \ell+1}$ is a partitioning of G . This partitioning is illustrated in Figure 1 for a 17×17 grid, with $\ell = 4$ and $\hat{i} = \hat{j} = 1$.

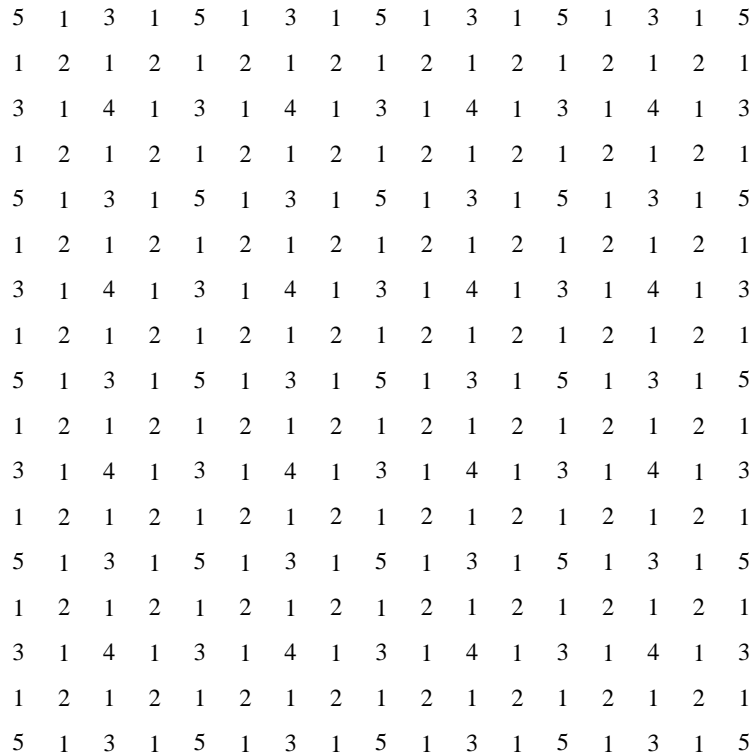


Figure 1. Level structure associated with a four-step RRB ordering defined on a 17×17 grid with node $(1, 1)$ in L_5 ; each node is represented by the index of the level to which it belongs

An $(\ell$ step) RRB ordering is an ordering consistent with that partitioning, i.e., an ordering for which the first $\#(L_1)$ nodes are those of L_1 , the next $\#(L_2)$ those of L_2 , etc. The definitions above allow us to implement quickly an algorithm which sets the permutation vector for such an ordering.

Note that the definition in [13] is similar, except that it is restricted to the case $\hat{i} = \hat{j} = 0$, i.e. no parameter is considered to allow to choose, at each step, which of the nodes are red and which of the nodes are black. With our definition, one sees in particular that, letting

$$\hat{L} = \left\{ (i, j) \in G \mid i = \hat{i} \pmod{2^{\frac{\hat{\ell}}{2}}} \text{ and } j = \hat{j} \pmod{2^{\frac{\hat{\ell}}{2}}} \right\} \tag{2.10}$$

where

$$\hat{\ell} = \begin{cases} \ell & \text{if } \ell \text{ is even} \\ \ell + 1 & \text{if } \ell \text{ is odd} \end{cases}$$

one has

$$\hat{L} \subset L_{\ell+1} \tag{2.11}$$

Assume then that the actual grid results from $m \geq \frac{\hat{\ell}}{2}$ successive refinements steps applied to a coarse (or semi-coarse) grid, or, equivalently, that each coarse grid cell has been divided into $m_1 2^{\frac{\hat{\ell}}{2}} \times m_2 2^{\frac{\hat{\ell}}{2}}$ fine cells. These relations imply that by selecting for (\hat{i}, \hat{j}) any of the coarse grid nodes, all of them will be left in $L_{\ell+1}$. In other words, the RRB ordering becomes a true multilevel ordering (except that in standard multilevel orderings one generally groups together the nodes in L_k and in L_{k+1} , $k = 1, 3, \dots$).

In Section 5, we shall see that much better results are obtained for these ‘multilevel’ RRB orderings.

3. The approximate factorization process

Once the ordering has been chosen, one has still to define the set of positions where fill-in will be permitted during the factorization.

In the context of this paper, considering a node (i, j) in L_k , $1 \leq k \leq \ell$, the corresponding row in the upper triangular factor may have at most four non-zero entries, connecting it to four nodes in $L_{k+1} \cup \dots \cup L_{\ell+1}$ that are nearest to it in the grid, namely the nodes $(i \pm 2^{\frac{k-1}{2}}, j)$, $(i, j \pm 2^{\frac{k-1}{2}})$ if k is odd and the nodes $(i \pm 2^{\frac{k}{2}-1}, j \pm 2^{\frac{k}{2}-1})$ if k is even. These non-zero entries correspond indeed to a five-point connection scheme between the nodes in L_k and those in $L_{k+1} \cup \dots \cup L_{\ell+1}$.

An interesting observation in [13] is that such a non-zero pattern is obtained when discarding fill-in between the nodes in the same level L_k , $1 \leq k \leq \ell$, which is much simpler to implement than a thorough verification of the condition above.

Indeed, consider for instance in Fig. 1 the nodes in level L_1 , and assume that they are connected to only their four nearest neighbours in $L_2 \cup \dots \cup L_{\ell+1}$. For a typical node i , it is seen that two of them, say j_1, j_2 , are in L_2 , one of them, say j_3 , is in L_3 , and the remaining, say j_4 is in $L_4 \cup \dots \cup L_{\ell+1}$. The elimination of i creates a connection $j_1 j_2$ which has to be discarded (j_1 and j_2 are in a same level), four connections $j_1 j_3, j_1 j_4, j_2 j_3, j_2 j_4$ which correspond to permitted connections between L_2 and $L_3 \cup \dots \cup L_{\ell+1}$, and one connection $j_3 j_4$ which also corresponds to a permitted connection between j_3 and one of its nearest neighbours in $L_4 \cup \dots \cup L_{\ell+1}$.

An induction argument shows that essentially the same happens when eliminating the nodes in the subsequent levels.

We now recall for completeness an algorithm which implements the computation of the upper triangular factor $U = (u_{ij})$ of the modified incomplete factorization preconditioner $B = U^t P^{-1} U$, where $P = \text{diag}(U)$. Fill-in is permitted when coupling nodes in different levels and inside the last level. One should thus keep in mind that, with this algorithm, we (implicitly) assume that the system matrix has initially non-zero entries only in positions where fill-in is expected according to the considerations above. Otherwise, the resulting fill-in may be unpredictable.

```

initialize:  $u_{ij} := a_{ij} \quad \forall 1 \leq i \leq j \leq n$ 
for  $k = 1, \dots, \ell + 1$  do
  for all  $i$  in  $L_k$  do
    for all  $j_1 > i$  such that  $u_{ij_1} \neq 0$  do
       $u_{j_1 j_1} := u_{j_1 j_1} - \frac{u_{ij_1}^2}{u_{ii}}$ 
      for all  $j_2 > j_1$  such that  $u_{ij_2} \neq 0$  do

        if  $j_1 \in L_{k_1}, j_2 \in L_{k_2}$  with  $k < k_1 < k_2$  or  $k_1 = k_2 = \ell + 1$ 
           $u_{j_1 j_2} := u_{j_1 j_2} - \frac{u_{ij_1} u_{ij_2}}{u_{ii}}$ 

        otherwise:
           $u_{j_1 j_1} := u_{j_1 j_1} - \frac{u_{ij_1} u_{ij_2}}{u_{ii}}$ 
           $u_{j_2 j_2} := u_{j_2 j_2} - \frac{u_{ij_1} u_{ij_2}}{u_{ii}}$ 

      end( $j_2$  loop)
    end( $j_1$  loop)
  end( $i$  loop)
end( $k$  loop)

```

If A is a (not strictly) diagonally dominant Stieltjes matrix, it is proved in [20] that this algorithm cannot break down, and produces positive diagonal entries u_{ii} , $1 \leq i \leq n$ provided that one has at least one off-diagonal non-zero u_{ij} in each row but the last, which always holds for the kind of factorization considered here.

From a practical point of view, note that no dynamic data structure is needed to create space for the fill entries. The off-diagonal part of U related to the nodes in $L_1 \cup \dots \cup L_\ell$ may indeed be stored in a vector of length $4n$ with which an integer vector for column indexes is associated. For the part related to the nodes in $L_{\ell+1}$, one may use a band storage scheme. After factorization, these vectors may be compressed in a general sparse matrix format which stores only the actual non-zero entries. Some useful routines are provided in the SPARSKIT package for the so-called compressed sparse row format [23]. The program which served for the numerical tests in Section 5 also makes use of some reordering routines from SPARSKIT.

To complete the description, we have still to discuss the choice of ℓ , that is, the size $\#L_{\ell+1}$ of the block that one will have to factorize exactly. This is also important for the conditioning because, as it will be seen, κ increases with ℓ .

In [13] it is proposed to use ℓ such that the factorization cost is kept $\mathcal{O}(n)$, which in addition ensures that the cost of the part of the triangular solutions related to the nodes in $L_{\ell+1}$ is fairly negligible. For square grids with nodes in $L_{\ell+1}$ naturally ordered, this means using ℓ such that

$$\#(L_{\ell+1}) \approx \sqrt{n} \tag{3.1}$$

that is, since $\#(L_{\ell+1}) \approx 2^{-\ell}n$, ℓ such that

$$2^\ell \approx \sqrt{n} \tag{3.2}$$

One may however think that slightly smaller values for ℓ realize a better compromise between factorization and solution cost.

In [9], where a somewhat similar method (but based on hierarchical finite elements) is discussed, it is proposed to exchange the exact factorization inside level $\ell + 1$ for an (inner) iterative solution of the corresponding system. Avoiding the factorization cost, one may then use still smaller values for ℓ . Since, as will be seen, increasing ℓ by 1 unity increases the number of iterations by about 10%, the optimal ℓ is such that the computational cost of the operations related to the nodes in $L_{\ell+1}$ represents less than 10% of the total cost of each iteration.

In the remainder of this paper, we shall consider the rule (3.2), which will allow a fair comparison with previous results, and for which we will be able to prove sufficiently nice conditioning properties. Note however in this respect that the bound $\kappa \leq \mathcal{O}(h^{-\frac{1}{3}})$ is obtained in [15] by combining the analysis in [13] with $\ell \approx \frac{2}{3} \log_2 h^{-1}$, which is practically viable only when one has to solve many systems differing only by the right-hand side. In such cases, one can of course combine the same rule with our analysis to derive still stronger estimates.

4. Conditioning analysis

In view of the theoretical analysis, we need explicit relations on the entries of B and U .

First, discarded fill entries are added to the diagonal, whence

$$Be = Ae \tag{4.1}$$

where $e = (1 \ 1 \ \dots \ 1)^t$. As is well known, since these discarded fills are non-positive, $A - B$ is then a symmetric M -matrix, whence

$$v_{\min}(B^{-1}A) = 1 \tag{4.2}$$

Therefore,

$$\kappa(B^{-1}A) = v_{\max}(B^{-1}A), \tag{4.3}$$

and we need to analyse only the largest eigenvalue.

In view of this, we note that

$$\beta * \text{offdiag}(B) = \text{offdiag}(A) \tag{4.4}$$

where $*$ stands for the Hadamard multiplication $((\beta * C)_{ij} = \beta_{ij} C_{ij})$ and where

$$\beta_{ij} = \begin{cases} 0 & \text{if } i, j \in L_k, 1 \leq k \leq \ell \\ 1 & \text{otherwise} \end{cases} \tag{4.5}$$

(the off-diagonal entries of B may differ from those of A only in positions where fill-in was forbidden).

Now, let $U = P - F$ where $P = \text{diag}(U)$, $F = (f_{ij})$ and define, for $k = 1, \dots, \ell + 1$

$$F_k : (F_k)_{ij} = \begin{cases} f_{ij} & \text{if } i \in L_k \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$E_k = F_k^t$, and, for $k = 1, \dots, \ell$

$$F_k^{(1)} : (F_k^{(1)})_{ij} = \begin{cases} (F_k)_{ij} & \text{if } i \in L_k, j \in L_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

$$F_k^{(2)} : (F_k^{(2)})_{ij} = \begin{cases} (F_k)_{ij} & \text{if } i \in L_k, j \in \cup_{r=k+2}^{\ell+1} L_r \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

$$E_k^{(1)} = F_k^{(1)t}, E_k^{(2)} = F_k^{(2)t}$$

Note that

$$F_k = F_k^{(1)} + F_k^{(2)} \quad k = 1, \dots, \ell \tag{4.9}$$

(no connection in F between nodes in a same level L_k , $1 \leq k \leq \ell$)

Therefore, with $E = F^t$,

$$\begin{aligned} B &= (P - E) P^{-1} (P - F) \\ &= P + \sum_{k=1}^{\ell-1} \left(-E_k - F_k + E_k^{(1)} P^{-1} F_k^{(1)} + E_k^{(1)} P^{-1} F_k^{(2)} \right. \\ &\quad \left. + E_k^{(2)} P^{-1} F_k^{(1)} + E_k^{(2)} P^{-1} F_k^{(2)} \right) \\ &\quad - E_\ell - F_\ell + E_\ell P^{-1} F_\ell - E_{\ell+1} - F_{\ell+1} + E_{\ell+1} P^{-1} F_{\ell+1} . \end{aligned}$$

Obviously, for $k = 1, \dots, \ell - 1$, $\beta * E_k^{(1)} P^{-1} F_k^{(1)} = 0$ and

$$\beta * \left(E_k^{(1)} P^{-1} F_k^{(2)} + E_k^{(2)} P^{-1} F_k^{(1)} \right) = E_k^{(1)} P^{-1} F_k^{(2)} + E_k^{(2)} P^{-1} F_k^{(1)}$$

while $(E_{\ell-1}^{(2)} P^{-1} F_{\ell-1}^{(2)})_{ij}$, $(E_\ell P^{-1} F_\ell)_{ij}$ and $(E_{\ell+1} P^{-1} F_{\ell+1})_{ij}$ are non-zero only for $i, j \in L_{\ell+1}$, so that the Hadamard multiplication by β leaves these terms unchanged.

On the other hand, it was observed in Section 3 that there are at most two non-zero entries connecting a node in L_k , $1 \leq k \leq \ell - 2$ to the nodes in $L_{k+2} \cup \dots \cup L_{\ell+1}$, with at most one in L_{k+2} and at most one in $L_{k+3} \cup \dots \cup L_{\ell+1}$. Hence, all fills corresponding to $E_k^{(2)} P^{-1} F_k^{(2)}$ are accepted and contribute to $F_{k+2} + E_{k+2}$.

Thus,

$$\begin{aligned} \beta * B &= \beta * P - E_1 - F_1 - \left(E_2 + F_2 - E_1^{(1)} P^{-1} F_1^{(2)} - E_1^{(2)} P^{-1} F_1^{(1)} \right) \\ &\quad - \sum_{k=3}^{\ell} \left(E_k + F_k - E_{k-1}^{(1)} P^{-1} F_{k-1}^{(2)} - E_{k-1}^{(2)} P^{-1} F_{k-1}^{(1)} \right. \\ &\quad \left. - \text{offdiag} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right) \right) \\ &\quad - \left(E_{\ell+1} + F_{\ell+1} - E_{\ell-1}^{(2)} P^{-1} F_{\ell-1}^{(2)} - E_\ell P^{-1} F_\ell - E_{\ell+1} P^{-1} F_{\ell+1} \right) . \end{aligned}$$

Letting $-F^{(A)}$ be the strictly upper triangular part of A , and decomposing $F = \sum_{k=1}^{\ell+1} F_k^{(A)}$ similarly as in (4.6), we then obtain, equating each term in the upper triangular part of $\beta * B$ with the corresponding part of A :

$$\begin{aligned} F_1 &= F_1^{(A)} \\ F_2 &= F_2^{(A)} + E_1^{(1)} P^{-1} F_1^{(2)} \\ F_k &= F_k^{(A)} + E_{k-1}^{(1)} P^{-1} F_{k-1}^{(2)} + \text{upp} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right) \quad k = 3, \dots, \ell \\ F_{\ell+1} &= F_{\ell+1}^{(A)} + \text{upp} \left(E_{\ell-1}^{(2)} P^{-1} F_{\ell-1}^{(2)} + E_{\ell} P^{-1} F_{\ell} + E_{\ell+1} P^{-1} F_{\ell+1} \right) \end{aligned} \tag{4.10}$$

Finally, we note that (4.1), together with $Ae \geq 0$, implies

$$Pe \geq Fe \tag{4.11}$$

Relations (4.1), (4.6)–(4.11) form the basic assumptions of the next theorem which contains our main algebraic result. The resulting upper eigenvalue bound is a function of some parameters, to be chosen such that auxiliary matrices are non-negative definite. The practical computation of these parameters will be addressed immediately after the proof of the theorem.

Theorem 4.1. *Let $A = (a_{ij})$ be a Stieltjes matrix, $P = (p_{ii}\delta_{ij})$ a diagonal matrix with positive diagonal entries, $F = (f_{ij})$ a non-negative strictly upper triangular matrix and $E = F^t$.*

Let $\mathcal{L} = (L_k)_{k=1, \dots, \ell+1}$ be a partitioning of $[1, n]$ such that the indices in L_{k+1} follow those in L_k , $1 \leq k \leq \ell$.

Define, for $k = 1, \dots, \ell + 1$

$$F_k : (F_k)_{ij} = \begin{cases} f_{ij} & \text{if } i \in L_k \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

$$F_k^{(A)} : (F_k^{(A)})_{ij} = \begin{cases} -a_{ij} & \text{if } i \in L_k, \text{ and } j > i \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

$E_k = F_k^t$, $E_k^{(A)} = F_k^{(A)t}$, and, for $k = 1, \dots, \ell$

$$F_k^{(1)} : (F_k^{(1)})_{ij} = \begin{cases} (F_k)_{ij} & \text{if } i \in L_k, j \in L_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

$$F_k^{(2)} : (F_k^{(2)})_{ij} = \begin{cases} (F_k)_{ij} & \text{if } i \in L_k, j \in \cup_{r=k+2}^{\ell+1} L_r \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

$$E_k^{(1)} = F_k^{(1)t}, E_k^{(2)} = F_k^{(2)t}$$

Assume that, for $k = 1, \dots, \ell$

$$F_k = F_k^{(1)} + F_k^{(2)} \tag{4.16}$$

and that

$$\begin{aligned}
 F_1 &= F_1^{(A)} \\
 F_2 &= F_2^{(A)} + E_1^{(1)} P^{-1} F_1^{(2)} \\
 F_k &= F_k^{(A)} + E_{k-1}^{(1)} P^{-1} F_{k-1}^{(2)} + \text{upp} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right) \quad k = 3, \dots, \ell \\
 F_{\ell+1} &= F_{\ell+1}^{(A)} + \text{upp} \left(E_{\ell-1}^{(2)} P^{-1} F_{\ell-1}^{(2)} + E_{\ell} P^{-1} F_{\ell} + F_{\ell+1} P^{-1} F_{\ell+1} \right)
 \end{aligned} \tag{4.17}$$

Define further, for $k = 3, \dots, \ell - 1$

$$\hat{F}_k = \text{upp} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right) \tag{4.18}$$

$$\hat{F}_k^{(1)} : \left(\hat{F}_k^{(1)} \right)_{ij} = \begin{cases} \left(\hat{F}_k \right)_{ij} & \text{if } i \in L_k, j \in L_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

$$\hat{F}_k^{(2)} : \left(\hat{F}_k^{(2)} \right)_{ij} = \begin{cases} \left(\hat{F}_k \right)_{ij} & \text{if } i \in L_k, j \in \cup_{r=k+2}^{\ell+1} L_r \\ 0 & \text{otherwise} \end{cases} \tag{4.20}$$

$\hat{E}_k = \hat{E}_k^t, \hat{E}_k^{(1)} = \hat{F}_k^{(1)t}, \hat{E}_k^{(2)} = \hat{F}_k^{(2)t}$, and let $\hat{E}_1^{(m)} = \hat{F}_1^{(m)} = \hat{E}_2^{(m)} = \hat{F}_2^{(m)} = 0, m = 1, 2$.

Let $T_k^{(m)}, k = 1, \dots, \ell - 1, m = 1, 2$, be the matrices with zero row-sum, such that

$$\begin{aligned}
 \text{offdiag} \left(T_k^{(m)} \right) &= \text{offdiag} \left(E_k^{(m)} P^{-1} F_k^{(m)} \right) - \tau_k \left(E_k^{(m)} + F_k^{(m)} \right) \\
 &\quad - (1 - \tau_k) \left(\hat{E}_k^{(m)} + \hat{F}_k^{(m)} \right)
 \end{aligned} \tag{4.21}$$

where $\tau_k, 0 \leq \tau_k < 1$, are real numbers such that $T_k^{(m)}$ is non-negative definite for $m = 1, k = 1, \dots, \ell - 1$ and $m = 2, k = 1, \dots, \ell - 3$.

If, letting $e = (1 \ 1 \ \dots \ 1)^t$,

$$P e \geq F e \tag{4.22}$$

and if $B = (P - E)P^{-1}(P - F)$ is such that

$$B e \geq (1 - \alpha) A e \tag{4.23}$$

where

$$\alpha = \prod_{k=1}^{\ell-1} (1 - \tau_k) \tag{4.24}$$

then

$$\nu_{\max}(B^{-1}A) \leq \alpha^{-1} \tag{4.25}$$

Proof

First, with (4.17)

$$\text{upp}(B) = -F + \text{upp} \left(E P^{-1} F \right)$$

$$\begin{aligned}
 &= -\sum_{k=1}^{\ell+1} F_k + \sum_{k=1}^{\ell-1} \text{upp} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right) + \sum_{k=1}^{\ell-1} E_k^{(1)} P^{-1} F_k^{(2)} \\
 &\quad + \sum_{k=1}^{\ell-1} \text{upp} \left(E_k^{(2)} P^{-1} F_k^{(2)} \right) + \text{upp} \left(E_\ell P^{-1} F_\ell + E_{\ell+1} P^{-1} F_{\ell+1} \right) \\
 &= -\sum_{k=1}^{\ell+1} F_k^{(A)} + \sum_{k=1}^{\ell-1} \text{upp} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right)
 \end{aligned}$$

and thus

$$\text{offdiag}(B - \alpha A) = \sum_{k=1}^{\ell-1} \text{offdiag} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right) + (1 - \alpha) \text{offdiag}(A)$$

Next, we note that by (4.16), there are no non-zero entries $(F_k)_{ij}$ in F_k such that $j \in L_k$. This is *a fortiori* true for \hat{F}_k , whence

$$\hat{F}_k = \hat{F}_k^{(1)} + \hat{F}_k^{(2)}, \quad k = 3, \dots, \ell - 1$$

Letting

$$\alpha_k = \prod_{i=k}^{\ell-1} (1 - \tau_i), \quad k = 1, \dots, \ell - 1$$

$$\alpha_\ell = 1$$

one has, since $\alpha_k/\alpha_{k+1} = 1 - \tau_k$, $k = 1, \dots, \ell - 1$,

$$\begin{aligned}
 &\sum_{k=1}^{\ell-1} \alpha_{k+1} \text{offdiag} \left(T_k^{(1)} \right) + \sum_{k=1}^{\ell-3} \alpha_{k+1} \text{offdiag} \left(T_k^{(2)} \right) \\
 &= \sum_{k=1}^{\ell-1} \alpha_{k+1} \text{offdiag} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right) \\
 &\quad - \sum_{k=1}^{\ell-1} (\alpha_{k+1} - \alpha_k) \left(E_k^{(1)} + F_k^{(1)} \right) \\
 &\quad - \sum_{k=3}^{\ell-1} \alpha_k \left(\hat{E}_k^{(1)} + \hat{F}_k^{(1)} \right) + \sum_{k=1}^{\ell-3} \alpha_{k+1} \text{offdiag} \left(E_k^{(2)} P^{-1} F_k^{(2)} \right) \\
 &\quad - \sum_{k=1}^{\ell-3} (\alpha_{k+1} - \alpha_k) \left(E_k^{(2)} + F_k^{(2)} \right) - \sum_{k=3}^{\ell-3} \alpha_k \left(\hat{E}_k^{(2)} + \hat{F}_k^{(2)} \right) \\
 &\geq \sum_{k=1}^{\ell-1} \alpha_{k+1} \text{offdiag} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right)
 \end{aligned}$$

$$\begin{aligned}
 & - \sum_{k=1}^{\ell-1} (\alpha_{k+1} - \alpha_k) (E_k + F_k) \\
 & - \sum_{k=3}^{\ell-1} (\alpha_k - \alpha_{k-1}) \text{offdiag} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right)
 \end{aligned}$$

On the other hand, (4.22) implies $Pe \geq F_k e$ for $1 \leq k \leq \ell + 1$. The matrices W_k such that

$$\begin{aligned}
 W_k e &= 0 \\
 \text{offdiag} (W_k) &= \text{offdiag} \left((P - E_k) P^{-1} (P - F_k) \right)
 \end{aligned}$$

are therefore non-negative definite by virtue of Lemma 5.1 in [22]. One has, since $\alpha_\ell = 1$,

$$\begin{aligned}
 & \sum_{k=1}^{\ell-1} (1 - \alpha_{k+1}) \text{offdiag} (W_k) \\
 &= \sum_{k=1}^{\ell-1} (1 - \alpha_{k+1}) \left(\text{offdiag} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right) - (E_k + F_k) \right) \\
 &+ \sum_{k=2}^{\ell-1} (1 - \alpha_k) \left(E_{k-1}^{(1)} P^{-1} F_{k-1}^{(2)} + E_{k-1}^{(2)} P^{-1} F_{k-1}^{(1)} \right) \\
 &+ \sum_{k=3}^{\ell} (1 - \alpha_{k-1}) \text{offdiag} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right) \\
 &= \sum_{k=1}^{\ell-1} (1 - \alpha_{k+1}) \left(\text{offdiag} \left(E_k^{(1)} P^{-1} F_k^{(1)} \right) - (E_k + F_k) \right) \\
 &+ \sum_{k=1}^{\ell-1} (1 - \alpha_k) \left(E_k + F_k - E_k^{(A)} - F_k^{(A)} \right) \\
 &+ \sum_{k=3}^{\ell} (\alpha_k - \alpha_{k-1}) \text{offdiag} \left(E_{k-2}^{(2)} P^{-1} F_{k-2}^{(2)} \right)
 \end{aligned}$$

These relations (together with $\alpha = \alpha_1 \leq \alpha_k \leq 1, 1 < k \leq \ell$) show that

$$\text{offdiag} \left(B - \alpha A - \sum_{k=1}^{\ell-1} (1 - \alpha_{k+1}) W_k - \sum_{k=1}^{\ell-1} \alpha_{k+1} T_k^{(1)} - \sum_{k=1}^{\ell-3} \alpha_{k+1} T_k^{(2)} \right) \leq 0 \tag{4.26}$$

On the other hand, this matrix has non-negative row-sum by (4.23). It is thus a symmetric M -matrix, hence non-negative definite. The conclusion then readily follows because $W_k, T_k^{(1)}$ and $T_k^{(2)}$ are non-negative definite. ■

Concerning the numbers $\{\tau_k\}$, note first that they may be computed by considering separately in $T_k^{(m)}$ the contribution of each row. Indeed, let $F_{k(i)}^{(m)} \left(\hat{F}_{k(i)}^{(m)} \right)$ be the matrix

with its i th row equal to the i th row in $F_k^{(m)} (\hat{F}_k^{(m)})$ and the other rows zero, i.e., $(F_{k(i)}^{(m)})_{rs} = \delta_{ri} (F_k^{(m)})_{rs}$ and $(\hat{F}_{k(i)}^{(m)})_{rs} = \delta_{ri} (\hat{F}_k^{(m)})_{rs}$; let also $T_{k(i)}^{(m)}$ be the matrix with zero row-sum such that

$$\text{offdiag} (T_{k(i)}^{(m)}) = \text{offdiag} (E_{k(i)}^{(m)} P^{-1} F_{k(i)}^{(m)} - \tau_k (E_{k(i)}^{(m)} + F_{k(i)}^{(m)}) - (1 - \tau_k) (\hat{E}_{k(i)}^{(m)} + \hat{F}_{k(i)}^{(m)}))$$

where $E_{k(i)}^{(m)} = F_{k(i)}^{(m)t}$, $\hat{E}_{k(i)}^{(m)} = \hat{F}_{k(i)}^{(m)t}$.

Obviously,

$$T_k^{(m)} = \sum_{i \in L_k} T_{k(i)}^{(m)} \quad (m = 1, 2) \tag{4.27}$$

and it is sufficient to check that each $T_{k(i)}^{(m)}$ is non-negative definite.

In view of this, we note that, as observed above, there are at most two non-zero entries in each row of $F_k^{(m)}$ (and thus of $\hat{F}_k^{(m)}$ since $F_k^{(m)} \geq \hat{F}_k^{(m)} \geq 0$). Hence, $T_{k(i)}^{(m)}$ is zero except in a 3×3 diagonal block. Letting f', f'' be the two non-zero entries in $F_{k(i)}^{(m)}$, and \hat{f}', \hat{f}'' the corresponding entries in $\hat{F}_{k(i)}^{(m)}$, this block writes, with $\tau = \tau_k$ and $p = p_i$

$$T(\tau) = \begin{pmatrix} \tau(f' + f'' - \hat{f}' - \hat{f}'') + \hat{f}' + \hat{f}'' & -\tau(f' - \hat{f}') - \hat{f}' & -\tau(f'' - \hat{f}'') - \hat{f}'' \\ -\tau(f' - \hat{f}') - \hat{f}' & \tau(f' - \hat{f}') + \hat{f}' - \frac{f' f''}{p} & \frac{f' f''}{p} \\ -\tau(f'' - \hat{f}'') - \hat{f}'' & \frac{f' f''}{p} & \tau(f'' - \hat{f}'') + \hat{f}'' - \frac{f' f''}{p} \end{pmatrix}$$

Now, $T(\tau) = T(0) + \tau(T(1) - T(0))$. With $f' \geq \hat{f}' \geq 0$, $f'' \geq \hat{f}'' \geq 0$, it is easily seen that $T(1) - T(0)$ is non-negative definite. Hence, if $T(\tau)$ is non-negative definite for some τ , it will be *a fortiori* non-negative definite for any $\bar{\tau} \geq \tau$.

On the other hand, one can easily check ² that $T(0)$ is non-negative definite if and only if $\gamma \geq 0$, where

$$\gamma = \hat{f}' \hat{f}'' - (\hat{f}' + \hat{f}'') \frac{f' f''}{p}$$

whereas, if $\gamma < 0$, $T(\tau)$ is non-negative definite if and only if τ is not less than the positive root of

$$P(\tau) = \alpha \tau^2 + \beta \tau + \gamma$$

where

$$\alpha = (f' - \hat{f}') (f' - \hat{f}''), \quad \beta = \hat{f}' (f'' - \hat{f}'') + (f' - \hat{f}') \hat{f}'' - \frac{f' f''}{p} (f' - \hat{f}' + f'' - \hat{f}'').$$

Note that this root will always be less than one since $T(1)$ is readily found positive definite when $p > f' + f''$, which always holds in practice because of (4.16), (4.22).

² $\begin{pmatrix} b+c & -b & -c \\ -b & b-a & a \\ -c & a & c-a \end{pmatrix}$ with $a, b, c \geq 0$ is non-negative definite if and only if $bc - a(b+c) \geq 0$.

Taking the maximum of this root over the polynomials corresponding to the different $T_{k(i)}^{(m)}$, $m = 1, 2, i \in L_k$, gives thus a value $\tau_k < 1$ for which both $T_k^{(1)}$ and $T_k^{(2)}$ are non-negative definite as required in Theorem 4.1. Note that in particular τ_k will be zero when $\gamma \geq 0$ for all concerned matrices. The model problem analysis below shows that this happens at some levels (more precisely, $\tau_3 = 0$ in that case). It is, however, generally not true that this may happen at every level, because then τ_k would be zero for all k , whence $\kappa(B^{-1}A) = 1$, which would mean that $B = A$. It is nevertheless possible, when $\hat{f}'\hat{f}'' \neq 0$ (that is for level three and higher) to enforce $\gamma \geq 0$ (or, more generally, to control τ_k) by setting the corresponding diagonal entries p_{ii} sufficiently large, which amounts to perturbing the approximate factorization algorithm by adding a non-negative diagonal matrix to the matrix being factorized. The discussion of such a perturbation technique, which implies a decrease of the lowest eigenvalues, lies however beyond the scope of the present paper.

Model problem analysis

We consider here the matrix resulting from the five-point finite difference discretization of the Laplacian on the unit square with Dirichlet boundary conditions and uniform mesh size h in both directions. Hence, all diagonal and non-zero off-diagonal entries are equal to 4 and -1 , respectively.

Considering an interior node i in L_k , $1 \leq k \leq \ell$, all of whose nearest neighbours in $L_{k+1} \cup \dots \cup L_{\ell+1}$ are also inside the domain, it follows from a symmetry argument that the four non-zero entries in the corresponding row of F are equal, and further, depend only on the considered level. Let f_k denote the value of these off-diagonal entries for such nodes in L_k , and let p_k be the corresponding diagonal entry in P .

With the help of Fig. 1, one may check that each such non-zero in F_k receives two contributions from the term $E_{k-1}^{(1)}P^{-1}F_{k-1}^{(2)}$ (for $k \geq 2$), and one from the term $\text{upp}(E_{k-2}^{(2)}P^{-1}F_{k-2}^{(2)})$ (for $k \geq 3$). Thus,

$$f_1 = 1, \quad f_2 = 2 \frac{f_1^2}{p_1},$$

$$f_k = \frac{2 f_{k-1}^2}{p_{k-1}} + \frac{f_{k-2}^2}{p_{k-2}}, \quad k = 3, \dots, \ell$$

On the other hand, again because we restrict ourselves to interior nodes, $(Be)_i = (Ae)_i = 0$, implying $(Pe)_i = (Fe)_i$, i.e., $p_k = 4 f_k$. Therefore

$$f_1 = 1, \quad f_2 = \frac{f_1}{2} = \frac{1}{2}$$

$$f_k = \frac{f_{k-1}}{2} + \frac{f_{k-2}}{4} \quad k = 3, \dots, \ell$$

One may check that the solution of a recursion $\mu_k = a\mu_{k-1} + b\mu_{k-2}$ is any linear combination of r_1^k and r_2^k , where r_1, r_2 are the solutions of $r^2 - ar - b = 0$. Taking into

account the given values for f_1, f_2 , we then obtain

$$\begin{aligned} f_k &= \frac{2\sqrt{5}}{5} \left(\left(\frac{1+\sqrt{5}}{4} \right)^k - \left(\frac{1-\sqrt{5}}{4} \right)^k \right) \\ &= \frac{2\sqrt{5}}{5} \left(\frac{1+\sqrt{5}}{4} \right)^k \left(1 - (-1)^k \left(\frac{3-\sqrt{5}}{2} \right)^k \right), \quad k = 1, \dots, \ell \end{aligned}$$

In particular, $f_1 = 1, f_2 = \frac{1}{2}, f_3 = \frac{1}{2}, f_4 = \frac{3}{8}$ and $f_k \approx \frac{2\sqrt{5}}{5} \left(\frac{1+\sqrt{5}}{4} \right)^k$ with a relative error less than 1% for $k \geq 5$.

Considering the computation of τ_k , one readily finds, with $f' = f'' = f_k, \hat{f}' = \hat{f}'' = 0$ for $k = 1, 2$ and $\hat{f}' = \hat{f}'' = \frac{f_{k-2}}{4}, k = 3, \dots, \ell - 1$:

$$\begin{aligned} \tau_1 &= \frac{1}{2}, \quad \tau_2 = \frac{1}{2} \\ \tau_k &= \frac{\frac{2f_k}{4} - \frac{f_{k-2}}{4}}{f_k - \frac{f_{k-2}}{4}} \quad k = 3, \dots, \ell - 1 \end{aligned}$$

Using $\frac{f_{k-2}}{4} = f_k - \frac{f_{k-1}}{2}$, the last equation writes $\tau_k = (f_{k-1} - f_k) / f_{k-1}$, which, incidentally, also holds for $k = 2$. Therefore,

$$\frac{1}{1 - \tau_k} = \frac{f_{k-1}}{f_k} \quad k = 2, \dots, \ell - 1$$

and, for $\ell \geq 2$

$$\alpha^{-1}(\ell) = \prod_{k=1}^{\ell-1} \frac{1}{1 - \tau_k} = \frac{1}{1 - \tau_1} \frac{f_1}{f_{\ell-1}} \tag{4.28}$$

i.e.,

$$\kappa \leq \frac{\sqrt{5}(\sqrt{5} - 1)^{\ell-1}}{1 + (-1)^\ell \left(\frac{3-\sqrt{5}}{2} \right)^{\ell-1}} \tag{4.29}$$

To be completely rigorous, we should include a thorough verification that the above values for τ_k are sufficient to make $T_{k(i)}^{(m)}$ non-negative definite for the nodes near boundaries too. This is not so easy however, although we have done it for the first levels. We prefer to refer the reader to the next section, where the actual computation of the bound of Theorem 4.1. reveals that the value (4.28) is obtained, not only for the model problem, but for a quite large range of applications.

Assuming $h^{-1} = 2^\ell$ (see (3.2)), the estimate (4.29) gives

$$\begin{aligned} \kappa &\leq \frac{\sqrt{5}}{\sqrt{5} - 1} \frac{h^{-\log_2(\sqrt{5}-1)}}{1 - \frac{3+\sqrt{5}}{2} h^{\log_2\left(\frac{3+\sqrt{5}}{2}\right)}} \\ &< \frac{\sqrt{5}}{\sqrt{5} - 1} \frac{h^{-\log_2(\sqrt{5}-1)}}{1 - h} \quad \text{for } h \leq \frac{1}{16} \end{aligned} \tag{4.30}$$

Table 1. Our bound (4.29) and the estimate $(\frac{\ell}{2} + 1)^2$ as a function of ℓ

ℓ	$2^\ell \approx \sqrt{n}$	Bound(4.29)	$(\frac{\ell}{2} + 1)^2$
2	4	2.00	4
3	8	4.00	
4	16	4.00	9
5	32	5.33	
6	64	6.40	16
7	128	8.00	
8	356	9.85	25
9	512	12.19	
10	1 024	15.06	36
11	2 048	18.62	
12	4 096	23.01	49
13	8 192	28.44	
14	16 384	35.16	64
15	32 768	43.46	
16	65 536	53.72	81
17	131 072	66.40	
18	262 144	82.07	100
19	524 288	101.45	
20	1 048 576	125.40	121

where $\log_2(\sqrt{5} - 1) = 0.306$ and $\frac{\sqrt{5}}{\sqrt{5}-1} = 1.81$.

On the other hand, $j = \frac{\ell}{2}$ corresponds just to the number of successive refinements between the fine grid and a coarse grid with sufficiently few nodes to allow the use of a direct solver. It is then interesting to compare our bound (4.29) with $(\frac{\ell}{2} + 1)^2$, an estimate frequently met in the analysis of (*V*-cycle) multilevel methods, and often called ‘nearly optimal’ (see [24,25] for instance). These methods present, in addition, some similarities with the method considered here, although they are based on hierarchical finite elements.

Table 1 makes this comparison, from which it turns out that our bound is better for any realistic problem size. The improvement is not dramatic, but sufficient to rank the considered method among the category of ‘nearly optimal’ preconditioners.

5. Numerical results

We first consider the five-point finite difference approximation of $\Delta u = 1$ on the unit square with uniform mesh size $h = 2^{-m}$ in both directions and homogeneous Dirichlet boundary conditions.

We consider ℓ step RRB orderings as defined in Section 2 with $\hat{i} = \hat{j} = 0$. This choice corresponds to the definition in [13] and, for the considered problem, leads also to a multilevel RRB ordering, i.e., leaves in the last level the nodes corresponding to the coarsest meshes.

Our numerical bound, the actual conditioning, and some iteration counts are given in Table 2 for various mesh sizes, using $\ell = m = \log_2 h^{-1}$, and for various ℓ and fixed mesh size $h = \frac{1}{64}$.

For the bound, we obtain in each case exactly the predicted value (4.29), which depends

Table 2. Results for the Poisson problem

h^{-1}	ℓ	Bound(4.25)	κ	# it ($\epsilon = 10^{-3}$)	# it ($\epsilon = 10^{-6}$)
16	4	4.00	1.95	5	9
32	5	5.33	2.39	6	10
64	6	6.40	3.00	8	13
128	7	8.00	3.73	9	15
256	8	9.85	4.63	11	18
512	9	12.19	5.73	13	21
64	4	4.00	1.99	6	10
64	5	5.33	2.44	7	11
64	6	6.40	3.00	8	13
64	7	8.00	3.62	8	14
64	8	9.85	4.33	9	14
64	9	12.19	4.33	9	14

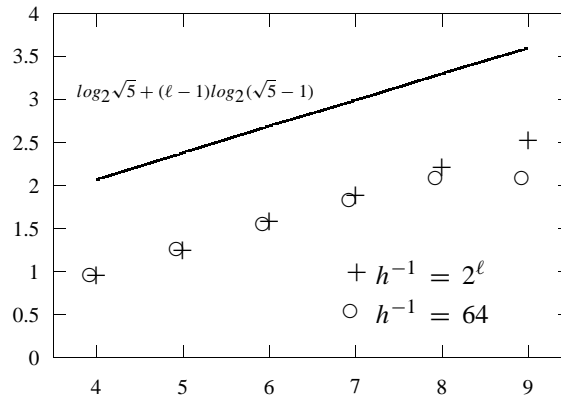


Figure 2. $\log_2 \kappa$ as a function of ℓ for increasing and fixed problem size

only on ℓ . It seems to predict correctly the behaviour of the actual conditioning, which is overestimated by a factor about 2.1 only. This is confirmed in Fig. 2, where we have plotted $\log_2 \kappa$ as a function of ℓ together with the line $\log_2 \sqrt{5} + (\ell - 1) \log_2 (\sqrt{5} - 1)$ which represents the leading asymptotic term in (4.29).

For $h^{-1} = 64$, the conditioning does not increase for ℓ larger than eight because eight RRB steps reduce the 63×63 grid to a 3×3 grid for which one hardly sees the difference between an approximate and an exact factorization.

The given numbers of iterations are those necessary to reduce the norm of the residual $\| r_k \|$ below $\epsilon \| b \|$ when using the preconditioned conjugate gradient algorithm with zero initial approximation. They are fairly small, especially when only a modest precision is required. This is much different than with most incomplete factorization preconditioners, which present small isolated eigenvalues and for which fast convergence may be observed only after some delay [21,22].

As expected from the formula (1.3) and our conditioning estimate, the number of iterations increases by about 10% when ℓ is increased by one unity for fixed h . If h^{-1} increases together

with ℓ , the number of iterations grows more quickly, in our opinion because superlinear convergence effects are more dramatic for modest problem sizes.

We next consider the PDE

$$-\nabla a \nabla u = f \quad \text{in } \Omega = (0, 1) \times (0, 1) \quad (5.1)$$

with

5.1. Problem A

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

$$f = \begin{cases} 100 & \text{in } (\frac{1}{4}, \frac{3}{4}) \times (\frac{1}{4}, \frac{3}{4}) \\ 0 & \text{elsewhere} \end{cases}$$

$$\begin{cases} u = 0 & \text{for } 0 \leq x \leq 1, y = 0 \\ \partial_n u = 0 & \text{on the remaining part of the boundary} \end{cases}$$

5.2. Problem B

$$a = \begin{cases} 10^{-3} & \text{in } (\frac{1}{12}, \frac{1}{2}) \times (\frac{1}{12}, \frac{1}{2}) \\ 1 & \text{elsewhere} \end{cases}$$

$$f = \begin{cases} 1 & \text{in } (\frac{1}{12}, \frac{1}{2}) \times (\frac{1}{12}, \frac{1}{2}) \\ 0 & \text{elsewhere} \end{cases}$$

$$\begin{cases} u = 0 & \text{for } 0 \leq x \leq 1, y = 1 \text{ and } 0 \leq y \leq 1, x = 1 \\ \partial_n u = 0 & \text{on the remaining part of the boundary} \end{cases}$$

In each case, we use the five-point finite difference mesh box integration scheme [16] with uniform mesh size h in both directions, and $h^{-1} = 4 \cdot 2^m$ for Problem A, $h^{-1} = 12 \cdot 2^m$ for Problem B (where m is a non-negative integer).

We consider ℓ step RRB orderings where, according to the rule (3.2), $\ell = m + 2$ for Problem A and $\ell = m + 3$ for Problem B. In both cases, two RRB orderings were considered: a 'non-multilevel RRB' ordering obtained by letting $\hat{i} = \hat{j} = 0$ in our Definitions (the node (1,1) being the first node in the grid corresponding to an unknown, i.e., the node at $x = 0, y = h$ for Problem A and the node at $x = 0, y = 0$ for Problem B), and a 'multilevel RRB' ordering obtained by letting $\hat{i} = 1, \hat{j} = 0$ for Problem A and $\hat{i} = \hat{j} = 1$ for Problem B (so as to leave the coarse grid nodes, for instance the node at $x = y = \frac{1}{2}$, in the last level). Note that, in Problem B, the nodes at $(\frac{1}{12}, \frac{1}{12})$, $(\frac{1}{12}, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{12})$ which one may consider as belonging to the coarsest mesh on which the problem is defined, are effectively in the last level only for $\ell \leq 2m$ (cf., end of Section 2), i.e., for $m \geq 3$ or $h^{-1} \geq 96$.

The results are given in Table 3 for Problem A and in Table 4 for Problem B. The iteration counts refer to the preconditioned conjugate gradient algorithm with zero initial approximation and the same stopping criterion as for the Poisson problem.

Table 3. Results for Problem A

h^{-1}	ℓ	Bound(4.25)	κ	#it ($\epsilon = 10^{-3}$)	#it ($\epsilon = 10^{-6}$)
Multilevel RRB					
16	4	4.00	2.00	7	10
32	5	5.33	2.43	8	13
64	6	6.40	3.016	10	15
128	7	8.00	3.74	12	18
256	8	9.85	4.63	14	20
Non-multilevel RRB					
16	4	216.29	3.11	8	13
32	5	613.42	2.99	9	14
64	6	2589.0	5.14	12	19
128	7	5467	5.14	13	20
256	8	37746	8.43	17	26

Table 4. Results for Problem B

h^{-1}	ℓ	Bound(4.25)	κ	#it ($\epsilon = 10^{-3}$)	#it ($\epsilon = 10^{-6}$)
Multilevel RRB					
24	4	4.00	2.00	5	8
48	5	5.33	2.44	6	10
96	6	6.40	3.031	6	11
192	7	8.00	3.75	8	13
Non-multilevel RRB					
24	4	4004	7.99	8	14
48	5	17.29	2.67	6	11
96	6	103.19	4.45	9	15
192	7	209.44	5.75	10	16

It is seen that the multilevel RRB ordering leads to a very robust method: the computation of the bound delivers exactly the value (4.29) obtained for the model problem, while the actual conditionings and iteration counts are nearly identical. These conclusions are supported by extensive experiments made on various PDEs of the type (5.1).

More generally, since the matrix coefficients in the successive levels are such that the computation of the bound delivers the same result (4.29) for various values and locations of the discontinuities, one may be convinced that this result is essentially independent of these parameters, and therefore the method is robust in the presence of Neumann boundary conditions and/or discontinuities of the type considered here.

The results obtained with non multilevel RRB orderings reveal that the use of these orderings may lead to some difficulties. Clearly, the elimination of the coarse grid nodes perturbs the computation of the bound which is no longer relevant. At the same time, the actual conditioning and the iteration counts deteriorate. This deterioration is not dramatic at all, but indicates a potential lack of robustness. Hence, we advise the use of multilevel RRB orderings whenever it is possible.

Acknowledgements

The first author is supported by the Belgian 'Fonds National de la Recherche Scientifique' (Chercheur qualifié).

This work presents research results of the Belgian Incentive Program 'Information Technology'—Computer Science of the Future, initiated by the Belgian State—Prime Minister's Service—Federal Office for Scientific, Technical and Cultural Affairs (Contract No. IT/IF/14). The Scientific responsibility is assumed by the authors.

This work was also supported by IBM through a research contract between ULB and IBM. We thank Professor Robert Beauwens and anonymous referees for useful comments.

REFERENCES

1. O. Axelsson. A multilevel solution method for nine-point difference approximations, in *Parallel Supercomputing : Methods, Algorithms and Applications*, G. Carey, editor, pp. 191–205. Wiley, 1989.
2. O. Axelsson. The method of diagonal compensation of reduced matrix entries and multilevel iteration. *Journal of Computational and Applied Mathematics*, 38, 31–43, 1991.
3. O. Axelsson. *Iterative Solution Methods*. University Press, Cambridge, 1994.
4. O. Axelsson and V. A. Barker. *Finite Element Solution of Boundary Value Problems. Theory and Computation*. Academic Press, New York, 1984.
5. O. Axelsson and V. Eijkhout. The nested recursive two level factorization for nine-point difference matrices. Technical Report 8936, Department of Mathematics, Catholic University, Nijmegen, The Netherlands, 1989.
6. O. Axelsson and V. Eijkhout. Analysis of recursive 5-point/9-point factorization method, in *Preconditioned Conjugate Gradient Methods*, O. Axelsson and L. Kolotilina, editors, pp. 154–173. Lectures Notes in Mathematics No. 1457, Springer-Verlag, 1990.
7. O. Axelsson and V. Eijkhout. The nested recursive two level factorization for nine-point difference matrices. *SIAM J. Sci. Stat. Comput.*, 12, 1373–1400, 1991.
8. O. Axelsson and M. Neytcheva. Algebraic multilevel iterations for Stieltjes matrices. *Numer. Lin. Alg. Appl.*, 1, 213–236, 1994.
9. O. Axelsson and M. Neytcheva. The short length AMLI method. 1. Technical Report 9417, Department of Mathematics, Catholic University, Nijmegen, The Netherlands, 1994.
10. O. Axelsson and P. S. Vassilevski. Algebraic multilevel preconditioning methods. I. *Numer. Math.*, 56, 157–177, 1989.
11. R. Beauwens and R. Wilmet. Conditioning analysis of positive definite matrices by approximate factorizations. *J. Comput. Appl. Math.*, 26, 257–269, 1989.
12. A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.
13. C. W. Brand. An incomplete factorization preconditioning using repeated red–black ordering. *Numer. Math.*, 61, 433–454, 1992.
14. C. W. Brand and Z. E. Heinemann. A new iterative solution technique for reservoir simulation equations on locally refined grids. *SPE Reservoir Eng.*, 5, 555–560, 1990.
15. P. Ciarlet. Repeated Red Black ordering: a new approach. *Numerical Algorithms*, 7, 295–324, 1994.
16. S. Nakamura. *Computational Methods in Engineering and Science*. John Wiley & Sons, Inc., New York, 1977.
17. M. Neytcheva. Arithmetic and communication complexity of preconditioning methods. PhD thesis, Department of Mathematics, Catholic University, Nijmegen, The Netherlands, 1995.
18. Y. Notay. Polynomial acceleration of iterative schemes associated with subproper splittings. *J. Comput. Appl. Math.*, 24, 153–167, 1988.
19. Y. Notay. Incomplete factorization of singular linear systems. *BIT*, 29, 682–702, 1989.
20. Y. Notay. Solving positive (semi)definite linear systems by preconditioned iterative methods, in *Preconditioned Conjugate Gradient Methods*, O. Axelsson and L. Kolotilina, editors, pp. 105–125. Lectures Notes in Mathematics No. 1457, Springer-Verlag, 1990.

21. Y. Notay. On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math.*, 65, 301–317, 1993.
22. Y. Notay. DRIC : a dynamic version of the RIC method. *Num. Lin. Alg. Appl.*, 1, 511–532, 1994.
23. Y. Saad. SPARSKIT: a basic tool kit for sparse matrix computations. Technical report, University of Minnesota, Minneapolis, 1994.
24. P. Vassilevski. Hybrid V-cycle algebraic multilevel preconditioners. *Math. Comp.*, 58, 489–512, 1992.
25. H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49, 379–412, 1986.