ON THE ITERATIVE SOLUTION OF FINITE ELEMENT SYSTEMS OF EQUATIONS

O. Axelsson
Department of Mathematics
University of Nijmegen, The Netherlands

1. Introduction

For large sparse systems of linear equations it is mostly too timeconsuming and sometimes even not possible to use direct solution methods. Such problems arise for instance from discretized partial differential equations, in particular on a three dimensional body. In recent years however,very effective iterative solvers, based on certain preconditioned conjugate gradient methods have been developed.

2. Approximate factorizations

We consider here sparse approximate factorizations in triangular factors of a given sparse matrix A of order N and with symmetric nonzero structure. There are several such methods but the most effective seems to be the one based on a generalization (by allowing for incomplete factorizations) of the classical Gaussian elimination algorithm (see [8],[7],[6],[3]).

We recall that the envelope S of A is the set of indices

$$S = \{(i,j) \cup (j,i);\ i_j \le i \le j,\ 1 \le j \le N\}$$

where $i_j = \min\{i,\ 1 \le i \le j;\ a_{ij} \ne 0\}$, $j = 1,\dots,N$.

Let now $J \subset S$ be a subset of S. We shall describe the (modified) incomplete factorization LU of A corresponding to the index set J.

During the generalized Gaussian elimination algorithm we construct a sequence of matrices $A^{(r)}$ of order $N-r+1$ with $A^{(1)} = A$ defined in the following way:

If $(i,j) \notin J$ then we put $a_{ij}^{(r+1)} = 0$ but (in the modified algorithm only) before that we add this entry to the current value of the diagonal entry $a_{ii}^{(r+1)}$. In other words, for $r = 1,2,\dots,N-1$ put

$$(2.1)\qquad \ell_{ir} = a_{ir}^{(r)};$$

$$a_{ij}^{(r+1)} = \begin{cases} a_{ij}^{(r)} - \ell_{ir} a_{ij}^{(r)} & ,\ (j = r+1,\dots,N) \wedge (i,j) \in J) \wedge (i \ne j), \\ 0 & ,\ (j = r+1,\dots,N) \wedge (i,j) \notin J), \\ a_{ij}^{(r)} - \ell_{ir} a_{rj}^{(r)} + \sum\limits_{\substack{k \ge r+1 \\ (i,k) \notin J}} (a_{ik}^{(r)} - \ell_{ir} a_{rk}^{(r)}), & i = j, \end{cases}$$

where $i = r+1,\dots,N$.

In this way we avoid the growth of the number of fill-in entries, which is a wellknown disadvantage of the full factorization algorithm, where $J = S$. Since when J is a proper subset of S, in general we only perform an approximate factorization we have to couple the method with an iterative method. This shall be discussed later. At first we discuss the stability of the algorithm for a special class of matrices.

Definition 2.1. An N×N matrix A is said to be an $\hat{M}$-matrix if

$$a_{ii} > 0,\ i = 1,2,\dots,N-1,\ a_{NN} \ge 0$$

$$a_{ij} \le 0,\ i,j = 1,2,\ldots,N,\ i \ne j$$

and for $i = 1,2,\ldots,N-1$ there is an entry $a_{i,j_0(i)} \ne 0$, where $i < j_0(i) \le N$.

We now introduce the <u>growth factor</u>

$$q := \max_{\substack{i\le j\le N\\ r+1\le i\le N-1\\ 1\le r\le N-1}} |a_{ij}^{(r)} / a_{ii}^{(r)}|$$

(note that the terms $|\ell_{ir} a_{rj}^{(r)}|$ in the Gaussian elimination are bounded by $q|a_{ir}^{(r)}|$) and the rowsums of $A^{(r)}$

$$s_i^{(r)} = \sum_{j=r+1} a_{ij}^{(r)},\ i = r+1,\ldots,N.$$

<u>Theorem 2.1</u>. Let A be a diagonally dominant $\hat{M}$-matrix. Then $A^{(2)},\ldots,A^{(N)}$ are also diagonally dominant $\hat{M}$-matrices. Further

$$a_{ij}^{(r+1)} \le a_{ij}^{(r)},\quad i,j = r+1,\ldots,N$$
$$s_i^{(r+1)} \ge s_i^{(r)},\quad i = r+1,\ldots,N. \qquad \square$$

This follows from algorithm (2.1) by an easy calculation.

The diagonal dominance implies $s_i^{(r)} \ge 0$. Hence

<u>Corollary 2.1</u>. Modified incomplete factorization of a diagonally dominant $\hat{M}$-matrix is a stable process in the respect that $q = 1$. $\square$

It may happen that the final diagonal entry $a_{NN}^{(N)}$ in U is zero. It is easily seen that this cannot happen however, if we add two properties.

<u>Theorem 2.2</u>. Let A be a diagonally dominant $\hat{M}$-matrix with symmetric structure and suppose that J is symmetric and that at least one rowsum is positive. Then

(i) The matrix $A^{(r)}$ $(r = 1,\ldots,N-1)$ has at least one positive rowsum

(ii) $a_{NN}^{(N)} > 0$. $\square$

The example $A^{(1)} = A = \begin{bmatrix} 3 & -1 & 0 & -2 \\ -2 & 4 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$, $J = \{(i,j),\ a_{ij} \ne 0\}$

for which we get

$$U = U_J = \begin{bmatrix} 3 & -1 & 0 & 2 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

shows that it is essential for $a_{NN}^{(N)} > 0$ that the nonzero structure is symmetric.

3. <u>The spectral conditionnumber</u>

For the solution of $\underline{A}\underline{x} = \underline{b}$ we use some iterative method like the generalized conjugate gradient method on the form

$$\begin{aligned} &\underset{\sim}{x}^{\ell+1} = \underset{\sim}{x}^{\ell} + \tau_\ell \underset{\sim}{d}^{\ell}, \quad LU\underset{\sim}{g}^{(\ell)} = \underset{\sim}{r}^{\ell} \\ &\underset{\sim}{d}^{\ell+1} - \underset{\sim}{d}^{\ell} = \beta_\ell \underset{\sim}{g}^{\ell}, \quad \underset{\sim}{r}^{\ell} = A\underset{\sim}{x}^{\ell} - \underset{\sim}{b} \end{aligned} \qquad \ell = 0,1,\ldots$$

where $\underset{\sim}{x}^0$ is a given initial approximation; $d^0 = -r^0$ and τ_ℓ, β_ℓ are calculated from certain innerproducts (see [1]).

In general, the rate of convergence depends on the distribution of eigenvalues of $B = (LU)^{-1}A$. For symmetric positive definite matrices however, the spectral conditionnumber of B, $\varkappa(B) = \lambda_0(B)/\lambda_1(B)$ where λ_0, λ_1 are the extreme eigenvalues, is often a proper measure. Hence we now consider $\varkappa(B)$. We have

$$A = LU+R$$

where
$$R = \sum_{r=1}^{N-1} R^{(r+1)}, \quad r_{ij}^{(r+1)} = \begin{cases} 0 & (i,j) \in J,\ i \neq j, \\ \tilde{a}_{ij}^{(r+1)} = a_{ij}^{(r)} - \ell_{ir} a_{rj}^{(r)}, & (i,j) \notin J, \\ \sum\limits_{\substack{k \geq r+1 \\ (i,k) \notin J}} - \tilde{a}_{ik}^{(r+1)} & , i = j. \end{cases}$$

Assume now that A is a symmetric diagonally dominant $\hat{M}$-matrix with at least one positive rowsum. Then (with J being a symmetric set) $C := LU = LDL^T$, $D = \text{diag}(a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{NN}^{(N)})$, $A = C+R$. C is symmetric and positive definite and R is symmetric and positive semidefinite.

By similarity the eigenvalues of $B = C^{-1}A$ and $\tilde{B} = D^{-\frac{1}{2}}L^{-1}AL^{-T}D^{-\frac{1}{2}}$ are equal. Since $\tilde{B}$ is symmetric and positive definite the eigenvalues are real and positive. Further

$$\lambda_1 = \min_{\underline{x} \in \mathbb{R}^N} \frac{\underline{x}^t A \underline{x}}{\underline{x}^t C \underline{x}} = 1 + \min_{\underline{x}} \frac{\underline{x}^t R \underline{x}}{\underline{x}^t C \underline{x}} = 1.$$

In order to estimate λ_0 we shall now consider a specific class of problems.

4. Finite element matrices

Consider a sequence of finite element matrices $\{K_h\}$ constructed from an original coarse triangular mesh Ω_0 by uniformly subdividing all triangles a number of times. Let the meshparameter h be defined as the ratio of the edges in the triangles in the resulting mesh Ω_h to the corresponding edges in the original triangles.

Definition 4.1. Let $\{\Omega_h\}$ be a sequence of meshes as defined above and let $N = N(h)$ be the number of nodes in Ω_h, excluding Dirichlet nodes. Let $\{K_h^{(1)}\}, \{K_h^{(2)}\}$ be two classes of positive definite matrices of order N(h). They are said to be spectrally equivalent if there exist positive constants α, β such that

$$\alpha \leq \frac{\underline{x}^t K_h^{(2)} \underline{x}}{\underline{x}^t K_h^{(1)} \underline{x}} \leq \beta \qquad \forall \underline{x} \in \mathbb{R}^N. \qquad \square$$

Note that for spectrally equivalent classes of matrices, the spectral conditionnumber

$$\varkappa(K_h^{(1)} K_h^{(2)}) \leq \beta/\alpha = O(1), \quad h \to 0.$$

We shall now present three examples of spectrally equivalent classes of matrices.

Example 1. Let $K = K_h$ be a diagonally dominant symmetric $\hat{M}$-matrix and let

(i) $\nu = \mu_1^{\frac{1}{2}}$, where μ_1 is the smallest eigenvalue of $D_K^{-1}K$, $D_K = \text{diag}(K)$,

(ii) N be a set of disjoint points in the set of meshpoints $(1,\ldots,N)$ such that from every point $i \in N$ in the connectivity graph of the matrix K, there is a path, $P(i) = \{j_0(i) = i, j_1(i), \ldots, j_{\tilde{p}}(i)\}$ of length $\tilde{p} = p(i)$, where any same point appears only once in any path and in only one path. Let $p = \min\limits_{i \in N} p(i)$.

Example 4.1. A cube with "brick-elements". Here $p = O(h^{-1})$ of N is contained in a fixed number of planes (i.e. a number independent on h).

Theorem 4.1. Let ν, N and the path $\{j_0(i), \ldots, j_{p(i)}\}$ be defined as above and let

$$K_h^{(1)} = \tilde{K}_h := K_h + (\zeta_1 \nu)^2 D_{K_h} + \zeta_2 \nu D'_{K_h},$$

where ζ_1, ζ_2 are nonnegative constants (independent on h) and where

$$(D_K')_{ii} = \begin{cases} (D_K)_{ii} & , \quad i \in N, \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Then
$$\mathcal{K}(K_h^{-1}\tilde{K}_h) \le 1 + \zeta_1^2 + \zeta_2\left(\frac{2a_0}{\nu p} + \frac{\nu p}{a_1}\right)$$
where
$$a_1 = \min_{i \in N_1} \min_{0 \le \ell \le p-1} |K_{j_\ell, j_{\ell+1}} / K_{ii}|$$
$$a_0 = \min_{i \in N_1} \max_{0 \le \ell \le p-1} K_{ii}/K_{j_\ell(i), j_\ell(i)}$$

<u>Corollary 4.1</u>. For a 2'nd order problem with the set N so defined that $p = O(h^{-1})$, then
$$\mathcal{K}(K_h^{-1}\tilde{K}_h) = O(1), \quad h \to 0.$$

<u>Proof</u>. This follows from Theorem 4.1, because $\nu = \mu_1^{\frac{1}{2}} = O(h)$, $h \to 0$ for second order problems.

<u>Example 2</u>. Let $a(.,.)$ be a symmetric coercive bilinear form. Let
$$K_h^{(1)} = [a(\lambda_j^{(h)}, \lambda_i^{(h)})]$$
where $\{\lambda_i^{(h)}\}_{i=1}^{N(h)}$ is the set of piecewise linear finite element basis functions and let
$$K_h^{(2)} = [a(\phi_j^{(h)}, \phi_i^{(h)})]$$
where $\{\phi_i^{(h)}\}_{i=1}^{N(h)}$ is the set of piecewise quadratic (or cubic....)finite element basis functions.

Then $\{K_h^{(1)}\}$, $\{K_h^{(2)}\}$ are spectrally equivalent (see [2]). As an example, let
$$a(u,v) = \int_\Omega \nabla u \nabla v d\underline{x}, \quad \Omega = [0,1]\times[0,1],$$
and use isosceles triangles and linear and quadratic basis functions. Then
$$\mathcal{K}(K_h^{(1)^{-1}} K_h^{(2)}) = 4/3.$$
Note that $K_h^{(1)}$ is a diagonally dominant $\tilde{M}$-matrix. Hence a "good" preconditioning C_h of MIC-type (as described in Section 2) may be constructed for $K_h^{(1)}$ and this is then also a "good" preconditioning for $K_h^{(2)}$. With MIC(0), i.e. $J = \{(i,j);\ K_{ij} \ne 0\}$ one finds in fact
$$\mathcal{K}(C_h^{-1} K_h^{(h)}) \le 4/3\lambda_0 \le 4/3(2 + \frac{2}{\pi h}),$$
where λ_0 is the largest eigenvalue of $C_h^{-1}K_h^{(1)}$. (Hence, if $h = 1/16$, which is a reasonable value in practice, the condition number is not larger than 16.)

In this case $\zeta_1 = \pi$ and $\zeta_2 = 0$ in Theorem 4.1. If we have a problem with discontinuous materials, then also $\zeta_2 > 0$ and N is the set of node points on the surface of intersecting materials.

<u>Example 3</u>. Let $a(u,v) = \int_\Omega [\sum_{i,j} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv] d\Omega$, $\underset{\sim}{x} \in \Omega \subset \mathbb{R}^n$, $[a_{ij}(x)]_{i,j=1}^n$ uniformly positive definite and $c \ge 0$ on $\bar{\Omega}$.

Consider the extended Cauchy inequality (see [5] and [4])
$$|a(u,v)| \le \gamma\{a(u,u)a(v,v)\}^{\frac{1}{2}} \quad \forall u \in U_{2h},\ v \in V_h$$
where $0 < \gamma < 1$ is independent on h and
$$U_{2h} = \text{SPAN}\ \{\lambda_i^{(2h)}\},$$
$$V_h = \text{SPAN}\ \{\phi_i^{(h)}\}\ ,\ u \equiv 1 \notin V_h$$
$$U_{2h} \cap V_h = 0,\ U_{2h} \oplus V_h = W_h$$

and

W_h = {all piecewise quadratic (or cubic ...) polynomials on elements of Ω_h}.

Note that $N(2h) = N(h)/r^n$, where $r=2$ for quadratic, $r=3$ for cubic etc.

Let $A_{2h} = [a(\lambda_j^{(2h)},\lambda_i^{(2h)})]$, $B_{2h} = [a(\phi_j^{(h)},\phi_i^{(h)})]$

$C_h = [a(\lambda_j^{(2h)},\phi_i^{(h)})]$.

Then the finite element matrix corresponding to a(.,.) is $K_h = \begin{bmatrix} A_{2h} & C_h \\ C_h^t & B_h \end{bmatrix}$

We let $K_h^{(1)} = \begin{bmatrix} A_{2h} & 0 \\ 0 & B_h \end{bmatrix}$. Then $K_h^{(1)}$ is spectrally equivalent with $K_h^{(2)} = K_h$ (see [5],[4]) and

$$\mathcal{K}(K_h^{(1)^{-1}} K_h^{(2)}) \le \frac{1+\gamma}{1-\gamma}.$$

For the special case considered in Example 2 one now finds $\mathcal{K} \simeq 9.9$ $(r=2)$. In practice A_{2h} and B_h in $K_h^{(1)}$ is approximated by (modified) incomplete factorization. Note that B_h has a spectral condition number that is independent on h (see [4]). The resulting computational complexity is O(N) if the linear systems with preconditioning matrix $K_h^{(1)}$ are solved exactly and behaves in this way for a wide range of values of N even if we use incomplete factorizations for A_{2h}(and B_h). For details see [4].

References

1. O. Axelsson (1980), Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations, Linear Algebra and its applications 29,1-16.
2. O. Axelsson, I. Gustafsson (1980), A preconditioned conjugate gradient method for finite element equations, which is stable for rounding errors, IFIP Proceedings 80, (E. Lavingtond ed.), pp.723-728.
3. O. Axelsson, I. Gustafsson (1981), Preconditioning and two-level multigrid methods of arbitrary degree of approximation, to appear.
4. O. Axelsson, N. Munksgaard (1979), A class of preconditioned conjugate gradient methods for the solution of a mixed finite element discretization of the biharmonic operator, Int.J.Numer.methods in Engin. 14, 1001-1019.
5. R. Bank, T. Dupont (1980), Analysis of a two-level scheme for solving finite element equations, Report CNA-159, Center for Numerical Analysis, The University of Texas at Austin.
6. D. Braess, The contraction number of a multigrid method for solving the Poisson equation, to appear.
7. I. Gustafsson (1979), Stability and rate of convergence of modified incomplete Cholesky factorization methods, 7902R, Department of Computer Sciences, Chalmers University of Technology, Gothenborg, Sweden.
8. J.A. Meijerink, H.A. van der Vorst (1977), An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, Math. Comp. 31, 148-162.
9. R.S. Varga (1960), Factorization and normalized iterative methods, in Boundary Problems in Differential Equations (R.E. Langer, ed.), University of Wisconsin Press, Madison, pp. 121-142.