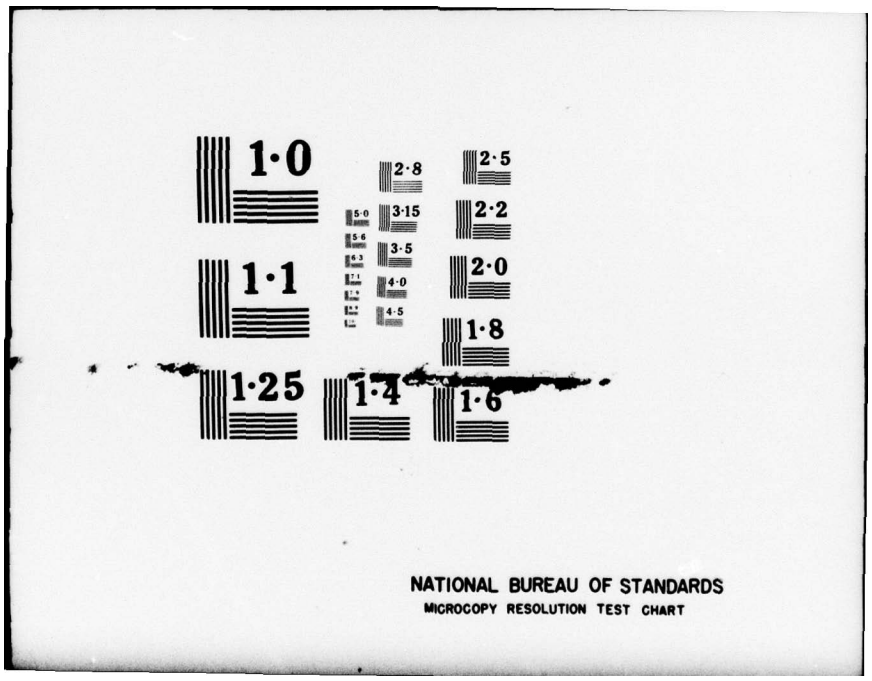
AD-A039 309 CALIFORNIA UNIV BERKELEY ELECTRONICS RESEARCH LAB F/G 12/1
ROUNDOFF ERROR IN THE SOLUTION OF FINITE ELEMENT SYSTEMS,(U)
JUN 76 B PARLETT N00014-76-C-0013
UNCLASSIFIED ERL-M593 NL

| OF |
ADA
039309

END

DATE
FILMED
6-77


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
5·0
5·6
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

(12) FC

(6) ROUNDOFF ERROR IN THE SOLUTION OF FINITE ELEMENT SYSTEMS

by

(10) Beresford Parlett

(14) Memorandum No. ERL-M593

(11) 22 June 1976 (12) 22p.

Contract No. N00014-76-C-0013
NR-044-324
(15)

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

12755D

ROUNDOFF ERROR IN THE SOLUTION OF FINITE ELEMENT SYSTEMS[†]

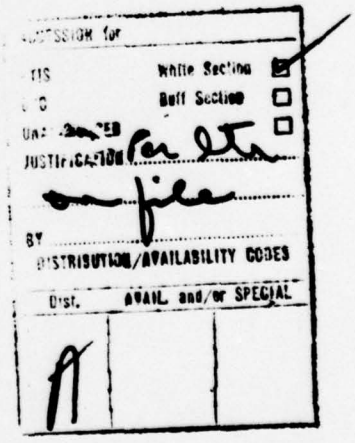
Beresford Parlett

Department of Mathematics and
Computer Science Division
Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley

An address to the U.S.-Germany Symposium entitled
"Formulations and Computational Algorithms in Finite Element Analyses"
M.I.T., August 1976

Abstract

The paper considers the effect of roundoff error in the solution process of equilibrium FEM equations The use of the equivalent perturbation matrix, the misuse of scaling, and the estimation of the condition number are discussed.



code 432
0414-324

[†]Research sponsored by Office of Naval Research Contract N00014-76-C-0013.

Contents

1. Introduction

The Finite Element Method (FEM) is a successful blend of clever engineering, fine mathematics, and ingenious programming. When put to use, the product seems to be slightly tarnished by roundoff errors. It would be nice to overlook them.

Roundoff afflicts (1) the numerical integration used to compute the element stiffness matrices, (2) the assemblage of the global stiffness matrix, (3) the load, or force, vector f which forms the right hand side of the familiar equation of virtual work, (4) the numerical solution of the equation to obtain a vector u. Consequently this u will not represent the approximate displacement of FEM theory for which all those nice error bounds have been proved.

However the same warning applies to the exact solution of the system of equations delivered by the exact evaluation of the numerical integration formulas. Indeed these formulas usually make a greater change in the larger elements of the true stiffness matrix than do all the sources of roundoff in (4) put together. The big difference is that the changes in the matrix elements due to numerical integration are highly correlated whilst those due to roundoff in (4) are not.

At present the favorite method for solving the system of linear equations generated by the FEM is a refined version of the familiar elimination process. As a result of twenty-five years of intensive study the problem of the influence of roundoff error is well understood. The main facts will be presented here, with emphasis on the positive definite systems that occur in equilibrium problems.

2. Basic Misconceptions About Gaussian Elimination

The remarks in this section are not confined to finite element problems.

> MISCONCEPTION 1. The tiny roundoff errors which occur in each of the basic arithmetic operations required to solve systems with hundreds of unknowns by Gaussian elimination may accumulate enough to spoil the results completely.

This fear was widespread in the late 1940's.

REPLY. When roundoff does lead to unacceptable error in the solution this misfortune is attributable to a mere handful (perhaps only one) of the errors, those which happen to occur at sensitive places. This is in contrast to the propagated error which afflicts initial value problems for differential equations where accumulation does occur.

The standard example is the Hilbert segment ($h_{ij} = 1/(i+j-1)$) which is discussed in detail in [4]. The surprising fact is that the initial errors made in rounding off 1/3, 1/6, 1/7, etc. to single precision do more damage to the solution than all the errors committed during the subsequent triangular factorization and backsolving. This result is typical for graded positive definite matrices. A matrix is graded if its elements decline (or do not increase) as their indices increase,

$$|a_{ij}| \leq |a_{k\ell}| \quad \text{if} \quad i \geq k, \quad j \geq \ell \, .$$

> MISCONCEPTION 2. In Gaussian Elimination small pivots cause large errors in the computed triangular factors.

REPLY. The actual size of a pivot is irrelevant. A small one may or may not accompany large errors. By scaling perversely any pivot can be made to look small. An initial scaling of a positive definite matrix so that all diagonal elements are unity does not prevent the occurrence of completely harmless small pivots. For example,

$$\begin{pmatrix} 1 & .98 & .01 \\ .98 & 1 & .01 \\ .01 & .01 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & .04 & -.01 \\ 0 & -.01 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & .04 & 0 \\ 0 & 0 & .96 \end{pmatrix}$$

MISCONCEPTION 3. Large multipliers in Gaussian Elimination provoke large errors in the factorization.

REPLY. As in No. 2 the multipliers are not the relevant quantities. For arbitrary matrices large multipliers may or may not accompany large errors in the computed triangular factors. However for symmetric positive definite matrices large multipliers cannot be blamed for large errors because in any subsequent computation the large multiplier is always multiplied by other elements so that the product is less than the corresponding diagonal element. For example,

$$\begin{pmatrix} 10^{-8} & 10^{-4} \\ 10^{-4} & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 10^{+4} & 1 \end{pmatrix} \begin{pmatrix} 10^{-8} & 0 \\ 0 & 2-\underline{(10^{4})(10^{-4})} \end{pmatrix} \begin{pmatrix} 1 & 10^{4} \\ 0 & 1 \end{pmatrix}$$

3. The Equivalent Perturbation Matrix (EPM)

Let us consider the standard Gaussian elimination process when applied to a sparse positive definite $n \times n$ matrix K. Some previous experience with the matrix interpretation of elimination is assumed.

From the original $K = K^{(1)}$ the algorithm implicitly derives a sequence of reduced matrices $K^{(j)}$, $j = 2,\ldots,n$, where $K^{(j)}$ is of order $n+1-j$, is also positive definite, and is written over $K^{(j-1)}$. Upon completion the array K holds an upper triangular matrix U which is related to the $K^{(j)}$ by

$$(1) \qquad u_{ij} = (K^{(i)})_{ij}, \quad \text{for } j = i,i+1,\ldots,n .$$

The rows of $K^{(i)}$ run from i to n to conform to their actual positions in the array K. The multipliers may have been discarded but, in any case they are the nontrivial elements of a unit ($\ell_{ii} = 1$) lower triangular matrix L. In exact arithmetic

$$(2) \qquad U = DL^T .$$

Because of roundoff errors LU (the exact product) does not equal K and the so-called equivalent perturbation matrix E is defined by

$$(3) \qquad K - E = LU .$$

Note that the exact triangular factors of K are completely ignored. E can be computed but all we really want to know is that E is, in some sense, small compared to K.

It would be nice to show that $|e_{ij}/k_{ij}|$ is tiny for all i, j but experience with "fill in" shows that this is not true. Interestingly enough it is true for the Hilbert segment referred to in an earlier section. It will be shown later that $|e_{ii}/k_{ii}|$ is always tiny for FEM problems.

It can happen that roundoff destroys positive definiteness (i.e., $K - E$ is not positive definite) but this property is so desirable that many codes will make small perturbations in diagonal elements of K to ensure that every element of D, the diagonal part of U, is actually positive. We shall assume that D is positive in this paper.

In exact arithmetic the reduced matrices are related as follows: partition $K^{(j)}$ as indicated, δ_j is its (1,1) element,

$$(4) \qquad K^{(j)} \equiv \begin{pmatrix} \delta_j & u_j^T \\ u_j & M_j \end{pmatrix} ; \qquad K^{(j+1)} \equiv M_j - u_j \delta_j^{-1} u_j^T .$$

It follows from (4) that, in exact arithmetic, for $i = j+1,\ldots,n$

$$(5) \qquad k_{ii}^{(j+1)} \leq k_{ii}^{(j)} \qquad \text{(equality only if } u_{ij} = 0) .$$

Under the assumption that D remains positive (5) will also hold in practice. It is this monotonic decrease in each diagonal element which makes it unnecessary to rearrange rows and columns. The argument is worth reviewing. The dreaded element growth is defined as

$$(6) \qquad g_n \equiv \left| \max_{m,i,j} k_{ij}^{(m)} / \max_{\alpha,\beta} k_{\alpha\beta}^{(1)} \right| .$$

For positive definite matrices

$$(7) \qquad |k_{ij}| \leq \sqrt{k_{ii} k_{jj}} \leq \max(k_{ii}, k_{jj})$$

and so, by (5),

$$(8) \qquad g_n = \left(\max_{m,i} k_{ii}^{(m)} / \max_{\alpha} k_{\alpha\alpha} \right) \leq 1 .$$

Note that one permutation of K (to $P^T K P$) might yield a smaller E (term by term) than another but all permutations produce equally satisfactory E's in the sense that $\|E\|/\|A\|$ is tiny. This possibility has not received much attention because, when used at all, permutations are selected for the more important goal of keeping down the number of nonzero elements as the reduction proceeds.

As long as no overflow occurs the elements of L may be arbitrarily large: it is the size of the $k_{ij}^{(m)}$, $m = 1,\ldots,\min(i,j)$, which affect e_{ij} as we now show.

The History of the (i,j) Element, $i \le j$.

A typical element k_{ij} will be modified in only a few of the $n-1$ steps of the reduction. Let the m^{th} step be one of them. First the multiplier ℓ_{im}, $m < i$, is calculated. The division $k_{mi}^{(m)}/k_{mm}^{(m)}$ will not be done correctly but the error affects e_{im} and not e_{ij}. The key calculation replaces $k_{ij}^{(m)}$ by $k_{ij}^{(m+1)}$ which satisfies

$$k_{ij}^{(m+1)} = k_{ij}^{(m)} - \ell_{im}k_{mj}^{(m)} - e_{ij}^{(m)} \tag{9}$$

where $e_{ij}^{(m)}$ is the error incurred in the multiplication and the subtraction. In order to relate u_{ij} $(= k_{ij}^{(i)})$ to k_{ij} one writes down (9) for $m = 1,2,\ldots,i-1$ and adds:

$$\begin{aligned} k_{ij}^{(2)} &= k_{ij}^{(1)} - \ell_{i1}k_{1j} - e_{ij}^{(1)}, \\ k_{ij}^{(3)} &= k_{ij}^{(2)} - \ell_{i2}k_{2j}^{(2)} - e_{ij}^{(2)}, \\ &\cdots\cdots \\ k_{ij}^{(i)} &= k_{ij}^{(i-1)} - \ell_{i,i-1}k_{i-1,j}^{(i-1)} - e_{ij}^{(i-1)}. \end{aligned} \tag{10}$$

Cancelling $\sum_{m=2}^{i-1} k_{ij}^{(m)}$ from each side yields

$$k_{ij}^{(i)} = k_{ij}^{(1)} - \sum_{m=1}^{i-1} \ell_{im}k_{mj}^{(m)} - \sum e_{ij}^{(m)}. \tag{11}$$

The crucial observation is that since $k_{mj}^{(m)} = u_{mj}$ (11) can be rewritten in the illuminating form

$$(12) \qquad (LU)_{ij} = k_{ij} - e_{ij} , \qquad e_{ij} = \sum_m e_{ij}^{(m)} .$$

This is equivalent to equation (3) which defined E. It shows that e_{ij} is simply the sum of the errors made in modifying the (i,j) element of K.

Note that $e_{ij}^{(m)} = 0$ if either $k_{mi}^{(m)} = 0$ or $k_{mj}^{(m)} = 0$. For sparse matrices the sum over m contains few nonzero terms compared with the order n.

It remains to consider $e_{ij}^{(m)}$ and the actual roundoff errors made in the calculation. A careful analysis is given in [4, p. 100] and it is not relevant to our purposes to reproduce that material. Instead of the true product $\ell_{im}u_{mj}$ the algorithm computes $f_{ij}^{(m)}$ satisfying

$$(13) \qquad f_{ij}^{(m)} = \ell_{im}u_{mj} + \mu_{ij}^{(m)}\ell_{im}u_{mj} .$$

Instead of the difference $k_{ij}^{(m)} - f_{ij}^{(m)}$ the algorithm computes

$$(14) \qquad k_{ij}^{(m+1)} = k_{ij}^{(m)} - f_{ij}^{(m)} - \alpha_{ij}^{(m)}k_{ij}^{(m+1)}$$

where $\mu_{ij}^{(m)}$ and $\alpha_{ij}^{(m)}$ are complicated but satisfy the simple bounds

$$(15) \qquad |\mu_{ij}^{(m)}| \leq \varepsilon , \qquad |\alpha_{ij}^{(m)}| \leq \varepsilon$$

where ε is the relative precision of the machine; for current computations $\varepsilon \leq 10^{-14}$. In passing we must acknowledge that (14) does not hold for all North American computers; the appropriate modifications complicate the analysis without changing its essential features. On substituting (13) into (14) and comparing with (9) the desired expression is obtained

$$(16) \qquad e_{ij}^{(m)} = \mu_{ij}^{(m)}\ell_{ij}u_{mj} + \alpha_{ij}^{(m)}k_{ij}^{(m+1)} .$$

Equations (12) and (16) give an <u>exact</u> formula for e_{ij}, no approximations have been made. If $k_{ij}^{(m+1)}$ is huge compared to k_{ij}, say $10^{14}\,k_{ij}$, then

$e_{ij}^{(m)}$ will (except on rare occasions) be as big as k_{ij}.

Before proceeding we note:

Case 1: $f_{ij}^{(m)} = 0$. Then $e_{ij}^{(m)} = 0$.

Case 2: $k_{ij}^{(m)} = 0$, fill-in. Then $\alpha_{ij}^{(m)} = 0$.

Case 3: Exponent($k_{ij}^{(m)}$) = Exponent($f_{ij}^{(m)}$), cancellation. Then $\alpha_{ij}^{(m)} = 0$.

From (16), (12) and (15) comes a reasonable bound on e_{ij}, $i \leq j$, namely

$$|e_{ij}| \leq \varepsilon\{\sum_{m}^{i-1} |\ell_{im}u_{mj}| + \sum_{m}^{i-1} |k_{ij}^{(m+1)}|\} \tag{17}$$

where the sum is over all m not corresponding to Case 1. If K has bandwidth $2w+1$ then the sum has at most $w-j+i$ nonzero terms. The matrix E is not quite symmetric because of the final division in "$\ell_{ji} = u_{ij}/u_{ii}$"; it is necessary to add $\varepsilon|u_{ij}|$ to the bound in (17) to account for it, yielding

$$|e_{ji}| \leq \varepsilon\{\sum_{m}^{i} |\ell_{im}u_{mj}| + \sum_{m}^{i-1} |k_{ij}^{(m+1)}|\} . \tag{18}$$

Many inferences can be drawn from (17) and (18). This is done in the next section.

Roundoff error also afflicts the forward elimination and the backsubstitution required to complete the solution of the equations. However these errors are dominated by those in the triangular factorization. To incorporate them into the analysis one writes

$$\begin{aligned} (L + \delta L)y &= f + \delta f , \\ (U + \delta U)u &= y \end{aligned} \tag{19}$$

where δL and δU have the desirable property that they are tiny compared to L and U, element to element. Consequently

$$[K - E + (\delta L)U + L(\delta U) + (\delta L)(\delta U)]u = f + \delta f \ . \tag{20}$$

Having said this we shall confine our remarks to E.

To sum up

> The effect of all the roundoff errors in the direct solution of $Kv = f$ is to produce a vector u which satisfies the equation
>
> $$(K + \delta K)u = f + \delta f$$
>
> with δK, δf given by (20), (19), (16), and (12). ∎

The advantage of this formulation is that it puts the errors in the solution process on the same footing as the other three parts of the FEM method which are afflicted by roundoff and were mentioned in the introduction. Each of these effectively makes a perturbation of K, of f, or of both. Now it is reasonable to ask, for example, whether the errors made in assembling the stiffness matrix are more damaging than those made in reducing the equations to triangular form or less so.

4. Implications of the Equivalent Perturbation Formulation

The stiffness matrix K is symmetric positive definite and of bandwidth $2w+1$. Not all zero elements within the band will be filled in during the reduction process. Various inferences can be drawn from the form of the expressions (12) and (16) for the elements of E.

A. Each element e_{ij} within the band is a sum of at most $2(w-j+i)$ roundoff errors. So much for the accumulation effect which was so dreaded

in the 1940's.

B. The Hilbert segment phenomenon is explained as follows. It happens that all the elements decay during elimination. So the sum in (17) is dominated by the first term (m = 1); $|e_{ij}| \doteq \varepsilon|\ell_{i1}u_{1j} + k_{ij}^{(2)}| \doteq \varepsilon k_{ij}$, since $k_{ij}^{(2)} = k_{ij} - \ell_{i1}u_{1j}$. This is the size of the error incurred when rounding numbers to working precision.

C. In general, whenever the (i,j) element shrinks below its original value and remains below it then e_{ij} is of the order of a unit in the last place of k_{ij}.

D. The elements ℓ_{im} and u_{mj} $(= d_m\ell_{jm})$ do not occur separately in the analysis. In the absence of overflow or underflow d_m (a pivot) may be tiny or ℓ_{im} (a multiplier) may be huge. Only their product counts and, in the positive definite case, $|\ell_{im}u_{mj}| \leq (k_{ii}k_{jj})^{1/2} \leq \max_i k_{ii}$.

E. Bunch [ ] has given exact, but complicated expressions for the number of nonzero terms in the sum in (12), (18), and (19), in terms of the number of nonzeros in each row of K. Our cruder expression (w-j+i) is not ridiculously pessimistic, see H.

F. For full positive definite matrices in which the reduced matrices $K^{(j)}$ remain positive definite Wilkinson [11] shows

$$\|E\| \leq 2.5\, n^{3/2}\varepsilon\|K\|$$

for the spectral norm, $\|M\| = \sqrt{\lambda_{max}(M^TM)}$. Such results throw away information in the sense that there are many matrices E which satisfy this inequality but could not possibly be equivalent perturbations accounting for errors in triangular factorization. The reason for being content with such norm bounds is given in G below.

G. Equivalent perturbations δK are very useful for specialists who wish to compare methods. However the user wishes to know the effect of these

changes on the solution. Perturbation Theory (see [4]) yields

Theorem. Let $Ax = b$, $(A+\delta A)(x+\delta x) = b + \delta b$. If A and $A+\delta A$ are invertible and $\|A^{-1}\delta A\| < 1/2$ then

$$\frac{\|\delta x\|}{\|x\|} \leq 2\kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right) ,$$

where $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ is the condition number of A in the given norm.

Note that this is not simply an asymptotic result holding only for infinitessimal δA and δb.

Since $\|\delta K\|/\|K\|$ and $\|\delta b\|/\|b\|$ are small, provided that $n^2\varepsilon < 1$, the computed solution must be accurate unless $\kappa(A)$ is large. In the FEM case much better bounds than Wilkinson's general one can be given.

H. For positive definite matrices

$$|\ell_{im}u_{mj}| = |k_{im}^{(m)}k_{mj}^{(m)}/k_{mm}^{(m)}| \leq (k_{ii}^{(m)}k_{jj}^{(m)})^{1/2} \leq (k_{ii}k_{jj})^{1/2} ,$$
$$|k_{ij}^{(m+1)}| \leq (k_{ii}^{(m+1)}k_{jj}^{(m+1)})^{1/2} \leq (k_{ii}^{(m)}k_{jj}^{(m)})^{1/2} \leq (k_{ii}k_{jj})^{1/2} .$$

A simpler but cruder bound than (17) on e_{ij} for matrices of bandwidth $2w+1$ is

$$|e_{ij}| \leq 2\sum_{m}(k_{ii}^{(m)}k_{jj}^{(m)})^{1/2} \leq 2(w-j+i)(k_{ii}k_{jj})^{1/2} .$$

Thus e_{ij} is always tiny compared with the geometric mean of the associated diagonal elements. In fact E can be majorized, element by element, by the special matrix W defined by

$$w_{ij} = w_{ji} = \begin{cases} (w-j+i)(k_{ii}k_{jj})^{1/2} , & j-i \leq w , \\ 0 , & j-i > 0 . \end{cases}$$

Hence $\|E\| \leq 2\varepsilon\|W\| \leq 2\varepsilon w^2 \max_i k_{ii}$, using the spectral norm.

I. In one important respect, which is rarely mentioned, the equivalent perturbation δK from the solution process is <u>not</u> similar to the perturbations corresponding to numerical integration, formation of the global stiffness matrix, etc. The other sources all give small <u>relative</u> perturbations of the elements. With, fill-in during the elimination we have e_{ij}'s which are tiny relative to $\|K\|$ but enormous, or ∞, compared to k_{ij}. As shown in G these fill-in errors do not impair $\|\delta u\|/\|u\|$ but some of them do not have a reasonable FEM interpretation. They correspond to wiggles in the trial functions far outside the elements to which they belong.

There is a device for removing this anomaly. δK is split into a sum $\delta S + \delta T$ where $|(\delta S)_{ij}| \leq \varepsilon w |k_{ij}|$, say, and $|(\delta T)_{ij}| > \varepsilon w |k_{ij}|$. The equivalent perturbation relation can now be rewritten

$$(K + \delta S)u = f + \delta f - \delta T u \ .$$

All that remains is to attribute $(\delta Tu)_i$ to one or several of the larger elements among the $|k_{im}u_m|$ or to f_i itself. For example if $(\delta Tu)_i$ is less than the uncertainty in f_i, for each i, then δTu is best regarded as a further perturbation of f.

J. Often, but not always, the δK due to errors in the solution process are less than the inherent uncertainties in the k_{ij} or, more likely, in f. In such cases the computed u is as good as the data warrants and the only legitimate question concerns the inherent sensitivity of the solutions to perturbations in K from any source.

5. Estimating the Condition Number

Current programs for solving general nonsingular systems do not deliver any estimates of the accuracy of the computed solution. A major reason for this is that, at present, the cost of estimating the condition number is at least equal to the cost of the solution, either in time or in storage or both. This is a price which users appear to be unwilling to pay. No one knows how far the cost of estimating κ can be reduced.

For FEM problems Fried [5,6] have given very nice, almost computable, a priori bounds. In particular, for energies involving m^{th} derivatives, as $h \to 0$,

$$\kappa = O(h^{-2m})$$

where h is the minimal diameter of an element.

Here we mention some a posteriori techniques, each of which has its drawbacks. By convention we use the spectral norm. Roundoff will be ignored here.

At the completion of the calculation there is available

(i) α such that $\|K\| = \lambda_{max}(K) \le \alpha \le (2w+1) \max_i k_{ii}$

(ii) L, D such that $LDL^T = K$

The bandwidth of L is $w+1$, $I = (e_1, e_2, \ldots, e_n)$.

A. Inverse Iteration (a lower bound on κ). Solve $Lw = e$, $e_j = \pm 1$, and $Uw = w$. The sign is chosen to maximize $|w_j|$ for each j. Then $\|K^{-1}\| \ge \|v\|/\|e\| = \|v\|/\sqrt{n}$ and $\kappa(K) \ge (\max k_{ii})\|v\|/\sqrt{n}$.

Cost: $2nw$ multiplications, 1 n-vector for w and v.

B. Majorizing L^{-1} (an upper bound on κ). Suppose that $U = DL^T$. Then $\|K^{-1}\| \le \|D^{-1}\| \cdot \|L^{-1}\|^2$. The following observation [8] yields an inexpensive bound on $\|L^{-1}\|_\infty$ = max row sum of L^{-1}. Given any lower triangular

matrix T define a new lower triangular matrix $\tilde{T}$ by

$$\tilde{t}_{ii} = |t_{ii}| \; ; \quad \tilde{t}_{ij} = -|t_{ij}| \; , \quad i > j \; .$$

It can be shown that $\tilde{T}^{-1}$ is nonnegative and, element for element $T^{-1} \leq \tilde{T}^{-1}$. So let $e = (1,1,\ldots,1)^T$ and solve $\tilde{L}y = e$ for y. Then

$$\|L^{-1}\|_\infty \leq \|\tilde{L}^{-1}\|_\infty = \|y\|_\infty$$

$$\kappa_\infty(K) = \|K\|_\infty \|K^{-1}\|_\infty \leq \alpha \cdot \|y\|_\infty^2 / \min_j u_{jj}$$

Cost: nw multiplications, 1 n-vector for y.

Karasalo [8] points out that for a full triangular matrix these bounds can sometimes be pessimistic by a factor of 2^{n-1}. For banded LU factors the bound cannot be that crude. Anderson [1] offers more complicated but sometimes better bounds based on the same idea.

C. Iterating on the Residuals [4, p. 109]. This process was invented for improving the accuracy of computed solutions to ill-conditioned systems. If the matrix is too ill-conditioned for Gaussian elimination, with the given precision to arithmetic, to be meaningful then the sequence $\{x_i\}$ will almost certainly not settle down at all. That information is worth having and the process can be stopped after four or five steps. Now suppose that

convergence does occur. It is linear and the convergence factor depends on the condition number. If convergence is slow (6 or 7 steps) then κ must be large and can be estimated from

$$\kappa(A) \doteq 2^t \|x_3 - x_2\| / \|x_2 - x_1\|$$

where x_1, x_2, x_3 are the three previous iterates, x_3 being the latest. If convergence is immediate (1 or 2 steps) then κ may or may not be large.

Cost: $(2w+1)n$ double precision operations per iteration. However K and x_2 must be saved.

D. Diagonal Energy Criterion [7]. The user is not content to know E, even in principle, but needs to know the effect of E upon his system. One measure is the relative change in the strain energy for a displacement vector x, i.e.

$$x^T E x / x^T K x \ .$$

From (H) in Section 4 we have $|e_{ij}| \leq 2\varepsilon w \sqrt{k_{ii} k_{jj}}$. Using the Cauchy-Schwarz Inequality we have

$$(21) \qquad |x^T E x| \leq 2\varepsilon w \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{k_{ii} k_{jj}}\, x_i x_j = 2\varepsilon w \left(\sum_{i=1}^{n} \sqrt{k_{ii}}\, x_i \right)^2$$

$$\leq 2\varepsilon w \sum_{i=1}^{n} k_{ii} x_i^2 = 2\varepsilon w x^T K_d x \ .$$

By using standard statistical estimates for the expected value of $e_{ij}^{(m)}$ Irons replaces w by $\sqrt{w}$. He also suggests that wk_{ii} can be replaced by $\left(\sum_m k_{ii}^{(m)2} \right)^{1/2}$ where the sum extends over those m for which k_{ii} changes. However this estimate is relevant only to perturbations E attributable to

Gaussian elimination whereas (21) covers all perturbations satisfying $|e_{ij}| \leq 2\varepsilon w|k_{ij}|$. For example, it would be appropriate to replace $2w$ by p to account for the error in assembling the global stiffness matrix. Here p is the maximum number of elements meeting at a node.

When K is equilibrated so that $K_d = I$ and when x is an eigenvector for K's smallest eigenvalue λ_1 then Irons' bound is $\varepsilon\sqrt{w}/\lambda_1$ which is a very good estimate of $\varepsilon c(K)$ where c is the optimal condition number of K. In practice we have u, the computed solution, and the estimate becomes $\varepsilon\sqrt{w}\ u^T K_d u/u^T f$.

Cost: $3n$ operations, 1 vector to save f.

6. Optimal Condition Numbers are Irrelevant

A well known result, due to Bauer [2], establishes that Gaussian elimination, roundoff errors included, is invariant under scaling. More precisely, if the sequence of pivotal elements is fixed and if the scaling is done solely with the powers of the number base of the arithmetic unit, then the fractional parts of the all the floating point numbers which occur in the algorithm are unchanged if the calculation is repeated on a scaled system.

A convenient practical consequence of this fact is that, if one is prepared to risk that no under/overflow occurs, then there is no need to bother with scaling positive definite matrices. A more subtle corollary is that discussion of elimination should employ scaling invariant terms. Now the condition number with respect to a given norm,

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| \ ,$$

is certainly not invariant. The nearest qualified candidate for a measure of near singularity is the optimal condition number

$$c(A) = \min_{D} \kappa(DAD)$$

over all positive diagonal matrices.

This is a beautiful example of solving a problem by fiat! In fact it can be argued that c is simply the measure which makes the results look their best. There is an alternative approach.

For general linear systems, where it does influence pivot selection, scaling is regarded as a vexing and difficult problem. This is because the difficulty is not mathematical and cannot be resolved without more information. The task of solving a linear system in noisy arithmetic is not properly set until the user specifies a suitable norm for measuring the right hand side (the load vector) and a suitable, possibly different, norm for measuring the solution (the displacements). Ideally these norms should reflect the users interest in the problem and they should have the property that vectors of equal norm should represent states that have equal impact on the user.

Here lies the rub. Most users are not in the habit of formulating their knowledge of their problem in this formal manner. Nevertheless such specifications do resolve the difficulties.

Let $\|\cdot\|_d$ and $\|\cdot\|_\ell$ be the given norms for displacements and loads, respectively. The corresponding norm for a stiffness matrix K is then given by

$$\|K\|_d^\ell \equiv \max_{u\neq 0} \|Ku\|_\ell / \|u\|_d$$

and the correct norm of a flexibility matrix $K^{-1} = F$ is

$$\|F\|_\ell^d \equiv \max_{v\neq 0} \|Fv\|_d / \|v\|_\ell \ .$$

Finally the sensitivity of the model must be measured by

$$\kappa_{d,\ell}(K) = \|K\|_d^{\ell} \|K^{-1}\|_{\ell}^{d} \quad .$$

This measure is also scaling invariant by decree but in a more relevant way than is c. If K were to be scaled (in fact there is no point in doing so) this would correspond to a scaling of loads and displacements. This is fine but it would be meaningless not to change the norms in a corresponding way. When this is done properly the new condition number turns out to be identical to the old condition number, as it should if this measures the nearness to singularity of K for the user's purpose.

The formal verification of these remarks is left as an exercise for the interested reader.

Although this viewpoint clears up the difficulty of choosing the right condition number it is very much a pure mathematician's solution. The proper matrix norms are well defined but how are they to be constructed? We have no answer to that question.

Another reasonable scaling is to choose D so that the absolute uncertainty in each element of DKD or of Df is constant.

References

1. N. Anderson, "On Majorizing the Inverse of Triangular Matrices, with Applications," submitted to BIT.

2. F.L. Bauer, "Optimal Scaling and the Importance of the Minimal Condition," In Information Processing, ed. Popplewell, North Holland, 1962.

3. J.R. Bunch, "Analysis of Sparse Elimination," SIAM Numerical Analysis 11, 5 (1974), 847-873.

4. G.E. Forsythe and C.B. Moler, Computer Solution of Linear Algebraic Systems, Prentice Hall Inc., New Jersey, 1967.

5. I. Fried, "Condition of Finite Element Matrices Generated from Nonuniform Meshes," AIAA J. 10 (1971), 219-221.

6. I. Fried, "The ℓ_2 and ℓ_∞ Condition Numbers," In The Mathematics of Finite Elements and Applications, ed. J.R. Whiteman, Academic Press, New York, 1973.

7. B. Irons, "Round-Off Criteria in Direct Stiffness Solutions," AIAA J. 6 (1968), 1308-1312.

8. I. Karasalo, "A Criterion for the Truncation of the QR Decomposition Algorithm for the Singular Linear Least Squares Problem," BIT 4 (1974), 156-166.

9. J. von Neumann and H.H. Goldstine, "Numerical Inverting of Matrices of High Order," Bull. Amer. Math. Soc. 53 (1947), 1021-1099.

10. A. van der Sluis, "Condition, Equilibration, and Pivoting in Linear Algebraic Systems," Numer. Math. 15 (1970), 74-86.

11. J.H. Wilkinson, "A Priori Error Analysis of Algebraic Processes," Proceedings of the International Congress of Mathematicians, Moscow 1966, 629-640.