Int. J. Solids Structures, 1970, Vol. 6, pp. 195 to 207. Pergamon Press. Printed in Great Britain

THREE-DIMENSIONAL ANALYSIS OF ELASTIC SOLIDS—II THE COMPUTATIONAL PROBLEM†

Y. R. RASHID

Gulf General Atomic Incorporated, San Diego, California

Abstract—A method of analysis of nonhomogeneous elastic solids involving general three-dimensional states of stress was presented in Part I. The displacement equations of equilibrium were based on the finite-element variational procedure. The element shape considered was a tetrahedron with linear displacement approximations. The main feature of that paper was the alternating component iterative method. The general procedure of the method was presented in detail without reference to the computational problem. In the present paper, we deal primarily with the computational characteristics of the method and discuss the roundoff problem and its influence on the rate of convergence of the iterative process and the accuracy of the computed solution.

INTRODUCTION

IN ANY given computational problem, it is normally necessary to replace the exact values of the functions involved by digital numbers. Since these digital numbers can only be expressed to a finite accuracy by terminating fractions with a given base, each operation injects a small perturbation into the calculation. The cumulative effect of these perturbations on the final answer is commonly known as roundoff error.

One of the most important characteristics that distinguish iterative methods from elimination methods is the repetitive structure of the former, which tends to make these methods self-correcting. This feature is often credited with minimizing roundoff errors. The magnitude and effectiveness of this self-correction is strongly dependent on the conditioning of the system being solved. Ill-conditioned systems are generally the normal targets for roundoff, and it is not unusual to encounter problems that are so ill-conditioned that a meaningful solution by either method is unattainable. For problems of this kind, iterative methods are inherently slowly convergent, and, due to roundoff, they may even break down or converge to the wrong solution.

Iterative schemes involve, in one form or another, the repetitive application of a simple algorithm that consists mainly of the accumulation of inner products and the subsequent multiplication by the inverse of a single number or of a small matrix. In the performance of those inner products, although one continues to work with the original matrix, roundoff error is introduced in two ways: first, during the accumulation of inner products, and, second, in the subsequent matrix division. It is generally incorrect, therefore, to assume that iterative methods are protected from rounding errors. Because they involve the direct solution (by elimination or by matrix inversion) of lower-order systems, block-iterative methods are generally more susceptible to roundoff error than point-iterative methods. On the other hand, block methods for positive definite symmetric matrices converge much

† Work supported by Union Carbide Corporation, Nuclear Division, under Subcontract 2848 (Prime Contract W-7405-eng-26).

faster than point methods. In making a choice between the two-solution techniques, we find that roundoff problems are not decisive factors.

The alternating component iterative method [1] is essentially a block method that involves the direct solution of m block-tridiagonal systems of order $(1/m)$th of the total system. In practical problems, the order of these subsystems may exceed 5000, and rounding errors will therefore be present. In this part we discuss the influence of rounding errors on the final answer and on the convergence properties of the iteration matrix.

SYSTEM OF GOVERNING EQUATIONS

The governing system of equations of an elastic solid can be written in the following partitioned matrix form [1], equation (24):

$$F_i(P) = \sum_{j=1}^{m} K_{ij}(P, Q)V_j(Q) \qquad i = 1, 2, \ldots, m, \tag{1}$$

where P and Q are lists of field and source points, respectively; $F_i(P)$ and $V_i(P)$ are load and displacement subvectors, respectively; $K_{ij}(P, Q)$ are the stiffness coefficient matrices; and m is the number of components. Applying the over-relaxation iterative method to equation (1) results in the following form:

$$V_i(P)^{(s+1)} = V_i(P)^{(s)} + \omega K_{ii}^{-1}(P, Q)\left[F_i(P) - \sum_{j=1}^{i-1} K_{ij}(P, Q)V_j(Q)^{(s+1)} - \sum_{j=i}^{m} K_{ij}(P, Q)V_j(Q)^{(s)} \right]$$
$$i = 1, 2, \ldots, m, \tag{2}$$

where $1 \leq \omega < 2$ is the over-relaxation factor and s is the iteration cycle. The solution of equation (1) in the manner indicated by (2) was presented in detail in Ref. 1.

CONVERGENCE CRITERION

If we refer to equation (48) of Ref. 1 and rename K_D, K_L and K_U, calling them D, L and U, respectively, equation (2) can be written as

$$(D + \omega L)V^{(s+1)} = [(1 - \omega)D - \omega U]V^{(s)} + \omega F, \tag{3}$$

where D is a block-diagonal matrix, L and U are strictly block-lower-triangular and block-upper-triangular matrices, respectively, and $V = (V_1(P), V_2(P), \ldots, V_m(P))$ and $F = (F_1(P), F_2(P), \ldots, F_m(P))$ are the displacement and load vectors, respectively. The subvectors $V_i(P)$ and $F_i(P)$ are of order $n = N/m$, where m is the number of displacement components and N is the total number of unknowns.

To develop a convergence criterion that accounts for rounding errors, it is necessary to deal with the computational form of (3).

It can be shown [2] that the computed solution of equation (1) is the exact solution of a perturbed equation of the form

$$(K + E_K)V = F + E_F, \tag{4}$$

where E_K and E_F are roundoff error quantities. Similarly, $V^{(s+1)}$ in equation (3) can be shown to be the exact vector iterate of the equation

$$(D+\omega L)V^{(s+1)}+(E_D+\omega E_L)^{(s+1)}V^{(s+1)} = [(1-\omega)D-\omega U]V^{(s)}+[(1-\omega)E_D-\omega E_U]^{(s+1)}V^{(s)}+\omega F+\omega E_F^{(s+1)}, \quad (5)$$

where E_D, E_L and E_U are the roundoff error matrices in D, L and U, respectively, and E_F is the roundoff error vector in F. These error quantities, which change from cycle to cycle, may include initial truncation errors also. By defining $V^{(s)}$ in this manner we are justified in treating all subsequent derivations as exact mathematical operations. When the expression

$$\varepsilon^{(s)} = V^{(s)}-V \quad (6)$$

is introduced into equation (5), and after simple manipulation, the following expression for the error vector iterate is obtained:

$$\varepsilon^{(s+1)} = [I+M_1^{(s+1)}]^{-1}[A+M_2^{(s+1)}]\varepsilon^{(s)}, \quad (7)$$

where

$$\left.\begin{aligned} M_1^{(s+1)} &= [I+\omega D^{-1}L]^{-1}[D^{-1}E_D+\omega D^{-1}E_L]^{(s+1)} \\ M_2^{(s+1)} &= [I+\omega D^{-1}L]^{-1}[(1-\omega)D^{-1}E_D-\omega D^{-1}E_U]^{(s+1)} \\ A &= [I+\omega D^{-1}L]^{-1}[(1-\omega)I-\omega D^{-1}U] \end{aligned}\right\} \quad (8)$$

It follows by successive substitution in (7) that

$$\varepsilon^{(s)} = \left(\prod_{i=1}^{s} A^{(i)}\right)\varepsilon^{(0)}, \quad (9)$$

where

$$A^{(i)} = [I+M_1^{(i)}]^{-1}[A+M_2^{(i)}]. \quad (10)$$

In the absence of roundoff, i.e. $M_i^{(i)} = M_2^{(i)} = 0$ for all $i \geq 1$, equation (9) reduces to the exact form

$$\varepsilon^{(s)} = A^s\varepsilon^{(0)}. \quad (11)$$

Conditions for convergence of equation (11) were discussed in Ref. 1. Those conditions that were stated for the exact iteration matrix must also hold for the computational form of that matrix, namely, $A^{(i)}$ $(i = 1, 2, \ldots)$.

At this point, it is necessary to introduce the definitions of matrix and vector norms as follows:

$$\|A\| = \sup_{X\neq 0}\frac{\|AX\|}{\|X\|} \quad (12)$$

and

$$\|X\|_q = (|x_1|^q+|x_2|^q+\cdots+|x_n|^q)^{1/q} \qquad q = 1, 2, \infty, \quad (13)$$

where $\|X\|_\infty$ is interpreted as $\max |x_i|$.

The three matrix norms that correspond to the three vector norms are given by

$$\left.\begin{aligned} \|A\|_1 &= \max_j \sum_i |a_{ij}| \\ \|A\|_\infty &= \max_i \sum_j |a_{ij}| \\ \|A\|_2 &= \max_{1 \le i \le n} [\lambda_i(A^T A)]^{\frac{1}{2}}, \end{aligned}\right\} \tag{14}$$

where λ_i is an eigenvalue of $A^T A$. The matrix and vector norms defined above satisfy the usual relations for norms.

If we ignore initial truncation errors, equation (5) serves as the computational equivalent of equation (3). We seek a finite sequence of vectors $\{V^{(1)}, V^{(2)}, \ldots V^{(S)}\}$ in which $V^{(S)}$ is close enough to the limit vector V in the sense

$$\lim_{s \to S} \|V^{(s)} - V\| = \lim_{s \to S} \|\varepsilon^{(s)}\| \le \delta, \tag{15}$$

where $\|\varepsilon^{(s)}\|$ is some suitable norm of the error vector and δ is considered small enough for our purposes.

System (9) converges in the sense of (15) if, and only if, any norm of $A^{(i)}$, for all $i = 1, 2, \ldots$, is less than unity.

From (10) we have

$$\|A^{(i)}\| \le \|(I + M_1^{(i)})^{-1}\| \, \|A + M_2^{(i)}\| \tag{16}$$

or

$$\|A^{(i)}\| \le \frac{\|A + M_2^{(i)}\|}{1 - \|M_1^{(i)}\|} \le \frac{\|A\| + \|M_2^{(i)}\|}{1 - \|M_1^{(i)}\|}, \tag{17}$$

provided $\|M_1^{(i)}\| < 1$. Equation (17) can be written as

$$\|A^{(i)}\| \le \beta\left(\frac{1 + \alpha_2^{(i)}}{1 - \alpha_1^{(i)}}\right) < 1, \tag{18}$$

where

$$\left.\begin{aligned} \beta &= \|A\| \\ \alpha_1^{(i)} &= \|M_1^{(i)}\| \\ \alpha_2^{(i)} &= \frac{\|M_2^{(i)}\|}{\|A\|} \end{aligned}\right\} \tag{19}$$

It is not uncommon to encounter exact iteration matrices of type A whose largest norms are very close to unity. For this class of matrices the contribution of rounding errors may be of such magnitude that (18) will not be satisfied for some value of i. If we assume that $\alpha_1^{(i)}$ and $\alpha_2^{(i)}$ are of the same order of magnitude, then the permissible upper limit for $\alpha_1^{(i)}$ and $\alpha_2^{(i)}$ is given by

$$\alpha_2^{(i)} \le \alpha_1^{(i)} < \frac{1 - \beta}{1 + \beta}. \tag{20}$$

We cannot guarantee convergence if $\alpha_1^{(i)}$ and $\alpha_2^{(i)}$ exceed the strict upper bound in (20). From (8) and (19) we have for $\alpha_1^{(i)}$

$$\alpha_1^{(i)} \leq \|(I+\omega D^{-1}L)^{-1}D^{-1}\|\,\|E_D^{(i)}+\omega E_L^{(i)}\|. \tag{21}$$

We can consider the case where $\omega = 1$, without loss of generality; then (21) simplifies to

$$\alpha_1^{(i)} \leq \|(D+L)^{-1}\|\,\|E_D+E_L\|. \tag{22}$$

The rounding error matrix (E_D+E_L) comes from solving a triangular set of equations with matrix $(D+L)$. Following Wilkinson [2], the norm of this error matrix is bounded as follows:

$$\|E_D+E_L\| \leq N2^{-t}\|D+L\|, \tag{23}$$

where N is the order of the matrix and t is the number of precision binary digits. Combining (22) and (23), we obtain

$$\alpha_1^{(i)} \leq \|(D+L)^{-1}\|\,\|D+L\|N2^{-t}. \tag{24}$$

The quantity $(\|D+L\|\,\|(D+L)^{-1}\|)$ can be regarded as condition number. For ill-conditioned matrices the upper bound in (24) can be quite large; therefore, we cannot guarantee convergence, in the practical sense, for that upper bound of $\alpha_1^{(i)}$. Convergence is guaranteed, however, if the quantities $\alpha_1^{(i)}$, $\alpha_2^{(i)}$, and $(\|D+L\| \cdot \|(D+L)^{-1}\|)$ satisfy the following inequality:

$$\alpha_2^{(i)} \leq \alpha_1^{(i)} \leq \|(D+L)^{-1}\|\,\|D+L\|N2^{-t} < \frac{1-\beta}{1+\beta}. \tag{25}$$

If β falls in the range $0.99 \leq \beta < 1$, which implies ill-conditioning, inequality (25) represents a very severe restriction on the value of condition number $(\|(D+L)^{-1}\|\,\|D+L\|)$. Since condition numbers of ill-conditioned matrices can be appreciably greater than unity, it is entirely possible that for large N (of order 10^4) the error buildup may cause the iterative process to diverge.

RATE OF CONVERGENCE

The rate of convergence of the iterative process may be defined for an s-cycle process as follows [3]:

$$R\left(\prod_{i=1}^{s} A^{(i)}\right) = -\frac{1}{s}\ln\frac{\|\varepsilon^{(s)}\|}{\|\varepsilon^{(0)}\|}. \tag{26}$$

Recognizing that

$$\left(\frac{\|\varepsilon^{(s)}\|}{\|\varepsilon^{(0)}\|}\right)^{1/s}$$

is the average reduction factor of the error norm per iteration, it is clear from (26) that the reciprocal of

$$R\left(\prod_{i=1}^{s} A^{(i)}\right)$$

gives a measure of the number of iterations required to reduce the norm of the initial error vector by a factor e. At first glance, definition (26) does not seem to be useful since in a practical problem the error vector is not usually known. However, we shall derive a measure of

$$R\left(\prod_{i=1}^{s} A^{(i)}\right)$$

in the form of an upper bound expressed in terms of known quantities. From equation (6) we obtain

$$\varepsilon^{(s-1)} - \varepsilon^{(s)} = V^{(s-1)} - V^{(s)} = -\Delta V^{(s)} \tag{27}$$

and

$$\varepsilon^{(s-1)} + \varepsilon^{(s)} = V^{(s-1)} + V^{(s)} - 2V. \tag{28}$$

From the fundamental properties of vector norms we have

$$\|\varepsilon^{(s-1)} + \varepsilon^{(s)}\| \leq \|\varepsilon^{(s-1)}\| + \|\varepsilon^{(s)}\| \tag{29}$$

and

$$\|\varepsilon^{(s-1)}\| - \|\varepsilon^{(s)}\| \leq \|\varepsilon^{(s-1)} - \varepsilon^{(s)}\|. \tag{30}$$

Since $\|\varepsilon^{(s-1)}\| - \|\varepsilon^{(s)}\| > 0$, we can add (29) and (30), and after simple manipulation obtain

$$\|\varepsilon^{(s-1)} + \varepsilon^{(s)}\| - \|\varepsilon^{(s-1)} - \varepsilon^{(s)}\| \leq 2\|\varepsilon^{(s)}\|. \tag{31}$$

If we substitute in (31) from (27) and (28), we have

$$\begin{aligned} 2\|\varepsilon^{(s)}\| &\geq \|2V - V^{(s-1)} - V^{(s)}\| - \|\Delta V^{(s)}\| \\ &\geq 2\|V\| - \|V^{(s)}\| - \|V^{(s-1)}\| - \|\Delta V^{(s)}\|, \end{aligned} \tag{32}$$

but

$$\|\Delta V^{(s)}\| = \|V^{(s)} - V^{(s-1)}\| \geq \|V^{(s)}\| - \|V^{(s-1)}\|. \tag{33}$$

When we add the last two inequalities, we obtain

$$2\|\varepsilon^{(s)}\| + \|\Delta V^{(s)}\| \geq 2\|V\| - 2\|V^{(s-1)}\| - \|\Delta V^{(s)}\| \tag{34}$$

or

$$\|\varepsilon^{(s)}\| \geq \|V\| - \|V^{(s-1)}\| - \|\Delta V^{(s)}\|. \tag{35}$$

Without loss of generality we may take the initial displacement to be identically zero; in other words,

$$\|\varepsilon^{(0)}\| \equiv \|V\| \neq 0. \tag{36}$$

Dividing inequality (35) by definition (36) gives

$$\frac{\|\varepsilon^{(s)}\|}{\|\varepsilon^{(0)}\|} \geq 1 - \frac{1}{\|V\|}(\|V^{(s-1)}\| + \|\Delta V^{(s)}\|). \tag{37}$$

It remains to express the generally unknown $\|V\|$ in terms of known quantities. Remembering that V is the exact limit vector, it is necessary that the computed bound for $\|V\|$ be as

sharp as possible. For this we return to equation (1),

$$F = KV. \tag{38}$$

In order to obtain a sharp bound for V, we introduce, purely on intuitive basis, the following relation between V and $V^{(s)}$:

$$\lim_{s \to S} V^{(s)} = V^{(S)} = \lambda V \tag{39}$$

for sufficiently large S. Here, λ is a positive number slightly less than unity. Equation (39) implies that toward the end of the iteration $V^{(s)}$ will have the correct shape but the wrong amplitude. From (38) and (39) we obtain for λ

$$\lambda = \frac{V^{(S)T} K V^{(S)}}{V^{(S)T} F}, \tag{40}$$

and for $\|V\|$,

$$\|V\| = \frac{1}{\lambda} \|V^{(S)}\|. \tag{41}$$

If $V^{(S)}$ is sufficiently close to V, then (40) is an energy balance that, incidentally, is insensitive to roundoff. Equation (41) represents a very good estimate for $\|V\|$.

Substituting for $1/(\|V\|)$ in (37) from (41) gives

$$\frac{\|\varepsilon^{(s)}\|}{\|\varepsilon^{(0)}\|} \geq 1 - \frac{\lambda}{\|V^{(S)}\|} (\|V^{(s-1)}\| + \|\Delta V^{(s)}\|). \tag{42}$$

Finally, by virtue of (26),

$$R\left(\prod_{i-1}^{s} A^{(i)} \right) \leq -\frac{1}{s} \ln \left[1 - \frac{\lambda}{\|V^{(S)}\|} (\|V^{(s-1)}\| + \|\Delta V^{(s)}\|) \right] \tag{43}$$

in which $0 < s \leq S$.

Equation (43) serves as a computational sharp upper bound for the rate of convergence

$$R\left(\prod_{i=1}^{s} A^{(i)} \right),$$

and it can also be used to compare the convergence properties of different iteration matrices. In particular, we wish to compare

$$R\left(\prod_{i=1}^{s} A^{(i)} \right)$$

and $R(A^s)$ of equations (9) and (11), respectively. Since it is virtually impossible to find the exact value of $R(A^s)$, for the purpose of this comparison one may compute the upper bound for

$$R\left(\prod_{i=1}^{s} A^{(i)} \right)$$

from (43) by the two computational schemes discussed in Ref. 1.

RELATIVE ACCURACY OF THE SOLUTION

The second major problem area of rounding errors is their influence on the final results. If the coefficient matrix is ill-conditioned with respect to the solution, it usually exhibits poor convergence properties. Therefore, the behavior of the iterative process is a very good index of the magnitude of roundoff.

A correction procedure, which was equivalent to performing the last few cycles in double precision, was presented in Ref. 1. From a practical standpoint, this treatment of the problem is quite sufficient. It would be helpful, however, if one were able to determine the relative accuracy of the solution through some sort of error analysis. Unfortunately, rigorous error analysis usually leads to establishing upper bounds for the relative accuracy of the computed solution in terms of generally noncomputable quantities. Those upper bounds may be functions of the size of the matrix, the type of arithmetic used, and the norms of the exact inverses of the original matrix or the triangular decompositions. Although such bounds can be helpful in studying the general properties of the matrices involved, and perhaps can point out the areas where and when to expect trouble, they tend to overestimate the true situation by emphasizing the worst conditions.

The two general criteria that are used to gauge the accuracy of the computed solution as iteration progresses are the change in the solution $\Delta V^{(s)}$ and the size of the residual $R^{(s)}$ expressed as norm quantities. Both criteria when considered separately can be misleading, especially for ill-conditioned problems. On the one hand, a large $\|R^{(s)}\|$ may not necessarily indicate an inaccurate solution. On the other hand, although a large $\|\Delta V^{(s)}\|$ implies large error in $V^{(s)}$, a small $\|\Delta V^{(s)}\|$ may be indicative of poor convergence and not good accuracy. Therefore, one must examine both quantities, namely, the correction in the solution, as well as the residual, in order to be able to determine closely the relative accuracy of the final results.

In the alternating-component iterative method presented in Ref. 1, one deals in each iteration with N inner products of order N and with the direct solution by triangular decomposition of m systems of equations. The computation procedure of that method is summarized as follows:

1. The residual vector $R(P)$ is computed from

$$R_i(P)^{(s)} = F_i(P) - \sum_{j=1}^{i-1} K_{ij}(P, Q)V_j(Q)^{(s)} - \sum_{j=i}^{m} K_{ij}(P, Q)V_j(Q)^{(s-1)} \qquad i = 1, 2, \ldots, m. \tag{44}$$

2. The displacement increment vector $\Delta V(P)$ is then computed from

$$\Delta V_i(P)^{(s)} = U_i^{-1} L_i^{-1} R_i(P)^{(s)}, \tag{45}$$

where U_i and L_i are, respectively, upper and lower block triangular matrices such that

$$L_i U_i = K_{ii} \qquad i = 1, 2, \ldots, m \tag{46}$$

3. Finally, the displacement vector iterate is obtained from

$$V_i^{(s)} = V_i^{(s-1)} + \omega \Delta V_i^{(s)}. \tag{47}$$

From Ref. 1, equation (58), the convergence criterion is established by

$$\lim_{s \to \infty} \|R^{(s)}\|_1 = 0. \tag{48}$$

Since the main concern is the problem of roundoff, we assume that iteration has progressed to the point where no further reduction in $\|R^{(s)}\|_1$ is obtained, which indicates that R consists mainly of rounding errors. Assuming this stalemate condition has been reached after S cycles, then we can continue for q additional correction (double precision) cycles in the manner outlined in Ref. 1. Since the triangular decomposition of K_{ii} [equation (46)] is independent of the iteration cycle s, it need not be recomputed. However, (44) and (45) will have to be computed in double precision if any further improvement in the solution is to be obtained. Denoting the double precision equivalents of R and ΔV, as defined in Ref. 1, by $\bar{r}$ and $\delta\bar{V}$, the correction $\delta\bar{V}_i^{(S+q)}$ then satisfies the following equation:

$$[K_{ii}+E_{ii}]\delta\bar{V}_i^{(S+q)} = \bar{r}_i^{(S+q)} \qquad i = 1, 2, \ldots, m, \tag{49}$$

where E_{ii} is the rounding error matrix that comes from the single-precision computation of L_i and U_i. In other words, the computed L_i and U_i satisfy

$$L_iU_i = K_{ii}+E_{ii}. \tag{50}$$

Equation (49) holds if no further rounding errors are introduced in the solution of the two triangular equations

$$L_i\bar{r}_i^{*(S+q)} = \bar{r}_i^{*(S+q)} \tag{51}$$

and

$$U_i\delta\bar{V}_i^{(S+q)} = \bar{r}_i^{*(S+q)}, \tag{52}$$

and we naturally assume this. This assumption is valid approximately, since in the correction cycle we solve (51) and (52) in double precision. The residual is defined by

$$\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}, \tag{53}$$

and from (49) we see that

$$\|\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}\| \leq \|E_{ii}\|\,\|\delta\bar{V}_i^{(S+q)}\|. \tag{54}$$

Following Wilkinson [2], $\|E_{ii}\|_\infty$ may have the estimate

$$\|E_{ii}\|_\infty \leq 2g\,2^{-t}(n/2+1)(n-1), \tag{55}$$

where n in our case is N/m and g is the order of magnitude of the largest element in U_i. The inequality (54) becomes

$$\|\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}\|_\infty \leq 2g2^{-t}(n/2+1)(n-1)\|\delta\bar{V}_i^{(S+q)}\|_\infty. \tag{56}$$

Ignoring the value of n in comparison with n^2, from (56), we have

$$\|\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}\|_\infty \leq g2^{-t}n^2\|\delta\bar{V}_i^{(S+q)}\|_\infty. \tag{57}$$

If $\|\delta\bar{V}_i^{(S+q)}\|_\infty$ after q cycles falls in the range

$$2^{-[z(q)+1]} < \|\delta\bar{V}_i^{(S+q)}\|_\infty \leq 2^{-z(q)}, \tag{58}$$

then from (57), replacing n by N/m, we have

$$\|\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}\|_\infty \leq \left(\frac{N}{m}\right)^2 g2^{-[z(q)+t]}. \tag{59}$$

If $z(q) = t$, then the displacement vector becomes that of the single-precision, correctly rounded solution. In this case we have

$$\|\bar{r}_i^{(S+q)} - K_{ii}\delta\bar{V}_i^{(S+q)}\|_\infty \leq g(N/m)^2 2^{-2t}. \tag{60}$$

If g is $O(10^9)$, $N = 10^4$, $m = 3$, and we are working to eight-decimal-place accuracy, the bound in (60) is $O(1)$. Such a bound is attainable in practice.

By virtue of the definition of the ∞-norm, equation (59) can be generalized to the total residual vector; hence,

$$\|\bar{r}^{(S+q)} - D\delta\bar{V}^{(S+q)}\|_\infty \leq g(N/m)^2 2^{-[z(q)+t]} \tag{61}$$

The significance of this equation is that if the residual satisfies (61) we can guarantee the accuracy of $V^{(S+q)}$ to $z(q)$ binary digits, provided that after the stalemate condition $\|R\|_1^{(s)} =$ constant has been reached, q correction (double-precision) iterations were performed.

EXAMPLES

Example 1

This example was given in Ref. 1; it serves our purposes to discuss it again here. The problem being investigated consists of a cantilever beam of rectangular cross section and a span-to-depth ratio of 10 loaded with end shear. The applied shear stresses were distributed consistently with the three-dimensional beam theory. Although this example is of no real practical importance, it presents an interesting computational problem.

The example contains 10,530 displacement unknowns, grouped into 3 components, and has a total bandwidth of 1050. The quantities to be investigated are the following vector norms: the 1-norm of the residual force vector $\|R\|_1$, the ∞-norm of the displacement vector $\|V\|_\infty$. the ∞-norm of the displacement increment vector $\|\Delta V\|_\infty$, and the rate of convergence $R(\prod_{i=1}^{s} A^{(i)})$ computed from equation (43). If we consider, for illustration purposes, that these vector norms are continuous functions of the number of iteration cycles, we may then plot these quantities as shown in Fig. 1. Ignoring the initial apparent divergence as reflected in the $\|R\|_1$ curve, the iterative process seems to converge very slowly at a decreasing rate. At the 27th cycle, the displacement increment is of the order of 1% of the exact solution and about 3% of the computed 27th vector iterate. The rate of convergence was computed to be 0·019019. The reciprocal of this quantity is a measure of the number of cycles required to reduce the norm of the error in V by a factor 2·71828. Therefore, to reduce the norm $\|\varepsilon^{(27)}\| = \|V^{(27)} - V\|$ by one order of magnitude, 121 cycles are needed. The computed $\|\varepsilon^{(27)}\|_\infty$ is approximately equal to 57% of the exact $\|V\|_\infty$. Therefore, from the above computed rate of convergence, at least 121 cycles are needed to reduce the error norm $\|\varepsilon\|_\infty$ to an acceptable value of 0·57%.

By applying the extrapolation formula (68) in Ref. 1, the computed norms at the end of the first cycle after extrapolation indicated substantial improvement in the iterative process. Although the norm $\|R\|_1$ increased by more than one order of magnitude, the rapid rate of convergence that followed brought $\|R\|_1$ down to its previous value in two cycles. The rate of convergence increased by over six times. The computed $\|V^{(27)}\|_\infty$ increased by about three times to within 2% of the exact value. At the end of the 36th cycle, $\|V^{(36)}\|_\infty$ was brought to within 0·5% of the exact solution. The computed rate of convergence increased

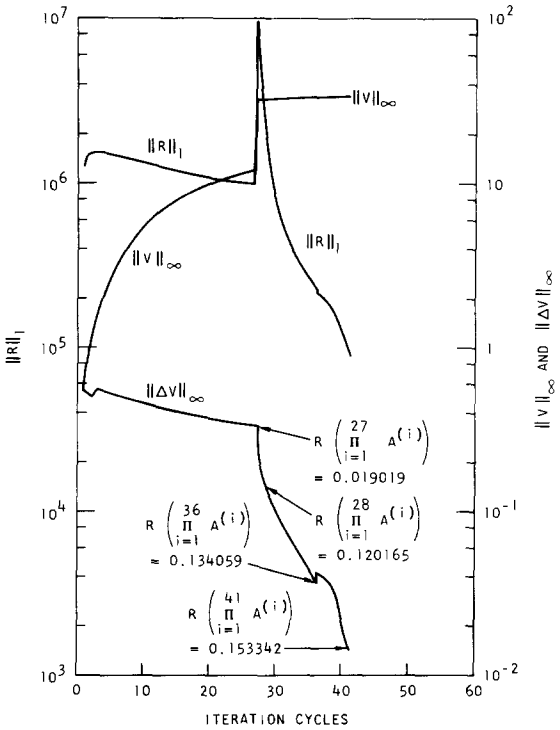

FIG. 1. Example 1, rate of convergence of the iterative solution.

monotonically from 0·120165 to 0·134059. The largest displacement increment $\|\Delta V\|_\infty$ was close to 0·1% of the exact solution at the end of the 36th cycle.

Further improvement of the solution was effected by carrying out cycles 37 through 41 in double precision. As expected, a slight increase in $\|\Delta V\|_\infty$ occurred, but it was reduced to a value about half the value anticipated if double precision were not used. This implies that at this stage a large percentage of $\|\Delta V\|_\infty$ is contributed by roundoff.

It is of interest to compare the values at the end of iteration of $\|\Delta V\|_\infty$ and $\|R\|_1$ of Example 2 in Ref. 1, the pressure vessel, with those of this example. Although $\|R\|_1$ divided by the total number of unknowns, which is a measure of the average error, is of the same order of magnitude in both problems, the normalized norm quantity $\|\Delta V\|_\infty/\|V\|_\infty$ is 0·04% and 0·0009% for the beam and the pressure vessel, respectively. Therefore, if one judges the accuracy of the solution on the basis of $\|R\|_1$ alone, both solutions have the same relative accuracy. However, it is clear from the above results that the degree of accuracy is about 200 times higher in the pressure vessel. This type of behavior typically distinguishes the well-conditioned system from the ill-conditioned ones.

Example 2

In this example, we investigate the influence of roundoff on the computed solution. Although the problem was discussed briefly in the previous beam problem, the solution was not carried out far enough to permit a quantitative error analysis. For the present purposes, we selected a pressure vessel problem with 7092 displacement unknowns and a total bandwidth of 648.

The single-precision (S.P.) vector norms, $\|R\|_1$ and $\|\Delta V\|_\infty$, and the corresponding double-precision (D.P.) vector norms, $\|\bar{r}\|_1$ and $\|\delta \bar{V}\|_\infty$, are plotted in Fig. 2. At first, the iteration was carried out in single precision until the condition $\|R\|_1 = \text{constant}$ was attained, indicating roundoff error contribution. The calculations were then repeated in double precision, starting with cycle 11. Although the roundoff error component was always present, it was not detectable until the 14th cycle. At the end of the 20th cycle, the condition $\|\bar{r}\|_1 = \text{constant}$ was attained. At that stage, the value of $\|\bar{r}\|_1$ was about four times smaller than the corresponding single-precision quantity $\|R\|_1$. A more significant comparison, however, is that of $\|\Delta V\|_\infty$ with its double-precision counterpart $\|\delta \bar{v}\|_\infty$. The former reached an oscillatory stage, at the limit of single-precision representation, at the end of the 18th cycle, whereas the latter maintained its monotonic decrease. At the end of the 20th cycle, the two quantities were one order of magnitude apart.

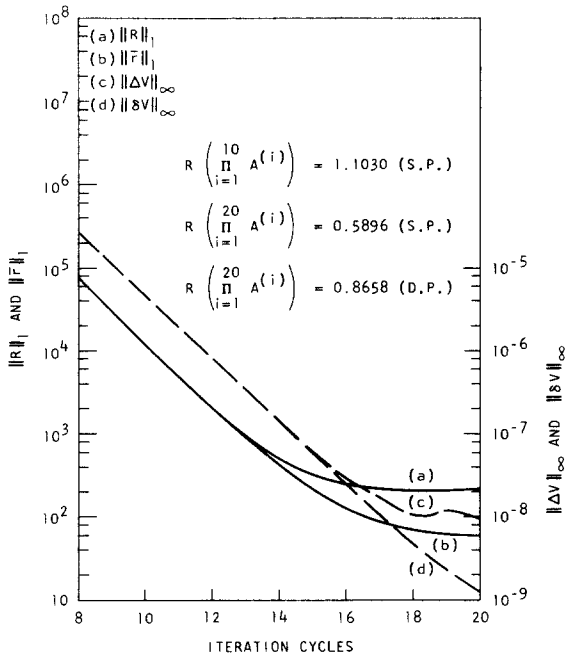

FIG. 2. Example 2, accuracy and rate of convergence of the iterative solution.

The stalement condition $\|R\|_1 = \text{constant}$ implies that any further iteration involves roundoff error manipulation only. By changing the mode of computation to double precision, equation (61) allows us to evaluate quantitatively the relative accuracy of the solution. In our case, $z(q) = t = 26$, g is $O(10^9)$, and N/m is 2364. The theoretical upper bound of the residual, from equation (61), is approximately equal to 0·56. Therefore, for the solution to be of at least 8 decimal-place accuracy, the computed residual at the end of the 20th cycle must not exceed that value. This computed value was found to be 0·26611, which satisfies the condition stipulated above. This example provides at least one verification of the error analysis presented earlier. Therefore, one may use equation (61) to establish the relative accuracy of the solution provided the conditions on which (61) is based are satisfied. In practice, however, the level of accuracy given by (61) is seldom required.

The second influence of roundoff is in the rate of convergence of the iterative process. It was demonstrated in Example 1 that the rate of convergence can be increased by using double-precision arithmetic. However, roundoff error was not isolated from other effects; therefore, the influence of roundoff error on the rate of convergence was not conclusively established by that example. In the present example, the double-precision and single-precision rates of convergence [equation (43)] were computed at the end of the 20th cycle and are shown in Fig. 2. Since double-precision iteration after cycle 18 dealt mainly with the roundoff error component in the solution, the difference in the rates of convergence at the end of the 20th cycle is attributed to roundoff error. However, at that stage in the iteration, the difference of 50% between the two values, as indicated in Fig. 2, is of little practical value. In earlier cycles, the contribution of roundoff to the rate of convergence and the solution accuracy is not detectable. In well-conditioned systems, the effect of roundoff becomes conspicuous at a stage beyond the range of practical interest. For ill-conditioned systems, since double-precision iteration is more effective during the latter stages of iteration, the last few cycles only may be computed in double precision.

REFERENCES

[1] Y. R. RASHID, Three-dimensional analysis of elastic solids Part I—Analysis procedure. *Int. J. Solids Struct.* **5**, 1311–1331 (1969).
[2] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*. Prentice-Hall (1963).
[3] R. W. VARGA, *Matrix Iterative Analysis*. Prentice-Hall (1962).

(*Received* 3 *January* 1969)

Абстракт—В части I, дается метод расчета неоднородных упругих тел, в общем трехмерном напряженном состоянии. Приводятся уравнения в перемещениях для равновесия на основе вариационного метода конечного элемента. Рассматривается форма элемента в виде тетраэдра с линейными приближениями для перемещений. Главной особенностью предыдущей работы является итерационный метод переменного компонента. Указнно в ней подробно общий процесс метода, без отномения к задаче численного решения. В настоящей работе, рассматривается главным образом вопрос расчетных характеристик метода и обсуждается задача округлечения, далее ее влияние на скорость сходимости итерации а также точность решения.