View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Journal of Computational and Applied Mathematics 24 (1988) 265–275
North-Holland

Preconditioned conjugate gradients for solving singular systems

E.F. KAASSCHIETER
TNO, DGV Institute of Applied Geoscience, 2600 AG Delft, The Netherlands

Received 8 February 1988
Revised 18 June 1988

Abstract: In this paper the preconditioned conjugate gradient method is used to solve the system of linear equations $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$, where $\boldsymbol{A}$ is a singular symmetric positive semi-definite matrix. The method diverges if $\boldsymbol{b}$ is not exactly in the range $R(\boldsymbol{A})$ of $\boldsymbol{A}$. If the null space $N(\boldsymbol{A})$ of $\boldsymbol{A}$ is explicitly known, then this divergence can be avoided by subtracting from $\boldsymbol{b}$ its orthogonal projection onto $N(\boldsymbol{A})$.

As well as analysing this subtraction, conditions necessary for the existence of a nonsingular incomplete Cholesky decomposition are given. Finally, the theory is applied to the discretized semi-definite Neumann problem.

Keywords: Preconditioned conjugate gradients, symmetric positive semi-definite matrices, incomplete Cholesky decomposition, semi-definite Neumann problem.

1. Introduction

In this paper the system of linear equations

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \tag{1.1}$$

is considered, where $\boldsymbol{A}$ is a symmetric positive semi-definite matrix. Two cases can be distinguished: the case where $\boldsymbol{A}$ is nonsingular and consequently positive definite, and the case where $\boldsymbol{A}$ is singular.

Much is known about the first case (see e.g. [5]). If $\boldsymbol{A}$ is a large and sparse matrix, then iterative methods for the approximate solution of (1.1) are often to be preferred over direct methods, because iterative methods help to reduce both memory requirements and computing time. The conjugate gradient method is a successful iterative method (see [5, section 10.2] and [8]).

The convergence rate of the conjugate gradient method is determined by the spectrum of eigenvalues of the matrix $\boldsymbol{A}$ (see [8]). An acceleration of the convergence rate can often be achieved by replacing the system (1.1) by the preconditioned system

$$\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{M}^{-1}\boldsymbol{b}. \tag{1.2}$$

The symmetric positive definite matrix $\boldsymbol{M}$ must be chosen in such a way that the system $\boldsymbol{M}\boldsymbol{z}=\boldsymbol{r}$ can be solved with less computational work than the original system (1.1) for every vector $\boldsymbol{r}$ on the right-hand side of the equation, and so that the matrix $\boldsymbol{M}^{-1}\boldsymbol{A}$ has a more 'favourable' spectrum of eigenvalues than $\boldsymbol{A}$.

0377-0427/88/$3.50 © 1988, Elsevier Science Publishers B.V. (North-Holland)

Numerical experiments indicate that in many situations the construction of the preconditioning matrix $\boldsymbol{M}$ by a suitable incomplete Cholesky decomposition of $\boldsymbol{A}$ is a good choice (see [6,7]). If $\boldsymbol{A}$ is a symmetric $\boldsymbol{M}$-matrix, then every incomplete Cholesky decomposition exists. However, this condition is not necessary.

If $\boldsymbol{A}$ is singular, then the system (1.1) has a solution if, and only if, $\boldsymbol{b}$ is in the range $R(\boldsymbol{A})$ of $\boldsymbol{A}$. In that case the solution is not unique. Nevertheless, a solution can be determined by the preconditioned conjugate gradient method, because only those eigenvalues and eigenvectors of $\boldsymbol{M}^{-1}\boldsymbol{A}$ that are represented in the right-hand side of (1.2) participate in the conjugate gradient process (see e.g. [8, section 2.2]).

However, the method diverges if $\boldsymbol{b} \notin R(\boldsymbol{A})$, e.g. as a result of perturbation of domain errors. This divergence can usually be avoided by eliminating the singularity of $\boldsymbol{A}$, i.e. by fixing some entries of the solution $\boldsymbol{x}$ (as many as the dimension of the null space $N(\boldsymbol{A}) = R(\boldsymbol{A})^{\perp}$ of $\boldsymbol{A}$), deleting the corresponding rows and columns of $\boldsymbol{A}$, adjusting the right-hand side and solving the resulting system $\hat{\boldsymbol{A}}\hat{\boldsymbol{x}} = \hat{\boldsymbol{b}}$ by the preconditioned conjugate gradient method.

If $N(\boldsymbol{A})$ is explicitly known, then there is another way of avoiding the divergence mentioned above. It is then obvious to subtract from $\boldsymbol{b}$ its orthogonal projection onto $N(\boldsymbol{A})$, thereby yielding the vector $\boldsymbol{b}_R$, and to solve the adjacent $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_R$. In many situations this results in a faster convergence rate than when solving the nonsingular system $\hat{\boldsymbol{A}}\hat{\boldsymbol{x}} = \hat{\boldsymbol{b}}$. This approach is discussed in Section 2.

The construction of an incomplete Cholesky decomposition of $\boldsymbol{A}$ may fail. Conditions for the existence of a nonsingular incomplete Cholesky decomposition of a symmetric positive semi-definite matrix are given in Section 3.

Finally, an important application, the discretized semi-definite Neumann problem, is dealt with in Section 4. The results are illustrated by a numerical experiment.

2. The preconditioned conjugate gradient method

Consider the system of linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{2.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a singular symmetric positive semi-definite matrix and $\boldsymbol{b} \in \mathbb{R}^n$.

The system (2.1) has a solution if, and only if, $\boldsymbol{b} \in R(\boldsymbol{A})$ where $R(\boldsymbol{A}) = \{ \boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} = \boldsymbol{A}\boldsymbol{z}$ for $\boldsymbol{z} \in \mathbb{R}^n\}$ is the range of $\boldsymbol{A}$. If the system (2.1) has a solution, then it is not unique. Indeed, Let $\boldsymbol{x} \in \mathbb{R}^n$ be a solution of (2.1), then $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{y}$ is a solution for every $\boldsymbol{y} \in N(\boldsymbol{A})$, where $N(\boldsymbol{A}) = \{\boldsymbol{z} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}\}$ is the null space of $\boldsymbol{A}$ (note that $N(\boldsymbol{A}) = R(\boldsymbol{A})^{\perp}$).

Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a suitable symmetric positive definite preconditioning matrix; then the corresponding preconditioned conjugate gradient method (cg-method) (see e.g. [5, chapter 10]) generates a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, starting with a vector $\boldsymbol{x}_0 \in \mathbb{R}^n$, according to

Algorithm 1

```
$r_0 := b - Ax_0$
for $i = 0, 1, \ldots$
    $z_i := M^{-1} r_i$
    if $r_i = 0$ then stop
```

$$\begin{aligned}
&\beta_{i-1} := z_i^{\mathrm{T}} r_i / z_{i-1}^{\mathrm{T}} r_{i-1} \ (\beta_{-1} := 0)\\
&p_i := z_i + \beta_{i-1} p_{i-1} \ (p_0 := z_0)\\
&\alpha_i := z_i^{\mathrm{T}} r_i / p_i^{\mathrm{T}} A p_i\\
&x_{i+1} := x_i + \alpha_i p_i\\
&r_{i+1} := r_i - \alpha_i A p_i.
\end{aligned}$$

Since M is symmetric positive definite, there is a nonsingular matrix $C \in \mathbb{R}^{n \times n}$, such that $M = CC^T$. The preconditioned cg-method is equivalent to the ordinary cg-method for solving the preconditioned system

$$\tilde{A}\tilde{x} = \tilde{b}, \tag{2.2}$$

where $\tilde{A} = C^{-1}AC^{-\mathrm{T}}$, $\tilde{x} = C^{\mathrm{T}}x$ and $\tilde{b} = C^{-1}b$ (choose $\tilde{x}_0 = C^{\mathrm{T}}x_0$). For the analysis of the preconditioned cg-method we will occasionally switch between these two viewpoints.

Corresponding to $\tilde{r}_0 = C^{-1}r_0$ there are uniquely determined eigenvalues $0 = \mu_0 < \mu_1 < \ldots < \mu_m$ and normalized eigenvectors $u_0, \ldots, u_m$ of $\tilde{A}$, such that $\tilde{r}_0 = \Sigma_{j=0}^m \xi_j u_j$, where $\xi_0 \geqslant 0$ and $\xi_j > 0$ for $j = 1, \ldots, m$ (see [8, section 2.2]). Note that $\xi_0 = 0$ if, and only if, $\tilde{r}_0 \in R(\tilde{A})$, i.e. $b \in R(A)$. These eigenvalues and eigenvectors are the active ones; in view of [8, section 2.1] the other eigenvalues and eigenvectors do not participate in the conjugate gradient process.

If $b \notin R(A)$, then (2.1) does not have an exact solution. In practice this situation may arise because of perturbation of domain errors (see [2]). Using the preconditioned cg-method we can then still generate a sequence x_1, $x_2, \ldots$. However, from numerical experiments it appears that the Euclidean norm of the residual $\tilde{r}_i$ initially tends to decrease, but at some stage suddenly increases. It seems that the orthogonal projection of the vector $\tilde{x}_i$ onto $R(\tilde{A})$ converges to a certain vector $\tilde{x}$, before it suddenly diverges. Three questions arise:

- In what sense does $x = C^{-\mathrm{T}}\tilde{x}$ represent a solution?
- Can we understand the sudden divergence?
- How can we preclude this divergence?

The last question will be answered in this section; the first two will be discussed in Section 4.

If $b \notin R(A)$, then one often resorts to a least squares solution of (2.1) (which always exists), i.e. a vector x for which $\| b - Ax \|_2$ is minimal (see [5, section 6.1]). Since A is singular, there is an infinite number of least squares solutions. In this whole set of least squares solutions there is a unique vector x whose Euclidean norm is minimal. This is referred to as the minimum norm least squares solution of (2.1). Note that x is a least squares solution of (2.1) if, and only if, x is a solution of the projected system

$$Ax = b_R, \tag{2.3}$$

where b_R is the orthogonal projection of b onto $R(A)$.

If $R(A)$ is explicitly known, then we can prove that a solution x of (2.3) can be determined using the preconditioned cg-method. For the preconditioned starting residual $\tilde{r}_{0,R} = \tilde{b}_R - \tilde{A}\tilde{x}_0$, where $\tilde{b}_R = C^{-1}b_R$, we have $\tilde{r}_{0,R} = \Sigma_{j=1}^m \tilde{\xi}_j u_j$, where $\tilde{\xi}_j > 0$ for $j = 1, \ldots, m$ (note that in general $\tilde{\xi}_j \neq \xi_j$, because of the non-orthogonality of the projection of $\tilde{r}_0$ onto $R(\tilde{A})$, resulting in $\tilde{r}_{0,R}$). From this it follows that the cg-method for solving the system

$$\tilde{A}\tilde{x} = \tilde{b}_R \tag{2.4}$$

generates a sequence $\tilde{x}_1$, $\tilde{x}_2, \ldots$, starting with a vector $\tilde{x}_0 = C^{\mathrm{T}}x_0$. This sequence has the following property

$$\| \tilde{x} - \tilde{x}_i \|_{\tilde{A}} = \min_{y - \tilde{x}_0 \in K_i} \| \tilde{x} - y \|_{\tilde{A}}, \tag{2.5}$$

where $K_i = \text{span}\,\{\tilde{r}_{0,R}, \tilde{A}\tilde{r}_{0,R}, \ldots, \tilde{A}^{i-1}\tilde{r}_{0,R}\}$ is the ith Krylov subspace of $\tilde{A}$ with respect to $\tilde{r}_{0,R}$ ($\| z \|_{\tilde{A}} = (z^{\mathrm{T}}\tilde{A}z)^{1/2}$ for all $z \in \mathbb{R}^n$). For the basic relations of the cg-method, which also hold in this case, see e.g. [5, section 10.2] and [8, section 2]. Since $\| \cdot \|_{\tilde{A}}$ is a norm in $R(\tilde{A})$ and $K_i \subset R(\tilde{A})$ for $i = 1, 2\ldots$, it follows from (2.5) that $\tilde{x}_i$ converges to a solution $\tilde{x}$ of (2.4), and thus $x_i = C^{-\mathrm{T}}\tilde{x}_i$ converges to a solution $x = C^{-\mathrm{T}}\tilde{x}$ of (2.3).

This solution is not necessarily the minimum norm solution of (2.3), i.e. the minimum norm least squares solution of (2.1). With the popular choice $x_0 = \mathbf{0}$ (and thus $\tilde{x}_0 = C^{\mathrm{T}}x_0 = \mathbf{0}$) it follows from (2.5) that $\tilde{x}_i \in R(\tilde{A})$ for $= 0, 1,\ldots$. Therefore $\tilde{x}_i$ converges to the minimum norm solution of (2.4) (note that $\tilde{x}$ is the minimum norm solution of (2.4) if, and only if, $\tilde{x} \in R(\tilde{A})$). An approximation to the minimum norm least squares solution of (2.1) can be determined by subtracting from $x_i = C^{-\mathrm{T}}\tilde{x}_i$ its orthogonal projection onto $N(A)$.

3. Incomplete Cholesky decompositions

A symmetric positive definite preconditioning matrix $M = CC^{\mathrm{T}}$, where C is a lower triangular matrix, may be determined by an incomplete Cholesky decomposition of the symmetric positive semi-definite matrix A (see [6], [7]). The most general form of an incomplete Cholesky decomposition is indicated in [7, section 1], where it is suggested that a Cholesky decomposition of A be made, during which elimination corrections are partly ignored in C in appropriate places. The ignoration factors will be given by a symmetric matrix $\Theta \in \mathbb{R}^{n \times n}$, where $0 \leqslant \theta_{ij} \leqslant 1$ for i, $j = 1,\ldots, n$. In this way we obtain

Algorithm 2

```
for $i = 1,\ldots, n$
    for $j = 1,\ldots, i-1$
        $c_{ij} := \left( a_{ij} - \theta_{ij} \sum_{k=1}^{j-1} c_{ik}c_{jk} \right) / c_{jj}$
    $c_{ii} := \left( a_{ii} - \theta_{ii} \sum_{k=1}^{i-1} c_{ik}^2 \right)^{1/2}$
```

(we define $0/0 = 0$). If Algorithm 2 does not fail, i.e. if $a_{ii} - \theta_{ii}\Sigma_{k=1}^{i-1}c_{ik}^2 \geqslant 0$ for $i = 1,\ldots, n$ and $c_{jj} > 0$ if $a_{ij} - \theta_{ij}\Sigma_{k=1}^{j-1}c_{ik}c_{jk} > 0$, then we will denote by $C = C(A, \Theta)$ the lower triangular matrix C, constructed by the incomplete Cholesky decomposition of A with respect to the ignoration matrix Θ. Thus the matrix C constructed by the complete Cholesky decomposition of A, which exists if A is symmetric positive semi-definite (see [5, chapter 5]), will be denoted by $C = C(A, \mathbf{1})$ where every entry of $\mathbf{1} \in \mathbb{R}^{n \times n}$ is equal to 1. Note that $C(A, \Theta_1) = C(A, \Theta_2)$ might be possible for $\Theta_1 \neq \Theta_2$. Henceforth we will say that $C(A, \Theta)$ exists for a matrix A and an ignoration matrix Θ, when Algorithm 2 is executable.

Note that the executability of Algorithm 2 implies that $c_{ii} \geqslant 0$ for $i = 1,\dots,n$. However, to solve the system of linear equations $\boldsymbol{Mz} = \boldsymbol{r}$ in each iteration of Algorithm 1 according to:

Algorithm 3

```
for $i = 1,\dots,n$
    $z_i := \left(r_i - \sum_{j=1}^{i-1} c_{ij} z_j\right) / c_{ii}$
for $i = n,\dots,1$
    $z_i := \left(r_i - \sum_{j=i+1}^{n} c_{ij} z_j\right) / c_{ii}$,.
```

it is necessary that $c_{ii} > 0$ for $i = 1,\dots,n$.

It has been proved that $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists for every ignoration matrix $\boldsymbol{\Theta}$, if $\boldsymbol{A}$ is a symmetric $\boldsymbol{M}$-matrix (see [6, theorem 2.4; 7, section 1]). In this case, $c_{ii} > 0$ for $i = 1,\dots,n$. $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an $\boldsymbol{M}$-matrix, if $a_{ij} \leqslant 0$ for all $i \neq j$, $\boldsymbol{A}$ is nonsingular and $\boldsymbol{A}^{-1} \geqslant 0$. Note that $\boldsymbol{A}$ is a symmetric $\boldsymbol{M}$-matrix if, and only if, $\boldsymbol{A}$ is a Stieltjes matrix, i.e. $a_{ij} \leqslant 0$ for all $i \neq j$ and $\boldsymbol{A}$ is symmetric positive definite (see [9, p.85]).

Before deriving a necessary condition for the existence of an incomplete Cholesky decomposition of a singular symmetric positive semi-definite matrix $\boldsymbol{A}$, a definition need to be given.

Definition 3.1. A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a singular Stieltjes matrix if $\mathrm{a}_{ij} \leqslant 0$ for all $i \neq j$ and $\boldsymbol{A}$ is singular and symmetric positive semi-definite.

Theorem 3.2. *If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an irreducible singular Stieltjes matrix, then $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists for every ignoration matrix $\boldsymbol{\Theta}$. In this case $c_{ii} > 0$ for $i = 1,\dots,n$ if, and only if, $\boldsymbol{C} \neq \boldsymbol{C}(\boldsymbol{A}, \mathbf{1})$.*

Proof. Let $\boldsymbol{\Theta} \in \mathbb{R}^{n \times n}$ be a certain ignoration matrix and consider the incomplete Cholesky decomposition of an irreducible singular Stieltjes matrix $\boldsymbol{A}$ with respect to $\boldsymbol{\Theta}$. Since the leading principal submatrix that is obtained from $\boldsymbol{A}$ by omitting the last row and column is a nonsingular Stieltjes matrix (see [4, section 5]) the first $n-1$ loops of Algorithm 2 are executable and $c_{ii} > 0$ for $i = 1,\dots,n-1$.

Assume that Algorithm 2 is not executable, i.e. $a_{nn} - \theta_{nn}\Sigma_{k=1}^{n-1} c_{nk}^2 < 0$, and let $\boldsymbol{A}^{(\epsilon)} = \boldsymbol{A} + \epsilon \boldsymbol{e}_n \boldsymbol{e}_n^{\mathrm{T}}$, where $\epsilon > 0$ and $\boldsymbol{e}_n = (0,\dots,0,1)^{\mathrm{T}}$ is the nth unity vector, then $\boldsymbol{A}^{(\epsilon)}$ is a nonsingular Stieltjes matrix (see [4, (5, 11)]) and thus $\boldsymbol{C}(\boldsymbol{A}^{(\epsilon)}, \boldsymbol{\Theta})$ exists. If $\epsilon > 0$ is small enough we have

$$a_{nn}^{(\epsilon)} - \theta_{nn} \sum_{k=1}^{n-1} c_{nk}^2 = a_{nn} + \epsilon - \theta_{nn} \sum_{k=1}^{n-1} c_{nk}^2 < 0.$$

This gives a contradiction, thus Algorithm 2 is executable, i.e. $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta})$ exists.

($\Rightarrow$) Suppose that $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{A}, \mathbf{1})$, then $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ (see [5, chapter 5]) and thus $\Pi_{i=1}^n c_{ii}^2 = (\det \boldsymbol{C})^2 = \det \boldsymbol{A} = 0$, i.e. $c_{nn} = 0$.

($\Leftarrow$) Suppose that $\boldsymbol{C} \neq \boldsymbol{C}(\boldsymbol{A}, \mathbf{1})$.

Assume that

$$\max\left\{ i \mid c_{ii} > \left(a_{ii} - \sum_{k=1}^{i-1} c_{ik}^2 \right)^{1/2} \right\} \geqslant \max\left\{ i \mid c_{ij} > \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right) / c_{jj} \text{ for some } j \right\}. \tag{3.1}$$

Define in this case

$$i_0 = \max\left\{ i \mid c_{ii} > \left(a_{ii} - \sum_{k=1}^{i-1} c_{ik}^2 \right)^{1/2} \right\},$$

i.e. i_0 agrees with the last partial ignoration in Algorithm 2.

If $i_0 = n$, then $c_{nn} > (a_{nn} - \Sigma_{k=1}^{n-1} c_{nk}^2)^{1/2} \geqslant 0$. Thus, assume that $i_0 < n$ and define $\boldsymbol{C}' = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta}')$, where

$$\theta'_{ij} = \begin{cases} 1 & \text{if } i = j = i_0, \\ \theta_{ij} & \text{otherwise.} \end{cases}$$

Since $\boldsymbol{A}$ is irreducible, there is a row of integers $\{i_s\}_{s=0}^{r}$, such that $i_r = n$ and $a_{i_{s-1} i_s} < 0$ for $s = 1, \ldots, r$ (see [9, p.20]). A subrow $\{i'_\sigma\}_{\sigma=0}^{\rho}$ of the row $\{i_s\}_{s=0}^{r}$ exists, such that $i_0 = i'_0 < i'_1 < \cdots < i'_\rho = n$. By complete induction it follows that

$$\begin{aligned} c_{i'_\sigma i'_{\sigma-1}} &= \left(a_{i'_\alpha i'_{\sigma-1}} - \sum_{k=1}^{i'_{\sigma-1}-1} c_{i'_\sigma k} c_{i'_{\sigma-1} k} \right) \Big/ c_{i'_{\sigma-1} i'_{\sigma-1}} \\ &< \left(a_{i'_\sigma i'_{\sigma-1}} - \sum_{k=1}^{i'_{\sigma-1}-1} c'_{i'_\sigma k} c'_{i'_{\sigma-1} k} \right) \Big/ c'_{i'_{\sigma-1} i'_{\sigma-1}} = c'_{i'_\sigma i'_{\sigma-1}} \leqslant 0, \end{aligned}$$

$$c_{i'_\sigma i'_\sigma} = \left(a_{i'_\sigma i'_\sigma} - \sum_{k=1}^{i'_\sigma - 1} c_{i'_\sigma k}^2 \right)^{1/2} > \left(a_{i'_\sigma i'_\sigma} - \sum_{k=1}^{i'_\sigma - 1} (c'_{i'_\sigma k})^2 \right)^{1/2} = c'_{i'_\sigma i'_\sigma} \geqslant 0$$

for $\sigma = 1, \ldots, \rho$.

Thus, in particular $c_{nn} > c'_{nn} \geqslant 0$.

Assume that (3.1) does not hold. Define in this case

$$i_0 = \max\left\{ i \mid c_{ij} > \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right) \Big/ c_{jj} \text{ for some } j \right\},$$

i.e., i_0 agrees with the last loop in Algorithm 2, in which an elimination correction is partly ignored. Define $\boldsymbol{C}' = \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{\Theta}')$, where

$$\theta'_{ij} = \begin{cases} 1 & \text{if } j < i = i_0, \\ \theta_{ij} & \text{otherwise.} \end{cases}$$

Now we have

$$c_{i_0 i_0} = \left(a_{i_0 i_0} - \sum_{k=1}^{i_\sigma - 1} c_{i_0 k}^2 \right)^{1/2} > \left(a_{i_0 i_0} - \sum_{k=1}^{i_\sigma - 1} (c'_{i_0 k})^2 \right)^{1/2} = c'_{i_0 i_0} \geqslant 0.$$

The rest of the proof is analogous. □

If a symmetric matrix $\boldsymbol{A}$ is reducible, then a permutation matrix $\boldsymbol{P}$ exists, such that

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}^T \boldsymbol{A} \boldsymbol{P} = \begin{pmatrix} \tilde{\boldsymbol{A}}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{\boldsymbol{A}}_N \end{pmatrix}, \tag{3.2}$$

where every submatrix $\tilde{A}_i \in \mathbb{R}^{p_i \times p_i}$ is irreducible or equal to the 1×1 null matrix, $0 < p_i < n$ and $\sum_{i=1}^{N} p_i = n$. We say that (3.2) is the normal form of A (see [9, p.46]).

The normal form (3.2) is unique up to permutations in and of submatrices $\tilde{A}_i$. In the folllowing we choose $\boldsymbol{P}$ such that the rows and columns of every submatrix $\tilde{A}_i$ correspond to successive rows and columns of A. The normal form (3.2) is then unique up to permutations of submatrices $\tilde{A}_i$.

Define $\tilde{\boldsymbol{\Theta}} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{P}$. It follows from Algorithm 2, that $\boldsymbol{C} = \boldsymbol{C}(A, \boldsymbol{\Theta})$ exists if, and only if, $\tilde{\boldsymbol{C}} = \boldsymbol{C}(\tilde{A}, \tilde{\boldsymbol{\Theta}})$ exists. In this case we have

$$\tilde{\boldsymbol{C}} = \boldsymbol{P}^T\boldsymbol{C}\boldsymbol{P} = \begin{pmatrix} \tilde{C}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{C}_N \end{pmatrix}, \tag{3.3}$$

where $\tilde{C}_i = \boldsymbol{C}(\tilde{A}_i, \tilde{\boldsymbol{\Theta}}_i)$ and $\tilde{\boldsymbol{\Theta}}_i$ is the principal submatrix of $\tilde{\boldsymbol{\Theta}}$ corresponding to $\tilde{A}_i$.

At this stage we can prove:

Theorem 3.3. *If $A \in \mathbb{R}^{n\times n}$ is a reducible singular Stieltjes matrix, then $\boldsymbol{C} = \boldsymbol{C}(A, \boldsymbol{\Theta})$ exists for every ignoration matrix $\boldsymbol{\Theta}$. Let $\tilde{A} = \boldsymbol{P}^T A \boldsymbol{P}$ be the normal form (3.2) of A, where the rows and columns of every submatrix $\tilde{A}_i$ correspond to successive rows and columns of A. In this case $c_{ii} > 0$ for $i = 1, \ldots, n$ if, and only if, $\tilde{C}_i \neq \boldsymbol{C}(\tilde{A}_i, \boldsymbol{1}_i)$, where $\tilde{C}_i$ and $\boldsymbol{1}_i$ are the principal submatrices of $\tilde{\boldsymbol{C}}$ and $\boldsymbol{1}$ corresponding to $\tilde{A}_i$.*

Proof. Follows directly from Theorem 3.2 and (3.3). □

4. The semi-definite Neumann problem

An important practical example of a system (2.1) is obtained after the discretization of the semi-definite Neumann boundary value problem

$$-\nabla \cdot (A\nabla u) = f \quad \text{in } \Omega, \qquad -n \cdot (A\nabla u) = g \text{ on } \partial\Omega, \tag{4.1}$$

where $\Omega \subset \mathbb{R}^d$ is an open, bounded and connected domain with a piecewise smooth boundary $\partial\Omega$. Further, let $A \in L_\infty(\Omega, \mathbb{R}^{d\times d})$, where $A(\boldsymbol{x})$ is symmetric positive definite for almost every $\boldsymbol{x} \in \Omega$, and $f \in L_2(\Omega)$, $g \in L_2(\partial\Omega)$ satisfying the compatibility condition $\int_\Omega f \,\mathrm{d}\boldsymbol{x} = \int_{\partial\Omega} g \,\mathrm{d}s$ (see [3, section 1.2]).

The discretization of (4.1) by a suitable finite difference or finite element method, leads to a system (2.1), where A is a singular Stieltjes matrix (for details see [1], [9]). If the discretization grid is connected (see [9, p. 20]), then A is irreducible. Note that $N(A) = \mathrm{span}\{e\}$, where $e = (1, \ldots, 1)^{\mathrm{T}}$, because the solution u of (4.1) is unique up to a constant factor. As a result of perturbation of domain errors (Ω is approximated by a polygon $\tilde{\Omega}$) the system (2.1) may not have a solution, i.e. $\boldsymbol{b} \notin R(A)$ (see [2], where $\boldsymbol{b}$ is projected onto $R(A)$ to overcome this problem).

As an illustration we take the Laplace equation on $\Omega = (0,1)^2$ with Neumann boundary conditions:

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad -\partial u / \partial n = g \text{ on } \partial\Omega, \tag{4.2}$$

with g such that $u(\boldsymbol{x}) = x_1 + x_2 - 1$ for $\boldsymbol{x} = (x_1, x_2)^{\mathrm{T}} \in \Omega$ (it then follows that $\int_{\partial\Omega} g \,\mathrm{d}s = 0$). We

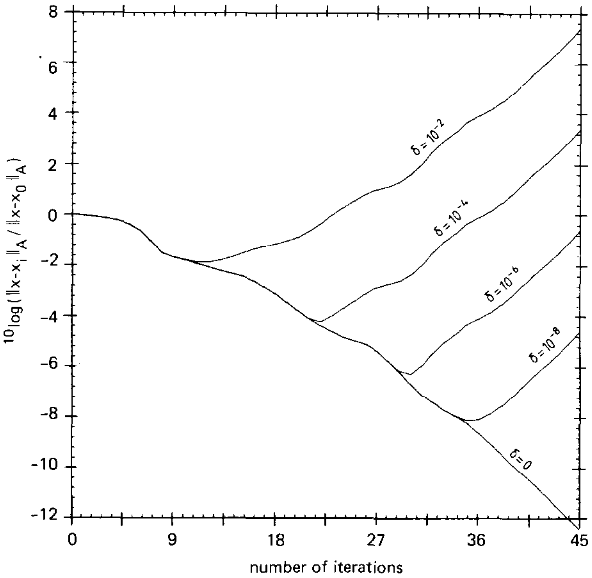

Fig. 1. Experimental results for different perturbations (with $x_{i+1} := x_i + \alpha_i p_i$).

choose a five-point finite difference discretization with step sizes of 1/29. This results in an irreducible singular Stieltjes matrix $A \in \mathbb{R}^{900\times 900}$ (see [9, section 6.3]). The resulting system (2.1) is solved by the preconditioned cg-method with the preconditioning matrix $M = CC^{\mathrm{T}}$, where the lower triangular matrix C is constructed by the incomplete Cholesky decomposition of A with respect to the ignoration matrix $\Theta \in \mathbb{R}^{n\times n}$, where

$$\theta_{ij} = \begin{cases} 1 & \text{if } a_{ij} \neq 0, \\ 0 & \text{if } a_{ij} = 0, \end{cases} \tag{4.3}$$

(see Section 3). This is the so-called ICCG(1,1) preconditioning (see [7, section 2.1.2]). From Theorem 3.2 it follows that $C = C(A, \Theta)$ exists and $c_{ii} > 0$ for $i = 1,\ldots, n$. In Algorithm 2 we choose the starting vector $x_0 = \mathbf{0}$.

To simulate a perturbation of the right-hand side we choose the vector $b_R = Ax$ as an unperturbed right-hand side, where $x \in R(A)$ corresponds to the solution of (4.2) ($b_R \in R(A)$). Next to this system we consider the perturbed systems $Ax = b$, where $b = b_R + \gamma e$ and $\gamma = \| b_R \|_2 \delta / \sqrt{n(1-\delta^2)}$ for $0 < \delta < 1$. Note that $b_R = b - (b^{\mathrm{T}}e/n)e$ is the orthogonal projection of b onto $R(A)$ (see Section 2). A good measure for the perturbation of a system $Ax = b$ is the angle θ between b and $R(A)$. We find

$$\sin\theta = \| b - b_R \|_2 / \| b \|_2 = \gamma\sqrt{n} / \| b \|_2 = \delta. \tag{4.4}$$

The preconditioned cg-method for solving the unperturbed system $Ax = b_R$, i.e. $\delta = 0$, converges monotonically (see Fig. 1). The preconditioned cg-method for solving a perturbed system $Ax = b$, i.e. $0 < \delta < 1$, initially seems to converge monotonically to the minimum norm solution of the unperturbed system $Ax = b_R$, but then suddenly starts to diverge (see Fig. 1 for $\delta = 10^{-2}$, 10^{-4}, 10^{-6}, 10^{-8}). The smaller $\delta > 0$, the longer it takes before the preconditioned cg-method starts to diverge. Two questions remain:

- In what sense does the preconditioned cg-method for solving a perturbed system $\boldsymbol{Ax}=\boldsymbol{b}$ initially converge to a solution of the unperturbed system $\boldsymbol{Ax}=\boldsymbol{b}_R$?
- Can we understand the sudden divergence of the preconditioned cg-method for solving the perturbed system $\boldsymbol{Ax}=\boldsymbol{b}$?

In order to answer the first question, note that the results in Fig. 1 are not influenced by the component of $\tilde{\boldsymbol{x}}_i$ orthogonal to $R(\tilde{\boldsymbol{A}})$. Thus the preconditioned cg-method for solving a perturbed system $\boldsymbol{Ax}=\boldsymbol{b}$ and the unperturbed system $\boldsymbol{Ax}=\boldsymbol{b}_R$ would generate the same results, if the constants α_i and β_i were equal in both cases. However, since $\boldsymbol{r}_i \neq \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_i)$ in the perturbed case, these constants are not equal in both cases. If $\delta > 0$ is small, then initially roughly the same constants α_i and β_i are computed in both cases and thus roughly the same results are generated, i.e. the orthogonal projections onto $R(\tilde{\boldsymbol{A}})$ of the approximations $\tilde{\boldsymbol{x}}_i$ generated by the cg-method for solving a perturbed system $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}=\tilde{\boldsymbol{b}}$ converges initially to a solution of the unperturbed system $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}=\tilde{\boldsymbol{b}}_R$.

In order to answer the second question note that the constants α_i in Algorithm 2 are chosen in accordance with the property

$$\|\boldsymbol{x}-\boldsymbol{x}_{i+1}\|_{\boldsymbol{A}} = \|\boldsymbol{x}-\boldsymbol{x}_i-\alpha_i\boldsymbol{p}_i\|_{\boldsymbol{A}} = \min_{\alpha\in\mathbb{R}} \|\boldsymbol{x}-\boldsymbol{x}_i-\alpha\boldsymbol{p}_i\|_{\boldsymbol{A}}, \tag{4.5}$$

at least if $\boldsymbol{b}\in R(\boldsymbol{A})$ (see [5, section 10.3]). If $\boldsymbol{b}\notin R(\boldsymbol{A})$, then (4.5) is not true and can be replaced by

$$\|\boldsymbol{x}-\boldsymbol{x}_i-\hat{\alpha}_i\boldsymbol{p}_i\|_{\boldsymbol{A}} = \min_{\alpha\in\mathbb{R}} \|\boldsymbol{x}-\boldsymbol{x}_i-\alpha\boldsymbol{p}_i\|_{\boldsymbol{A}}, \tag{4.6}$$

where $\hat{\alpha}_i = \boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_i)/\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{p}_i$. Since $\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{r}_i=\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{r}_i \neq \boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_i)$ in the perturbed case, we have $\alpha_i \neq \hat{\alpha}_i$. If $\delta > 0$ is small, then initially $0 < \alpha_i < 2\hat{\alpha}_i$ and thus it follows from (4.6) that $\|\boldsymbol{x}-\boldsymbol{x}_{i+1}\|_{\boldsymbol{A}} < \|\boldsymbol{x}-\boldsymbol{x}_i\|_{\boldsymbol{A}}$, i.e. the sequence $\|\boldsymbol{x}-\boldsymbol{x}_1\|_{\boldsymbol{A}}, \|\boldsymbol{x}-\boldsymbol{x}_2\|_{\boldsymbol{A}},\ldots$ converges, though not optimally. If the cg-process is perturbed too much, then $\alpha_i < 0$ or $\alpha_i > 2\hat{\alpha}_i$ and $\|\boldsymbol{x}-\boldsymbol{x}_i\|_{\boldsymbol{A}}$ starts diverging.

If the computation of $\boldsymbol{x}_{i+1}$ in Algorithm 2 is replaced by $\boldsymbol{x}_{i+1} := \boldsymbol{x}_i + \hat{\alpha}_i\boldsymbol{p}_i$ (note that $\tilde{\boldsymbol{A}}(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}_i)$ is the orthogonal projection of $\tilde{\boldsymbol{r}}_i$ onto $R(\tilde{\boldsymbol{A}})$, thus $\hat{\alpha}_i$ can be computed without knowing the solution $\boldsymbol{x}$ of (2.3)), then the sequence $\|\boldsymbol{x}-\boldsymbol{x}_1\|_{\boldsymbol{A}}, \|\boldsymbol{x}-\boldsymbol{x}_2\|_{\boldsymbol{A}},\ldots$ converges (see Fig. 2). The sudden divergence in the perturbed case is replaced by a stagnation of $\|\boldsymbol{x}-\boldsymbol{x}_i\|_{\boldsymbol{A}}$. This stagnation can be explained by realizing that (2.5) is not true, if $\boldsymbol{b}\notin R(\boldsymbol{A})$. This is not

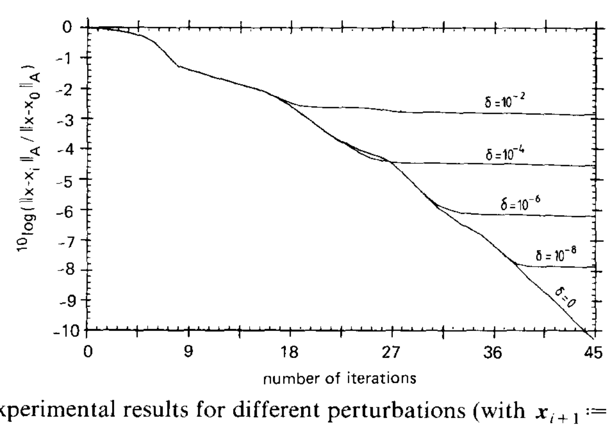

Fig. 2. Experimental results for different perturbations (with $\boldsymbol{x}_{i+1} := \boldsymbol{x}_i + \hat{\alpha}_i\boldsymbol{p}_i$).

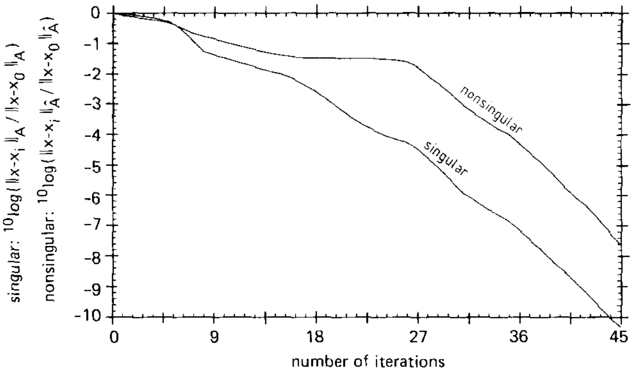

Fig. 3. Experimental results for the singular and nonsingular case.

caused by a loss of orthogonality (because from Algorithm 1 it follows that $z_i^T r_j = 0$ and $p_i^T A p_j = 0$ if $i \neq j$) but is the result of $r_i \neq A(x - x_i)$.

To get rid of the stagnation of $\| x - x_i \|_A$ it suffices to project b on $R(A)$, resulting in the vector $b_R = b - (b^T e/n)e$, and to solve the adjacent system $Ax = b_R$, resulting in a least squares solution of the perturbed system $Ax = b$ (see Section 2). Note that the convergence of the preconditioned cg-method for solving the projected system can be disturbed by rounding errors, if the matrix A is ill conditioned. In this case it may be advisable to project $\tilde{x}_i$ and $\tilde{r}_i$ on $R(\tilde{A})$ repeatedly, which is not a very expenasive process by itself.

In conclusion, note that the classic approach for eliminating the singularity of the matrix A is to fix an entry in the solution x, to delete the corresponding row and column of A, to adjust the right-hand side and to solve the resulting system $\hat{A}\hat{x} = \hat{b}$. Though the matrix $\hat{A}$ is nonsingular, the convergence rate of the precondition cg-method appears to be slower than in the nonsingular case (see Fig. 3 for the results of the experiment, where $x(900)$, which corresponds to the value $u(1,1)$ of the solution u of (4.2), is fixed). This experiment motivated the use of the preconditioned cg-method for the original singular system itself, as is described in this paper.

Acknowledgement

I wish to thank Henk van der Vorst for his criticisms and inspiring ideas.

References

[1] O. Axelsson and V.A. Barker, *Finite Element Solution of Boundary Value Problems* (Academic Press, New York, 1984).

[2] J.W. Barrett and C.M. Elliott, A practical finite element approximation of a semi-definite Neumann problem on a curved domain, *Numer. Math.* **51** (1987) 23–36.

[3] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam, 1978).

[4] M. Fiedler and V. Pták, On matrices with non-positive off-diagonal elements and positive principal minors, *Czechoslovakian Math. J.* **12** (1962) 382–400.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations* (North Oxford Academic, Oxford, 1983).

[6] 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. Comput.* **31** (1977) 148–162.

[7] J.A. Meijerink and H.A. Van der Vorst, Guidelines for the usage of incomplete decompositions in solving sets of linear systems as they occur in practical problems, *J. of Comput. Phys.* **44** (1981) 134–155.

[8] A. Van der Sluis and H.A. Van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.* **48** (1986) 543–560.

[9] R.S. Varga, *Matrix Iterative Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1962).