Repeated Red–Black ordering: a new approach

P. Ciarlet, Jr

Commissariat à l'Energie Atomique, Centre d'Etudes de Limeil–Valenton, 94195 Villeneuve-Saint-Georges Cedex, France

Received 22 June 1993; revised 14 December 1993
Communicated by C. Brezinski

Hereafter, we describe and analyze, from both a theoretical and a numerical point of view, an iterative method for efficiently solving symmetric elliptic problems with possibly discontinuous coefficients. In the following, we use the Preconditioned Conjugate Gradient method to solve the symmetric positive definite linear systems which arise from the finite element discretization of the problems. We focus our interest on sparse and efficient preconditioners. In order to define the preconditioners, we perform two steps: first we reorder the unknowns and then we carry out a (modified) incomplete factorization of the original matrix. We study numerically and theoretically two preconditioners, the second preconditioner corresponding to the one investigated by Brand and Heinemann [2]. We prove convergence results about the Poisson equation with either Dirichlet or periodic boundary conditions. For a meshsize h, Brand proved that the condition number of the preconditioned system is bounded by $O(h^{-1/2})$ for Dirichlet boundary conditions. By slightly modifying the preconditioning process, we prove that the condition number is bounded by $O(h^{-1/3})$.

Keywords: Conjugate gradients, sparse modified preconditioners, ordering strategies.

1. Introduction

Modified preconditioners for solving elliptic problems discretized by finite differences have been introduced by Dupont et al. [8]. Gustafsson [12–14] on the one hand and Meijerink and Van der Vorst [16, 17] on the other hand then studied modified incomplete Cholesky factorization methods for solving problems with discontinuous coefficients. More recently, Brand and Heinemann [2] investigated a modified preconditioner based on a reordering of the nodes they called the Repeated Red Black ordering. Moreover, Brand [1] proved some theoretical results concerning this new ordering. The main advantage of a modified preconditioner over its unmodified version is that, generally, even though they require the same amount of work to be built, the condition number of the preconditioned linear system improves when the former is used. In the following, we use the Preconditioned Conjugate Gradient method to solve the

symmetric positive definite discretization of elliptic problems with possibly discontinuous coefficients. We study two preconditioners based on the Repeated Red–Black ordering. The second preconditioner corresponds to the one investigated by Brand and Heinemann. We consider LDL^{T}-incomplete factorizations of the original problem, which we study both numerically and theoretically. In particular, for the modified factorizations, we prove convergence results about the Poisson equation with either a Dirichlet or periodic boundary conditions. For Dirichlet boundary conditions and a meshsize h, Brand proved that the condition number of the preconditioned system is bounded by $O(h^{-1/2})$. By slightly modifying the preconditioning process, we prove that the condition number is bounded by $O(h^{-1/3})$.

The paper is organized as follows. We define the elliptic problems to be solved and their discretization in section 2. In section 3, we introduce the Repeated Red–Black ordering and briefly recall some results about incomplete factorizations in section 4. Then we define the two preconditioners in sections 5 and 6, before studying them numerically (section 7) and theoretically (section 8).

2. The continuous and discrete problems

The aim of this paper is to solve numerically the following problem:

$$-div\,(\mathscr{A}\,grad\,u) = g \quad \text{in } \Omega, \text{ where } \mathscr{A}(x,y) = \begin{pmatrix} a(x,y) & 0 \\ 0 & b(x,y) \end{pmatrix}, \tag{2.1}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{2.2}$$

where $\Omega =]0,1[\times]0,1[$ and a, b and g are given functions, a and b being positive over the domain. The coefficients of $\mathscr{A}$ can have jumps over Ω. Problems with other boundary conditions can be handled without difficulty.

We discretize the problem by using the standard finite element method with isosceles right triangles. All hypotenuses are parallel to the $x+y=1$ diagonal of the domain. The length of the horizontal and vertical edges is a constant number called the meshsize $h = 1/(m+1)$. Here m is an integer equal to the number of nodes (or vertices) in each direction parallel to the x- or y-axis. Denote by n the number of nodes ($n = m^2$). The nodes are labeled sequentially by row.

Then, by using the usual P1 approximation leading to the classical five-point scheme, we obtain a sparse linear system with n equations and n unknowns, where x is approximating the node values of u:

$$Ax = f. \tag{2.3}$$

Definition 2.1
Denote by $\text{tridi}_m(a_i, b_i, c_i)$ the tridiagonal matrix

$$\text{tridi}_m(a_i, b_i, c_i) = \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{m-1} & b_{m-1} & c_{m-1} \\ & & & a_m & b_m \end{pmatrix}.$$

If $a \equiv 1$ and $b \equiv 1$, then

$$A = \begin{pmatrix} T & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T \end{pmatrix}$$

is of order n where $T = \text{tridi}_m(-1, 4, -1)$ and I is the identity matrix of order m.

Properties on a and b ensure that the matrix A is symmetric positive definite. So we use the Preconditioned Conjugate method to solve (2.3). In the following, we first reorder the nodes and then define sparse preconditioners of A.

3. Repeated Red–Black ordering

In this section we introduce a numbering which can be derived from the classical Red–Black numbering (see [2]) or the Alternating Diagonal numbering (see [7]). These two numberings are defined in the following way: if the nodes are associated with the squares of a chessboard, then the Red unknowns are labeled diagonally before the Black ones, which are also labeled diagonally. The main difference with the Repeated Red–Black ordering (denoted by RRB) is that the Black nodes are no longer labeled sequentially.

From now on, we use the Brand and Heinemann terminology (see again [2]). The RRB ordering principle is based on a recursive process. At a given step, k, the set of remaining nodes, that is, the set of "not yet" labeled nodes, is split into two halves respectively called $R^{[k]}$ and $B^{[k]}$. The nodes of $R^{[k]}$ are then labeled sequentially and the process is reiterated for the set of nodes $B^{[k]}$.

We suppose that the number of nodes is of the form $n = 2^l$. Then the process stops naturally when the set of remaining nodes is reduced to only one element. We can also choose to stop the process at step K and then the nodes of $B^{[K]}$ are also labeled sequentially. The first strategy is called the complete RRB ordering and the second one the K-step RRB ordering. They are respectively denoted by RRB(c) and RRB(K).

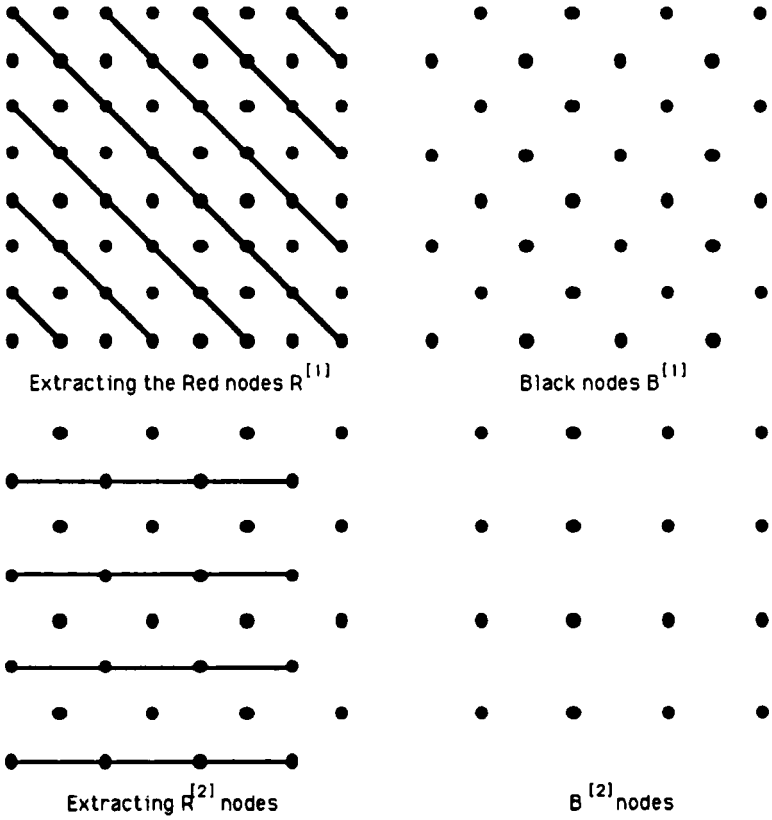

Fig. 1. Steps 1 and 2 of the numbering process.

We give in this paragraph the splitting criterion for the set of remaining nodes. If during step k the splitting occurred along diagonals, then it is done along parallels to the axis at step $k+1$. Conversely, if the splitting occurred along parallels to the axes, it is done along diagonals. More precisely, if at a given step the splitting occurs in the direction of angle θ, then the next one is done in the direction of angle $\theta + \pi/4$.

An RRB(2) ordering example is given in figures 1 and 2.

Remark 3.1

This method applies for a number of nodes of the form $n = 2^l$. As we label half of the remaining nodes at each step, the maximum number of steps is equal to l. If n is not a power of 2, the same method still applies, but the number of nodes labeled at each step is no longer equal to half the number of the remaining nodes.

4. Factorization of a matrix

4.1. *Complete factorization*

One can think of two ways to achieve the complete factorization of a symmetric

29	61	30	62	31	63	32	64
45	25	46	26	47	27	48	28
21	57	22	58	23	59	24	60
41	17	42	18	43	19	44	20
13	53	14	54	15	55	16	56
37	9	38	10	39	11	40	12
5	49	6	50	7	51	8	52
33	1	34	2	35	3	36	4

Fig. 2. RRB(2) ordering.

positive definite matrix $A = \tilde{L}\tilde{D}\tilde{L}^{\mathrm{T}}$, where $\tilde{L}$ is lower triangular with a unit diagonal and $\tilde{D}$ is diagonal. The first method is "algebraic", whereas the second one is "geometrical".

The first approach is Cholesky's factorization algorithm for matrix A. It relies on a column by column or row by row construction of the matrix. We consider the row by row construction. At step i of the algorithm, the entries of row i of $\tilde{L}$ are computed in increasing order before the diagonal entry of $\tilde{D}$, as a function of the corresponding entry of A and entries of $\tilde{L}$ and $\tilde{D}$ previously computed (see for example [10] or [4]).

The second approach is to remove successively the vertices of the graph associated to A and to reason on the edges (between vertices) of the graph. We use here the notations of [15] (see the appendix). In that case, the construction of the matrices $\tilde{L}$ and $\tilde{D}$ is column by column like. Let $A^0 = A$ and $\underline{A}^1 = A^0$. We perform the removal of a vertex in figure 3.

Vertex 1 is adjacent to its neighbors 2, 3, 4, and 5 (and to itself). Therefore, in column 1 of $\tilde{L}$, only $\tilde{L}_{11}, \tilde{L}_{21}, \tilde{L}_{31}, \tilde{L}_{41}$ and $\tilde{L}_{51}$ are non-zero entries. Moreover, the removal of vertex 1 produces new edges between vertices 2, 3, 4 and 5 (all vertices are adjacent to one another).

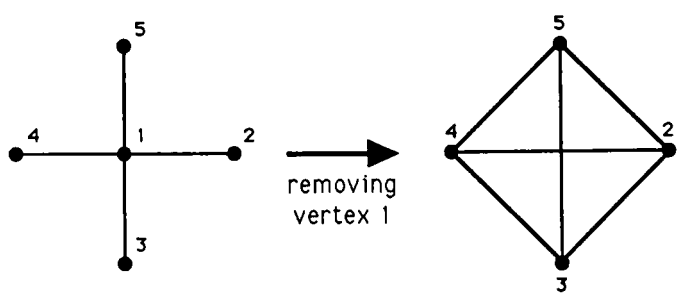

Fig. 3. Removing a vertex.

It induces a modification of the restriction of our starting point $\underline{A}^1$ to the nodes 2, 3, 4, 5.... We call A^1 the new matrix and $A^{(1)}$ its restriction:

$$A^1 = \begin{pmatrix} \tilde{D}_{11} & (0 & \dots & 0) \\ \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} & & A^{(1)} & \end{pmatrix}.$$

To remove vertex 2, our starting point is now $\underline{A}^2 = A^1$. The removal process is then carried out iteratively on the vertices: at step i, columns i of $\tilde{L}$ and the ith entry of $\tilde{D}$ are computed by removing vertex i and a new matrix, called A^i (and its submatrix $A^{(i)}$) is obtained; the starting point of step $i+1$ is $\underline{A}^{i+1} = A^i$.

4.2. *Incomplete factorization*

A way of designing preconditioners M for the matrix A is to perform an incomplete factorization of A according to the RRB ordering.

Denote by P the permutation matrix allowing to shift from the row ordering to the RRB ordering. In the following, we rename A the matrix PAP^{-1}.

We know that if M is a nonsingular matrix, then $A = M - R$ represents a *splitting* of the matrix A and that if M is symmetric positive definite (therefore equal to $M = LDL^T$), then M is an (LDL^T-)incomplete factorization of A. Moreover, the incomplete factorization is said to be *modified* if the row sums of R are equal to zero.

In the following, the basic idea is to perform an incomplete factorization with a prescribed structure of L and to set $M = LDL^T$. We propose two preconditioners, called M_1 and M_2. Brand and Heinemann investigated the latter [2]. Our factorizations are either unmodified or modified.

This kind of incomplete factorization was first investigated by Dupont et al. [8, 9], although the term modified does not appear in these papers.

We have previously introduced the complete factorization process. It is done by iteratively removing nodes to get $A = \tilde{L}\tilde{D}\tilde{L}^T$. The process can be changed quite simply to give an incomplete factorization of A ($= LDL^T - R$). At step i of the process, only prescribed non-zero entries of A^{i-1} are kept and a new matrix $\underline{A}^i$ is obtained:

$$\underline{A}^i = A^{i-1} + R^i, \quad \text{where } R^i \neq 0.$$

The new process computes an unmodified incomplete factorization of the matrix. To produce a modified incomplete factorization, we add the entries of A^{i-1} that were neglected in the previous case to the corresponding diagonal entry. This is done by modifying consequently the ith diagonal entry of R^i. We provide a thorough description of the method in the appendix.

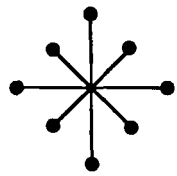

Fig. 4. Nine-point scheme in $B^{[l]}$.

5. The first preconditioner

In this section, our goal is to construct a matrix L_1 to get an incomplete factorization $M_1 = L_1 D_1 L_1^T$. This matrix M_1 is designed to be an efficient preconditioner of A for the Preconditioned Conjugate Gradient method. The factorization process is recursive, as for the construction of the RRB ordering. At a given step, k, the nodes of $R^{[k]}$ are removed in the factorization process of $\underline{A}_1^k$. Here, the removal of a single node is considered as an inner step. The corresponding columns of L_1 are built and so are A_1^k and its submatrix $A_1^{(k)}$, defined on $B^{[k]}$. If an RRB(c) ordering was performed, then the process is carried out until the end. In the case of an RRB(K) ordering, the process is carried out until the last level $B^{[K]}$ is reached, and then the complete factorization of $\underline{A}_1^{(K+1)}$ is performed.

The process is completely defined by the structure of $\underline{A}_1^{k+1}$. From the whole graph, only edges corresponding to a five-point scheme in $B^{[k]}$ are kept in $\underline{A}_1^{k+1} = A_1^k + R_1^{k+1}$. We now describe the construction of the submatrix $A_1^{(1)} + R_1^{(2)}$. When the nodes of $R^{[1]}$ are removed, it is easy to see that a node of $B^{[1]}$ is adjacent to its eight neighbors (and to itself), according to figure 4.

This nine-point scheme rules the structure of $A_1^{(1)}$ if the complete factorization process takes place. For $A_1^{(1)} + R_1^{(2)}$, only entries corresponding to the five-point scheme in $B^{[1]}$ are kept (see figure 5).

Therefore, we have entirely described the structures of both submatrices $A_1^{(1)}$ and $A_1^{(1)} + R_1^{(2)}$ which respectively correspond to the left- and right-hand figures in figure 6.

Remark 5.1

There are two differences between (sub)matrices $A_1^{(0)} + R_1^{(1)} = A$ and $A_1^{(1)} + R_1^{(2)}$. On the one hand, the meshlines are horizontal instead of being diagonal. On the other hand, the meshsize is now equal to $\sqrt{2}h$. These two reasons justify the

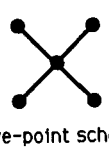

Fig. 5. Five-point scheme in $B^{[1]}$.

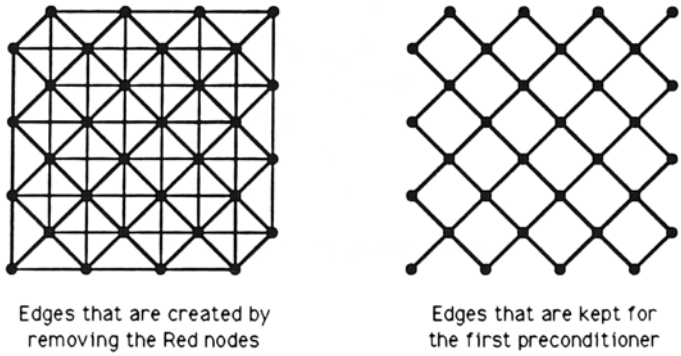

Fig. 6. Non-zero entries of $A_1^{(1)}$ and $A_1^{(1)} + R_1^{(2)}$.

alternating splitting directions, along diagonals or parallels to the axes, together with the multiplication by a factor $\sqrt{2}$ of the meshsize at each step.

The method previously described leads to a preconditioner M_{1u}, where the suffix u stands for unmodified.

Lemma 5.1
M_{1u} is a symmetric positive definite matrix.

Proof
See the appendix. □

Adding the entries that are neglected during the transition from the exact nine-point to the approximate five-point schemes to the corresponding diagonal entry leads to the modified preconditioner called M_{1m}.

Lemma 5.2
M_{1m} is a symmetric positive definite matrix.

Proof
See the appendix. □

6. The Brand and Heinemann preconditioner

Another preconditioner is considered in this section. This preconditioner has been investigated by Brand and Heinemann [2]. The factorization process is still recursive. The only difference with the construction of the preconditioner M_1 is the definition of the matrix R_2^{k+1} (or its submatrix $R_2^{(k+1)}$) at step k. Here, edges are kept if and only if they correspond to a five-point scheme in $B^{[k']}$, for $k' \geq k$.

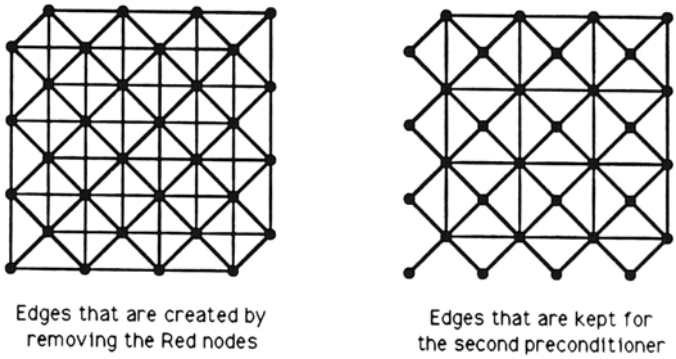

Fig. 7. Non-zero entries of $A_2^{(1)}$ and $A_2^{(1)} + R_2^{(2)}$.

In that case, the structure of submatrices $A_2^{(1)}$ (for a complete factorization) and $A_2^{(1)} + R_2^{(2)}$ are given in figure 7.

Remark 6.1
Two steps are necessary to build the preconditioner M_2. In the first place, the binary matrix is built: the default entry (false) is set to true if it corresponds to a five-point scheme edge in one of the $B^{[k]}$. In the second place, the factorization process is carried out and an entry of $\underline{A}_2^{k+1}$ is kept if the same entry of the binary matrix is true.

The method leads to the M_{2u} preconditioner.

Lemma 6.1
M_{2u} is a symmetric positive definitive matrix.

Proof
See the appendix. □

Adding the entries that are neglected during the factorization to the corresponding diagonal entry leads to the modified preconditioner called M_{2m}.

Lemma 6.2
M_{2m} is a symmetric positive definite matrix.

Proof
See the appendix. □

Note that the four preconditioners M_{1u}, M_{1m}, M_{2u} and M_{2m} have exactly the same structure. Actually, if we compare figures 6 and 7, the number of non-zero entries is greater for the second preconditioner. But we only described the first step of the factorization process. During the second step, these "extra" entries

are included in $\underline{A}_1^3 = A_1^2 + R_1^3$ because they correspond to the five-point scheme in $B^{[2]}$. The number of non-zero entries is therefore the same for all the preconditioners.

Remark 6.2
When it is necessary, we will denote by M_{ix}^K $(i \in \{1,2\}, x \in \{u,m\})$ the preconditioner derived from a RRB(K) ordering.

7. Numerical results

We now compare the Preconditioned Conjugate Gradient method with the four preconditioners to a reference iterative method, also based on the Preconditioned Conjugate Gradient method. The reference preconditioner is an ILU factorization of A according to the row ordering, called IC(1, 1) by Meijerink and van der Vorst [17]. The IC(1, 1) factorization is equal to $L^*D^*L^{*T}$, where L^* is lower triangular with the structure of the lower triangular part of A and a unit diagonal and D^* is diagonal. Figure 8 depicts the structure. This choice is motivated by Duff and Meurant [7]. They stressed that this preconditioner is often the most efficient to solve numerically the problem (1.1–2) on various examples. Moreover, we give the results obtained with a MILU factorization of A (here M stands for modified) with the tridiagonal structure depicted in figure 8.

We try the methods on two test problems. The test problems are defined by choosing the values of a and b, the diagonal entries of the operator $\mathcal{A}$. For problem #1, our model problem, the Poisson equation, $a = 1$ and $b = 1$ over the domain. For problem #2, $a = 100$ and $b = 1$ over the domain. The number of nodes n is equal to 256, 1024, 4096 or 16384. We choose the right-hand side f of the linear system as follows: define $\tilde{x}_{i,j} = u(x_i, y_i)$ where $u(x,y) = x(1-x)y(1-y)\,e^{xy}$, and compute $f = A\tilde{x}$.

All methods are based on the Preconditioned Conjugate Gradient method. In order to be able to use it, two parameters have to be fixed. The first one is the initial estimate x^0 of the solution which is uniformly set to 0. If r^k denotes the residual at

$$L^* = \begin{pmatrix} 1 & & & & \\ l_{2,1} & 1 & & & \\ & \ddots & \ddots & & \\ l_{m+1,1} & & l_{m+1,m} & 1 & \\ & \ddots & & \ddots & \ddots \end{pmatrix}$$

Fig. 8. Structure associated to IC(1, 1).

iteration k, then the stopping criterion is fulfilled when

$$\frac{||r^{k+1}||_\infty}{||r^0||_\infty} < \epsilon,$$

where $||y||_\infty = \max_i |y_i|$. The value of ϵ is set to 10^{-6} in the following.

This section is divided into two parts. In the first place, we focus our interest on a RRB(K) ordering with a fixed value of K. Then, in the second place, we consider a RRB(K) ordering where the value of K varies, in order to compensate for the number of non-zero entries of the preconditioners which increases too rapidly with respect to n.

We give two different kinds of results. On the one hand, the number of iterations of the Preconditioned Conjugate Gradient method to reach the prescribed decrease of the norm of the residual. On the other hand, the number of non-zero entries needed to store the preconditioners, which does not depend on the problem to be solved.

7.1. *The RRB(4) ordering*

We have chosen to set K to 4 because this value is a fair tradeoff between a reasonable number of iterations and an acceptable number of non-zero entries.

Remark 7.1

The smaller is the value of K, the smaller is the number of iterations and the bigger is the number of non-zero entries. In particular, for $K = 1$, the complete Cholesky decomposition of A is computed and the Preconditioned Conjugate Gradient method converges in one step.

We collect in tables 1 and 2 the number of iterations needed to reach the prescribed stopping criterion for each preconditioner and each problem.

Table 1
Problem #1: number of iterations.

n	ILU	M_{1u}	M_{2u}	MILU	M_{1m}	M_{2m}
256	14	14	12	13	16	8
1024	23	21	19	18	18	8
4096	41	37	32	28	19	8
16384	69	66	36	39	19	8

Table 2
Problem #2: number of iterations.

n	ILU	M_{1u}	M_{2u}	MILU	M_{1m}	M_{2m}
256	6	25	20	7	72	36
1024	9	49	39	11	104	44
4096	14	87	67	18	120	46
16384	23	110	117	28	130	46

We are able to draw some conclusions from these tables.

First of all, unmodified incomplete preconditioners behave rather similarly. The number of iterations is proportional to $1/h$. Concerning ILU, the result has previously been proved (see for example [11]).

The modified incomplete preconditioners are more efficient than their unmodified counterpart, with the exception of ILU for problem #2 although it should be noted that the row ordering favours the ILU preconditioner very much. Indeed, for this problem and this ordering, the matrix A is almost tridiagonal. Now the ILU factorization of a tridiagonal symmetric positive definite matrix is also the Cholesky factorization of the matrix. Thus, in that case, the incomplete factorization ILU is almost complete, which accounts for the good results.

Lastly, we focus on the modified incomplete preconditioners. For the model problem, Gustafsson [12] proved that the number of iterations is proportional to $1/\sqrt{h}$ for MILU. The numerical results are in accordance: for the model problem, the number of iterations behaves like $3.5/\sqrt{h}$. For both preconditioners M_{1m} and M_{2m}, the numerical results are even more interesting, as the number of iterations seems bounded, in each case, independently of the meshsize h (maybe with the exception of M_{1m} and problem #2). Moreover, the second preconditioner M_{2m} is always more efficient than the first preconditioner M_{1m}.

We now study the number of non-zero entries, respectively called $Z(n)$ and $ilu(n)$, required to store M and ILU (see table 3).

For ILU, the ratio goes to 3, as only two diagonals of L^* and the diagonal of D^* may have non-zero entries (not uniformly equal to 1). On the contrary, the ratio $Z(n)/n$ does not seem to be bounded for M. So we need to choose a new RRB ordering to be able to make objective comparisons.

7.2. *RRB(K) ordering, K varying*

To solve the unbounded ratio problem, we let K vary. The goal is twofold: to have a number of non-zero entries $Z(n)$ proportional to n and to keep the number of iterations (almost) independent of the meshsize.

We now define K as a function of n to get $Z(n)$ proportional to n. We need two lemmas.

Table 3
$Z(n)$ and $ilu(n)$: non-zero entries for M and ILU.

n	$Z(n)$	$Z(n)/n$	$ilu(n)$	$ilu(n)/n$
256	1182	4.62	751	2.93
1024	5178	5.06	3039	2.97
4096	23154	5.65	12223	2.98
16384	109794	6.70	49023	2.99

Lemma 7.1
During the removal of the nodes of $(R^{[k]})_{1 \leq k \leq K}$, the number of non-zero entries generated for the matrix M is less than $5n$.

Proof
In fact, during these K steps, two nodes i and j ($j \geq i$) are adjacent in the graph of M if node j belongs to the five-point scheme of node i. Now, a node i is related to at most five nodes (including himself) inside a five-point scheme graph. As the number of nodes of $\cup_{i=1}^{i=K} R^{[k]}$ is less than n, less than $5n$ non-zero entries are generated during these steps. □

Lemma 7.2
During the complete factorization of the matrix $\underline{A}^{(K)}$, the number of non-zero entries generated for the matrix M is in the order of $2^{-3K/2}n^{3/2}$.

Proof
The number of nodes of $B^{[K]}$ is equal to $2^{-K}n$. $\underline{A}^{(K)}$ is thus a $2^{-K}n \times 2^{-K}n$ matrix. Moreover, its average half-bandwidth is in the order of $2^{-K/2}\sqrt{n}$. Therefore, the number of non-zero entries generated during the complete factorization of $\underline{A}^{(K)}$ is in the order of $2^{-3K/2}n^{3/2}$. □

We are now able to state the following.

Theorem 7.1
If K is equal to $\lfloor \frac{1}{3}\log_2(n) + \frac{4}{3} \rfloor$, then $Z(n)$, the number of non-zero entries of M, is less than $6n$.

Proof
To find an upper bound of $Z(n)$, it is enough to get an upper bound for both steps of the removal process, the first one on $\cup_{i=1}^{i=K} R^{[k]}$ and the second one on $B^{[K]}$. We have immediately $Z(n) < 5.5n$, which proves the result. □

This condition on K is weaker than the one used by Brand in [1], where $K = \lfloor \frac{1}{2}\log_2(n) \rfloor$. This value of K gives a number of non-zero entries in the order of $n^{3/4}$ for the complete factorization of $\underline{A}^{(K)}$. On the other hand, Brand noted that for $K = \lfloor \frac{1}{2}\log_2(n) \rfloor$ the number of floating operations to build the preconditioner is proportional to n. In our case, this number is proportional to $n^{4/3}$. Both choices for K are therefore valid.

The values of K, $Z(n)$ and $Z(n)/n$ are given in table 4.

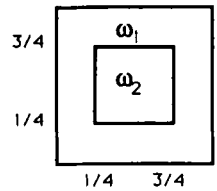

Fig. 9. Subdomains ω_1 and ω_2 for problem #3.

Table 4
$Z(n)$: non-zero entries for M.

n	K	$Z(n)$	$Z(n)/n$
256	4	1182	4.62
1024	4	5178	5.06
4096	5	21424	5.23
16384	6	84083	5.13

We now study the number of iterations as a function of n with the new values of K. To emphasize the results, we considered a third test problem, called problem #3: $a = b = 1$ in ω_1 and $a = b = 1000$ in ω_2 (see figure 9). In the next two tables, we give only the results obtained with the modified preconditioners.

In tables 5 and 6, the preconditioners M are defined by the RRB(K) orderings with values of K equal to those of table 4.

Remark 7.2
We provide a bound of the condition number of the preconditioned system for the model problem in the next paragraph for both preconditioners.

Table 5
Number of iterations: M_{1m} and M_{2m}.

n	problem #1		problem #2		problem #3	
256	16	8	72	36	16	9
1024	18	8	104	44	19	10
4096	27	9	163	46	28	10
16384	39	11	243	49	46	12

Table 6
FLOPs(n): M_{1m} and M_{2m}.

n	problem #1		problem #3	
256	1.84	0.92	1.83	1.03
1024	1.35	0.60	1.43	0.75
4096	1.17	0.39	1.20	0.43
16384	0.96	0.27	1.11	0.29

To end this section, we compare the number of floating operations required for solving problems #1 and #3, with the ILU and the modified preconditioners. As stated earlier, problem #2 is very anisotropic and we stressed the fact that the ILU factorization is the most efficient preconditioner. If we call $FLOPs(n)$ the ratio:

$$FLOPs(n) = \frac{FLOPs(M_{im})}{FLOPs(\mathrm{ILU})},$$

we have the results as given in table 6.

When the number of nodes n is multiplied by four, the ratio $FLOPs(n)$ is reduced by average factors of 1.5 for M_{2m} and 1.4 for M_{1m}. Therefore, the Preconditioned Conjugate Gradient methods with M_{1m} and M_{2m} as preconditioners are all the more efficient as the meshsize is small, with a clear advantage for the latter one.

8. Bounding the condition number for the Poisson equation

In the following, we bound the condition number of $(\{M_{im}^K\}^{-1}A)_{i=1,2}$, for the Poisson equation and $K = \lfloor \frac{1}{3}\log_2(n) + \frac{4}{3} \rfloor$. It is well known that the condition number $\kappa(M^{-1}A)$ is equal to the ratio $\lambda_{\max}(M^{-1}A)/\lambda_{\min}(M^{-1}A)$ when both matrices A and M are symmetric. Now, we prove the following.

Lemma 8.1
Let M_1, M_2 and M_3 be three symmetric positive definite matrices. Then $\kappa(M_1^{-1}M_3) \leq \kappa(M_1^{-1}M_2)\kappa(M_2^{-1}M_3)$.

Proof
The eigenvalues of $M_1^{-1}M_3$ and $M_1^{-1/2}M_3M_1^{-1/2}$ are equal. Moreover, we know (see for example [5]) that

$$\lambda_{\min}(M_1^{-1/2}M_3M_1^{-1/2}) = \inf_{x\neq 0} \frac{(M_1^{-1/2}M_3M_1^{-1/2}x, x)}{(x, x)}$$

and

$$\lambda_{\max}(M_1^{-1/2}M_3M_1^{-1/2}) = \sup_{x\neq 0} \frac{(M_1^{-1/2}M_3M_1^{-1/2}x, x)}{(x, x)}.$$

If we let $y = M_1^{-1/2}x$, then

$$\lambda_{\min}(M_1^{-1/2}M_3M_1^{-1/2}) = \inf_{y\neq 0} \frac{(M_3y, y)}{(M_1y, y)}$$

and

$$\lambda_{\max}(M_1^{-1/2}M_3M_1^{-1/2}) = \sup_{y\neq 0} \frac{(M_3y, y)}{(M_1y, y)}.$$

Thus

$$\kappa(M^{-1}M_3) = \kappa(M_1^{-1/2}M_3M_1^{-1/2}) = \frac{\lambda_{\max}(M_1^{-1/2}M_3M_1^{-1/2})}{\lambda_{\min}(M_1^{-1/2}M_3M_1^{-1/2})}$$

$$= \frac{\sup_{y\neq 0}\dfrac{(M_3y, y)}{(M_1y, y)}}{\inf_{y\neq 0}\dfrac{(M_3y, y)}{(M_1y, y)}}.$$

This leads immediately to the result, as the supremum (resp. the infimum) of a product is lower (resp. greater) than or equal to the product of the suprema (resp. infima). □

8.1. *Periodic boundary conditions*

We solve numerically

$$-\Delta u = g \text{ in }]0,1[^2,$$

$$u(0,y) = u(1,y), \quad u(x,0) = u(x,1).$$

The solution to this problem is unique up to a constant. In the following, we therefore neglect the zero eigenvalue. Here, we provide a bound of the condition number for the modified preconditioner M_{1m}^K. As we are considering a problem with periodic boundary conditions, we slightly modify its discretization. The mesh-size h is still equal to $1/(m+1)$, but the number of unknowns is now $(m+1)^2$, as we have to handle the periodic conditions. Therefore, A is a $(m+1)^2 \times (m+1)^2$ matrix. On each row, there is a 4 on the diagonal and four non-zero off-diagonal entries equal to -1. If we use the classical Red–Black (or RRB(1)) ordering, then we rewrite

$$A = \begin{pmatrix} 4I & -L^{\mathrm{T}} \\ -L & 4I \end{pmatrix},$$

where L is the matrix which corresponds to the five-point operator with stencil

$$\begin{array}{ccc} & 0 & \\ 1 & 0 & 1. \\ & 1 & \end{array}$$

If we remove the nodes of $R^{[1]}$, then we get

$$A = \begin{pmatrix} I & \\ -\frac{1}{4}L & I \end{pmatrix}\begin{pmatrix} 4I & \\ & S \end{pmatrix}\begin{pmatrix} I & -\frac{1}{4}L^{\mathrm{T}} \\ & I \end{pmatrix}$$

and

$$M_{1m}^1 = \begin{pmatrix} I & \\ -\frac{1}{4}L & I \end{pmatrix}\begin{pmatrix} 4I & \\ & S_1^1 \end{pmatrix}\begin{pmatrix} I & -\frac{1}{4}L^{\mathrm{T}} \\ & I \end{pmatrix},$$

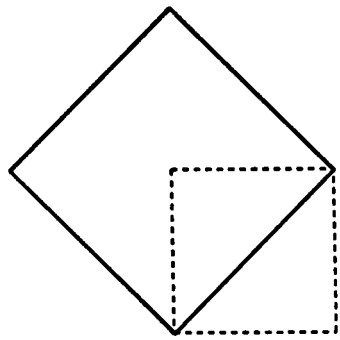

Fig. 10. The new domain.

where S $(= 4I - \frac{1}{4}LL^{\mathrm{T}})$ and S_1^1 respectively correspond to the following nine- and five-point operators in $B^{[1]}$ (even on the boundary):

$$\begin{array}{ccccc} & & -\frac{1}{4} & & \\ & -\frac{1}{2} & & -\frac{1}{2} & \\ -\frac{1}{4} & & 3 & & -\frac{1}{4} \\ & -\frac{1}{2} & & -\frac{1}{2} & \\ & & -\frac{1}{4} & & \end{array} \qquad \begin{array}{ccc} & & \\ -\frac{1}{2} & & -\frac{1}{2} \\ & 2 & \\ -\frac{1}{2} & & -\frac{1}{2} \\ & & \end{array}.$$

Therefore, the eigenvalues of $\{M_{1m}^1\}^{-1}A$ are either 1 or the eigenvalues of $\{S_1^1\}^{-1}S$. By using the periodic boundary condition, it is easy to see that S and S_1^1 can be extended (as nine- and five-point operators) over the domain in figure 10.

Now, we are able to compute the eigenvalues, following [3], and we find, for $(u, v) \in \mathscr{U} = \{(\cos 2\pi kh, \cos 2\pi lh), 0 \leq k, l \leq m, k \neq 0 \text{ or } l \neq 0\}$:

$$\lambda(S) = 3 - u - v - uv,$$

$$\lambda(S_1^1) = 2 - u = v.$$

Thus, the eigenvalues of $\{M_{1m}^1\}^{-1}A$ are either 1 or $(3 - u - v - uv)/(2 - u - v)$, for $(u, v) \in \mathscr{U}$. In particular, $\mathscr{U}$ is imbedded in $[-1, 1]^2 \setminus \{(1, 1)\}$. Now, let

$$f(u, v) = \begin{cases} \dfrac{3 - u - v - uv}{2 - u - v} & \text{if } (u, v) \in [-1, 1]^2\{(1, 1)\}, \\ 2 & \text{if } (u, v) = (1, 1). \end{cases}$$

Lemma 8.2

f is continuous over $[-1, 1]^2$ and continuously differentiable over $]-1, 1[^2$. Moreover, for all (u, v) in $]-1, 1[^2$, $\mathrm{d}f_{(u,v)} \neq 0$.

Proof

It uses standard calculus and is omitted here. □

Therefore, we have the following situation: f is continuous over the compact set $[-1,1]^2$, $\mathrm{d}f$ is continuous over $]-1,1[^2$ and $\mathrm{d}f_{(u,v)} \neq 0$ for all (u,v). Thus it is well known that f reaches its maximum (and its minimum) on the boundary. We find that

$$\max_{(u,v)\in[-1,1]^2} f(u,v) = 2.$$

We find the minimum value of f to be equal to 1. Finally, we have

$$\kappa(\{M_{1m}^1\}^{-1}A) \leq 2.$$

Moreover, as S_1^1 corresponds (apart from a scaling factor of $1/2$) to the original problem on the new domain (see figure 10), we also have

$$\kappa(\{M_{1m}^{k+1}\}^{-1}M_{1m}^k) \leq 2 \quad \text{for } 1 \leq k \leq K.$$

We are now able to prove the following.

Theorem 8.1
For the Poisson equation with periodic boundary conditions discretized on an $(m+1)\times(m+1)$ grid and $K = \lfloor \frac{1}{3}\log_2(n) + \frac{4}{3} \rfloor$, the condition number of the preconditioned system $\{M_{1m}^K\}^{-1}A$ is such that

$$\kappa(\{M_{1m}^K\}^{-1}A) \leq 2^{5/3}(m+1)^{2/3}.$$

Proof
It suffices to use the previous results together with lemma 8.1 for the prescribed value of K. □

8.2. *Dirichlet boundary conditions*

We solve

$$-\Delta u = g \text{ in }]0,1[^2,$$
$$u = 0 \text{ on the boundary.}$$

We provide a bound of the condition number for both modified preconditioners. With the RRB(1) ordering

$$A = \begin{pmatrix} 4I & -L^{\mathrm{T}} \\ -L & 4I \end{pmatrix},$$

where L is the matrix which corresponds to the five-point (except near the boundary) operator with stencil

$$\begin{matrix} & 1 & \\ 1 & 0 & 1. \\ & 1 & \end{matrix}$$

By removing the red nodes, we get

$$A = \begin{pmatrix} I & \\ -\frac{1}{4}L & I \end{pmatrix} \begin{pmatrix} 4I & \\ & S \end{pmatrix} \begin{pmatrix} I & -\frac{1}{4}L^{\mathrm{T}} \\ & I \end{pmatrix}$$

and

$$M_{im}^{1} = \begin{pmatrix} I & \\ -\frac{1}{4}L & I \end{pmatrix} \begin{pmatrix} 4I & \\ & S_i^1 \end{pmatrix} \begin{pmatrix} I & -\frac{1}{4}L^{\mathrm{T}} \\ & I \end{pmatrix}, \quad i \in \{1,2\}.$$

Here, S and S_1^1 correspond to their periodic counterpart of the previous section (except near the boundary). To describe S_2^1, we suppose that an RRB(2) ordering is performed, that is, $B^{[1]} = R^{[2]} \cup B^{[2]}$. Then, according to its definition, S_2^1 corresponds (except near the boundary) to a nine-point operator in $B^{[2]}$ and a five-point operator in $R^{[2]}$, like the ones already defined for S and S_1^1:

$$\begin{array}{ccccccc} & & -\frac{1}{4} & & & & \\ & -\frac{1}{2} & & -\frac{1}{2} & & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{4} & & 3 & & -\frac{1}{4} & & 2 \\ & -\frac{1}{2} & & -\frac{1}{2} & & -\frac{1}{2} & -\frac{1}{2} \\ & & -\frac{1}{4} & & & & \end{array}.$$

Remark 8.1

We have mentioned the fact that the expressions of the operators are valid except near the boundary. As a matter of fact, there are two differences. First, some off-diagonal terms may be equal to zero if the related node is on the boundary. Secondly, as Brant noticed in [1], the diagonal entry may actually be greater than 2 (when it is supposed to be 2). Indeed, near the boundary, some off-diagonal terms that are added to the diagonal entry when the modified factorization is performed may be equal to 0 instead of $-1/4$. This is not really a problem, because Brand also pointed out that considering that all diagonal entries are equal to 2 indeed gives an upper bound of the actual condition number.

The matrices S, S_1^1 and S_2^1 are defined on $B^{[1]} = R^{[2]} \cup B^{[2]}$. Blockwise, they can therefore be written

$$S = \begin{pmatrix} 3I - \frac{1}{4}N & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & 3I - \frac{1}{4}N \end{pmatrix},$$

$$S_1^1 = \begin{pmatrix} 2I & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & 2I \end{pmatrix},$$

$$S_2^1 = \begin{pmatrix} 2I & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & 3I - \frac{1}{4}N \end{pmatrix},$$

where N and M correspond to the following operators (except near the boundary)

$$\begin{array}{ccccccc} & & 1 & & & & \\ & & 0 & & & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & \quad 0 & 0 & 0. \\ & & 0 & & & 1 & 0 & 1 \\ & & 1 & & & & \end{array}$$

As in the case of periodic boundary conditions, the eigenvalues of $\{M_{im}^1\}^{-1}A$ are either 1 or the eigenvalues of $\{S_i^1\}^{-1}S$. Unfortunately, the S_i^1 matrices are no longer simply related to the original problem, unlike the case of periodic boundary conditions. So we can not deduce the condition number of $\{M_{im}^K\}^{-1}A$ from $\{M_{im}^1\}^{-1}A$ by iterating the process. We have to go one step further, to M_{im}^2. Although the nodes of $R^{[2]}$ are removed, the matrices can still be factorized in the following way

$$M_{im}^2 = \begin{pmatrix} I & \\ -\frac{1}{4}L & I \end{pmatrix}\begin{pmatrix} 4I & \\ & S_i^2 \end{pmatrix}\begin{pmatrix} I & -\frac{1}{4}L^{\mathrm{T}} \\ & I \end{pmatrix}, \quad i \in \{1,2\}.$$

Next, we have to find the expression of the S_i^2. For the first method, we pointed out that, using the partitioning $B^{[1]} = R^{[2]} \cup B^{[2]}$, S_1^1 is of the form

$$S_1^1 = \begin{pmatrix} 2I & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & 2I \end{pmatrix}.$$

When the removal of the nodes of $R^{[2]}$ is performed, this leads to the following complete and incomplete factorizations

$$S_1^1 = \begin{pmatrix} I & \\ -\frac{1}{4}M & I \end{pmatrix}\begin{pmatrix} 2I & \\ & 2I - \frac{1}{8}MM^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} I & -\frac{1}{4}M^{\mathrm{T}} \\ & I \end{pmatrix},$$

$$S_1^2 = \begin{pmatrix} I & \\ -\frac{1}{4}M & I \end{pmatrix}\begin{pmatrix} 2I & \\ & I - \frac{1}{4}N \end{pmatrix}\begin{pmatrix} I & -\frac{1}{4}M^{\mathrm{T}} \\ & I \end{pmatrix}$$

$$= \begin{pmatrix} 2I & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & I - \frac{1}{4}N + \frac{1}{8}MM^{\mathrm{T}} \end{pmatrix}.$$

Interestingly, S_1^2 is now a friendly five-point operator on $B^{[2]}$ (except near the boundary)

$$\begin{array}{ccc} & -\frac{1}{4} & \\ -\frac{1}{4} & 1 & -\frac{1}{4}, \\ & -\frac{1}{4} & \end{array}$$

equal to the original problem apart from a scaling factor of 1/4. Therefore, if we are able to bound the condition number of $\{S_1^2\}^{-1}S$, we can provide a bound for the condition number of $\{M_{1m}^K\}^{-1}A$.

For the second method, we find

$$S_2^1 = \begin{pmatrix} I & \\ -\frac{1}{4}M & I \end{pmatrix} \begin{pmatrix} 2I & \\ & 3I - \frac{1}{4}MM^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} I & -\frac{1}{4}M^{\mathrm{T}} \\ & I \end{pmatrix},$$

$$S_2^2 = \begin{pmatrix} I & \\ -\frac{1}{4}M & I \end{pmatrix} \begin{pmatrix} 2I & \\ & 2I - \frac{1}{2}N \end{pmatrix} \begin{pmatrix} I & -\frac{1}{4}M^{\mathrm{T}} \\ & I \end{pmatrix}$$

$$= \begin{pmatrix} 2I & -\frac{1}{2}M^{\mathrm{T}} \\ -\frac{1}{2}M & 2I - \frac{1}{2}N + \frac{1}{8}MM^{\mathrm{T}} \end{pmatrix}.$$

And S_2^2 corresponds to a five-point operator on $B^{[2]}$ (except near the boundary)

$$\begin{matrix} & -\frac{1}{2} & \\ -\frac{1}{2} & 2 & -\frac{1}{2}, \\ & -\frac{1}{2} & \end{matrix}$$

thus the same conclusion holds for $\{S_2^2\}^{-1}S$ and $\{M_{2m}^K\}^{-1}A$. A way to compute the eigenvalues of $\{S_i^2\}^{-1}S$ is to study those of $\{S_i^2\}^{-1}(S - S_i^2)$, because

$$\{S_i^2\}^{-1}S = I + \{S_i^2\}^{-1}(S - S_i^2).$$

$(S - S_i^2)_{i=1,2}$ are respectively equal to

$$(S - S_1^2) = \begin{pmatrix} I - \frac{1}{4}N & \\ & 2I - \frac{1}{8}MM^{\mathrm{T}} \end{pmatrix},$$

$$(S - S_2^2) = \begin{pmatrix} I - \frac{1}{4}N & \\ & I + \frac{1}{4}N - \frac{1}{8}MM^{\mathrm{T}} \end{pmatrix}.$$

Now, if λ is an eigenvalue of $\{S_i^2\}^{-1}(S - S_i^2)$, then there exists a non-zero vector x, defined on $B^{[1]}$, such that

$$(S - S_i^2)x = \lambda S_i^2 x.$$

We have $B^{[1]} = R^{[2]} \cup B^{[2]}$, so we can write $x = (x_r \quad x_b)^{\mathrm{T}}$. The eigenvalue problem for the first preconditioner is now

$$(I - \tfrac{1}{4}N)x_r = \lambda\{2x_r - \tfrac{1}{2}M^{\mathrm{T}}x_b\},$$

$$(2I - \tfrac{1}{8}MM^{\mathrm{T}})x_b = \lambda\{-\tfrac{1}{2}Mx_r + (I - \tfrac{1}{4}N + \tfrac{1}{8}MM^{\mathrm{T}})x_b\}.$$

Or, equivalently,

$$(2I - \tfrac{1}{8}MM^{\mathrm{T}})x_b = \lambda\left\{-\frac{\lambda}{4}M([2\lambda - 1]I + \tfrac{1}{4}N)^{-1}M^{\mathrm{T}} + (I - \tfrac{1}{4}N + \tfrac{1}{8}MM^{\mathrm{T}})\right\}x_b.$$

With a suitable x_b, i.e. an eigenvector of both matrices, we are able to find the explicit value of λ. This has to be done for the $n/4$ eigenvalues λ. Fortunately,

we can exhibit a vector x_b for each eigenvalue, for x_b is defined on a regular grid of meshsize $H = 2h$ and both matrices are very similar to five-point operators on this grid. Therefore, we take a classical set of $(n/4)$ vectors $(x_b^{kl})_{1 \le k,l \le \frac{1}{2}m}$, defined by

$$(x_b^{kl})_{ij} = \sin(ik\pi H)\sin(jl\pi H), \quad \text{for } 1 \le i, j \le \tfrac{1}{2}m.$$

Lemma 8.3
The matrix–vector products of N, MM^{T} or $M([2\lambda - 1]I + \frac{1}{4}N)^{-1}M^{\mathrm{T}}$ and x_b^{kl} are equal to

$$Nx_b^{kl} = 4(u_k^2 + v_l^2 - 1)x_b^{kl},$$

$$MM^{\mathrm{T}}x_b^{kl} = 16u_k^2 v_l^2 x_b^{kl},$$

$$M([2\lambda - 1]I + \tfrac{1}{4}N)^{-1}M^{\mathrm{T}}x_b^{kl} = \frac{16u_k^2 v_l^2}{2\lambda - 2 + u_k^2 + v_l^2}\, x_b^{kl},$$

where $u_k = \cos(\frac{1}{2}k\pi H)$ and $v_l = \cos(\frac{1}{2}l\pi H)$.

Proof
The result concerning Nx_b^{kl} is straightforward, because N is a five-point operator on the grid. For the remaining products, the proof is a little less obvious. As a matter of fact, the vector $M^{\mathrm{T}}x_b^{kl}$ is defined on $R^{[2]}$. However, it can be easily shown that it is equal to $4u_k v_l x_r^{kl}$, where the vector x_r^{kl} is such that its components are equal to those of x_b^{kl}. For the sake of brevity, the (technical) proof of this point is omitted. By the same trick, $Mx_r^{kl} = 4u_k v_l x_b^{kl}$. □

We have replaced the eigenvalue problem by

$$\{2 - 2u_k^2 v_l^2\}x_b^{kl} = \lambda\left\{\frac{-4\lambda u_k^2 v_l^2}{2\lambda - 2 + u_k^2 + v_l^2} + 2 + 2u_k^2 v_l^2 - u_k^2 - v_l^2\right\}x_b^{kl},$$

$$\text{for } 1 \le k,\ l \le \tfrac{1}{2}m.$$

For each value of k and l, we have thus a second degree equation in λ to solve:

$$p_1(\lambda) \equiv a_1(u_k^2, v_l^2)\lambda^2 + b_l(u_k^2, v_l^2)\lambda + c_1(u_k^2, v_l^2) = 0,$$

where

$$a_1(x, y) = 2x + 2y - 4,$$

$$b_1(x, y) = -2x^2 y - 2xy^2 + x^2 + y^2 + 2xy - 4x - 4y + 8,$$

$$c_1(x, y) = -2x^2 y - 2xy^2 + 4xy + 2x + 2y - 4.$$

Here, both $x = u_k^2$ and $y = v_l^2$ belong to]0, 1[. For the sake of clarity, the dependence of p_1 in (x, y) is omitted. Let us prove now that the roots of p_1 are greater than 0 and lower than 3. First, note that the roots are real, because they are also

eigenvalues of a symmetric matrix. Then, we have to check that

(i) $a_1 < 0$.
(ii) $p_1(0) < 0$ and $p_1'(0) > 0$.
(iii) $p_1(3) < 0$ and $p_1'(3) < 0$.

For (i) and (ii), we simply factorize the three polynomials in x and y to find that the inequalities hold for any (x, y) in $]0, 1[^2$:

$$a_1 = 2(x + y - 2) < 0,$$
$$p_1(0) = (x + y - 2)(2 - 2xy) < 0,$$
$$p_1'(0) = 8 + (x + y)(x + y - 2xy - 4) > 0.$$

To prove (iii), we use some standard calculus to study $p_1(3)$ and $p_1'(3)$ as continuously differentiable functions of x and y in the compact set $[0, 1]^2$, which yields the results. Using (i), (ii) and (iii) together with the fact that the roots of p_1 are real, we find

$$0 < \lambda < 3.$$

Therefore, the eigenvalues of $\{S_1^2\}^{-1}S$ range from 1 to 4. Finally,

$$\kappa(\{M_{1m}^2\}^{-1}A) \leq 4 \text{ and so on } \kappa(\{M_{1m}^{k+2}\}^{-1}M_{1m}^k) \leq 4.$$

We are now able to prove the following.

Theorem 8.2
For the Poisson equation with Dirichlet boundary conditions discretized on an $(m + 1) \times (m + 1)$ grid and $K = \lfloor \frac{1}{3}\log_2(n) + \frac{4}{3} \rfloor$, the condition number of the preconditioned system $\{M_{1m}^K\}^{-1}A$ is such that

$$\kappa(\{M_{1m}^K\}^{-1}A) \leq 3m^{2/3}.$$

Proof
See theorem 8.1. □

In the same way, the eigenvalue problem for the second preconditioner is

$$(I + \tfrac{1}{4}N - \tfrac{1}{8}MM^{\mathrm{T}})x_b = \lambda\Big\{ -\frac{\lambda}{4}M([2\lambda - 1]I + \tfrac{1}{4}N)^{-1}M^{\mathrm{T}}. + (2I - \tfrac{1}{2}N + \tfrac{1}{8}MM^{\mathrm{T}})\Big\}x_b.$$

The eigenvalues are the roots of

$$p_2(\lambda) \equiv a_2(u_k^2, v_l^2)\lambda^2 + b_2(u_k^2, v_l^2)\lambda + c_2(u_k^2, v_l^2) = 0,$$

where

$$a_2(x, y) = 4x + 4y - 8,$$
$$b_2(x, y) = -2x^2y - 2xy^2 + 2x^2 + 2y^2 + 4xy - 6x - 6y + 8,$$
$$c_2(x, y) = -2x^2y - 2xy^2 + x^2 + y^2 + 6xy - 2x - 2y.$$

By using the same technique, we find that

(i) $a_2 < 0$.
(ii) $p_2(0) < 0$ and $p_2'(0) > 0$.
(iii) $p_2(1) < 0$ and $p_2'(1) < 0$.

In this case

$$0 < \lambda < 1.$$

Therefore, the eigenvalues of $\{S_2^2\}^{-1}S$ range from 1 to 2. Finally,

$$\kappa(\{M_{2m}^2\}^{-1}A) \leq 2 \text{ and so on } \kappa(\{M_{2m}^{k+2}\}^{-1}M_{2m}^k) \leq 2.$$

We are now able to prove the following.

Theorem 8.3
For the Poisson equation with Dirichlet boundary conditions discretized on an $(m+1) \times (m+1)$ grid and $K = \lfloor \frac{1}{3}\log_2(n) + \frac{4}{3} \rfloor$, the condition number of the preconditioned system $\{M_{2m}^K\}^{-1}A$ is such that

$$\kappa(\{M_{2m}^K\}^{-1}A) \leq 2m^{1/3}.$$

Proof
See theorem 8.1. □

9. Conclusion

Based on the Repeated Red–Black ordering, we have studied two distinct preconditioners for solving symmetric positive definite discretizations of elliptic problems. Although the two preconditioners have the same structure, the second one performs better. On the other hand, they both compare quite favorably with the MILU factorization (for the row ordering). This is in particular true for the condition numbers of the preconditioned linear systems. For the Poisson equation with Dirichlet boundary conditions and a meshsize h, we proved that they are respectively bounded by $O(h^{-2/3})$ and $O(h^{-1/3})$.

Acknowledgement

The author thanks both Prof. Tony Chan and Dr. Gérard Meurant for their interest in this paper.

Appendix

We study here the modified and unmodified incomplete factorization processes. Among other results, we prove that the preconditioners M_{ix} ($i \in \{1,2\}$, $x \in \{u,m\}$) are symmetric positive definite. Let us begin by a description of the processes. They

can be defined by:

$$\left.\begin{aligned} A^0 &= A, \\ \underline{A}^k &= A^{k-1} + R^k \\ A^k &= (L^k)^{-1}\underline{A}^k(L^k)^{-\mathrm{T}} \end{aligned}\right\} \quad \text{for } k = 1, \ldots, n-1.$$

In the complete factorization case, the matrices R^k are set to zero. To be more precise, at step k, the structure of the matrices is the following (see [15]):

$$A^{k-1} = \begin{pmatrix} D_{k-1} & 0 \\ 0 & A^{(k-1)} \end{pmatrix},$$

where D_{k-1} is a $(k-1)\times(k-1)$ diagonal matrix and $A^{(k-1)}$ is $(n-k+1)\times(n-k+1)$. When processing an incomplete factorization, let:

$$R^k = \begin{pmatrix} 0 & 0 \\ 0 & R^{(k)} \end{pmatrix},$$

where $R^{(k)}$ is $(n-k+1)\times(n-k+1)$. The structure of $R^{(k)}$ depends on the factorization process (modified or not):

$$R_u^{(k)} = \begin{pmatrix} 0 & r_k^{\mathrm{T}} \\ r_k & \begin{pmatrix} 0 & \ddots \\ & & 0 \end{pmatrix} \end{pmatrix}$$

and

$$r_m^{(k)} = \begin{pmatrix} -\Sigma_j (r_k)_j & r_k^{\mathrm{T}} \\ r_k & \begin{pmatrix} -(r_k)_{k+1} & & \\ & \ddots & \\ & & (r_k)_n \end{pmatrix} \end{pmatrix}.$$

In the modified case, the row sums of the matrix are equal to zero. We have $\underline{A}^k = A^{k-1} + R^k$ then, if

$$A^{(k-1)} + R^{(k)} = \begin{pmatrix} d_k & b_k^{\mathrm{T}} \\ b_k & B^k \end{pmatrix} \quad \text{and} \quad L^k = \begin{pmatrix} I_{k-1} & & \\ & 1 & \\ 0 & \frac{1}{d_k} b_k & I_{n-k} \end{pmatrix},$$

we also have

$$A^k = \begin{pmatrix} D_k & 0 \\ 0 & A^{(k)} \end{pmatrix}, \quad D_k = \begin{pmatrix} D_{k-1} & 0 \\ 0 & d_k \end{pmatrix} \quad \text{and} \quad A^{(k)} = B^k - \frac{1}{d_k} b_k b_k^{\mathrm{T}}.$$

Theorem A.1

When the factorization process ends, $A = LDL^{\mathrm{T}} - R$, where

$$L = L^1 \cdots L^{n-1}, \quad D = A^{n-1} = \begin{pmatrix} D_{n-1} & 0 \\ 0 & A^{(n-1)} \end{pmatrix} \quad \text{and} \quad R = \sum_{j=1}^{j=n-1} R^j.$$

Proof
To prove the result, it suffices to note that, according to the respective structure of L^k and R^{k+1}, $(L^k)R^{k+1}(L^k)^T = R^{k+1}$. By induction, it follows that

$$\begin{aligned} A = A^0 &= \underline{A}^1 - R^1 = (L^1)A^1(L^1)^T - R^1 \\ &= (L^1L^2)A^2(L^1L^2)^T - R^1 - R^2 \\ &\vdots \\ &= (L^1 \cdots L^k)A^k(L^1 \cdots L^k)^T - \sum_{j=1}^{j=k} R^j \\ &\vdots \\ &= (L^1 \cdots L^{n-1})A^{n-1}(L^1 \cdots L^{n-1})^T - \sum_{j=1}^{j=n-1} R^j. \end{aligned}$$

□

Remark A.1
The structure of L is such that its kth column is equal to the kth column of L^k, i.e. 1 on the diagonal and b_k/d_k below.

The definition and theorems hereafter are helpful to prove the positive definiteness of the preconditioners previously introduced. We follow Meijerink and Van der Vorst [16] for the unmodified preconditioners and Brand [1] for the modified ones, although directly considering LDL^T factorizations instead of LU factorizations requires new theorems.

Definition A.1
A is a nonsingular M-matrix if $a_{ij} \leq 0$, $\forall i \neq j$, A is nonsingular and $A^{-1} \geq 0$. A nonsingular symmetric M-matrix is called a *Stieltjes* matrix.

Theorem A.2
Let A be a symmetric matrix such that $a_{ij} \leq 0$, $\forall i \neq j$. Then A is a Stieltjes matrix if and only if A is symmetric positive definite.

Proof
See [11]. □

Theorem A.3
If A is a Stieltjes matrix, then A^1 is also a Stieltjes matrix for the complete factorization process.

Proof
Let

$$A = A^{(0)} = \begin{pmatrix} d_1 & b_1^T \\ b_1 & B^1 \end{pmatrix}.$$

Then

$$A^1 = \begin{pmatrix} d_1 & 0 \\ 0 & A^{(1)} \end{pmatrix},$$

where

$$A^{(1)} = B^1 - \frac{1}{d_1} b_1 b_1^{\mathrm{T}}.$$

As $b_1 \leq 0$ and $d_1 > 0$, then $-(1/d_1)b_1 b_1^{\mathrm{T}} \leq 0$. Therefore all off-diagonal entries of $A^{(1)}$ are non-positive. Thus it remains only to prove that for a given non-zero vector x_1, $(A^{(1)}x_1, x_1)_{n-1} > 0$:

$$\begin{aligned}(A^{(1)}x_1, x_1)_{n-1} &= (B^1 x_1, x_1)_{n-1} - \frac{1}{d_1}(b_1^{\mathrm{T}}x_1, b_1^{\mathrm{T}}x_1)_1 \\ &= (B^1 x_1, x_1)_{n-1} + 2\left(b_1^{\mathrm{T}}x_1, -\frac{1}{d_1}b_1^{\mathrm{T}}x_1\right)_1 + d_1\left(-\frac{1}{d_1}b_1^{\mathrm{T}}x_1, -\frac{1}{d_1}b_1^{\mathrm{T}}x_1\right)_1 \\ &= (Ax, x)_n, \quad \text{where } x = \begin{pmatrix} -\frac{1}{d_1}b_1^{\mathrm{T}}x_1 \\ x_1 \end{pmatrix}. \qquad \square\end{aligned}$$

Remark A.2
By induction, it shows that, for the complete factorization process, if A^{k-1} is a Stieltjes matrix, then $\underline{A}^k = A^{k-1}$ is so, too. As this is particularly true for $k = 1$ (initial step), then $\tilde{L}\tilde{D}\tilde{L}^{\mathrm{T}}$ is symmetric positive definite. We already knew that, because the preconditioner is equal to the original matrix! We only emphasize this point because we use the same induction reasoning to prove the positive definiteness of the preconditioners.

Theorem A.4
If A is a Stieltjes matrix and B is such that:

$$a_{ij} \leq b_{ij} \leq 0 \quad \forall i \neq j \quad \text{and} \quad 0 < a_{ii} \leq b_{ii},$$

then B is also a Stieltjes matrix.

Proof
See [16]. $\square$

Lemmas 5.1 and 6.1
The unmodified preconditioners are symmetric positive definite.

Proof
We proceed by induction. Note particularly that the initial step is included hereafter for $k = 1$. At step k, we have

$$\underline{A}^k = A^{k-1} + R^k,$$

where R^k is defined by:

$$r_{ii}^k = 0, 1 \leq i \leq n$$

$$r_{ij}^k = \begin{cases} -a_{ij}^{k-1} & \text{if the entry is neglected,} \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i \neq j.$$

The induction assumption is that A^{k-1} is a Stieltjes matrix. Therefore, $a_{ij}^{k-1} \leq 0$, $\forall i \neq j$. So

$$a_{ij}^{k-1} \leq \underline{a}_{ij}^k \leq 0 \quad \forall i \neq j \quad \text{and} \quad 0 < a_{ii}^{k-1} - \underline{a}_{ii}^k.$$

By theorem A.3, $\underline{A}^k$ is a Stieltjes matrix. By induction, it follows that during the unmodified factorization process the matrices $\underline{A}^k$ are Stieltjes matrices, thus leading to symmetric positive definite preconditioners. □

Theorem A.5
If A is a weakly diagonally dominant Stieltjes matrix, then A^1 is also a weakly diagonally dominant Stieltjes matrix for the complete factorization process.

Proof
We know that A^1 is a Stieltjes matrix. Therefore $a_{ii}^1 > 0$, $\forall i$, and $a_{ij}^1 \leq 0$, $\forall i \neq j$. We still have to prove that it is weakly diagonally dominant, that is,

$$\sum_j a_{ij}^1 \geq 0, \quad \forall i.$$

This is immediately true for $i = 1$. For $i, j \geq 2$, we have $a_{ij}^1 - (1/a_{11})a_{1i}a_{1j}$, so (for $i \geq 2$)

$$\begin{aligned} \sum_j a_{ij}^1 &= \sum_{j \geq 2} a_{ij} - \frac{a_{1i}}{a_{11}} \sum_{j \geq 2} a_{1j} \\ &= \sum_j a_{ij} - \frac{a_{1i}}{a_{11}} \left\{ a_{11} + \sum_{j \geq 2} a_{1j} \right\} \\ &\geq -\frac{a_{1i}}{a_{11}} \sum_j a_{1j} \geq 0. \end{aligned}$$

□

Theorem A.6
If A is a weakly diagonally dominant Stieltjes matrix and B is such that:

$$a_{ij} \leq b_{ij} \leq 0 \quad \forall i \neq j \quad \text{and} \quad \sum_j a_{ij} = \sum_j b_{ij} \quad \forall i,$$

and B is irreducible, then B is also a weakly diagonally dominant Stieltjes matrix.

Proof
See [6] or [1]. □

Lemmas 5.2 and 6.2
The modified preconditioners are symmetric positive definite.

Proof
It is similar to the proof of lemmas 4.1 and 4.3. The only assertion which remains to be checked is that $\underline{A}^k$ is irreducible. This is a consequence of its structure. As we keep a five-point scheme structure, the graph associated to the matrix is strongly connected. Thus $\underline{A}^k$ is irreducible (see [18]). □

A by-product is that, in the modified case, R is negative semidefinite. Indeed, its off-diagonal entries are non-negative and its rows sums are equal to zero. Thus the following holds.

Theorem A.7
For a modified incomplete factorization of A,

$$\lambda_{\min}(M^{-1}A) \geq 1.$$

Proof
We have $A = M - R$ and we know that R is negative semidefinite. Therefore

$$\begin{aligned}(Ax, x)_n &= (Mx, x)_n - (Rx, x)_n \\ &\geq (Mx, x)_n.\end{aligned}$$

This yields the result. □

References

[1] C.W. Brand, An incomplete-factorization preconditioning using repeated red–black ordering, Numer. Math. 61 (1992) 433–454.

[2] C. Brand and Z.E. Heinemann, A new iterative solution technique for reservoir simulation equations on locally refined grids, SPE 18410 (1989).

[3] T.F. Chan and H.C. Elman, Fourier analysis of iterative methods for elliptic problems, SIAM Rev. 31 (1989) 20–49.

[4] P. Ciarlet, Jr, Etude de préconditionnements parallèles pour la résolution d'équations aux dérivées partielles elliptiques. Une décomposition de l'espace $L^2(\Omega)^3$, Thèse, Université Pierre et Marie Curie (1992).

[5] P. Ciarlet, *Introduction à l'Analyse Numérique Matricielle et à l'Optimisation* (Masson, 1982).

[6] P. Ciarlet, B. Miara and J.-M. Thomas, *Exercises d'Analyse Numérique Matricielle et d'Optimisation avec Solutions* (Masson, 1986).

[7] I.S. Duff and G. Meurant, The effect of ordering on preconditioned conjugate gradients, BIT 29 (1989) 635–657.

[8] T. Dupont, R.P. Kendall and H.H. Rachford, An approximate factorization procedure for solving self-adjoint elliptic difference equations, SIAM J. Numer. Anal. 5 (1968) 559–573

[9] T. Dupont, H.L. Stone and H.H. Rachford, Factorization techniques for elliptic difference equations, SIAM Proc. 2 (1970) 168–174.

[10] A. George and J.W.H. Liu, *Computer Solution of Large Sparse Positive Definite Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1981).
[11] G.H. Golub and G. Meurant, *Résolution Numérique des Grands Systèmes Linéaires* (Eyrolles, Paris, 1983).
[12] I. Gustafsson, A class of first order factorization methods, BIT 18 (1978) 142–156.
[13] I. Gustafsson, On first and second order symmetric factorization methods for the solution of elliptic difference equations, Computer Sciences, 78.01R, Chalmers University of Technology, Sweden (1978).
[14] I. Gustafsson, On modified incomplete Cholesky factorization methods for the solution of problems with mixed boundary conditions and problems with discontinuous material coefficients, J. Numer. Meth. Eng. 14 (1979) 1127–1140.
[15] P. Lascaux and R. Théodor, *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur* (Masson, 1986).
[16] J.A. Meijerink and H.A. Van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, Math. Comp. 31 (1977) 148–162.
[17] J.A. Meijerink and H.A. Van der Vorst, Guidelines for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems, J. Comp. Phys. 44 (1981) 135–155.
[18] R.S. Varga, *Matrix Iterative Analysis* (Prentice-Hall, 1962).