Impulse Constraint

The deformation associated with plastic strain occurs across the c_2 characteristic wave front. This plastic deformation within the bar results in a displacement δ of the end of the rod‡

$$\delta = \sum_{m=0}^{M} \epsilon_m{}^P[l_m - l_{m-1}] \tag{5}$$

where, by definition, δ is positive when the rod is lengthened and $l_{-1} = 0$. Substituting into Eq. (5) from Eqs. (1–4) we obtain a nondimensional ratio of displacement to impulse applied at the end of the rod

$$\delta\rho c_2/P\tau = [(\alpha + 1)/\alpha]\{1 + 2M/(\alpha + 1) - (\sigma_y/P)[(\alpha + 1)/(\alpha - 1)]^M\} \tag{6}$$

This ratio of displacement to impulse is shown in Fig. 2.

With this idealized material, the impulse response curves are not very sensitive to changes in α. The family of curves is closely approximated by

$$\delta\rho c_2/P\tau = \ln(P/\sigma_y) \tag{7}$$

with the variation about this function increasing as α decreases.§ The impulse response curves show an increasing ratio of displacement to impulse with increasing P/σ_y. Consequently, for this class of loads, an impulsive pressure (Dirac delta function) results in the maximum displacement from any given impulse.

Energy Constraint

The energy per unit area, denoted by E, that is imparted to the bar will be

$$E = \int_0^\infty p(t)v(0,t)dt \tag{8}$$

where $v(0,t)$ is the particle velocity at the end of the rod. Since $p(t)$ is a constant during $[0,\tau]$ and then vanishes

$$E = (P\tau/\rho c_1)[P\alpha - \sigma_y(\alpha - 1)] \tag{9}$$

This equation, together with Eq. (6), results in a nondimensional ratio of displacement to energy

$$\delta\sigma_y/E = \{(\alpha + 1)\sigma_y/P/[\alpha - (\alpha - 1)\sigma_y/P]\}\{1 + 2M/(\alpha + 1) - (\sigma_y/P)[(\alpha + 1)/(\alpha - 1)]^M\} \tag{10}$$

that is shown in Fig. 3. From these curves, the value of P that will maximize the displacement-energy ratio can be determined. A corresponding load duration for maximal deformation, τ_{opt}, is shown in Fig. 4. In contrast to the case where impulse was limited, the optimal value depends on α.

Conclusion

Results have been obtained for a system that represents some metal forming operations where either the impulse or the energy imparted to the system on each stroke is limited. The results for the cases of an impulse constraint and an energy constraint are distinct. With a constraint on impulse, the deformation is maximized by applying the greatest pressure in the shortest possible time. With a constraint on energy, a longer time results in the maximal deformation; the corresponding applied pressure is not much larger than the yield stress. In general, these results indicate that if one wishes to obtain maximal deformations from some loading system, the constraints peculiar to that system must be identified.

‡ If the pressure is either increased or decreased slightly during the loading period, this displacement magnitude is only changed slightly.

§ The limiting case, $\alpha \to \infty$, results from either the rigid-strain hardening or the elastic-perfectly plastic material descriptions. In the first case, $E_1 \to \infty$ and the plastic deformation is confined to the length l_0. In the second case, $E_2 = 0$ and the plastic deformation is limited to the end of the rod. The latter is necessarily a large deformation result.

References

[1] Hopkins, H. G., "The Method of Characteristics and its Application to the Theory of Stress Waves in Solids," *Engineering Plasticity*, edited by J. Heyman and F. A. Leckie, Cambridge University Press, 1968, pp. 277–315.

[2] Lee, E. H. and Tupper, S. J., "Analysis of Plastic Deformation in a Steel Cylinder Striking a Rigid Target," *Journal of Applied Mechanics*, Vol. 21, 1954, pp. 63–70.

[3] Cristescu, N., *Dynamic Plasticity*, North-Holland, Amsterdam, 1967, pp. 63–65.

Discretization and Computational Errors in High-Order Finite Elements

Isaac Fried*
Massachusetts Institute of Technology,
Cambridge, Mass.

1. Introduction

DISCRETIZATION of an elliptic equation of the $2m$th order ($m = 1$ harmonic, $m = 2$ biharmonic) by finite elements of diameter h, inside which the interpolation polynomials include a complete set of degree p, assures[1] that the error in the energy is decreased as $h^{2(p+1-m)}$. Thus, by reducing the mesh size h sufficiently, one theoretically should be able to obtain any desired accuracy in the solution. Unfortunately, roundoff errors coupled with the particular nature of the difference matrices generated by the finite element method drastically alter this prediction.

The roundoff errors affecting the accuracy of x in solving the linear system $Kx = b$ come from two sources: from truncating or rounding the initial data in K and b, and from the accumulation of errors during the solution process. It has been shown[2–3] that with current methods of solution these two sources give rise to errors of similar magnitude. In the case where the data is given exactly, the solution can be improved iteratively.[4] But since in practice the data is never given exactly, it is the most serious source of error. Basically, this error is a function of the accuracy of the computer and of the condition of the global (stiffness) matrix K.

The difference matrices generated by either the finite element or finite difference method inevitably become ill-conditioned as the mesh of elements is refined. In fact, the spectral condition number of K, $Cn(K)$, increases as ch^{-2m}, where c is a function of various mesh parameters. Hence as the mesh is refined the discretization errors decrease, but simultaneously the effect of roundoff errors increases until at a certain h they become dominant. In this respect the higher order (say bending where $m = 2$) problems are more prone to round off error perturbation than the lower order ones. Now since p, the order of the element, appears in the exponent of h in the expression for the discretization errors but not in the expression for the condition number, by using higher order elements one expects to reduce the relative effect of roundoff

Received March 24, 1971; revision received June 7, 1971. This research was supported in part by the Office of Scientific Research of the USAF under Contract F44620-67-C0019. The author wishes to express his appreciation to T. H. H. Pian for his support.

Index category: Structural Static Analysis.

* Post Doctoral Research Fellow, Department of Aeronautics and Astronautics.

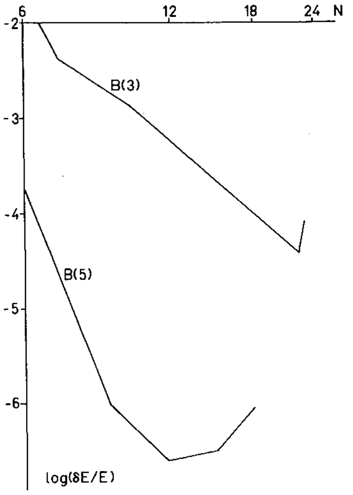

Fig. 1 Variation of the total relative error in the energy vs N for the case of a cantilever beam loaded by a quadratically varying load. Computations carried out in single precision.

errors and hence to achieve greater accuracy in the finite element solution. This possibility is explored in the present Note.

At the time this note was being revised Kelsey[5] et al. published an interesting paper on the condition number of the finite element stiffness matrix.

2. Discretization Errors

Let the variables be interpolated inside the element by functions which include a complete polynomial of the pth degree. Then by Taylor's theorem the finite element solution can approximate the true solution such that the difference between them is of the order h^{p+1}. For problems of the $2m$-th order the energy expression involves derivatives of the m-th order. These derivatives can be approximated by the finite-element interpolation scheme up to an error of the order h^{p+1-m}. It then follows immediately that the error in the energy is of the order $h^{2(p+1-m)}$. Therefore by using higher order interpolation schemes inside the element (i.e., more degrees of freedom per element), the error in the energy is made to decrease exponentially.

Computational errors

Consider the linear algebraic system

$$Kx = b \tag{1}$$

where the matrix K is assumed to be positive definite and symmetric. If K arises from the discretization of elliptic problems of the $2m$th order, its minimal eigenvalue is $O(1)$, while its maximal eigenvalue[1] is $O(h^{-2m})$, where h denotes the diameter of the element. In practice, both x and b are near the direction of the eigenvector corresponding to the minimal eigenvalue of K. It requires an exceedingly complicated set of forces b to produce an x in the direction of the eigenvector corresponding to the maximal eigenvalue of K. Hence $\|b\| = \|x\|$ and it is assumed that $\|b\| = 1$. In the computer, Eq. (1) is perturbed into

$$(K + \delta K)(x + \delta x) = b + \delta b \tag{2}$$

where $\|\delta K\| = \|K\|10^{-s}$ and $\|\delta b\| = \|b\|10^{-s}$, in which s denotes the number of decimals in the computer. From Eq. (2) it results that the error δx in x is given by

$$\delta x = K^{-1}\delta b - K^{-1}\delta K x \tag{3}$$

Even though x is usually near the direction of the minimal eigenvalue of K, since δK is random there is no reason why x should not have an appreciable component in the direction of the maximal eigenvalue of δK. Hence $\|K^{-1}\| = O(1)$, $\|\delta b\| = O[1]10^{-s}$, $\|\delta K\| = O(h^{-2m})10^{-s}$ and Eq. (3) yields

$$\|\delta x\| = 10^{-s}O(h^{-2m}) \tag{4}$$

or, since $\|x\| = 1$

$$\|\delta x\|/\|x\| = c10^{-s}Cn(K) \tag{5}$$

where $Cn(K)$, the spectral condition number of K, is the ratio between the maximal and minimal eigenvalues of K. It has been shown[2] for full low-order matrices, that the coefficient c in Eq. (5) is nearly equal to 1.

To see the effect of scaling,[6] at least from the perturbation point of view, consider the system

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \left|\begin{matrix} x_1 \\ x_2 \end{matrix}\right| = \left|\begin{matrix} 1 \\ 1 \end{matrix}\right| \tag{6}$$

where $(x_1, x_2) = (1,1)$ and the condition number of the matrix, $Cn(K)$, is equal to 3. Now let x_1 be transformed by $x_1 = \epsilon\hat{x}_1$, $\epsilon \ll 1$, and for reasons of symmetry let the first equation in Eq. (6) also be multiplied by ϵ. System (6) then becomes

$$\begin{pmatrix} 2\epsilon^2 & -\epsilon \\ -\epsilon & 2 \end{pmatrix} \left|\begin{matrix} \hat{x}_1 \\ x_2 \end{matrix}\right| = \left|\begin{matrix} \epsilon \\ 1 \end{matrix}\right| \tag{7}$$

where $(\hat{x}_1, x_2) = (1/\epsilon, 1)$ and $Cn(K) = 4/3\epsilon^2$. System (7) grows ill-conditioned as $\epsilon \downarrow 0$, however, small perturbations in the initial data will lead to only small perturbations in x_2. The resulting perturbations in $\hat{x}_1$ will be large, but since $\hat{x}_1$ itself is large, the relative error in $\hat{x}_1$ is again small. This suggests that scaling should not have a great effect on δx, and that the condition number governing the effect of roundoff errors is actually given[2] by

$$\hat{Cn}(K) = \text{Min}_D\, Cn(DKD)$$

where D is diagonal. For matrices generated from regular meshes, however, D is equal to 1. When iterative methods are used to solve the linear system (1), roundoff errors not only produce a loss of significant digits, but also slow down convergence.[7] Hence the end effect of the roundoff errors may depend not on $\hat{Cn}(K)$, but on the true condition number $Cn(K)$.

Condition Number

Let K and M denote the global (stiffness) and Gram (mass) matrices, respectively; then from Rayleigh's principle one readily obtains

$$(\lambda_N/\lambda_1)\lambda_1{}^M/\lambda_N{}^M \leq Cn(K) \leq (\lambda_N/\lambda_1)\lambda_N{}^M/\lambda_1{}^M \tag{8}$$

where λ_1 and λN are the extremal eigenvalues of $Kx_r - \lambda_r M x_r = 0$ $r = 1,2,\ldots,N$. The mass matrix M is, at least for uniform meshes of finite elements, well conditioned, its condition number being independent of N. Consider for instance the global mass matrix for a string. It has the form

$$M = \begin{pmatrix} 4 & 1 & & \\ 1 & 4 & 1 & \\ & 1 & 4 & 1 \\ & & & \ddots \end{pmatrix} \tag{9}$$

and by the Gerschgorin theorem it can readily be seen that the eigenvalues of M are enclosed between 2 and 6. Thus according to Eq. (8), $Cn(K)$ varies for uniform meshes as λ_N/λ_1. For a sufficiently large number of elements, λ_1 can be replaced for all practical purposes by its corresponding exact value. It can be shown[1] that the relative error in λ_N is of the order 1, and therefore for the present purposes λ_N can also be replaced by its corresponding exact value. By so doing, the study of the behavior of $Cn(K)$ has been reduced to the study of the dynamical behavior of the discretized structure.

From previous experience with eigen problems one concludes that:

a) For problems of the $2m$th order the condition number $Cn(K)$ of K varies as h^{-2m} or as Ne^{2m}, where Nes denotes the number of elements per side.

b) The condition number depends primarily on the nature of the problem and not on the order of the finite element. Therefore, it is expected that by using high-order elements greater total accuracy would be achieved in the finite element solution. This will be demonstrated numerically in the next section.

Numerical experiments

Higher order elements were applied to a beam problem to test the possibility of a reduction of the relative effect of roundoff error accumulation in the solution of the global system of algebraic equations. For comparison, two elements were used. One was denoted by $B(3)$, since the lateral deflection w was interpolated inside it by cubic polynomials ($p = 3$). The other was denoted by $B(5)$, since inside it w was interpolated by quintic polynomials ($p = 5$). From the numerical calculations it follows that the condition number can be expressed by $Cn(K) = 8Ne^4$ in the case where $B(3)$ elements were used and by $Cn(K) = 15Ne^4$ in the case where $B(5)$ elements were used. In any case the condition number varies only slightly with p.

To study the behavior of roundoff error effects for matrices generated by the $B(3)$ and $B(5)$ elements, the displacement of a cantilever beam loaded at the tip by a single force was calculated using Gauss elimination in single (24 bits, $s = 7.2$) precision. Both with $B(3)$ and $B(5)$ elements there are no discretization errors, the sole errors being numerical. It has been found that the error δx_t in the tip deflection x_t (this error is proportional to the error in the energy) can be expressed by

$$\log(\delta x_t/x_t) = -7.53 + 1.16 \log(Cn) \qquad (10)$$

when the $B(3)$ element was used and by

$$\log(\delta x_t/x_t) = -8.01 + 1.16 \log(Cn) \qquad (11)$$

when the $B(5)$ element was used. This is remarkably close to estimate (5).

The fact that by using higher order elements one may be able to achieve greater accuracy is confirmed in Fig. 1. It shows the variation of the relative total error in the energy vs the number of degrees of freedom N for the case of a cantilever beam loaded by a quadratically varying load. Computation was carried out in single precision. Even though roundoff errors become dominant earlier with the $B(5)$ element (at $N = 12$) than with the $B(3)$ element (at $N = 22$), it was possible to obtain a greater total accuracy with the former element than with the latter.

These were one-dimensional experiments. The precise manner in which both c in $Cn(K) = ch^{-2m}$ and the roundoff errors depend on the dimension in higher dimensional problems is still to be determined.

References

[1] Fried, I., "Discretization and Round-Off Errors in the Finite Element Analysis of Elliptic Boundary Value and Eigenvalue Problems," Ph.D. thesis, May 1971, Dept. of Aeronautics and Astronautics, MIT.

[2] Bauer, F. L., "Optimal Scaling and the Importance of the Minimal Condition," *Information Processing 1962*, edited by C. M. Popplewell, North-Holland Publishing, Amsterdam, 1963, pp. 198–200.

[3] Melosh, R. J. and Palacol, E. L., "Manipulation Errors in Finite Element Analysis of Structures," CR-1385, Aug. 1969, NASA.

[4] Fitzgerald, B. K. E., "Error Estimates for the Solution of Linear Algebraic Systems," *Journal of Research*, National Bureau of Standards-B. Mathematical Sciences, Vol. 74B, Oct.-Dec. 1970, pp. 251–310.

[5] Kelsey, S., Lee, K. N., and Mak, C. K. K., "The Condition of Some Finite Element Coefficient Matrices," *Computer-Aided Engineering, Proceedings of the Symposium held at the University of Waterloo, Canada*, May 11–13, 1971, edited by G. M. L. Gladwell, pp. 267–283.

[6] Sluis, A. Van der, "Condition Equilibration and Pivoting in Linear Algebraic Systems," *Numerische Mathematik*, Vol. 15, 1970, pp. 74–86.

[7] Fox, L. R. and Stanton, E. L., "Development in Structural Analysis by Direct Energy Minimization," *AIAA Journal*, Vol. 6, No. 6, June 1968, pp. 1036–1042.

Calculation of a Slender Body Moving through Air at Supersonic and Subsonic Velocities

M. L. Wilkins*
University of California, Lawrence Radiation Laboratory, Livermore, Calif.

The finite difference solutions of the equations of gas dynamics formulated in Lagrange coordinates are capable of accuracies of 1 part in 10^3 for problems with a reasonable number of Lagrange zones. Finite difference techniques permit physical phenomena to be simulated in a time-dependent manner, starting from actual initial conditions. Also, there are no restrictions on models used to describe the behavior of materials. The main strengths of the Lagrange formulation are the ease of applying boundary conditions and high accuracy. The major limitation of the Lagrange formulation is that nearest neighbors of the Lagrange zones must remain nearest neighbors, except for specified lines of sliding.

In the calculations presented here, a slender body of revolution defined by a set of Lagrange coordinates moves into another set of Lagrange points that have associated with them an equation of state appropriate for air. The specified line of sliding is the set of Lagrange coordinates that define the exterior boundary of the figure of revolution.

The slender body moving in the air, which is initially at rest, is shown in Fig. 1. The equation of state used for air is described by a perfect gas, although any equation of state could be used. The velocity of the slender body is given in units of the sound speed of the undisturbed air ahead of the body (Mach number). The shocks develop automatically in the calculation by the artificial viscosity method originated by Von Neumann and Richtmyer.[1] The shock positions shown in Fig. 1 are plotted by a computer routine that scans the grid and plots a symbol at positions where the artificial viscosity is a maximun. A steady-state shock pattern is established after the body has completely entered the calculation region (Fig. 1b). A steady-state shock pattern means that shock angles with respect to the body remain constant in time although the shocks are sweeping through new material. At time $t = 200$ μsec, the body is programed to reduce velocity on a linear ramp with time. The shock angles are shown, in Fig. 1, c–f, to open as the velocity is reduced. The bow shock detaches when the velocity becomes subsonic (Fig. 1, d–f).

To a fixed observer, the bow and aft shocks are moving faster than the slender body for the subsonic velocities shown in Fig. 1, e and f. At $t = 800$ μsec the body is programed to increase velocity on a linear ramp with time (Fig. 1, g and h). New bow and aft shocks are formed. Figure 1g shows

Received April 12, 1971; revision received June 3, 1971. This work was performed under the auspices of the U.S. Atomic Energy Commission.
Index category: Subsonic and Transonic Flow.

* Physicist.