Ann. Inst. Statist. Math.
Vol. 43, No. 3, 597–603 (1991)

ANALYSIS OF ERROR PROPAGATION IN THE ABS CLASS FOR LINEAR SYSTEMS*

A. GALANTAI**

Department of Computer Science, The University of Agriculture, Godollo, 2103, Hungary

(Received May 15, 1989; revised March 12, 1990)

Abstract. Broyden's backward error analysis technique is applied to evaluate the numerical stability of the ABS class of methods for solving linear systems.

Key words and phrases: Backward error analysis, optimally stable method.

1. Introduction

We study the numerical stability of the ABS-class for linear systems of the form

$$A^T x = b \quad (A \in R^{n\times n}) \tag{1.1}$$

where the matrix A^T is nonsingular. The ABS-class was developed by Abaffy *et al.* (1983, 1984), Abaffy and Spedicato (1985) and Abaffy and Galantai (1986). Let $P = [p_1, p_2, \ldots, p_n]$ and $V = [v_1, v_2, \ldots, v_n]$ be $n \times n$ type nonsingular matrices with column vectors p_j and v_j $(j = 1, \ldots, n)$. Denote by I the unit matrix of $n \times n$ type. The ABS class has the following form:

Let $x_0 \in R^n$ be arbitrary.

For $k = 1, \ldots, n$

Compute

$$a_k = v_k^T(A^T x_{k-1} - b)/(p_k^T A v_k) \tag{1.2}$$

$$x_k = x_{k-1} - a_k p_{k-1} \tag{1.3}$$

end for

where the ABS-update algorithm is given by

Set $H_1 = I$

* Part of this work was done during a stage at the University of Bergamo supported by CNR (Programma Professori Visitatori).

** Now at Institute of Mathematics, University of Miskolc, Miskolc-Egyetemváros, 3515, Hungary.

For $k = 1, \ldots, n$
Compute

$$p_k = H_k^T z_k \qquad (p_k^T A v_k \neq 0) \tag{1.4}$$

$$H_{k+1} = H_k - H_k A v_k w_k^T H_k / (w_k^T H_k A v_k) \qquad (w_k^T H_k A v_k \neq 0) \tag{1.5}$$

end for.

Algorithms (1.2)–(1.5) finitely terminate in n steps (Abaffy *et al.* (1984), Abaffy and Spedicato (1985) and Abaffy and Galantai (1986)). The matrix V is scaling the system $A^T x = b$. The pair (P, V) is said to be the A^T-conjugate (Stewart (1973)) if $V^T A^T P = L$ is the *lower triangular*. The ABS algorithm generates all A^T-conjugate directions for suitable choices of parameters (Abaffy and Galantai (1986)). Therefore the ABS class of methods coincides with those studied by Stewart (1973). Results on the numerical stability of conjugate direction methods are given by Broyden (1985) and Wozniakowski (1980). A stability analysis for descent methods is given in Bollen (1984).

We investigate the stability of the ABS-class by the backward error analysis technique due to Broyden (1974, 1985). Some basic results given by Broyden (1985) are extended here. We show, for example, that the error is proportional to $k(A)k(V)$ for the whole class (1.2) and (1.3), where $k(A)$ and $k(V)$ denote the condition number of matrices A and V, respectively. The condition for the residual perturbation to be minimal is also given.

2. Backward error analysis of the conjugate direction methods

The basic idea in Broyden's backward error analysis is the following. For the solution of some problems we consider any finite algorithm in the form

$$\boldsymbol{X}_{k+1} = \boldsymbol{\Psi}_k(\boldsymbol{X}_k) \qquad (k = 0, \ldots, n) \tag{2.1}$$

where $\boldsymbol{X}_{n+1}$ is the solution of the problem. Assume that an error $\boldsymbol{\epsilon}_j$ occurs at step j and that this error propagates further. It is also assumed that no other source of error occurs. The exact solution $\boldsymbol{X}_{n+1}$ is given by

$$\boldsymbol{X}_{n+1} = \boldsymbol{\Psi}_n\{\boldsymbol{\Psi}_{n-1}\{\cdots\{\boldsymbol{\Psi}_j(\boldsymbol{X}_j)\}\cdots\}\} = \boldsymbol{\Omega}^{n-j+1}(\boldsymbol{X}_j) \tag{2.2}$$

while the perturbed solution $\boldsymbol{X}'_{n+1}$ is given by

$$\boldsymbol{X}'_{n+1} = \boldsymbol{\Omega}^{n-j+1}(\boldsymbol{X}_j + \boldsymbol{\epsilon}_j). \tag{2.3}$$

If the quantity $\|\boldsymbol{X}_{n+1} - \boldsymbol{X}'_{n+1}\|$ is large, then algorithm (2.1) must be very unstable. Broyden argues that it must be small for stable algorithms. Consequently, $\|\boldsymbol{X}_{n+1} - \boldsymbol{X}'_{n+1}\|$ is a measure of stability for algorithms. Broyden applied this idea to the class (1.2) and (1.3) under the condition of A^T-conjugacy. For Huang's method he showed that

$$\|x_{n+1} - x'_{n+1}\| \leq k(A)\|\epsilon_j\| \tag{2.4}$$

holds, where x'_{n+1} denotes the perturbed solution.

Next we investigate the stability of the conjugate direction methods of the forms (1.2) and (1.3) with a special emphasize on the ABS-update (1.4) and (1.5). It is noted again that (P, V) is an A^T-conjugate pair. We use a projector technique due to Stewart (1973). Let us introduce the notation

$$P_k = p_k v_k^T A^T / (v_k^T A^T p_k). \tag{2.5}$$

The matrix P_k is a projector of rank one with $R(P_k) = R(p_k)$. $N(P_k) = R^{\perp}(Av_k)$. The notations R and N stand for the *range* and *nullspace*, respectively. With the notation P_k the algorithm (1.2) and (1.3) has the form

$$x_k = (I - P_k)x_{k-1} + d_k \quad (d_k = p_k v_k^T b / (v_k^T A^T p_k)). \tag{2.6}$$

Denote by x^* the solution of the linear system. Let $e_k = x^* - x_k$. Then we have the recursion

$$e_k = (I - P_k)e_{k-1}$$

with the solution

$$e_k = (I - P_k) \cdots (I - P_1)e_0.$$

Introduce the notation

$$(I - P_k) \cdots (I - P_j) = Q_{k,j}$$

for $k \geq j$. Let us suppose that an error occurs at the $(k-1)$-th step and only at it. The perturbed result of step $k-1$ is denoted by x'_{k-1}. The perturbed results of further steps are denoted by x'_j $(j = k, \ldots, n)$. Then we have

$$x'_n = \prod_{j=0}^{n-k} (I - P_{n-j})x'_{k-1} + \sum_{j=k}^{n} \left(\prod_{t=0}^{n-j-1} (I - P_{n-t}) \right) d_j$$

from which it follows that the error occuring in the final step

$$x^* - x'_n = x_n - x'_n = \prod_{j=0}^{n-k} (I - P_{n-j})(x_{k-1} - x'_{k-1}) = Q_{n,k}(x_{k-1} - x'_{k-1}).$$

The matrix $Q_{n,k}$ can be considered as the error matrix. Hence we have the error bound

$$\|x_n - x'_n\| \leq \|Q_{n,k}\| \, \|x_{k-1} - x'_{k-1}\|. \tag{2.7}$$

A method of the class (1.2) and (1.3) is considered to be *optimal* in the sense of Broyden (1985) if $\|Q_{n,k}\|$ is minimal for all k.

First we characterize $Q_{n,k}$. Using the A^T-conjugacy property, one can show in order (similarly to the proof of Theorem 2.6 in Stewart (1973)), that $P_t P_j = 0$ $(t < j)$, $P_t Q_{nk} = 0$ $(k \leq t \leq n)$, $(I - P_t)Q_{nk} = Q_{nk}$ $(k \leq t \leq n)$ implying that

$Q_{nk}Q_{nk} = Q_{nk}$, which means that $Q_{n,k}$ is a projector. Note that the projectors $P_1, \ldots, P_n$ are called *conjugate* by Stewart (1973), if $P_tP_j = 0$ $(t < j)$. By observing that $R(I - P_k) = N(P_k)$ and $N(I - P_k) = R(P_k)$ we easily find that

$$R(Q_{n,k}) = \bigcap_{j=k}^{n} N(P_j), \qquad N(Q_{n,k}) = \sum_{j=k}^{n} R(P_j). \tag{2.8}$$

It is also easy to see that

$$R(Q_{n,k}) = \bigcap_{j=k}^{n} R^{\perp}(Av_j) = R^{\perp}(AV^{n-k+1|})$$

and

$$N(Q_{n,k}) = \sum_{j=k}^{n} R(P_j) = R(P^{n-k+1|}) = R^{\perp}(AV^{|k-1}).$$

Here we used the Householder notations defined as follows. Given any matrix A the matrices $A^{\bar{k}}$, $A^{|k}$, $A^{\underline{k}}$ and $A^{k|}$ denote the submatrices consisting of respectively the first k rows, the first k columns, the last k rows and the last k columns of A.

For the sake of completeness we show that the symmetric projectors have a minimum property both in the Frobenius and in the spectral norm.

LEMMA 2.1. *If $A^2 = A$ and $A \neq 0$ then $\|A\|_F \geq \sqrt{m}$ $(m = \operatorname{rank}(A))$ and $\|A\| = \sqrt{m}$ if and only if A is symmetric.*

PROOF. We use the facts that the Frobenius norm is invariant under unitary transformations and it is possible to choose an orthogonal matrix U for which

$$U^TAU = B = \begin{bmatrix} I_m & B_2 \\ 0 & 0_{n-m} \end{bmatrix}$$

where B_2 is some matrix. Consequently $\|A\| = \|B\| \geq \|I\| = \sqrt{m}$. As B is symmetric if and only if A is so, the equality relation holds only for $B_2 = 0$ and therefore only for a symmetric A.

LEMMA 2.2. *If $A^2 = A$ and $A \neq 0$ then $\|A\|_{sp} \geq 1$ and $\|A\| = 1$ if and only if $A = A^T$.*

PROOF. The spectral norm is also invariant under unitary transformations. Thus $\|A\|_{sp} = \|B\|_{sp} = [\rho(B^TB)]^{1/2}$. By elementary calculations one has

$$B^TB = \begin{bmatrix} I_m & B_2 \\ B_2^T & B_2^TB_2 \end{bmatrix}.$$

Making use of the similarity transformation

$$\begin{bmatrix} I_m & 0 \\ -B_2^T & I_{n-m} \end{bmatrix} B^TB \begin{bmatrix} I_m & 0 \\ B_2^T & I_{n-m} \end{bmatrix} = \begin{bmatrix} I_m + B_2B_2^T & B_2 \\ 0 & 0 \end{bmatrix}$$

we find that B^TB has $n - m$ zero eigenvalues and m eigenvalues given by $\det\{B_2B_2^T - (\lambda - 1)I_m\} = 0$. If $B_2 \neq 0$ then $B_2B_2^T$ is positive semidefinite, implying that its eigenvalues are nonnegative and there exists at least one positive eigenvalue. Hence there is an eigenvalue of B^TB which is greater than 1 implying that $\|B^TB\|_{sp} > 1$. For a symmetric B (or A) it is obvious that $\|B\|_{sp} = 1$.

A projector P is symmetric if and only if $R(P) = N^{\perp}(P)$. Hence $Q_{n,k}$ is symmetric (and has minimal norm) if and only if

$$R(AV^{|k-1}) = R^{\perp}(AV^{n-k+1|}). \tag{2.9}$$

A method is optimal in Broyden's sense if and only if (2.9) is satisfied for all k. The latter condition is equivalent to

$$Av_i \perp Av_j \quad (i \neq j). \tag{2.10}$$

In matrix formulation it means that $V^TA^TAV = D$, where D is diagonal.

THEOREM 2.1. *The method of class* (1.2) *and* (1.3) *is optimal in the sense of Broyden if and only if* (2.10) *or equivalently* $V^TA^TAV = D$ *holds with a diagonal matrix* D.

This result was originally obtained by Broyden (1985) in a different way. The projector technique gives us a much deeper inside look at the structure of the error matrix $Q_{n,k}$ resulting from the following estimation of the whole class (1.2) and (1.3)

$$\|Q_{n,k}\| \leq k(A)k(V) \tag{2.11}$$

in spectral or Frobenius norm. It simply follows from the fact that $Q_{n,k}$, which is a projector onto $R^{\perp}(AV^{n-k+1|})$ along $R^{\perp}(AV^{|k-1})$, can be represented in the form

$$Q_{n,k} = (A^{-T}V^{-T})^{|k-1}(V^TA^T)^{\overline{k-1}}. \tag{2.12}$$

Since $\|B^{|k}\| \leq \|B\|$ and $\|B^{\bar{k}}\| \leq \|B\|$ are both in Frobenius and spectral norms, we may bound Q_{nk} by $\|Q_{nk}\| \leq \|A^{-T}V^{-T}\|\|V^TA^T\| \leq k(A)k(V)$, which is exactly (2.11).

THEOREM 2.2. *For the error propagation model* (2.1)–(2.3) *and* (2.7) *the class* (1.2) *and* (1.3) *yields the bound*

$$\|x_n - x'_n\| \leq k(A)k(V)\|x_{k-1} - x'_{k-1}\|. \tag{2.13}$$

If V is a unitary matrix then $k(V) = 1$ and $\|Q_{n,k}\| \leq k(A)$. For the original ABS class, of which Huang's method is a special case, the matrix $V = I$. Consequently the error propagation is proportional to $k(A)$ for that class. The same is

valid for the generalized ABS class (Algorithm (1.2)–(1.5)) with a unitary V. It is noted again that V can be considered as a scaling of the linear system $A^T x = b$ in the form $V^T A^T x = V^T b$. In this context we just refer to the well-known scaling techniques (Golub and Van Loan (1983)).

Defining the residual perturbation as $r'_k = A^T(x_k - x'_k)$, we have

$$r'_n = A^T Q_{nk} A^{-T} r'_k \tag{2.14}$$

for the model (2.1)–(2.3) and (2.7). Using the relation $(AB)^{|k}(CD)^{\bar{k}} = A\{B^{|k}C^{\bar{k}}\}D$ and (2.12) we find that

$$A^T Q_{nk} A^{-T} = (V^{-T})^{|k-1}(V^T)^{\overline{k-1}} \tag{2.15}$$

is a projector onto $R((V^{-T})^{|k-1})$ along $R((V^{-T})^{n-k+1|})$. The bound $\|r'_n\| \leq \|A^T Q_{nk} A^{-T}\| \|r'_k\|$ is minimal if and only if

$$R((V^{-T})^{|k-1}) = R^{\perp}((V^{-T})^{n-k+1}). \tag{2.16}$$

An algorithm of the class (1.2) and (1.3) can be called *optimal for the residual perturbation* r'_k if (2.15) holds for all k. This condition is obviously satisfied if and only if $(V^{-T})^T(V^{-T}) = D^{-1}$ for a suitable diagonal matrix D from which the condition $V^T V = D$ follows.

THEOREM 2.3. *The residual error r'_k is minimal for all k if and only if $V^T V = D$ is satisfied for a diagonal matrix D.*

As a result of Theorem 2.3, we can see that for a unitary V the ABS-class is optimal for the residual perturbation.

The structure of algorithm (1.2) and (1.3) yields the following simple extension of Broyden's model. Assume that instead of (1.2) and (1.3) we have the following recursion

$$x'_k = (I - P_k)(x'_{k-1} + \epsilon_{k-1}) + d_k \qquad (k = 1, \ldots, n) \tag{2.17}$$

where ϵ_{k-1} denotes the error which occured at the $(k-1)$-th step. Then we have

$$x_n - x'_n = \sum_{k=1}^{n} Q_{n,k} \epsilon_{k-1} \tag{2.18}$$

from which the bound

$$\|x_n - x'_n\| \leq \sum_{k=1}^{n} \|Q_{n,k}\| \, \|\epsilon_{k-1}\| \leq k(A)k(V) \sum_{k=1}^{n} \|\epsilon_{k-1}\| \tag{2.19}$$

follows. For the optimal method, $k(A)k(V)$ is obviously replaced by $\sqrt{m}$ or by 1 depending on the norm chosen.

THEOREM 2.4. *For the extended error propagation model* (2.17), *the class* (1.2) *and* (1.3) *satisfies the inequality* (2.19).

Finally we show the result on an ABS update. First we need to recall the result of Egervary (1960) on the update H_k. Namely,

$$H_{k+j} = H_k - H_k A V_{k,j} (W_{k,j}^T H_k A V_{k,j})^{-1} W_{k,j}^T H_k \tag{2.20}$$

holds for $j \geq 0$ provided that H_k is of rank no less than j. The matrices $V_{k,j}$ and $W_{k,j}$ denote $[v_k, \ldots, v_{k+j-1}]$ and $[w_k, \ldots, w_{k+j-1}]$, respectively.

According to Broyden (1974), assume now that an error occurs in the calculation of H_k and no further errors occur in the procedure. Denote the perturbed H_k by H_k'. Furthermore, we assume that $\operatorname{rank}(H_k') = \operatorname{rank}(H_k)$ and the error $\psi = H_k' - H_k$ satisfies the inequality

$$\|(W_{k,j}^T H_k A V_{k,j})^{-1}\| \|W_{k,j}\| \|A V_{k,j}\| \|\psi\| < 1 - 1/K \qquad (K > 1).$$

We recall that for any regular matrix A, $(A + F)^{-1} = A^{-1} + \xi$ holds with ξ satisfying $\|\xi\| \leq \|A^{-1}\|^2 \|F\| / (1 - \|A^{-1}\| \|F\|)$ provided that $\|A^{-1}\| \|F\| < 1$. The error in H_k results in a perturbed H_{k+j} denoted by H_{k+j}'. Then by an elementary calculation we obtain the estimation

$$\|H_{k+j}' - H_{k+j}\| \leq (1 + 2\|H_k\|\Lambda + K\|H_k\|^2\Lambda^2)\|\psi\| + (\Lambda + 2K\Lambda^2)\|\psi\|^2 + K\Lambda^2\|\psi\|^3,$$

where $\Omega = (W_{k,j}^T H_k A V_{k,j})^{-1}$ and $\Lambda = \|\Omega\| \|W_{k,j}\| \|A V_{k,j}\|$. It is noted that in general no one can expect an estimation of H_{k+j}' like that in (2.19) because the update algorithm is nonlinear.

REFERENCES

Abaffy, J. and Galantai, A. (1986). Conjugate direction methods for linear and nonlinear systems of algebraic equations, Quaderni DMSIA, 1986-7, Univ. di Bergamo, Bergamo, Italy.

Abaffy, J. and Spedicato, E. (1985). A generalization of the ABS algorithm for linear system, Quaderni DMSIA, 1985-4, Univ. di Bergamo, Bergamo, Italy.

Abaffy, J., Broyden, C. G. and Spedicato, E. (1983). Numerical performance of the pseudo symmetric algorithm in the ABS class versus LU factorization with iterative refinement, Rapporto SOFTMAT 8/83, Univ. di Bergamo, Bergamo, Italy.

Abaffy, J., Broyden, C. G. and Spedicato, E. (1984). A class of direct methods for linear systems, *Numer. Math.*, **45**, 361–376.

Bollen, J. A. M. (1984). Numerical stability of descent methods for solving linear equations, *Numer. Math.*, **43**, 361–377.

Broyden, C. G. (1974). Error propagation in numerical processes, *Journal of the Institute of Mathematics and Its Applications*, **14**, 131–140.

Broyden, C. G. (1985). On the numerical stability of Huang's and related methods, *J. Optim. Theory Appl.*, **47**, 401–412.

Egervary, E. (1960). On rank-diminishing operations and their applications to the solution of linear equations, *Z. Angew. Math. Phys.*, **11**, 376–386.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Maryland.

Stewart, G. W. (1973). Conjugate direction methods for solving systems of linear equations, *Numer. Math.*, **21**, 285–297.

Wozniakowski, H. (1980). Roundoff error analysis of a new class of conjugate gradient algorithms, *Linear Algebra Appl.*, **29**, 507–529.