DEMONSTRATIO MATHEMATICA
Vol. XXVIII No 2 1995

Alicja Smoktunowicz

A NOTE ON THE STRONG COMPONENTWISE STABILITY OF ALGORITHMS FOR SOLVING SYMMETRIC LINEAR SYSTEMS

1. Introduction

This work was intended as an attempt to find analogues of the results obtained by Bunch, Demmel and Van Loan (see [6]). Instead of normwise approach we use a componentwise way of measuring the size of the perturbations in data.

An algorithm for solving linear equations $Ax = b$ is said to be **numerically stable** if it gives a computed solution $\widetilde{x}$ satisfying a relation $(A + E)\widetilde{x} = b$ with $\|E\|$ of order $\epsilon\|E\|$, where ϵ is the relative computer precision. If all $|e_{i,j}|$ are of order $\epsilon|a_{i,j}|$, then an algorithm is numerically stable in a componentwise sense.

Note that the componentwise stability property may be achieved by the use of iterative refinement techniques performed only in single precision. For more details we refer the reader to [1], [3], [4], [12] and [14].

We will say that an algorithm for solving linear equations is **strongly stable** for a class of matrices $\mathcal{A}$ if for each $A \in \mathcal{A}$, the computed solution $\widetilde{x}$ to $Ax = b$ satisfies $\widetilde{A}\widetilde{x} = b$, where $\widetilde{A} \in \mathcal{A}$ and $\widetilde{A}$ is close to A (see [5]).

Bunch, Demmel and Van Loan (see [6]) show that if A is symmetric and $(A+E)\widetilde{x} = b, \widetilde{x} \neq 0$, then there exists $F = F^T$ such that $(A+F)\widetilde{x} = b$, where $\|F\|_2 \leq \|E\|_2$ and $\|E\|_F \leq \sqrt{2}\|E\|_F$. In other words, any stable algorithm on the class of nonsingular symmetric matrices is also strongly stable on the same matrix class. D. J. Higham and N. J. Higham (see [11]) prove that no such result holds when the perturbations E are measured indyvidually, i.e. $|E| \leq \epsilon|A|$. See [11] where a detailed discussion of these concepts may be found. Note that a matrix $|A|$ is the matrix whose elements are $|a_{i,j}|$ and we write $|A| \leq |B|$ to mean that $|a_{i,j}| \leq |b_{i,j}|$ for all i, j.

Key words: Stability, symmetric matrices, perturbation analysis

In some numerical applications ([1], [6] and [11]) it is important that the perturbed matrix $A+E$ has the same structure as A. When the perturbations E are measured indyvidually, then zeros in A force zeros in the corresponding entries of the perturbed matrix $A+E$. Symmetric positive definite systems or diagonally dominant matrices are the most frequently occuring classes of structured linear systems. Bunch, Demmel and Van Loan (see [6]) prove that any stable algorithm on the class of symmetric positive definite matrices or on the class of symmetric diagonally dominant matrices is also strongly stable on the same matrix class under suitable assumptions. The goal of this paper is to obtain similar results using the componentwise approach.

Section 2 deals with the case of diagonally dominant matrices. In section 3 we prove that stability in a componentwise sense implies the strong componentwise stability on the class symmetric positive definite matrices.

2. Diagonally dominant matrices

A matrix $A \in \mathbb{R}^{n\times n}$ is diagonally dominant if

$$|a_{i,i}| \geq \sum_{j\neq i} |a_{i,j}| \quad \text{for } i = 1,\ldots,n. \tag{1}$$

THEOREM 2.1. *Assume that $A \in \mathbb{R}^{n\times n}$ is symmetric and diagonally dominant. If $(A+E)\widetilde{x} = b$, where $|E| \leq \epsilon|A|$ and $\widetilde{x} \neq 0$, then there exists a matrix $F = F^T \in \mathbb{R}^{n\times n}$ such that $(A+F)\widetilde{x} = b$ and $|F| \leq 3\epsilon|A|$.*

Proof. It is sufficient to show that there exists a symmetric matrix F such that $E\widetilde{x} = F\widetilde{x}$ and $|F| \leq 3\epsilon|A|$. Consider two cases.

Case (i): Assume that

$$|\widetilde{x}_1| \leq |\widetilde{x}_2| \leq \ldots \leq |\widetilde{x}_n|. \tag{2}$$

Let $f_{1,1} = e_{1,1}$ and $f_{i,j} = e_{i,j}$ for $i = 1,\ldots,n$ and $j = i+1,\ldots,n$. We need to determine $f_{i,i}$ for $i = 2,\ldots,n$ so that

$$f_{i,i}\widetilde{x}_i = e_{i,i}\widetilde{x}_i + \sum_{j=1}^{i-1}(e_{i,j} - e_{j,i})\widetilde{x}_j. \tag{3}$$

If $\widetilde{x}_i = 0$, let $f_{i,i} = 0$ (from (2) we have $\widetilde{x}_1 = \widetilde{x}_2 = \ldots = \widetilde{x}_{i-1} = 0$). Suppose $\widetilde{x}_i \neq 0$. Let

$$f_{i,i} = e_{i,i} + \sum_{j=1}^{i-1}(e_{i,j} - e_{j,i})\frac{\widetilde{x}_j}{\widetilde{x}_i}. \tag{4}$$

All that remains is to show that $|f_{i,i}| \leq 3\epsilon|a_{i,i}|$. Since $|e_{i,j}| \leq \epsilon|a_{i,j}|$ for all

i, j and $|\tilde{x}_j| \leq |\tilde{x}_i|$ for $j = 1, 2, \ldots, i-1$, hence

$$|f_{i,i}| \leq \epsilon|a_{i,i}| + 2\epsilon \sum_{j=1}^{i-1} |a_{i,j}|. \tag{5}$$

By (1) we finally get $|f_{i,i}| \leq 3\epsilon|a_{i,i}|$.

Case (ii): Assume that

$$|\tilde{x}_{p_1}| \leq |\tilde{x}_{p_2}| \leq \ldots \leq |\tilde{x}_{p_n}| \tag{6}$$

for some permutation $\{p_1, p_2, \ldots, p_n\}$ of $\{1, 2, \ldots, n\}$.

Let P be the permutation matrix such that $P^T = [e_{p_1}, \ldots, e_{p_n}]$.

Then $P[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n]^T = [\tilde{x}_{p1}, \tilde{x}_{p_2}, \ldots, \tilde{x}_{p_n}]^T$. Let

$$\hat{x} = [\tilde{x}_{p1}, \tilde{x}_{p_2}, \ldots, \tilde{x}_{p_n}]^T, \quad \hat{A} = PAP^T, \quad \hat{E} = PEP^T \quad \text{and} \quad \hat{b} = Pb.$$

Clearly the matrix $\hat{A}$ is symmetric and diagonally dominant and for all i, j

$$\hat{a}_{i,j} = a_{p_i, p_j}. \tag{7}$$

We see that $\hat{x}$ satisfies a relation $(\hat{A} + \hat{E})\hat{x} = \hat{b}$, where $|\hat{E}| \leq \epsilon|\hat{A}|$ and

$$|\hat{x}_1| \leq |\hat{x}_2| \leq \ldots \leq |\hat{x}_n|. \tag{8}$$

Consequently, we see that there exists a matrix $\hat{F} = \hat{F}^T \in \mathbb{R}^{n \times n}$ such that $(\hat{A} + \hat{F})\hat{x} = \hat{b}$, where $|\hat{F}| \leq 3\epsilon|\hat{A}|$.

Let $F = P^T\hat{F}P$. Then F is symmetric and $(A + F)\tilde{x} = b$. It is readily seen that $|F| \leq 3\epsilon|P|^T|\hat{A}||P| = 3\epsilon|A|$. This completes the proof. ∎

We can use the similar arguments to prove the following theorem.

THEOREM 2.2. *Assume that $A \in \mathbb{R}^{n \times n}$ is symmetric and satisfies*

$$|a_{i,j}| \leq \gamma|a_{i,i}|, \quad i, j = 1, \ldots, n. \tag{9}$$

If $(A + E)\tilde{x} = b$, where $|E| \leq \epsilon|A|$ and $\tilde{x} \neq 0$, then there exists a matrix $F = F^T \in \mathbb{R}^{n \times n}$ such that $(A + F)\tilde{x} = b$ and $|F| \leq (2(n-1)\gamma + 1)\epsilon|A|$.

Proof. All that remains is to show that $f_{i,i}$ defined by (4) satisfy the inequalities

$$|f_{i,i}| \leq (2(n-1)\gamma + 1)\epsilon|a_{i,i}|. \tag{10}$$

This is because $|f_{i,i}| \leq \epsilon|a_{i,i}| + 2\epsilon\gamma(i-1)|a_{i,i}|$ for all $i = 2, \ldots, n$, which obviously proves (10). ∎

COROLLARY. *In Theorem 2.1 suppose that*

$$|a_{i,i}|(1 - 3\epsilon) \geq (1 + 3\epsilon) \sum_{j \neq i} |a_{i,j}| \quad \text{for } i = 1, \ldots, n. \tag{11}$$

Then the perturbed matrix $A + F$ is symmetric and diagonally dominant. ∎

We see that stability in a componentwise sense implies the strong componentwise stability on the class symmetric diagonally dominant matrices.

3. Symmetric positive definite matrices

Suppose now that $A = A^T \in \mathbb{R}^{n\times n}$ is positive definite, i.e. $x^TAx > 0$ for all nonzero $x \in \mathbb{R}^{n\times n}$. The following theorem is an immediate consequence of Theorem 2.2.

THEOREM 3.1. *Assume that $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $(A+E)\widetilde{x} = b$, where $|E| \le \epsilon|A|$ and $\widetilde{x} \neq 0$, then there exists a matrix $F = F^T \in \mathbb{R}^{n\times n}$ such that $(A+F)\widetilde{x} = b$ and $|F| \le (2n-1)\epsilon|A|$. Moreover, if A is a band matrix with bandwith w then $|F| \le (2w-1)\epsilon|A|$.*

Proof. The proof is similar in spirit to [7]. It is well known that if $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, then

$$a_{i,j}^2 \le a_{i,i}a_{j,j} \quad \text{for } i,j = 1,\dots,n. \tag{12}$$

Let $D = \operatorname{diag}(a_{1,1}^{-1/2}, a_{2,2}^{-1/2}, \dots, a_{n,n}^{-1/2})$. We use the transformation:

$$\widehat{A} = DAD, \quad \widehat{E} = DED, \quad \widehat{b} = Db \quad \text{and} \quad \widehat{x} = D^{-1}\widetilde{x}.$$

Then the main diagonal elements of $\widehat{A}$ are all equal to 1 and $|\widehat{a}_{i,j}| \le 1$ for all i, j. We see that $\widehat{x}$ satisfies a relation $(\widehat{A}+\widehat{E})\widehat{x} = \widehat{b}$, where $|\widehat{E}| \le \epsilon|\widehat{A}|$. Consequently, from Theorem 2.2 we conclude that there exists a matrix $\widehat{F} = \widehat{F}^T \in \mathbb{R}^{n\times n}$ such that $(\widehat{A}+\widehat{F})\widehat{x} = \widehat{b}$, where $|\widehat{F}| \le (2n-1)\epsilon|\widehat{A}|$.

Let $F = D^{-1}\widehat{F}D^{-1}$. Then F is symmetric and $(A+F)\widetilde{x} = b$. It is readily seen that $|F| \le (2n-1)\epsilon|D^{-1}||\widehat{A}||D^{-1}| = (2n-1)\epsilon|A|$. The second part of Theorem 3.1 follows from Theorem 2.3. This completes the proof. ∎

One question still unanswered is: "When does a symmetric perturbation $A+F$ of a symmetric positive definite matrix A remain positive definite?" We first establish a relation between the eigenvalues of matrices A and $A+F$.

If $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, then we let $\lambda_i(A)$ denote the i-th largest eigenvalue of A. Thus,

$$\lambda_1(A) \ge \lambda_2(A) \ge \cdots \ge \lambda_n(A).$$

It is well known that each eigenvalue of A satisfies the following "min-max" characterization (see [9], [15]):

THEOREM 3.2 (Courant–Fischer). *If $A = A^T \in \mathbb{R}^{n\times n}$ then for $i = 1,2,\dots,n$*

$$\lambda_i(A) = \max_{\dim(\mathcal{X})=i} \min_{0\neq x\in\mathcal{X}} \frac{x^TAx}{x^Tx}. \tag{13}$$

Now we can prove the following theorem.

THEOREM 3.3. *If $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $F = F^T \in \mathbb{R}^{n\times n}$ then for $i = 1, 2, \ldots, n$*

$$\frac{|\lambda_i(A+F) - \lambda_i(A)|}{\lambda_i(A)} \leq \rho(A^{-1}F), \tag{14}$$

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

Proof. The Courant–Fischer theorem implies that

$$\lambda_i(A+F) = \max_{\dim(\mathcal{X})=i} \min_{0\neq x\in\mathcal{X}} \frac{x^T(A+F)x}{x^Tx}. \tag{15}$$

Since $x^TAx > 0$ for $0 \neq x \in \mathcal{X}$, we have

$$\frac{x^T(A+F)x}{x^Tx} = \frac{x^TAx}{x^Tx}(1 + \frac{x^TFx}{x^TAx}). \tag{16}$$

It is easy to check that

$$\frac{x^TFx}{x^TAx} = \frac{\widetilde{x}^T\widetilde{F}\widetilde{x}}{\widetilde{x}^T\widetilde{x}},$$

where we defined $\widetilde{x} = A^{1/2}x$ and $\widetilde{F} = A^{-1/2}FA^{-1/2}$. Note that $\widetilde{F}$ is symmetric, hence

$$\lambda_n(\widetilde{F}) \leq \frac{\widetilde{x}^T\widetilde{F}\widetilde{x}}{\widetilde{x}^T\widetilde{x}} \leq \lambda_1(\widetilde{F}).$$

We observe that $\widetilde{F}$ is similar to $A^{-1}F$, because $\widetilde{F}$ can be expressed as $\widetilde{F} = A^{1/2}(A^{-1}F)A^{-1/2}$, hence $\rho(\widetilde{F}) = \rho(A^{-1}F)$. From this we conclude that

$$\frac{x^TAx}{x^Tx}(1 - \rho(A^{-1}F)) \leq \frac{x^T(A+F)x}{x^Tx} \leq \frac{x^TAx}{x^Tx}(1 + \rho(A^{-1}F)).$$

This together with (1)–(2) imply

$$\lambda_i(A)(1 - \rho(A^{-1}F)) \leq \lambda_i(A+F) \leq \lambda_i(A)(1 + \rho(A^{-1}F)),$$

which completes the proof. ∎

COROLLARY. *In Theorem* 3.1 *suppose that*

$$(2n-1)\epsilon\rho(|A^{-1}||A|) < 1. \tag{17}$$

Then the Perron–Frobenius theorem implies that

$$\rho(A^{-1}F) \leq \rho(|A^{-1}||F|) \leq (2n-1)\epsilon\rho(|A^{-1}||A|) < 1. \tag{18}$$

In other words, if $(2n-1)\epsilon\rho(|A^{-1}||A|) < 1$ then the perturbed matrix $A+F$ is symmetric positive definite. ∎

We see that stability in a componentwise sense implies the strong componentwise stability on the class symmetric positive definite matrices.

Theorem 3.3 is very similar to a result of Demmel and Veselić (see [16], [17]). The bound in (14) is only little sharper.

References

[1] M. Arioli, J. W. Demmel and I. S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), 165–190.

[2] F. L. Bauer, *Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme*, ZAMM, 46 (1966), 667–684.

[3] Å. Björck, *Iterative refinement and reliable computing*, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, 1990, 249–266.

[4] Å. Björck, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), 238–244.

[5] J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), 49–66.

[6] J. R. Bunch, J. W. Demmel and C. V. Loan, *The strong stability of algorithms for solving symmetric linear systems*, SIAM J. Matrix Anal. Appl., 10 (1989), 494–499.

[7] J. Demmel, *The componentwise distance to the nearest singular matrix*, SIAM J. Matrix Anal. Appl., 13 (1992), 10–19.

[8] J. E. Dennis, Jr. and J. J. Moré, *Quasi-Newton methods, motivations, and theory*, SIAM Rev., 19 (1977), 46–89.

[9] G. H. Golub and C. F. Van Loan, *Matrix computations*, Second Edition, Johns Hopkins University Press, Baltimore Maryland, 1989.

[10] W. W. Hager, *Condition estimates*, SIAM J. Sci. Stat. Comput., 5 (1984), 311–316.

[11] D. J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), 162–175.

[12] N. J. Higham, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, BIT 31 (1991), 441–468.

[13] W. Oettli and W. Prager, *Compatibility of approximate solutions of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math. 6 (1964), 405–409.

[14] R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp. 35 (1980), 817–832.

[15] G. W. Stewart and Ji-guang Sun, *Matrix perturbation theory*, New York-London, Academic Press, 1990.

[16] J. W. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl. 13 (1992), 1204–1245.

[17] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra*, Numerical Analysis Report No. 241, University of Manchester (1994).

INSTITUTE OF MATHEMATICS
WARSAW UNIVERSITY OF TECHNOLOGY
Pl. Politechniki 1
00-661 WARSZAWA, POLAND

Received March 2, 1994.