A CLASS OF FIRST ORDER
FACTORIZATION METHODS

IVAR GUSTAFSSON

Abstract.

A class of first order factorization methods for the solution of large, symmetric, sparse systems of equations is introduced. Asymptotic results for the computational complexity are developed, results from numerical experiments are presented and comparisons with other iterative and direct methods are carried out.

1. Introduction.

Almost all methods to solve a symmetric, positive definite, sparse system of linear equations, $Ax=f$, which arises from finite difference or finite element approximation of a selfadjoint elliptic partial differential equation problem of second order can be seen as special cases of a general method, a so-called factorization method, which can be stated as

$$(1.1) \qquad Cx^{l+1} = Cx^l - \beta_l r^l, \quad l=0,1,\dots,$$

where $C=A+R$, x^0 is arbitrary, $r^l=Ax^l-f$ and β_l is an iteration parameter.

In the following we assume that $C=LL^T$ is symmetric and positive definite. An acceleration procedure like the Chebyshev semi-iterative method or the conjugate gradient method can then be used for choosing β_l in (1.1), in order to increase the rate of convergence.

In [2] it is shown that these methods converge to a relative error ε in at most

$$(1.2) \qquad \text{ent} \left[\tfrac{1}{2}\mathscr{H}(C^{-1}A)^{\frac{1}{2}} \ln (2/\varepsilon)+1\right]$$

number of iterations, where $\mathscr{H}(C^{-1}A)$ is the spectral condition number of the matrix $C^{-1}A$.

The method (1.1) includes the well-known Cholesky method, namely if $R=0$. Then (neglecting rounding errors) $\mathscr{H}(C^{-1}A)=1$ and only one or a few steps of iterative refinement are needed. On the other hand, when $R=I-A$, that is when $C=I$, (1.1) becomes a purely iterative method.

In between these extremes there is an infinity of other choices of C, leading to different factorization methods. For a general discussion of the choice of C see e.g. [3] and [1].

In this paper a class of factorization methods, that is, methods for the construction of the matrix C, is introduced, which represents incomplete Cholesky

factorizations of A, see also [4]. The idea in these methods is to let L have nonzero entries in certain positions chosen in advance. Various methods arise by choosing different positions. Some choices for special matrices are described in [1] and in Section 3 of this paper.

The methods can be seen as modifications of the methods described in [4]. The modification is made in order to obtain $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$, where h is the size of the mesh. The methods are of first order in the sense that each component of the vector Ru is of order h, when u is the nodal point vector corresponding to an once differentiable function which vanishes on the boundary (also see [5]). For such vectors u we use the notation $u \in C_0^1(\Omega)$. In [6] Stone introduced a second order factorization method, that is, each component of Ru is of order h^2 for $u \in C_0^2(\Omega)$, the so-called strongly implicit method (SIP). In this method, however, C is not symmetric and furthermore in [5] Saylor claims that no symmetric second order factorization method exists which is practically useful. In Section 2 we will state a necessary condition for $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$, a relation satisfied by the methods introduced in this paper and closely related to the property first order method. We will also see that the methods in [4] are not of first order and that they do not satisfy the necessary condition. In Section 4 it is proved that some of the methods presented in Section 3 satisfy a certain sufficient condition for $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$. For finite difference (5-point) approximation the simplest method in Section 3 is identical with a method described by Dupont, Kendall and Rachford Jr. [7], as well as with the generalized $SSOR$ method in Axelsson [8], apart from a slightly different choice of the preconditioning parameter. The well-known $SSOR$ method, see e.g. [2], where $C = (\tilde{D} + L)\tilde{D}^{-1}(\tilde{D} + L^T)$, $\tilde{D} = \omega^{-1}D$, $D = \text{diag}(A)$ and L strictly lower triangular, gives, with the preconditioning parameter ω properly chosen, the same rate of convergence, that is, $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$, for Dirichlet problems. This, however, is not true for problems with Neumann boundary conditions.

2. A class of first order factorization methods.

We consider certain finite difference or finite element approximations arising from discretization of the second order self-adjoint elliptic partial differential equation problem in two dimensions;

(2.1)
$$-(\partial/\partial x)(a_1(x,y)(\partial/\partial x)u(x,y)) - (\partial/\partial y)(a_2(x,y)(\partial/\partial y)u(x,y)) + q(x,y) = f(x,y)$$

with $a_j(x,y) > 0$, $j = 1, 2$, $q \geqq 0$, $(x,y) \in \Omega \subset R^2$ and with suitable boundary conditions on $\partial\Omega$.

For simplicity but without loss of generality we assume that $q \equiv 0$. We refer to problem (2.1) with $a_1(x,y) = a_2(x,y) = 1$, $q \equiv 0$, Ω the unit square and Dirichlet boundary conditions as *the model problem*.

Discretization of (2.1) leads to a system of linear equations, $Ax = f$, where $A = (a_{ij})$ is a symmetric, positive definite, sparse matrix of order $N = \mathcal{O}(h^{-2})$.

Furthermore, since the basis functions have local support, A is a "local" matrix so that the distance between two points in the mesh representing indices i and j is $\mathcal{O}(h)$ for $a_{ij} \neq 0$ and the number of indices j such that $a_{ij} \neq 0$ is $\mathcal{O}(1)$ for each i. In the following we assume that the elements a_{ij} are normalized to be of order $\mathcal{O}(1)$.

By an elementary summation by parts, see [1], we obtain

$$(2.2) \qquad (Ax, x) = -\sum_i \sum_{j>i} a_{ij}(x_i - x_j)^2 + \sum_i \sum_j a_{ij} x_i^2 .$$

For $u \in C_0^1(\Omega)$ we have $\sum_j a_{ij} u_i^2 = \mathcal{O}(h^2)$ since $\sum_j a_{ij} = 0$ except for i representing nodal points near the boundary. Further, since A is local, $u_i - u_j = \mathcal{O}(h)$ for i and j such that $a_{ij} \neq 0$. From (2.2) and the fact that $N = \mathcal{O}(h^{-2})$ we then get

$$(2.3) \qquad (Au, u) = \mathcal{O}(1), \ h \to 0, \ u \in C_0^1(\Omega) .$$

For the defect matrix $R = (r_{ij})$ the relation corresponding to (2.2) is

$$(2.4) \qquad (Rx, x) = -\sum_i \sum_{j>i} r_{ij}(x_i - x_j)^2 + \sum_i \sum_j r_{ij} x_i^2$$

and for $u \in C_0^1(\Omega)$ we have

$$(2.5) \qquad (Ru, u) = \sum_i \sum_j r_{ij} u_i^2 + \mathcal{O}(1), \ h \to 0 .$$

From (2.3) and the relation

$$(Ax, x)/(Cx, x) = 1/[1 + (Rx, x)/(Ax, x)]$$

valid for all $x \neq 0$ it is clear that a *necessary* condition for $\mathscr{K}(C^{-1}A) = \mathcal{O}(h^{-1})$ is

$$(2.6) \qquad -\mathcal{O}(1) \leq (Ru, u) \leq \mathcal{O}(h^{-1}), \ u \in C_0^1(\Omega) .$$

This condition is related to the property first order method in the sense that $(Ru, u) = \mathcal{O}(h^{-1})$ when R is a first order matrix, that is

$$(2.7) \qquad (Ru)_i = \mathcal{O}(h), \ \forall i .$$

Notice that from

$$(Ru)_i = \sum_{j \in M_i} r_{ij}(u_i + \mathcal{O}(h)), \ u \in C_0^1(\Omega) ,$$

where the number of indices in $M_i = \{j; r_{ij} \neq 0\}$ is $\mathcal{O}(1)$, we have

$$(2.8) \qquad (Ru)_i = \sum_j r_{ij} u_i + \mathcal{O}(h), \ u \in C_0^1(\Omega) .$$

Also observe that (2.7) is not sufficient for (2.6). For the methods described in this paper, however, we have $\sum_j r_{ij} = \mathcal{O}(h^2)$ for Dirichlet problems. Then it is obvious from (2.5) that (2.6) is satisfied. $(Ru)_i$, however, is still of order $\mathcal{O}(h)$, see (2.8), giving a first order method.

For the methods in [4] we have $r_{ij} \geq 0$ and $\sum_j r_{ij} = \mathcal{O}(1)$ (for almost all values of i). Then it is clear from (2.8) that these methods are not of first order. Furthermore

it is an easy matter to find (a positive) $u \in C_0^1(\Omega)$ for which $(Ru, u) = \mathcal{O}(h^{-2})$, that is, the necessary condition (2.6) is not fulfilled. For $(Ru, u) = \mathcal{O}(h^{-2})$ we get $(Au, u)/(Cu, u) = \mathcal{O}(h^2)$ while e.g. $e_1 = (1, 0, \ldots, 0)^T$ gives $(Ae_1, e_1)/(Ce_1, e_1) = \mathcal{O}(1)$. Then it is clear that these methods give a condition number of order (at least) $\mathcal{O}(h^{-2})$, which is actually the same order as for $\mathcal{H}(A)$.

We shall show that, at least for A an M-matrix (that is $a_{ij} \leq 0$ for $i \neq j$ and $A^{-1} \geq 0$), it is possible to construct a factorization method for which (2.6) is satisfied with the upper bound $\mathcal{O}(1)$.

To this end let

$$(2.9) \qquad C = LL^T = A + R = A + \hat{R} + D ,$$

where $\hat{R} = (\hat{r}_{ij})$ is negative semidefinite (that is, $(\hat{R}x, x) \leq 0$, $\forall x$) and $\sum_j \hat{r}_{ij} = 0$, $\forall i$, and the choice of the positive diagonal matrix D depends on the boundary conditions.

For the Dirichlet problems we will choose $D = \xi h^2 \operatorname{diag}(A)$, $\xi > 0$ being a parameter. For Neumann problems some elements of D, corresponding to points on the part of the boundary with Neumann conditions, must be of order $\mathcal{O}(h)$, see [9]. In the following we confine the study to Dirichlet problems. Similar results for Neumann problems are shown in [9].

From (2.5) it is obvious that R in (2.9) satisfies (2.6) (with the upper bound $\mathcal{O}(1)$). From (2.8) it is also clear that the methods in the class (2.9) are of first order.

In the following m_j, $j = 1, 2, \ldots$ are positive constants independent of h.

Since $m_1 h^2 \leq (Ax, x)/(x, x) \leq m_2$ we have $0 \leq (Dx, x)/(Ax, x) \leq m_3$ and furthermore

$$(2.10) \qquad (1 + m_3)^{-1} \leq (Ax, x)/(Cx, x) \leq 1/[1 + (\hat{R}x, x)/(Ax, x)] .$$

We will now state a *sufficient* condition to obtain a condition number $\mathcal{H}(C^{-1}A)$ of order $\mathcal{O}(h^{-1})$.

THEOREM 2.1. *Let A be factored as in* (2.9). *Then a sufficient condition to obtain* $\mathcal{H}(C^{-1}A) = \mathcal{O}(h^{-1})$ *is*

$$(2.11) \qquad -(\hat{R}x, x) \leq (1 + kh)^{-1}(Ax, x), \ \forall x ,$$

where $k > 0$ is independent of h.

PROOF. The result is immediate from (2.10) and the definition of $\mathcal{H}(C^{-1}A)$. ∎

In Section 4 we will prove (2.11) for some particular methods.

3. Description of some first order factorization methods for the finite difference approximation of a model problem.

Consider the problem (2.1) with $q \equiv 0$, Ω the unit square and Dirichlet boundary conditions. We will use a type of simple graph-theoretic tool, see also [7], to show which gridpoints are involved, and coefficient-notations for A, L, LL^T and $\hat{R}$, regarded as operators (or corresponding matrices) applied to grid functions. In this notation A is defined in Figure 3.1, where m is the band width of the matrix.
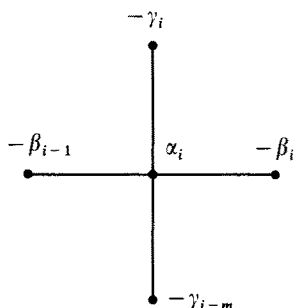
Figure 3.1. A.

As said above our methods can be treated as modifications of the methods (denoted $ICCG$ methods) in [4], and we will refer to them by $MICCG(n)$ to point out the relationship to a certain $ICCG(n)$ method.

The MICCG(0) method.

In this method the matrix L has nonzero elements in positions where the lower part of A has nonzero elements. We simply say that L contains no extra diagonal and we denote the method $MICCG(0)$ (*M*odified *I*ncomplete *C*holesky & *C*onjugate *G*radients with 0 extra diagonal).

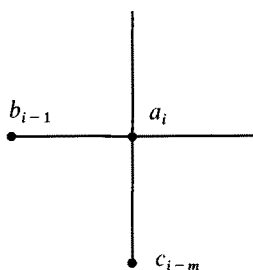For this method L, L^T, LL^T and $\hat{R}$ are defined in Figures 3.2–3.5.
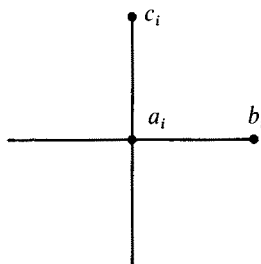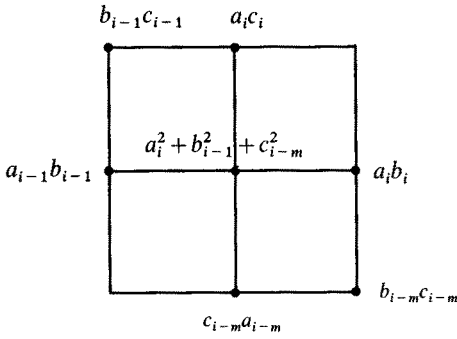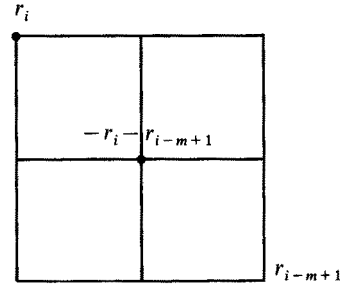
Figure 3.2. L.   Figure 3.3. L^T.

Figure 3.4. LL^T.

Figure 3.5. $\hat{R}$.

It is easy to see that (2.9) implies that the coefficients have to satisfy the formulas

(3.1)

$$\begin{cases} a_i^2 = \alpha_i(1+\delta) - r_i - r_{i-m+1} - b_{i-1}^2 - c_{i-m}^2 \\ b_i = -\beta_i/a_i \\ c_i = -\gamma_i/a_i \\ r_i = b_{i-1}c_{i-1} \\ \delta = \xi h^2, \ \xi > 0, \end{cases}$$

where elements not defined should be replaced by zeros.

This method is identical with that in [7] as well as with the generalized *SSOR* method in [8] apart from a slightly different choice of the parameter ξ.

For the (unmodified) $ICCG(0)$ method, for which diag $(R)=0$, the corresponding formulas are

$$a_i^2 = \alpha_i - b_{i-1}^2 - c_{i-m}^2$$

$$b_i = -\beta_i/a_i$$

$$c_i = -\gamma_i/a_i.$$

The MICCG(1) method.

A natural step to get a more accurate factorization is to allow L to have nonzero elements in positions where $\hat{R}$, in the $MICCG(0)$ method, has nonzero elements. This leads to the $MICCG(1)$ method defined in Figures 3.6–3.8 and the formulas (3.2).
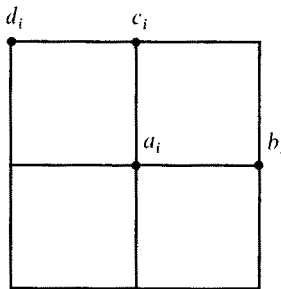
Figure 3.6. L^T.
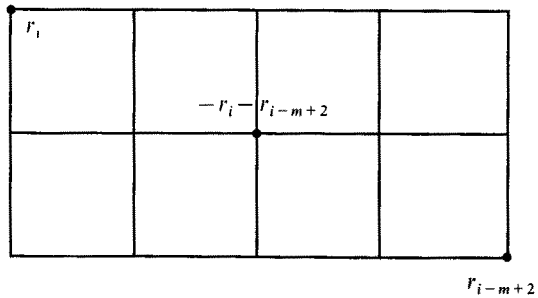
Figure 3.7. $\hat{R}$.
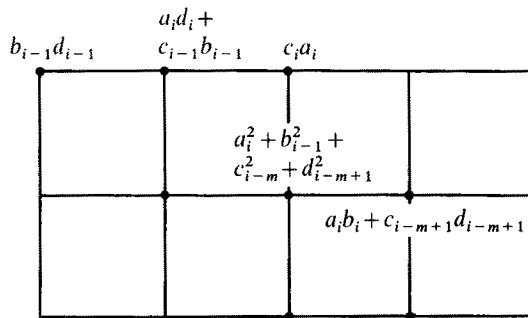
Figure 3.8. LL^T.

$$(3.2) \quad \begin{cases} a_i^2 = \alpha_i(1+\delta) - b_{i-1}^2 - c_{i-m}^2 - d_{i-m+1}^2 - r_i - r_{i-m+2} \\ b_i = -(\beta_i + c_{i-m+1}d_{i-m+1})/a_i \\ c_i = -\gamma_i/a_i \\ d_i = -c_{i-1}b_{i-1}/a_i \\ r_i = b_{i-1}d_{i-1} \\ \delta = \xi h^2, \ \xi > 0 . \end{cases}$$

Continuing in this way we first come to the $MICCG(2)$ method and then to the $MICCG(4)$ method. These and several other methods for finite element approximations of different problems (even three-dimensional) are described and investigated in [1] and [9].

In [4] Meijerink and van der Vorst claim that the $ICCG(3)$ method is a good method. The corresponding $MICCG(3)$ method, however, is not obtained by the general idea presented below to get more accurate first order factorization

methods. It is also confirmed in numerical tests that this method gives only a slightly more accurate factorization than the $MICCG(2)$ method. In order to compare with the unmodified methods, however, results for the $MICCG(3)$ method are given, among others, in Section 5.

Notice that with a simple modification of the formulas for deriving the elements in L one can avoid taking square-roots. This modification can be stated as $C = (\tilde{L} + \tilde{D}^2)(\tilde{D}^{-2})(\tilde{L}^T + \tilde{D}^2)$, where $\tilde{L}\tilde{D}^{-1} + \tilde{D} = L$ and $\tilde{L}$ is strictly lower triangular.

A general idea to obtain MICCG methods.

It is proved in [4] that an unmodified incomplete factorization of a general M-matrix gives $R \geq 0$, diag $(R) = 0$. In the same way it is easy to see that a modified method, as above, can be constructed for a general M-matrix corresponding to a finite difference or finite element operator to obtain $\hat{R}$ with nonnegative off-diagonal elements and negative diagonal elements so that (2.9) is satisfied.

For more general structured problems the idea to obtain a $MICCG$-method is to let L have nonzero elements in the same positions as A, form the product LL^T to see where $\hat{R}$ has nonzero elements, extend L to have nonzero elements in these positions to get a more accurate factorization and possibly continue in this manner a few steps more.

This general idea has been used in [1] and [9] to construct modified incomplete Cholesky factorization methods for various finite element approximations.

4. Bounds for the condition number $\mathscr{H}(C^{-1}A)$ for the finite difference approximation of a model problem.

For simplicity but without loss of generality we consider the model problem of Section 2. For smoothly variable coefficients and general domains the results of this section evidently can be obtained using the ideas in [7] or [8], see also [9].

In this section we will prove that the $MICCG(0)$ and $MICCG(1)$ methods, introduced in section 3, for the model problem give $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$. According to Theorem 2.1 we then have to prove relation (2.11).

To distinguish elements corresponding to different $MICCG(n)$ methods we sometimes use notations as $r_i^{(n)}$, $a_i^{(n)}$ etc.

The following lemma gives a bound for $r_i^{(0)}$ in the case of the model problem.

LEMMA 4.1. *Let $r_i^{(0)}$ be elements defined in (3.1) of the matrix $\hat{R}$ from the $MICCG(0)$ method concerning the 5-point approximation of the model problem. Then*

$$(4.1) \qquad r_i^{(0)} \leq 1/[2(1 + kh)], \text{ where } k > 0 \text{ is independent of } h.$$

PROOF. We first prove that a universal bound for $a_i^{(0)}$ is

$$(4.2) \qquad a_i^2 \geq 2(1 + kh), \forall i, k > 0 \text{ independent of } h.$$

Consider the formulas (3.1) for deriving the coefficients, where for the model problem $\alpha_i = 4$, $\beta_i \leqq 1$, and $\gamma_i \leqq 1$, $i = 1, 2, \ldots, N$. For $\xi = 0$ we can simply derive the bound $a_i^2 \geqq 2$, $i = 1, 2, \ldots, N$ by induction on i. Now $a_1^2 = 4 \geqq 2$ is obvious and we further suppose that $a_i^2 \geqq 2$ for $i = 1, 2, \ldots, p-1$. Then

$$a_p^2 = 4 - b_{p-1}(b_{p-1} + c_{p-1}) - c_{p-m}(c_{p-m} + b_{p-m})$$

$$\geqq 4 - 2/a_{p-1}^2 - 2/a_{p-m}^2 \geqq 4 - 1 - 1 = 2$$

and by induction

$$a_i^2 \geqq 2, \quad i = 1, 2, \ldots, N .$$

It is easy to see that this bound can be obtained by solving $\Psi(a) = 0$, where $\Psi(a) = a^2 + 4/a^2 - 4$.

In the same way, for $\xi > 0$, we obtain the bound by solving $\Psi(a) - 4\xi h^2 = 0$. This gives

$$a^2 - 2 = 2\xi^{\frac{1}{2}}ha ,$$

$$a = \xi^{\frac{1}{2}}h + (\xi h^2 + 2)^{\frac{1}{2}}$$

and

$$a^2 = \xi h^2 + \xi h^2 + 2 + 2h\xi^{\frac{1}{2}}(\xi h^2 + 2)^{\frac{1}{2}} \geqq 2(1 + kh) ,$$

where $k = (2\xi)^{\frac{1}{2}}$ and (4.2) is proved.

Since

$$r_i = b_{i-1}c_{i-1} ,$$

we obtain

$$r_i \leqq 1/a_{i-1}^2 \leqq 1/[2(1 + kh)]$$

and the lemma is proved. ∎

We also need the following lemma.

LEMMA 4.2. *Let c and d be positive and let a, b, and e be real. Then*

$$(c + d)^{-1}(a - b)^2 \leqq c^{-1}(a - e)^2 + d^{-1}(e - b)^2 .$$

PROOF. This is trivially seen from

$$(a - b)^2 \leqq (1 + \varepsilon)(a - e)^2 + (1 + \varepsilon^{-1})(e - b)^2$$

valid for all $\varepsilon > 0$ and in particular for $\varepsilon = d/c$. ∎

For the model problem and $x = [x_1, \ldots, x_N]^T$ (2.2) becomes

(4.3) $$(Ax, x) \geqq \sum_i [\beta_i(x_i - x_{i+1})^2 + \gamma_i(x_i - x_{i+m})^2]$$

where

(4.3′)
$$\beta_i = \begin{cases} 0 \text{ for } i = pm, \ p = 1, 2, \ldots, m \\ 1 \text{ otherwise} \end{cases}$$

and

(4.3″)
$$\gamma_i = \begin{cases} 0 \text{ for } i > N - m \\ 1 \text{ otherwise} . \end{cases}$$

Notice that since $r_{i+1} = \beta_i \gamma_i / a_i^2$ we have

(4.4)
$$r_{i+1} \neq 0 \quad \Leftrightarrow \quad \beta_i \gamma_i \neq 0 .$$

For the matrix $\hat{R}$ defined in Figure 3.5 we have from (2.4)

$$-(\hat{R}x, x) = \sum_{r_i \neq 0} r_i (x_i - x_{i+m-1})^2$$

and from (4.1)

$$-(\hat{R}x, x) \leq [2(1 + kh)]^{-1} \sum_{r_i \neq 0} (x_i - x_{i+m-1})^2 .$$

Using Lemma 4.2 with $c = d = 1$ for each gridpoint we get

$$-(\hat{R}x, x) \leq (1 + kh)^{-1} \sum_{r_i \neq 0} [(x_i - x_{i-1})^2 + (x_{i-1} - x_{i+m-1})^2]$$

$$= (1 + kh)^{-1} \sum_{r_{i+1} \neq 0} [(x_{i+1} - x_i)^2 + (x_i - x_{i+m})^2]$$

and from (4.3) and (4.4) we finally obtain

$$-(\hat{R}x, x) \leq (1 + kh)^{-1}(Ax, x) ,$$

which is the desired result (2.11).

Further straightforward calculations give $\mathcal{H}(C^{-1}A) \leq m_4 + m_5 h^{-1}$, where for the model problem $m_5 = m_5(\xi)$ is minimized for $\xi = \pi^2/8$ and then $\mathcal{H}(C^{-1}A) \leq 2 + 4(\pi h)^{-1}$.

For the $MICCG(1)$ method the proof is similar and details are left to the reader (however, see [1]). The following lemma gives a bound for $r_i^{(1)}$.

LEMMA 4.3. *Let* $r_i^{(1)}$ *be elements, defined in* (3.2), *of the matrix* $\hat{R}$ *from the* $MICCG(1)$ *method concerning the 5-point approximation of the model problem. Then*

(4.5) $\quad r_i^{(1)} \leq 1/[5(1 + kh)]$, *where* $k > 0$ *is independent of* h .

PROOF. Considering the formulas (3.2) for the model problem we see that a bound for $a_i^{(1)}$ when $\xi = 0$ is obtained by solving $\Psi(a) = 0$, where

$$\Psi(a) = a^2 + 1/a^2 + [a(1 - 1/a^2)]^{-2} - 4 .$$

$\Psi(a)$ has the positive solution $a = (1 + \sqrt{5})/2$ and therefore

$$a_i^2 \geq [(1 + \sqrt{5})/2]^2, \quad i = 1, 2, \ldots, N \ .$$

From (3.2) we can obtain corresponding bounds for the other coefficients for $\xi = 0$. This gives

$$-b_i \leq (5 + \sqrt{5})/10 \ ,$$

$$-d_i \leq (5 - \sqrt{5})/10$$

and finally

$$r_i = b_{i-1} d_{i-1} \leq 1/5 \ .$$

For $\xi > 0$ we have to solve $\Psi(a) - 4\xi h^2 = 0$ which leads to

$$a_i^2 \geq [(1 + \sqrt{5})/2]^2 (1 + kh)$$

and

$$r_i \leq 1/[5(1 + kh)], \quad i = 1, 2, \ldots, N \ ,$$

where $k > 0$ is independent of h, which proves the lemma. ∎

For the matrix $\hat{R}$ defined in Figure 3.7 we have, see (2.4),

$$-(\hat{R}x, x) = \sum_{r_i \neq 0} r_i (x_i - x_{i+m-2})^2 \ .$$

Using (4.5) and Lemma 4.2 twice, first with $c = 2, d = 3$ and then with $c = 1, d = 2$ we obtain

$$(4.6) \qquad -(\hat{R}x, x) \leq (1 + kh)^{-1} \sum_{r_i \neq 0} \tfrac{1}{5}(x_i - x_{i+m-2})^2$$

$$\leq (1 + kh)^{-1} \left\{ \sum_{r_{i+1} \neq 0} [\tfrac{1}{2}(x_{i+1} - x_i)^2 + \tfrac{1}{2}(x_i - x_{i+m})^2] \right.$$

$$\left. + \sum_{r_{i-m+1} \neq 0} [\tfrac{1}{2}(x_i - x_{i-1})^2 + \tfrac{1}{2}(x_i - x_{i-m})^2] \right\} \ .$$

From (4.3) we have

$$(4.7) \qquad (Ax, x) \geq \sum_i [\beta_{i-1}(x_i - x_{i-1})^2 + \sum_i \gamma_{i-m}(x_{i-m} - x_i)^2]$$

where elements not defined should be replaced by zeros.

From the formulas (3.2) it is clear that $r_{i+1} = r_{i-m+1} = 0$ for such an i that $\beta_i = 0$, $\beta_{i-1} = 0$, $\gamma_i = 0$ or $\gamma_{i-m} = 0$.

Comparing (4.6) and (4.7) we therefore obtain

$$-(\hat{R}x, x) \leq (1 + kh)^{-1}[\tfrac{1}{2}(Ax, x) + \tfrac{1}{2}(Ax, x)] = (1 + kh)^{-1}(Ax, x)$$

and (2.11) is proved.

Before summing up the results of this section in Theorem 4.2 we notice that the number of nonzero elements in L for these methods is $\mathcal{O}(N)$ and therefore each iteration in (1.1) can be carried out in $\mathcal{O}(N)$ arithmetic operations.

THEOREM 4.2. *The MICCG(0) and MICCG(1) methods for solving the model problem with 5-point approximation need $\mathcal{O}(N^{1.25}\ln(2/\varepsilon))$ operations to reduce the relative error by a factor ε.*

PROOF. This is an immediate consequence of (1.2) and the results in this section. ∎

This author believes that it is possible to prove relation (2.11) and thus obtain the result of Theorem 4.2 for the other methods described and investigated in [1] and [9] in the same way as above, at least when the matrix A is an M-matrix. This belief is supported by the numerical tests.

5. Comments to numerical experiments and conclusions.

Several testproblems with various approximations and boundary conditions, even problems with discontinuous material coefficients and three-dimensional problems, have been investigated in [1] and [9].

All the results properly reflect the theoretical result $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-1})$. Although this result is not yet covered by the theory for problems where A is not an M-matrix, the same rate of convergence has been observed for such problems.

The experiments show that the number of iterations is almost independent of the parameter ξ in a fairly wide range.

For the model problem $MICCG(n)$ methods with different values of n have been studied. The result was that it is worth computing nonzero elements in only a few subdiagonals of L to get a more accurate factorization of the matrix. In Figure 5.1 and Figure 5.2 the number of iterations and the total work, factorization and solution work, respectively, are given for the different methods when $N = 1600$ and $\varepsilon = 10^{-6}$.

In comparison with other iterative methods the best first order factorization methods need about 30% less arithmetic operations than the $SSOR$-method or the unmodified methods in [4], for average-sized ($N = 1000$–2000) Dirichlet problems. In Fig. 5.3 the number of iterations which was needed for some methods to solve the model problem, is given for different values of N and $\varepsilon = 10^{-6}$. It is observed that for large values of N the number of iterations for the unmodified $ICCG$ methods grows as $\mathcal{O}(N^{\frac{1}{2}})$, which is in agreement with the result $\mathscr{H}(C^{-1}A) = \mathcal{O}(h^{-2})$. For the modified methods as well as for the $SSOR$ method, however, the number of iterations grows as $\mathcal{O}(N^{\frac{1}{4}})$.
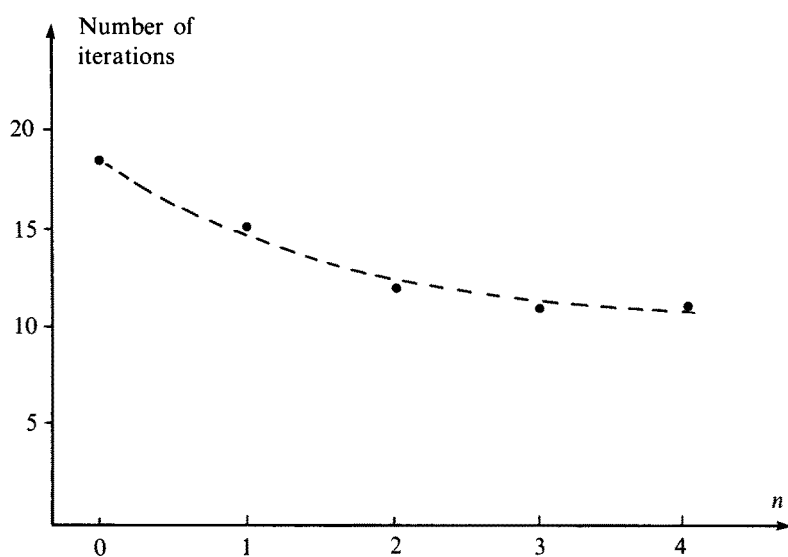
Figure 5.1. The number of iterations for the $MICCG(n)$ methods for the model problem with $N = 1600$ and $\varepsilon = 10^{-6}$.
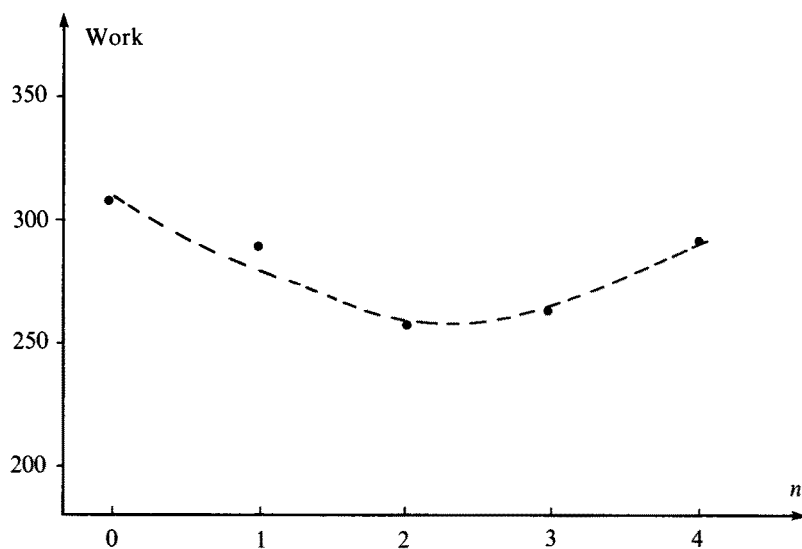

Figure 5.2. The number of arithmetic operations per unknown for the $MICCG(n)$ methods for the model problem with $N = 1600$ and $\varepsilon = 10^{-6}$.
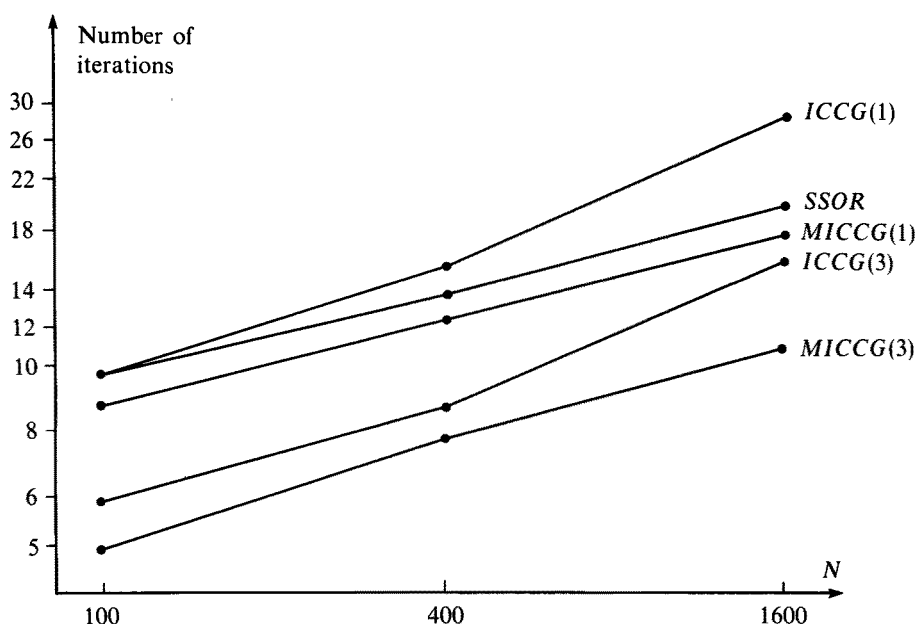
Figure 5.3. The number of iterations for some methods as a function of N for the model problem with ε $=10^{-6}$, logarithmic scale.

For Neumann problems and problems with discontinuous material coefficients the gain from the $SSOR$ method to the best first order factorization method often exceeds 60% for average-sized problems. This is in particular valid when the number of points with Neumann conditions is large in relation to the number of points with Dirichlet conditions or when the rate of discontinuity is important.

In [1] a comparison is also made between the first order factorization methods and some direct methods namely band-elimination, the nested dissection method [10], the one-way dissection method [11] and the minimum degree method [12]. The result was that the first order factorization methods compare with the best direct methods when one or a couple of righthand sides are present with the same matrix in two-dimensional problems of average size. For several righthand sides, however, a fast direct method is to prefer as long as the problem is not too large so that the limited capacity of the memory has to be taken into account.

For three-dimensional problems the iterative methods are superior to the direct methods even for an infinite number of righthand sides.

Other advantages with the iterative methods are that they need less storage, can benefit from a good initial approximation (as in time-dependent problems, in iterative design of boundary value problems and in quasi-linear problems), often have less accumulation of rounding (cancellation) errors, and are easy to program.

Finally it is of interest to mention that the first order factorization methods have successfully been used to factor even nonsymmetric matrices, see [13], and in problems where it is sufficient to have only an approximate factorization of a matrix (of current interest), for instance when the Navier equations of elasticity are solved by iterative methods, see [14].

Acknowledgements.

REFERENCES

1. I. Gustafsson, *A class of first order factorization methods*, Computer Sciences 77.04R, Chalmers University of Technology, Göteborg, Sweden, (1977).
2. O. Axelsson, *A class of iterative methods for finite element equations*, Comp.meth. in appl. mechanics and engineering 9 (1976), 123–137.
3. O. Axelsson, *On preconditioning and convergence acceleration in sparse matrix problems*, CERN 74-10, Genève, Switzerland (1974).
4. 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. of Comp. 31 (1977), 148–162.
5. P. Saylor, *Second order strongly implicit symmetric factorization methods for the solution of elliptic difference equations*, SIAM J. Numer. Anal. 11 (1974), 894–908.
6. H. L. Stone, *Iterative solution of implicit approximations of multidimensional partial differential equations*, SIAM J. Numer. Anal. 5 (1968), 530–558.
7. T. Dupont, R. Kendall and H. H. Rachford Jr., *An approximate factorization procedure for solving selfadjoint elliptic difference equations*, SIAM J. Numer. Anal. 5 (1968), 559–573.
8. O. Axelsson, *A generalized SSOR method*, BIT 12 (1972), 443–467.
9. I. Gustafsson, *On first order factorization methods for the solution of problems with mixed boundary conditions and problems with discontinuous material coefficients*, Computer Sciences 77.13R, Chalmers University of Technology, Göteborg, Sweden (1977).
10. A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal. 10 (1973), 345–363.
11. A. George, *Numerical experiments using dissection methods to solve n by n grid problems*, Research Report CS-75-07, University of Waterloo, Canada (1975).
12. H. M. Markowitz, *The elimination form of the inverse and its applications to linear programming*, Management Sci. (1957), 255–269.
13. O. Axelsson and I. Gustafsson, *A modified upwind scheme for convective transport equations and the use of a conjugate gradient method for the solution of non-symmetric systems of equations*, Computer Sciences 77.12R, Chalmers University of Technology, Göteborg, Sweden (1977).
14. O. Axelsson and I. Gustafsson, *Iterative methods for the solution of the Naviers equations of elasticity*, Computer Sciences 77.09R, Chalmers University of Technology, Göteborg, Sweden (1977).

CHALMERS UNIVERSITY OF TECHNOLOGY AND
THE UNIVERSITY OF GÖTEBORG
DEPARTMENT OF COMPUTER SCIENCES
FACK, S-402 20, GÖTEBORG
SWEDEN