Rev. Téc. Ing. Univ. Zulia. Vol. 20, Nº 3, 195-201, 1998

Numerical bounds for the solution of linear systems with symmetric positive definite matrices

H.-J. Dobner

University of Karlsruhe, D-76128 Karlsruhe, Germany
e-mail: dobner@ma2mar1.mathematik.uni-karlsruhe.de

Abstract

We consider linear systems with symmetric positive definite matrices. For those problems we present a new method for estimating a posteriori the error produced by a numerical solution. The method is especially intended for large sparse systems. This approach has been applied successfully to linear systems arising in process simulation with the conjugate gradient method as underlying scheme.

Key words: Linear systems, symmetric matrices, large sparse systems, conjugate gradient for estimating the error.

Límites numéricos para la solución de sistemas lineales con matrices simétricas definidas positivas

Resumen

Consideramos sistemas lineales con matrices simétricas definidas positivas. Para esos problemas presentamos un nuevo método para estimar a posteriori el error producido por una solución numérica. El método está dirigido especialmente a sistemas *sparse* grandes. El método propuesto ha sido aplicado exitosamente a sistemas lineales que aparecen en procesos de simulación con el método del gradiente conjugado como esquema fundamental.

Palabras clave: Sistemas lineales, matrices simétricas, sistemas *sparse* grandes, gradiente conjugado, estimación del error.

1. Introduction

We consider linear systems

$$Ax = b\,, \qquad (1.1)$$

with $x,b \in \Re^n$ and nonsingular $A \in \Re^{nxn}$. Throughout this paper we assume that A is symmetric and positive definite (SPD). Furthermore it is supposed that A is a sparse matrix of large dimension n without a band structure. If we know an approximate solution $\tilde{x}$ for the unknown x, then we get an a posteriori error bound by

$$\|x-\tilde{x}\| = \|A^{-1}(A\tilde{x}-b)\|\,. \qquad (1.2)$$

The utilization of (1.2) requires both the knowledge of the defect and of the inverse A^{-1}. While the precise computation of $A\tilde{x}-b$ is possible with an appropriate computer arithmetic (cf. Kulisch/Miranker[1]), the computation of a reliable inverse A^{-1} poses some difficult questions and is therefore often ignored in practice.

Numerical methods providing additionally, to a computed solution, safe error bounds are called validating schemes. There are only a few papers treating this aspect and most of them make use of a decomposition of A (see Rump [2], Cordes/Kaucher[3]).

Some schemes (see Kaucher/Rump [4]) for validating however require the explicit knowledge of an approximate inverse for A^{-1} and are therefore not practicable for large sparse matrices. Other papers for condition estimation (cf. Arioli et al [5], Guang-yao [6]) are yielding estimations rather than guaranteed bounds.

The paper is organized as follows. In the next section we present the basic theoretical results. These results are used in section 3 to derive a new method for determining bounds on $\|A^{-1}\|$ within a computational framework. In the last section we discuss aspects of realization and report about numerical tests.

2. Obtaining bounds on $\|A^{-1}\|$

Our goal is to compute a good bound for $\|x-\tilde{x}\|$ with a small amount of work and storage. For this purpose we define with a parameter $w > 0$ the Richardson operator

$$R_w := I - wA, \tag{2.1}$$

where I denotes the identity matrix.

Theorem 2.1

Let A be a SPD matrix and w be chosen such that

$$0 < w < \frac{2}{\|A\|} \tag{2.2}$$

then

$$\|R_w\| < 1 \tag{2.3}$$

Proof

Since A is a SPD matrix, there exists positive constants m,M with

$$m(x,x) \leq (Ax,x) \leq M(x,x), \tag{2.4}$$

where $(x,x) = x^T x$ is chosen as inner product and

$$\mathrm{M} = \|A\|$$

as corresponding matrix norm. Consider

$$m(R_w) = \inf_{\|x\|=1} (R_w x, x)$$

and

$$M(R_w) = \sup_{\|x\|=1} (R_w x, x),$$

together with (2.4) we deduce

$$m(R_w) = 1 - wM,$$

resp.

$$M(R_w) = 1 - wm.$$

From

$$\|R_w\| = \max\{|m(R_w)|, |M(R_w)|\}$$

it is easy to show that (2.3) is satisfied, if (2.2) holds.

The reverse of Theorem 2.1 is also true:

If (2.3) holds, then A is a SPD matrix and (2.2) is satisfied.

Since A is nonsingular the unique solutions of

$$x = R_w x + wb, \quad 0 < w < \frac{2}{\|A\|}, \tag{2.5}$$

and of (1.1) coincide, furthermore we derive from Theorem2.1 that the sequence

$$x^{(k+1)} = R_w x^{(k)} + wb, \quad k = 0, 1, \ldots, \tag{2.6}$$

converges to the solution x.

We note that (2.5) is an operator equation of the second kind, so its solution x is given, due to (2.3), (2.6) by the Neumann series

$$x = \sum_{\upsilon=0}^{\infty} R_w^{\upsilon} wb = w \sum_{\upsilon=0}^{\infty} R_w^{\upsilon} b.$$

This implies

$$A^{-1} = w \sum_{\upsilon=0}^{\infty} R_w^{\upsilon},$$

so that, by (2.3) and (2.5)

$$\|A^{-1}\| = \left\|w \sum_{\upsilon=0}^{r} R_w^\upsilon\right\| \le \frac{w}{1-\|R_w\|}. \qquad (2.7)$$

Combining (2.5) — (2.7) yields a result which serves as basis for an automatic error control:

Let $z \in \Re^n$ be arbitrary. Let $B_r(z)$ denote the ball of radius

$$r = \frac{w}{1-\|R_w\|}\|b - Az\| \qquad (2.8)$$

centered around z, then the solution x of (1.1) is contained in $B_r(z)$, that is

$$x \in B_r(z) \qquad (2.9)$$

Proof

Applying (2.7) to (1.2) gives

$$\|x - z\| \le \|A^{-1}(b - Az)\| \le \|A^{-1}\|\,\|b - Az\| \qquad (2.10)$$

which in turn implies (2.9).

Let denote $\lambda_{max}(A)$ and $\lambda_{min}(A)$ the largest and smallest eigenvalue of A. Choose the parameter w according to

$$w_{opt} = \frac{2}{\lambda_{max}(A) + \lambda_{min}(A)}, \qquad (2.11)$$

then the spectral radius of R_w is given by

$$\rho(R_{w_{opt}}) = \frac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)} < 1. \qquad (2.12)$$

Unfortunately the extreme eigenvalues of a matrix A are not known in general, therefore the results of this section must be formulated in a way, so that they are applicable for computational analysis.

3. An estimation theorem

In this section we demonstrate how to apply the theoretical results of the preceeding section in practice. It turned out, that intervals are an appropriate tool to describe numerical intolerances. The set of closed real intervals

$$[a] := [\underline{a}, \overline{a}] = \{x \in \Re \mid \underline{a} \le x \le \overline{a}\}$$

is denoted by $J(\Re)$. Analogously $J(\Re^n)$ and $J(\Re^{nxn})$ are defined. Intervals are written in square brackets. Set theoretic relations, such as =, ⊆, ∪ are explained as usual,for vectors and matrices componentwise.

For $[a] \in J(\Re)$ midpoint mid ([a]), diameter diam ([a]) and absolute value |[a]| are defined according to

$$\text{mid }([a]) = \frac{1}{2}(\underline{a} + \overline{a}),$$

$$\text{diam }([a]) = \overline{a} - \underline{a},$$

$$|[a]| = \max\{\,|\underline{a}|, |\overline{a}|\,\},$$

for interval vectors and matrices we write $\|.\|$ and again these definitions apply componentwise. For intervals the basic arithmetic operations $* \in \{+, -, \cdot, /\}$ are explained according to

$$[a] * [b] = [\min\{\underline{a} * \underline{b}, \underline{a} * \overline{b}, \overline{a} * \underline{b}, \overline{a} * \overline{b}\},$$
$$\max\{\underline{a} * \underline{b}, \underline{a} * \overline{b}, \overline{a} * \underline{b}, \overline{a} * \overline{b}\}\,],$$

with $0 \notin [b]$ in case of division.

With ∇, Δ we distinguish the downwardly and upwardly directed roundings. Of great importance is the scalar product of two vectors, e.g. for computing R_w in (2.1). This operation must be done with maximum accuracy, so that no floating point number lies between the exact and the rounded result of such an optimal dot product (cf. Kulisch/Miranker [1]).

For a given floating point quantity y we assign a corresponding interval $[y]$ quantity by

$$[y] = [\nabla y, \Delta y], \qquad (3.1)$$

so all imprecisions of a numerical process can be controlled. Using the notations from interval analysis the results of the previous section are summarized in

Theorem 3.1

Let $\tilde{x}$ denote an approximate solution of (1.1). Then the following error bound is guaranteed

$$\|x\| \leq \|[\tilde{x}]\| + \frac{\|[w]\|}{1-\|[R_w]\|}\|[A][\tilde{x}]-[b]\|. \qquad (3.2)$$

provided $\|[R_w]\| < 1$.

Proof

Combine Theorem 2.1, along with (2.5) — (2.7).

Let

$$C = \max_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad (3.3)$$

and α be a real number with $0 < \alpha < 1$. If the parameter w in (2.1) is chosen, so that

$$0 \frac{C}{\alpha} < \frac{1}{w} - a_{ii}, \quad i = 1, \ldots, n, \qquad (3.4)$$

holds, then R_w is strictly diagonal dominant. To see this, consider

$$\max_{i=1}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} \frac{|wa_{ik}|}{|1 - wa_{ii}|} \leq \frac{wC}{|1 - wa_{ii}|}$$

then (3.4) implies

$$0 < \frac{wC}{1 - wa_{ii}} < \frac{\alpha(1 - wa_{ii})}{1 - wa_{ii}} = \alpha .$$

Since A is selfadjoint, the matrix $R_w = I - wA$ is also selfadjoint, thus

$$\rho(R_w) = \|R_w\|,$$

where $\rho(R_w)$ denotes the spectral radius of R_w.

If A is additionally strictly diagonal dominant, we derive with Gerschgorin's circle Theorem the computable bound

$$\rho(R_w) \leq \max_{1 \leq i \leq n}\left\{1 - wa_{ii} + w\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\right\}, \qquad (3.5)$$

in this case (2.7) is immediate.

Once a floating point approximation $\tilde{x}$ has been computed, the next step is to set up the error bound (3.2). To this aim we realize the estimation by a computational process:

If $[z] \in J(\Re^n)$ is an interval vector with the property

$$wb + R_w[z] \subseteq [z] \qquad (3.6)$$

then the solution x of (1.1) lies within $[z]$.

Proof

The mapping on the left hand side of (3.6) is continuous, therefore the enclosure in (3.6) furnishes together with Brouwer's fixed point Theorem the assertion.

In practice, the set $[z]$ is determined iteratively, according to

$$\left[x^{(k+1)}\right] = [wb] + [R_w]\left[x^{(k)}\right], \quad k = 0, 1, \ldots, \qquad (3.7)$$

when starting with $[x^{(0)}]=[\tilde{x}]$.

As direct consequence of enclosure (3.6) we have:

If two iterates of (3.7) fulfill

$$\left[x^{(k+1)}\right] \subseteq \left[x^{(k)}\right], \qquad (3.8)$$

then the estimation

$$\|\tilde{x} - x\| = \|A^{-1}(A\tilde{x} - b)\| \leq \mathrm{diam}([x^{(k+1)}]) \qquad (3.9)$$

has been established.

Since by construction $\|R_w\| < 1$ is attained, the difficulty for explicitely bounding $\|A^{-1}\|$ resp. $\|R_w\|$ is transfered to the examination of the enclosure requirement (3.8), this condition however can be checked easily by computational means, such a computer assisted a posteriori error analysis for large sparse systems seems not available.

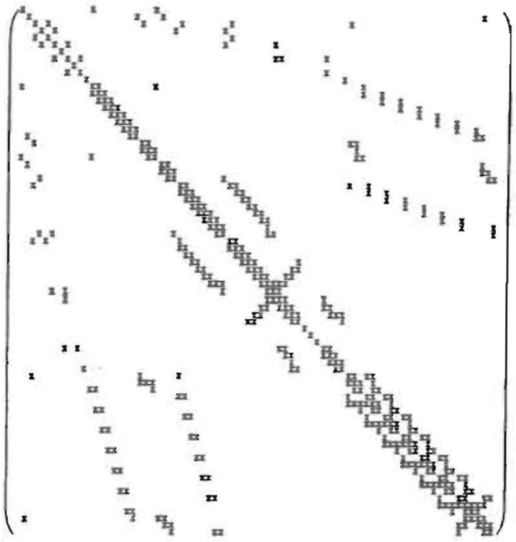

Figure 1. Matrix type 1.

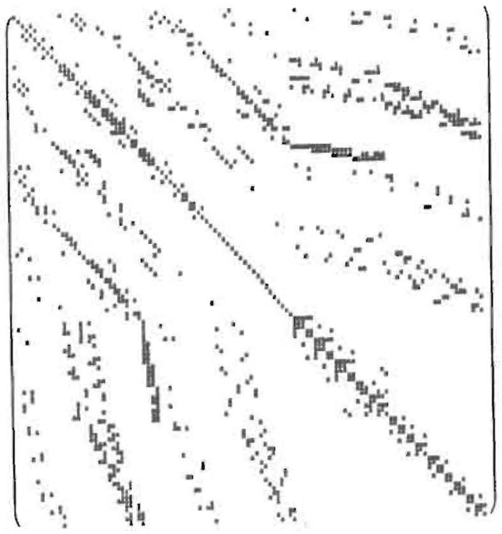

Figure 2. Matrix type 2.

4. Realization and numerical results

Here we report about numerical tests and discuss aspects of implementation. The examples treated are taken from process simulation (cf. Klettenheimer [7]). The matrices of the linear systems are sparse, symmetric and positive definite. There are considered two types of matrices, where the distribution of the non-nullelements is displayed in Figures 1 and 2.

The necessary steps are described below.

1. Compute a solution $\tilde{x}$ with the method of conjugate gradients ($\Re$)
2. Choose w and α, which are related by (3.9) and built up R_w ($\Re$).
3. Correction of the relaxation parameter w ($\Re$)
4. Validation step (I $\Re$)

We continue giving detailed comments on these steps.

Step 1

The matrices are stored with compressed row storage techniques. First a floating point approximation $\tilde{x}$ for (1.1) is computed by the following preconditioned *CG (PCG)* scheme, with M as preconditioner:

$x^{(0)}$ initial guess

$$r^{(0)} = b - Ax^{(0)} ,$$

$$i = 1, 2, \ldots$$

$$Mz^{(i-1)} = r^{(i-1)} ,$$

$$\beta_{i-1} = \begin{cases} 0 & , i = 1 \\ \dfrac{(r^{(i-1)}, z^{(i-1)})}{(r^{(i-2)}, z^{(i-2)})} & , i \neq 1 \end{cases}$$

$$p(i) = z^{(i-1)} + \beta_{i-1} p^{(i-1)} ,$$

$$\alpha_i = \frac{(r^{(i-1)}, z^{(i-1)})}{(p^{(i)}, Ap^{(i)})} ,$$

$$x^{(i)} = x^{(i-1)} + \alpha_i p^{(i)} ,$$

$$r^{(i)} = r^{(i-1)} - \alpha A p^{(i)} .$$

If M is positive definite and $x^{(i)}$ not the true solution, we derive $\alpha_i > 0$ and $\beta_i > 0$. After some manipulations we arrive at the recurrence relation

$$AM^{-1} r^{(i)} = -\frac{\beta_i}{\alpha_i} r^{(i-1)} + \left(\frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}} \right) r^{(i)} - \frac{1}{\alpha_{i+1}} r^{(i+1)}, \tag{4.1}$$

$$i = 1, \ldots, n-1,$$

in matrix form shortly written as

$$(AM^{-1})R = RB, \tag{4.2}$$

where

$$B = \begin{pmatrix} \frac{1}{\alpha_1} & -\frac{\beta_1}{\alpha_1} & & & \\ -\frac{1}{\alpha_1} & \frac{\beta_1}{\alpha_1}+\frac{1}{\alpha_2} & -\frac{\beta_2}{\alpha_2} & 0 & \\ & -\frac{1}{\alpha_2} & \frac{\beta_2}{\alpha_2}+\frac{1}{\alpha_3} & \ddots & \\ & & \ddots & \ddots & \\ & & -\frac{1}{\alpha_{n-1}} & \frac{\beta_{n-1}}{\alpha_{n-1}}+\frac{1}{\alpha_n} & \end{pmatrix} \tag{4.3}$$

and the $(i+1)$-th column of the matrix R is the residual vector $r^{(i)}$. Since AM^{-1} and B are similar matrices, they have identical eigenvalues. In general the iteration terminates after $k << n$ iterations, that is $\left\| r^{(k)} \right\|$ is very small. Then we treat $r^{(k)}$ as zero vector; define a submatrix B_k of B by using k instead of n and consider instead of R a $n \times k$-matrix R_k with columns $r^{(0)}, r^{(1)}, \ldots, r^{(k-1)}$, and obtain

$$(AM^{-1})R_k = R_k B_k \,. \tag{4.4}$$

In our computations M was taken as the identity matrix.

Step 2

This is a straightforward realization of inequality (3.4).

Step 3

The extreme eigenvalues of B_k are taken as approximation for the extreme eigenvalues of AM^{-1}. The computation of the eigenvalues of B_k with the QR method can be performed as a by product in the PCG scheme with additional computational and storage costs of order $O(c)$ only. Furthermore AM^{-1} must not be given explicitely. Now the parameter w is chosen due to (2.11).

Step 4

In a final validation step, the guaranteed error estimation (3.9) is established in an interval analytical framework.

Numerical results reported here, were obtained by a PASCAL-XSC code on an usual PC 486 DX-2/66. In the Table 1 we show some numerical results.

For a nonsingular and non-SPD matrix A we solve instead of (1.1) the equivalent linear system

$$A^T Ax = A^T b, \tag{4.6}$$

where A^T indicates the transposed matrix of A. Since (4.6) is now a system with a SPD matrix, the method discussed here is applicable to a wide class of problems.

Nomenclature

n dimension of the linear system.

NNE number of non-nullelements, compared to n^2 expressed as a percentage.

$iter$ number of iterations needed to compute a good approximation $\tilde{x}$, that is

Table 1

Type of matrix	1	1	1	2	2
N	28	1054	3102	113	257
NNE	17.4	0.62	0.42	8.75	3.85
iter	25	75	410	40	37
ε	1.0E-10	1.0E-10	1.0E-10	1.0E-10	1.0E-10
w	0.999	1.020	0.643	0.567	0.568
$\rho(R_w)$	0.979	0.979	0.996	0.992	0.985
diam (x)	2.564E-10	1.678E-3	8.616E-9	4.175E-8	2.835E-7

for an iterate of the CG-method the defect is sufficiently small

$$\|defect\| \leq \varepsilon \tag{4.5}$$

For simplicity as starting vector always the nullvector has been chosen.

ε stopping rule in (4.5).

w according to (2.11).

$\rho(R_w)$ estimation, according to (2.12).

Diam (x) diameter of the interval vector, enclosing the solution x, according to (3.9).

References

1. U. Kulisch und W.L. Miranker: Computer Arithmetic in Theory and Practice. Academic Press, New York (1981).
2. S.M. Rump: Validated Solution of Large Linear Systems. Computing, Suppl. 9, 191--212 (1993).
3. D. Cordes, E. Kaucher: Self-Validating Computation for Sparse Matrix Problems. Computerarithmetic: Scientific Computation and Programming Languages, Teubner, Stuttgart, 133-149 (1987).
4. E. Kaucher, S.M. Rump: E-Methods for Fixed Point Equations $f(x)=x$. Computing 28, 31-42 (1982).
5. M. Arioli, J.H. Demmel, I.S. Duff: Solving Sparse linear Systems with Backward Error. Siam J. Matrix Anal. Appl. 10(2), 165-190 (1989).
6. L. Guang-yao: On the Approximate Solution of Extreme Eigenvalues and the Condition Number of Nonsingular Matrices. Applied Mathematics and Mechanics 13(2), 199-204 (1992).
7. J. Klettenheimer: Verifizierte Einschlieβung der Lösung groβer schwach besetzter linearer Gleichungssysteme. Diploma Thesis, Karlsruhe (1995).

Recibido el 25 de Febrero de 1997
En forma revisada el 24 de Noviembre de 1997