MATRIX ERROR ANALYSIS FOR ENGINEERS

R. A. Rosanoff* & T. A. Ginsburg**
North American Aviation, Inc.
Downey, California

Topics from the mathematics of matrix error analysis are reviewed. It is demonstrated that numerically unstable equations do arise in physically stable problems. It is shown that unstable equations result from an absence of information so that preconditioning techniques or optimum inversion routines inherently must be of limited value. The analysis indicates that a new set of equations is the cure for ill-conditioned matrices. Thus, a great need is indicated for routine measurement of matrix conditioning numbers associated with various patterns of formulation. Some patterns, which arise in practice, are shown.

INTRODUCTION

It is sometimes true that the results of an operation with a particular matrix are changed greatly by small perturbations in the matrix. We then say that the matrix is ill conditioned with respect to this operation. Quite obviously, ill conditioning can be a direct reflection of a physical instability. It is less obvious that a perfectly stable physical problem can be represented by numerically unstable equations. This paper is concerned with the recognition, anticipation, and avoidance of ill conditioning.

The paper is divided into three sections. The first section is devoted to the mathematics of error bounds. For matrix inversion, it is shown how the ratio of the extreme eigenvalues is the dominant factor in conditioning. Vector and matrix norms are reviewed. Conditioning numbers for matrices are defined in terms of these norms, and error bounds are shown. An error estimate based on these bounds is given, and the results of numerical experiments are shown which indicate that, at least for the test matrices, the error estimates are quite satisfactory.

In the second section of the paper, simple examples of several numerical methods are examined. There are three considerations in this section: (1) some common sources of ill-conditioned equations are identified and catalogued, (2) application of concepts of the first section are exemplified, and (3) patterns which are common to some badly behaved problems are identified.

In the final section of the paper, practical considerations and general guides to cost of error analysis are considered from the viewpoint of modern computing equipment and efficient matrix schemes, and conclusions are drawn.

MATHEMATICAL FRAMEWORK OF ERROR ANALYSIS

The Real and Computer Number Sets

This paper is intended to present an engineering point of view of matrix conditioning. There are excellent numerical analysts who have compared errors in specific algorithms

*Sr. Engineer Structures, Structural Methods, Space & Information Systems Division, North American Aviation, Inc.

**Sr. Engineer Structures, Criteria and Structural Development, Space & Information Systems Division, North American Aviation, Inc.

for matrix inversion or the solution of equations (References 1 to 3). Our interest centers more in the avoidance of ill-conditioned equations that in optimizing routines for their solution. Experience has indicated that it is the engineer, as the originator of the problem, who dominates the form of the solution. A major communication problem arises if a numerical analyst, or programmer, takes responsibility for numerical stability after the fundamental decisions about the calculation have been made. We see no better path to good equations than to provide the engineer with adequate numerical insight.

The analysis begins with a discussion of the number system. The complex numbers, the real numbers, and the rational numbers are infinite sets possessing some fifteen properties which together are required to constitute a field (Reference 4). The set of numbers available for digital machine computation, however, is only a portion of the real number set; it is not a field. In particular, the computer number set has the following properties:

1. It is finite and bounded. It is not dense. It is full of holes through which the validity of a calculation readily may be lost.

2. It does not have the property of closure on either addition or multiplication

$$2^{127} + 2^{127} = \text{Undefined}$$

$$2^{100} \cdot 2^{100} = \text{Undefined}$$

3. It does not have the property of associativity in addition

$$(e^{-10} + e^{10}) - e^{10} = 0 \neq e^{-10} + (e^{10} - e^{10})$$

4. It does not have a unique zero

$$e^{10} + e^{-10} = e^{10}$$

5. It does not have the property of associativity in multiplication

$$2^{100} \cdot (2^{100} \cdot 2^{-100}) = 2^{100} \neq (2^{100} \cdot 2^{100})\, 2^{-100} = \text{Undefined}$$

Thus the algebra of the real numbers, which assumes the numbers to be elements of a field, does not really apply to the computer number set. To be specific, consider the important idea of linear independence. For two functions, say f(x) and g(x), if there exist constants a and b such that

$$af(x) + bg(x) = 0 \qquad 0 \leq x \leq \ell$$

$$a \neq 0 \quad \text{or} \quad b \neq 0$$

the functions are linearly dependent. If not, they are linearly independent. In the real number set, this is a clear yes or no situation. Two functions cannot be "more linearly independent" than two other functions. Zero is a unique number. The two functions cosh(x) and sinh(x), for example, are distinct. Specifically, their difference is precisely e^{-x} no matter how large x may be. Further, e^{-x} is never zero. Compare this situation in the real number set to the situation in the computer number set. In the latter, zero is not unique. Consider the two numbers u = cosh(x)-sinh(x) and $v = \ell^{-x}$. They are most certainly not equal except at x close to zero. When x = 1, u will already be one digit shorter than v. At x = 5, from the 6 place

tables (Reference 5), u = 74.2099 - 74.2032 = .0067 and v = .006738. Thus, although linear independence has not completely broken down, there are two fewer digits in the linear combination cosh (5) - sinh (5) than there are in ℓ^{-5}. Since this loss of digits will rapidly become total, it is asserted that the property of linear independence in a finite number system is not a clean cut dichotomy. In this example one may think of the functions e^{x} and e^{-x} as more linearly independent than cosh(x) and sinh(x). An analogous situation exists in matrices.

Conditioning and Eigenvalues

It will be shown in this section how the eigenvalues and eigenvectors are related to ill conditioning. Since this is the fundamental consideration of this paper, the reader's attention is directed to the ratio of the extreme eigenvalues of a matrix, $|\lambda_{max}|/|\lambda_{min}|$. This is called the conditioning number. It will be shown that the logarithm to the base ten of the conditioning number estimates the maximum number of significant figures lost in inversion or in the solution of simultaneous equations.

Matrices that are symmetrical, or that have distinct eigenvalues may be written as a linear combination of rank 1 matrices as: (References 6 and 7)

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{u}_i^T \qquad \text{and} \qquad \mathbf{A}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{u}_i^T \ldots . \tag{1}$$

$$\mathbf{v}_i^T \mathbf{u}_i = 1 \qquad i = 1, n$$

where

λ_i = i^{th} eigenvalue

$\mathbf{v}_i$ = i^{th} eigenvector of the matrix A

$\mathbf{u}_i$ = i^{th} eigenvector of the matrix A^T

$\mathbf{A}^T$ = transpose of the matrix A.

Notice how Expressions 1 expose the nature of possible difficulties. The first expression shows that the rank 1 matrices $\mathbf{v}_i \mathbf{u}_i^T$ of the eigenvectors enter the matrix **A** in amounts proportional to their respective eigenvalues. Just as e^{-x} becomes weakly represented in the hyperbolic functions when x becomes large, the lower modes of **A** become weakly represented as the ratio of the extremal eigenvalues becomes large. Specifically, as a first approximation, for each power of 10 in the ratio $|\lambda_{max}|/|\lambda_{min}|$ the lower mode will lose about 1 decimal digit in a finite computer number set representation of the matrix. But now look at $\mathbf{A}^{-1}$. The lower mode of **A** is the upper mode of $\mathbf{A}^{-1}$, because the coefficients of the linear combination, the eigenvalues, are inverted. Hence inverting matrices without some feel for their conditioning is directly analogous to determining e^{-x} as a linear combination of hyperbolic functions without consideration of the range of x.

The importance of this concept justifies a numerical example. Admittedly this is a contrived example (and not badly conditioned). The second portion of the paper however, will illustrate how some badly conditioned problems arise in practice. Consider a two by two matrix and an eight digit approximation to it.

$$A = \begin{bmatrix} \frac{1}{9} & \frac{1}{10} \\ \frac{1}{10} & \frac{1}{11} \end{bmatrix} \approx \begin{bmatrix} .11111111 & .10000000 \\ .10000000 & .090909090 \end{bmatrix}$$

Eight-digit approximations to the characteristic values and vectors of the true matrix **A** are:

$$\lambda_1 = .20151896 \qquad v_1 = u_1 = \begin{bmatrix} .74178794 \\ .67063452 \end{bmatrix}$$

$$\lambda_2 = .0005012437 \qquad v_2 = u_2 = \begin{bmatrix} .67063452 \\ -.74178794 \\ -.74178794 \end{bmatrix}$$

$$\frac{\lambda_1}{\lambda_2} = 402.0379 = 10^{2.604}$$

As stated above, the matrix A may be written as

$$A = \sum_{i=1}^{n} \lambda_i v_i u_i^T = \lambda_1 v_1 u_1^T + \lambda_2 v_2 u_2^T$$

$$= .20151896 \begin{bmatrix} .74178794 \\ .67063452 \end{bmatrix} \{ .74178794 \;\; .67063452 \}$$

$$+ .0005012437 \begin{bmatrix} .67063452 \\ -.74178794 \end{bmatrix} \{ .67063452 \;\; -.74178794 \}$$

$$= \begin{bmatrix} .11088567 & .10024935 \\ .10024935 & .090633285 \end{bmatrix} + \begin{bmatrix} .00022543467 & -.00024935298 \\ -.00024935298 & .00027580902 \end{bmatrix}$$

$$= \begin{bmatrix} .11111111 & .099999997 \\ .099999997 & 090909094 \end{bmatrix}$$

Notice that in forming this eight-digit approximation to the matrix, the componet matrix $\lambda_2 v_2 u_2^T$ which has three leading zeroes in its elements, was truncated to about five digits. Thus an eight-digit representation of the matrix A contains about five digits of information about the rank 1 matrix $v_2 u_2^T$.

Now consider A^{-1} . It is formed from the rank 1 matrices as

$$A^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} v_i u_i^T = \frac{1}{\lambda_1} v_1 u_1^T + \frac{1}{\lambda_2} v_2 u_2^T$$

$$= \begin{bmatrix} 2.7305090 & 2.4685944 \\ 2.4685944 & 2.2318030 \end{bmatrix} + \begin{bmatrix} 897.26951 & -992.46859 \\ -992.46854 & 1097.7681 \end{bmatrix}$$

$$= \begin{bmatrix} 900.00002 & -990.00000 \\ -990.00000 & 1099.9999 \end{bmatrix} \approx \begin{bmatrix} 900 & -990 \\ -990 & 1100 \end{bmatrix}$$

Notice that the rank 1 matrix $v_2 u_2^T$ which was only available to about five digits in the eight-digit approximation to A, is the largest component of A^{-1}. One should expect that five digits would be about the most one could obtain by numerically inverting the approximate matrix.

The true inverse can be obtained using rational number arithmetic and is shown to the right of the approximation sign above. Using eight-digit arithmetic the approximate matrix was inverted, giving:

$$\begin{bmatrix} .11111111 & .10000000 \\ .10000000 & .09090909 \end{bmatrix}^{-1} = \begin{bmatrix} 900.00089 & -990.00099 \\ -990.00099 & 1100.0011 \end{bmatrix}$$

The poorest terms in this approximate inverse are the off diagonal terms which have barely six significant digits. We have $\log_{10} |\lambda_{max}| / |\lambda_{min}| = \log_{10} 402 \cdot 0379 = 2 \cdot 604$. Our analysis would lead us to believe that the matrix $v_2 u_2^T$ would be represented to about 2.6 digits less than the number of digits in the arithmetic. We should expect from this analysis that the approximate inverse would be limited to 8.0 - 2.6 = 5.4 good digits. Thus for the two-by-two example shown, this estimator is amply realistic for engineering purposes. Experience shows it to be usually a digit on the conservative side.

For the real number set, matrices are singular or not singular. The usual criterion is that matrices are not singular if their determinant is nonzero. Since the determinant of a matrix is the product of its eigenvalues, a nonzero determinant is equivalent to no-zero eigenvalues. In the finite, non-dense computer number set, the first measure of ill conditioning—which is a tendency toward singularity—is the smallness of the smallest eigenvalue which, from Expression 1, is measured relative to the largest eigenvalue.

Additional experience with dynamic analyses of free-free systems has provided further insight. In such systems the rigid-body modes are associated with the eigenvalue zero. Thus, in the framework in which we are thinking, these modes are contained in the matrix to zero significant figures. The eigenvalues and vectors in our dynamic programs are obtained by a threshold Jacobi routine. We were encouraged in the concept that the log of the conditioning number measures the loss of digits in the lower modes when we repeatedly observed that the "zero" eigenvalues obtained were a bit less than 10^{-8} times the dominant eigenvalues.

One day an exception was encountered which turned out to be very instructive. A conditioning number of 10^{16} was observed. The problem was the modal analysis of a one-dimensional

free-free body, analysed with one axial, transverse and rotational degree of freedom at each mass point. The axial modes were not coupled to the transverse. This meant that the axial submatrix was independent. The smallest characteristic value was the axial rigid-body mode, and the conditioning number of the axial submatrix was about 10^8.

Before leaving this analysis, consider the inversion of a diagonal matrix. For a diagonal matrix, the modes are the identity matrix. The products $\mathbf{v}_i \mathbf{u}_i^T$ are each matrices with one nonzero element. Thus we have n one-element uncoupled matrices which may be inverted separately without loss of precision. This has led people to look to strong diagonals as a measure of conditioning. It will be shown later that this is fallacious. At this moment, it is emphasized that the conditioning number will always bound the error, and that the bound will be close except when the modes are weakly coupled. The conclusion is that the conditioning number is a good estimator for error analysis.

Conditioning and Norms

In this section we develop more general and rigorous bounds. First notice that the real number system has an ordering relationship. That is to say, if a and b are real numbers and $a \neq b$, then either $a>b$ or $b>a$. Complex numbers, vectors, and matrices do not have this property.

For error analysis (and also for study of convergence problems) it is convenient to define measures of magnitude which do have an ordering relationship. The most common choice for vectors and for complex numbers is the Euclidean length $\sqrt{\Sigma x_i^2}$. This is an example of a vector norm. Both in analysis and for computational convenience, it is sometimes useful to consider other norms. See Reference 3 for a more detailed discussion.

Vector norms, written $\|v\|$, are defined by their properties as follows:

$$\text{I} \quad \|v\| > 0 \qquad v \neq 0 \qquad \|0\| = 0$$

$$\text{II} \quad \|cv\| = |c|\,\|v\| \qquad c \text{ a scalar}$$

$$\text{III} \quad \|u + v\| \leq \|u\| + \|v\|$$

Let $v = x - u$ Then, from III

$$\|x\| = \|v + u\| \leq \|v\| + \|u\|$$

$$\|v\| = \|x - u\| \geq \|x\| - \|u\|$$

Matrix norms are also defined by the same three properties plus the fourth property.

$$\text{IV} \quad \|AB\| \leq \|A\| \cdot \|B\|$$

This fourth property can be used because the product of compatible matrices is again a matrix.

There is a family of vector norms called the Holder norms for which

$$\|x\|_K = \left(\sum_{i=1}^{n} |x_i|^K \right)^{1/K}$$

For K = 1 this is seen to be the sum of the absolute values of the elements. For K = 2 the Holder norm is the Euclidean length. For K infinite, the Holder norm is interpreted as the magnitude of the largest element. These three quantities do have the properties of norms.

Frequently matrices and vectors occur together, making it convenient to consider matrix norms which are rationally connected with vector norms. We therefore define:

$$\| A \|_K = \max_{\| x \|_K = 1} \| A x \|_K$$

The matrix norm is then said to be subordinate to the vector norm. For the three common Holder norms these become (Reference 3):

$$\| A \|_1 = \max_j \sum_{i=1}^{n} | a_{ij} |$$

$$\| A \|_2 = (\max \text{ eigenvalue } A^T A)^{1/2}$$

$$\| A \|_\infty = \max_i \sum_{j=1}^{n} | a_{ij} |$$

$\| A \|_2$, is called the spectral norm of **A**. The Euclidean norm defined by

$$\| A \|_E = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{1/2}$$

satisfies the inequalities

$$\| A x \|_2 \leq \| A \|_E \| x \|_2$$

$$\| A \|_2 \leq \| A \|_E \leq n^{1/2} \| A \|_2$$

Also we have

$$| \lambda(A)_{max} | \leq \| A \|_K \qquad \text{for all } K$$

Next, conditioning numbers and error bounds in terms of these norms will be established. First, however, let us see why this should be done, since it has already been shown that extreme eigenvalues provide a good estimator. In the first place, matrices can only be represented in the form of Expression 1 if they are symmetric or have distinct eigenvalues. In addition, for some matrices it will be much easier to obtain these norms than the extreme eigenvalues.

Following References 1 and 2, consider the equation:

$$A x = b$$

Because of the finite nature of the computer number set, representation of the matrix **A** will fail by an error matrix **E**. The equation which is actually solved is:

$$(A + E)(x + \delta x) = b$$

We seek bounds for the error vector δx. Define $S = A^{-1} E$; then multiplying by A^{-1}

$$(I + S)(x + \delta x) = A^{-1} b = x$$

and

$$\delta x = ((I + S)^{-1} - I)x$$

and for any norm

$$\| \delta x \| = \| ((I + S)^{-1} - I) x \| \tag{2}$$

Let

$$G = (I + S)^{-1} \text{ Then:}$$

$$\begin{aligned} (I + S) G &= I = G + SG \\ I - G &= SG \\ \| I - G \| &= \| SG \| \leq \| S \| \cdot \| G \| \end{aligned} \tag{3}$$

Also

$$I = G + SG = G - (-SG)$$

Then, except for the Euclidean matrix norm we have scalar inequality

$$\| I \| = 1 \geq \| G \| - \| SG \| \geq \| G \| - \| S \| \cdot \| G \|$$

and

$$\| G \| \leq \frac{1}{1 - \| S \|} \tag{4}$$

Then from Equations 2 and 3

$$\| \delta x \| \leq \| I - G \| \cdot \| x \| \leq \| S \| \cdot \| G \| \cdot \| x \|$$

So that from Equation 4

$$\frac{\| \delta x \|}{\| x \|} \leq \frac{\| S \|}{1 - \| S \|} = \frac{\| A^{-1} E \|}{1 - \| A^{-1} E \|} \leq \frac{\| A^{-1} \| \cdot \| E \|}{1 - \| A^{-1} \| \cdot \| E \|}$$

We now define the conditioning number K_n and a relative error ℓ as

$$\begin{aligned} K_n &= \| A \| \cdot \| A^{-1} \| \\ \ell &= \| E \| / \| A \| \end{aligned} \tag{5}$$

Then

$$\|A^{-1}\| \cdot \|E\| = K_n \ell$$

and finally the error is bounded by:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K_n \ell}{1 - K_n \ell} \tag{6}$$

In a norm sense, Equation 6 provides a relative error in x as a function of the conditioning number of A and the relative error in A. Notice that we neglected possible error in b, required that $\|S\| < 1$ so that $I + S$ is non-singular, and assumed that we could obtain norms of A^{-1}. What we have done, however, is correct for any norm except $\|A\|_E$. Notice, also, that our argument would be the same if x were the matrix A^{-1} and b the identity matrix.

To estimate the number of good digits, we may assume $\log_{10} 1/\ell = q$ where q is the number of digits in the arithmetic. Neglecting $1 - K_n \ell$, we may then estimate the number of significant digits from

$$p = \log_{10} \|x\| - \log_{10} \|\delta x\| = q - \log_{10} K_n$$

So that $\log_{10} K_n$ estimates the number of digits lost.

Since we have used norms of the matrix A^{-1} whose accuracy we are investigating, we must now inquire if this is permissible (Reference 8). Let C be any approximation to A^{-1} and define:

$$H = I - CA$$

$$CA = I - H \qquad \text{and if } \|H\| < 1, \ (I - H)^{-1} \text{ exists}$$

then

$$A^{-1} = (I - H)^{-1} C$$

$$\|A^{-1}\| \leq \|(I - H)^{-1}\| \cdot \|C\|$$

Let

$$D = (I - H)^{-1}$$

and

$$D - HD = I$$

$$1 = \|D - HD\| \geq \|D\| - \|HD\| \geq \|D\| - \|H\| \, \|D\|$$

$$\|D\| \leq \frac{1}{1 - \|H\|}$$

$$\|A^{-1}\| = \|DC\| \leq \|D\| \cdot \|C\| \leq \frac{\|C\|}{1 - \|H\|}$$

Thus if $\mathbf{C}$ is any matrix and $\|\mathbf{H}\| = \|\mathbf{I} - \mathbf{CA}\| < 1$ for some norm, then that norm of $\mathbf{C}$ will exceed $1 - \|\mathbf{H}\|$ times the corresponding norm of $\mathbf{A}^{-1}$. Thus the bound will remain conservative unless the process has broken down.

To verify these results by numerical experiment, three approaches are useful. The least specific is to check the overall solution to a physical problem for some special case where the answer is known. With this approach, we may say that no evidence has been observed in our work which invalidates the error criteria given here. The second approach is to generate and invert matrices at more than one precision. This is very flexible on a variable precision machine like the IBM 1620, and is also much easier on large machines since Fortran IV became available. This approach has been used on several problems, and has engendered additional confidence in the error bounds of this section. The third approach is numerical experimentation with test matrices. The n^{th} order Hilbert is a classic and useful test matrix. This is the n^{th} order matrix with the coefficients:

$$a_{ij} = \frac{1}{i + j - 1}$$

The exact inverse for the Hilbert matrices up to order 10 are given in Reference 9. High precision eigenvalues and vectors for the Hilbert matrices of order 3 through 6 were reported in Reference 10. The following table shows the results of experiments on the Hilbert matrices, orders 3 through 6. The inversion was accomplished in Fortran IV single precision on an IBM 7094. The inversion routine employed a full-pivot inversion scheme. Norms for the inverse were computed from the computed inverse. The last column shows the minimum number of good digits in the elements of the computed inverse.

TABLE I

N	*$\|H_n\|_1$	$\|H_n^{-1}\|_1$	K_1	8.3-Log K_1	$\|H\|_2$	$\|H^{-1}\|_2$	K_2	8.3-Log K_2	N.D.
3	11/6	408	748	5.4	1.408	372.1	524	5.6	6
4	$\frac{25}{12}$	13,620	28,375	3.8	1.5002	10341	15,513	4.1	6
5	$\frac{137}{60}$	413,280	943,656	2.3	1.567	304,202	476,684	2.6	3
6	$\frac{147}{60}$	**11,940,984	29,255,190	.8	1.619	9,230,000	$1.49 \cdot 10^7$	1.2	1

*For a symmetric matrix $\|\mathbf{H}_n\|_1 = \|\mathbf{H}_n\|_\infty$

**Computed from supposed inverse. Value from true inverse, 11,865,420.

APPLICATIONS OF ERROR ANALYSIS

In this Section examples of some numerical methods are examined in the light of the material contained in Section I, and it is shown that the characteristic values of matrices arising in physical problems may be influenced by the manner in which the equations are written.

Method of Least Squares

One of the most common sources of ill-conditioned equations which arises in practice is the fitting of data by the method of least squares. In seeing how this happens, additional insight of a general nature may be obtained.

Consider a group of m data points which are represented by

$$y_i \approx a x_i + b$$

To choose best values for a and b, the error vector ℓ_i is defined:

$$\ell_i = a x_i + b - y_i$$

a and b can now be chosen to minimize the sum $\sum_{i=1}^{m} \ell_i^2$. The resulting equations, called the normal equations, are:

$$m b + \sum_{i=1}^{m} x_i a = \sum_{i=1}^{m} y_i$$

$$\sum_{i=1}^{m} x_i b + \sum_{i=1}^{m} x_i^2 a = \sum_{i=1}^{m} x_i y_i$$

Notice that the second equation is formed from the first by multiplication under the summation sign.

Suppose now that the dispersion of x_i is small as compared to the mean of x_i. Then there should be very little difference between multiplication under the summation and multiplication outside the summation sign. Of course multiplication outside the summation sign would lead to exactly singular equations. To go further, define

$$\xi_i = x_i - \bar{x} \qquad \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Note that $\sum_{i=1}^{m} \xi_i = 0$ so that the matrix of the normal equations may now be written as

$$m \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2 \end{bmatrix}$$

This matrix differs from a rank 1 matrix with eigenvector

$$\left\{ \frac{1}{\sqrt{1+\bar{x}^2}} , \frac{\bar{x}}{\sqrt{1+\bar{x}^2}} \right\}$$

and eigenvalue $1+\bar{x}^2$ by the term

$$\frac{1}{m} \sum_{i=1}^{m} \xi_i^2$$

The characteristic polynomial may be written by inspection as

$$\lambda^2 - \left(1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2\right) \lambda + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2 = 0$$

so that

$$\lambda = \frac{1}{2} \left\{ 1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2 \pm \sqrt{\left(1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2\right)^2 - \frac{4}{m} \sum_{i=1}^{m} \xi_i^2} \right\}$$

And the conditioning number is

$$K_n = \frac{1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2 + \sqrt{\left(1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2\right)^2 - \frac{4}{m} \sum_{i=1}^{m} \xi_i^2}}{1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2 - \sqrt{\left(1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2\right)^2 - \frac{4}{m} \sum_{i=1}^{m} \xi_i^2}}$$

$$= \frac{1 + \sqrt{1-\epsilon}}{1 - \sqrt{1-\epsilon}} \qquad \epsilon = \frac{4 \sum_{i=1}^{m} \xi_i^2}{m\left(1 + \bar{x}^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2\right)^2}$$

Since we expect difficulties when ϵ is small relative to 1, we may write

$$K_n \approx \frac{4}{\epsilon}$$

Thus, when the standard deviation of the x coordinates of the data points becomes small, as compared to the square of the mean x coordinate, the equations become unsolvable.

This leads to an interesting idea. The information content of codes has been studied in information theory. From this the idea is borrowed that surprise is an element in information content. A message stating that President Johnson was elected in 1964 ceases to have information content because everyone already knows this. For this problem, the redundant information is $\bar{x}$. If, for example, the coordinate axis is chosen so that the range of x is 10^{-4} times $\bar{x}$, then 4 decimal digits essentially are not transmitting information. However, four digits of information about the dispersion in x may still be entered into the machine. When,

however, the sum of the squares of the x variables is formed, 16 digits are required for their full representation. All of the information about the dispersion of the data is included in the last eight of these 16 digits. In a single precision calculation, the last eight digits are truncated and all the information is lost.

Structural problems have a similar characteristic. Equations concerning the variation in deflection from one side of an element to the other are usually solved. The element, however, sees a rigid-body motion. As the element becomes smaller, the difference in the deflections, which is the basic variable of the equations, is represented by fewer digits. Thus one must expect eventual failure to obtain a solution. In addition, the inclusion of minor physical effects may also lead to such difficulties. This indicates that the rush to larger programs with greater generality must include numerical hazards that need exploration.

Finite Difference Methods

Finite difference methods are based on the assumption that the difference quotient will become a satisfactory approximation to the differential quotient, provided the increment is small enough. A source of ill conditioning in finite-difference problems will be described and some quantitative answers and guidelines for estimating error will be obtained. It will also be shown how some finite-difference formulations obviate the difficulties.

Consider the finite difference representation of the vibrating string problem. Using the second order finite-difference operator, we have

$$y_{i-1} - 2y_i + y_{i+1} = h^2 F_i$$

If the length of the string is L and if

$$h = \frac{L}{n+1} \qquad y_0 = 0 \qquad y_{n+1} = 0$$

we have a matrix of the form

$$\begin{bmatrix} -2 & 1 & 0 & & & & & & \\ 1 & -2 & 1 & 0 & & & & & \\ 0 & 1 & -2 & 1 & 0 & & & & \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & 0 & 1 & -2 & 1 \\ & & & & & & 0 & 1 & -2 \end{bmatrix} \tag{7}$$

It is well known (Reference 11) that the eigenvalues are given by

$$\lambda_j = 4 \sin^2 \frac{j\pi}{2(n+1)}$$

and the eigenvectors are

$$v_{i,j} = \sin \frac{ij\pi}{n+1}$$

From this solution, the conditioning number may be written immediately as

$$K_n = \frac{\lambda_{max}}{\lambda_{min}} = \frac{4 \sin^2 \frac{n\pi}{2(n+1)}}{4 \sin^2 \frac{\pi}{2(n+1)}} \approx n^2$$

Thus, with single precision arithmetic difficulty in obtaining more than three-digit accuracy should be expected if

$$8.3 - \log_{10} n^2 < 3 \quad \text{or} \quad n > 650$$

If the matrix is squared, we obtain the new matrix

$$\begin{bmatrix} 5 & -4 & 1 & 0 & & & & & & \\ -4 & 6 & -4 & 1 & 0 & & & & & \\ 1 & -4 & 6 & -4 & 1 & 0 & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & 0 & 1 & -4 & 6 & -4 \\ & & & & & \cdot & 0 & 1 & -4 & 5 \end{bmatrix} \tag{8}$$

which is the matrix of a beam of constant cross section simply supported at its ends. Since this matrix is the square of the previous matrix, it must have the same eigenvectors, and its eigenvalues must be the square of the eigenvalues of the previous matrix. Thus, the conditioning number of the beam matrix is:

$$K_n = \frac{16 \sin^4 \frac{n\pi}{2(n+1)}}{16 \sin^4 \frac{\pi}{2(n+1)}} \approx n^4$$

For this case difficulty in exceeding three-digit accuracy should be expected if

$$8.3 - \log_{10} n^4 > 0 \quad \text{or} \quad n > 21$$

At n = 50, the conditioning number is $6.25.10^6$, so that the number of good digits would be 8.3-6.8 = 1.5. Experience indicates that these estimates are not more than a digit or two on the conservative side.

The important conclusion is that formulations, such as the ones just illustrated, will numerically deteriorate approximately as n^p where p is the order of the difference operator. Thus there is a conflict. If the number of increments n is small, the difference quotient is a poor approximation to the differential quotient, and discretization error is large. On the other hand, if n is large, roundoff error becomes large. We observe that for each problem there must be an optimum n, and that problems may exist in which the optimum n is unsatisfactory.

Consider the matrices 7 and 8 from the point of view of strong diagonals. Both matrices are sparce, and both satisfy the inequality

$$|a_{ii}| \geq \sum_{j=1}^{n} |a_{ij}| \qquad i \neq j \qquad i = 1, n$$

The condition > holds twice for matrix 7 and four times for the matrix 8. There is a theorem which guarantees the nonsingularity of the matrices under these conditions.

The inadequacy of the strong diagonal criterion may now be seen clearly. Conditioning of these matrices is strongly dependent upon their order. Yet the strong diagonal criterion is independent of the order. Further, it is clear from consideration of the linear combination of rank 1 martices that it is the conditioning number and the coupling between the modes which determines stability. Yet the matrices 7 and 8 have the same modes but are not the same from the point of view of strength on the diagonal.

We conclude that careful thought is required for successful finite-difference calculations. Since one may not pass to the limit, approximations which are satisfactory in the limit may not be satisfactory in finite differences. In Reference 12, for example, finite difference solutions of a circular plate symmetrically loaded were investigated. It was found that finite difference approximations for the following equivalent differential expressions were not equivalent:

$$\frac{d^4w}{dr^4} + \frac{2}{r}\frac{d^3w}{dr^3} - \frac{1}{r^2}\frac{d^2w}{dr^2} + \frac{1}{r^3}\frac{dw}{dr} = \frac{9}{D}$$

and

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{d}{dr}\left(\frac{1}{r}\frac{d}{dr}\left(r\frac{dw}{dr}\right)\right)\right) = \frac{9}{D}$$

Expanding the first of these expressions with central difference operators led to an asymmetric matrix and unsatisfactory solutions. Expanding the second operator led to a symmetric matrix and satisfactory solutions. The conditioning number, in the latter case, was found to vary as n^4.

A most interesting and suggestive finite difference solution is presented in Reference 13. In this paper, a transformation is used to reduce a sixth-order differential equation to a pair of coupled second-order equations. The finite difference equations are finally posed as

$$a_{i,i-1}\psi_{i-1} + a_{ii}\psi_i + a_{i,i+1}\psi_{i+1} + \cdots\cdots + a_{i,i+n}\beta_i = f_i$$

$$a_{i+n,i}\psi_i + \cdots\cdots + a_{i+n,i+n-1}\beta_{i-1} + a_{i+n,i+n}\beta_i + a_{i+n,i+n+1}\beta_{i+1} = f_{i+1}$$

where β is the rotation of the shell about a circumferential element of the reference surface, and ψ is a function of the horizontal stress resultant. The coefficients are defined in the original paper. For our purposes, the following relationships are relevant:

For all geometrices:

$$a_{i,i-1} = a_{i+n,i+n-1}$$
$$a_{i,i+1} = a_{i+n,i+n+1}$$
$$a_{i,i+n} = -a_{i+n,i}$$

Thus, the matrix of coefficients from the foregoing equations may be partitioned as

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline -\mathbf{B} & \mathbf{C} \end{array}\right]$$

where **A** and **C** are tridiagonal and **B** is diagonal.

For the special case of a uniform cylinder these additional relations apply:

$$a_{i,i-1} = a_{i+n,i+n-1} = 1 = a_{i,i+1} = a_{i+n,i+n+1}$$

$$a_{i,i} = a_{i+n,i+n} = -2$$

$$a_{i,i+n} = -a_{i+n,i} \qquad \text{constant for all } i$$

Thus the matrix reduces to

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline -\mathbf{B} & \mathbf{A} \end{array}\right] = \mathbf{U}$$

where **A** is the same second-order finite difference operator matrix 7, and **B** is a scalar matrix. Now the spectral norm of this matrix and the conditioning number are readily obtained:

$$\mathbf{U}^T\mathbf{U} = \left[\begin{array}{c|c} \mathbf{A} & -\mathbf{B} \\ \hline \mathbf{B} & \mathbf{A} \end{array}\right]\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline -\mathbf{B} & \mathbf{A} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}^2+\mathbf{B}^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}^2+\mathbf{B}^2 \end{array}\right]$$

The eigenvalues of this matrix are

$$\lambda_j = 16 \sin^4 \frac{j\pi}{2(n+1)} + b^2$$

where

$$\mathbf{B} = b\,\mathbf{I}$$

Thus the spectral conditioning number is

$$K_n = \frac{\sqrt{16 \sin^4 \dfrac{n\pi}{2(n+1)} + b^2}}{\sqrt{16 \sin^4 \dfrac{\pi}{2(n+1)} + b^2}} \approx n^2$$

Inspection of the original paper shows that the coefficients depart slowly from this simple form as the derivatives of the shell radius and the rigidities increase. Further investigation of this problem should be most instructive, but even this simple analysis serves to explain the high numerical stability that has been observed on this and similar formulations. For numerical stability, the problem resembles a second-order problem more closely than a fourth.

At this time we only mention partial differential equations. In particular, these problems are strongly influenced by the shape of the domain. However, for Poisson's equation in a rectangular domain, it can be shown that the characteristic values are given by

$$\lambda_{\ell,k} = 4\left(\sin^2\frac{k\pi}{2(n+1)} + \sin^2\frac{\ell\pi}{2(m+1)}\right)$$

where

$$0 \le x \le (n+1)h$$

$$0 \le y \le (m+1)h$$

Thus, some similarity is shown to the ordinary difference equation, and conditioning troubles should again be expected to multiply rapidly for high-order operators.

Finite Element Problems

Consider a statically indeterminate structure in which it is desired to relate the forces $\mathbf{p}$ and displacements $\mathbf{x}$ at n generalized coordinates. A relation in the form $\mathbf{x} = \mathbf{F}\mathbf{p}$ can be written when $\mathbf{x}$ and $\mathbf{p}$ are n dimensional vectors of displacement and forces, and $\mathbf{F}$ is an $(n \times n)$ matrix called the flexibility matrix (Reference 14).

If the structure is now made statically determinate by introducing cuts at appropriate points, and if the m internal forces removed in the process are applied as external forces to the structure together with the original p forces, all the displacements and forces can be related by means of an $(n+m) \times (n+m)$ matrix as follows:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \hline x_{n+1} \\ \vdots \\ x_{n+m} \end{bmatrix} = \left[\begin{array}{c|c} F_{11} & F_{12} \\ \hline F_{12}^T & F_{22} \end{array}\right] \begin{bmatrix} p_1 \\ \vdots \\ p_n \\ \hline p_{n+1} \\ \\ p_{n+m} \end{bmatrix}$$

It can then be shown that the flexibility matrix F of the system is given by:

$$F = F_{11} - F_{12}\, F_{22}^{-1}\, F_{12}^T$$

In principle, $\mathbf{F}$ must be independent of the choice of coordinates. In a finite number system this is not necessarily so. Neglecting errors in $\mathbf{p}$, which are inherently not relevant to the present discussion, the error problem may be stated as follows:

$$\mathbf{x} + \delta\mathbf{x} = (\mathbf{F} + \mathbf{E})\,\mathbf{p}$$

i.e.,

$$\delta x = E p$$

we then have

$$F + E = F_{11} + E_{11} - (F_{12} + E_{12})(F_{22}^{-1} + E_{22}^{-1})(F_{12}^{T} + E_{12}^{T})$$

Neglecting higher order terms and taking norms on both sides, we obtain

$$\|E\| \leq \|E_{11}\| + \|E_{12} F_{22}^{-1} F_{12}^{T}\| + \|F_{12} E_{22}^{-1} F_{12}^{T}\| + \|F_{12} F_{22}^{-1} E_{12}^{T}\|$$

We can assume that in single precision arithmetic the norms of the error matrices are given by

$$\|E_{11}\| = 10^{-8.3} \|F_{11}\|, \quad \|E_{12}\| = 10^{-8.3} \|F_{12}\|, \|E_{22}^{-1}\| = 10^{-8.3} K_n$$

where K_n in the conditioning number of F_{22}. Then

$$\|E\| \leq 10^{-8.3} \left\{ \|F_{11}\| + (2 + K_n) \|F_{12}\| \|F_{22}\| \|F_{12}^{T}\| \right\}$$

Now, since in the real number system F is invariant over changes in redundants and

$$F = F_{11} - F_{12} F_{22}^{-1} F_{12}^{T}$$

it can be seen that the submatrices F_{ij} are related and that $\|F_{11}\|$ must increase with $\|F_{12} F_{22}^{-1} F_{12}^{T}\|$. One can see that the criterion "minimum $\|F_{11}\|$" is the same as the criterion "minimum $\|F_{12} F_{22}^{-1} F_{12}^{T}\|$". Thus, the above analysis confirms the well-known criteria which tell us to choose redundants so that their effect is local and the deflected shape of the released structure is as close as possible to that of the true structure. We have observed in limited experimentation that the conditioning of the F_{22} matrix was improved as the norms of F_{12} and F_{12}^{T} were reduced. At the same time, the norms of F_{22} and F_{22}^{-1} were also both reduced simultaneously. The relative error of F in a norm sense is:

$$\frac{\|E\|}{\|F\|} \leq 10^{-8.3} \frac{\|F_{11}\| + (2 + K_n) \|F_{12}\| \|F_{22}\| \|F_{12}^{T}\|}{\|F\|}$$

Thus, the number of good digits in single-precision arithmetic is estimated as:

$$\log_{10}\|F\| - \log_{10}\|E\| \approx 8.3 + \log_{10}\|F\| - \log_{10}\|F_{11}\|$$

The stiffness matrix formulation of the structural problem may be written as $Kx = p$ where K is the stiffness matrix (Reference 14). As is well known, the matrix becomes singular if the structure is inadequately supported. The physical meaning of this singularity is that the displacements are non-unique to the extent of one or more rigid-body motions. Since the equations of equilibrium are written in terms of the differences in displacements at joints connected by members, these equations break down if the differences are negligible portions of the displacements.

This interpretation of singularity also provides an adequate framework for understanding the problem of ill-conditioning. Most members in a structure which are not directly connected to a support experience a deformation and a rigid-body motion. As an analysis is refined by taking smaller elements, one may arrive at a condition where these rigid-body components interfere with the calculation of the deformations. Unfortunately, there are problems where one is interested in very local stresses. Dr. Przemieniecki's analysis of substructures (Reference 15) seems to offer an opportunity for obviating this difficulty. Briefly, this method considers the structure as a set of substructures, each containing interior and boundary points. The loaded substructure is then solved for its deformations as a function of the unknown deflections on the boundary. Then various substructures are combined, and an overall solution for the boundary deflections is obtained. The displacements in each substructure are then readily determined.

It is suggested that where conditioning is a problem, the rigid-body components of the substructure's boundary deflections may be identified and removed. Then one should be able to analyse the substructure to an accuracy determined principally by the conditioning number of its own submatrix.

Closed-Form Solutions

Finally a closed form solution to a boundary value problem is considered. For simplicity, consider the ordinary differential equation with constant coefficients:

$$\frac{d^4 y}{dx^4} = - \frac{K}{EI} y$$

which is the equation of a uniform beam on an elastic foundation. The foundation modulus is K and the rigidity is EI. The well-known solution is:

$$y = \sum_{j=1}^{4} a_j \phi_j(x)$$

The functions $\phi_j(x)$ must be linearly independent, and they must be solutions to the differential equation. The constants a_j must then be chosen to fit the boundary conditions. It is also known that the functions ϕ_j will be linearly independent, provided their Wronskian determinant is nonzero. The matrix of coefficients in the Wronskian determinant is given by:

$$w_{ij} = \frac{d^{i-1} \phi_j}{dx^{i-1}}$$

Now the equation for the constants a_j (Reference 16) may be written as

$$\left[B(0)\, W(0) + B(\ell)\, W(\ell) \right] a = P$$

where $B(o)$ is the matrix of coefficients of the linear combinations of derivative specified at $x = o$, $W(o)$ is the matrix of coefficients of the Wronskian determinant at $x = o$, $B(\ell)$ and $W(\ell)$ are the corresponding quantities at $x = \ell$, and P is a vector of boundary values. When these equations are solved for a, the state vector $y(x)$ may be determined from the transfer matrix equation:

$$y(x) = W(x)\, a$$

In Reference 17, a choice of basis functions is given which leads to an unstable solution if x becomes large. The problem is posed as an initial value problem so that $\mathbf{B}(\ell) = \mathbf{0}$, $\mathbf{y}$ is the vector

$$\left\{ y_0 , \theta_0 , -\frac{M}{EI} , -\frac{Q_0}{EI} \right\}$$

where y_0 is the deflection, θ_0 the rotation, M_0 the moment and Q_0 the shear all at $x = 0$. $\mathbf{\Phi}(o)$ will, in this case, be the identity matrix if

$$\phi_1 = F_1 = \cosh \beta x \cos \beta x$$

$$\phi_2 = F_2 = \frac{1}{2\beta} (\cosh \beta x \sin \beta x + \sinh \beta x \cos \beta x)$$

$$\phi_3 = F_3 = \frac{1}{2\beta^2} \sinh \beta x \sin \beta x$$

$$\phi_4 = F_4 = \frac{1}{4\beta^3} (\cosh \beta x \sin \beta x - \sinh \beta x \cos \beta x)$$

$$\beta = \sqrt{\frac{K}{4EI}}$$

The matrix $\mathbf{W}(x)$ is easy to write because

$$\frac{dF_1}{dx} = -4\beta^4 F_4 \qquad \frac{dF_2}{dx} = F_1 \qquad \frac{dF_3}{dx} = F_2 \qquad \frac{dF_4}{dx} = F_3$$

thus

$$\mathbf{W}(x) = \begin{bmatrix} F_1 & F_2 & F_3 & F_4 \\ -4\beta^4 F_4 & F_1 & F_2 & F_3 \\ -4\beta^4 F_3 & -4\beta^4 F_4 & F_1 & F_2 \\ -4\beta^4 F_2 & -4\beta^4 F_3 & -4\beta^4 F_4 & F_1 \end{bmatrix}$$

The inverse may be obtained by inspection because the operation of $\mathbf{W}(x)$ is "shift the state vector x units to the right." Its inverse, then, must be the matrix whose operation is "shift the state vector x units to the left." Hence $[\mathbf{W}(x)]^{-1} = \mathbf{W}(-x)$. An inspection of the functions shows that this will change the sign of the functions F_2 and F_4.

There are now a number of ways to see that the matrix becomes ill conditioned as x becomes large. For example, $\|\mathbf{W}\|_1 = \|\mathbf{W}^{-1}\|_1$, $\|\mathbf{W}\|_\infty = \|\mathbf{W}^{-1}\|_\infty$ because both of these norms involve only absolute values of the elements w_{ij}. If it is noticed that

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{W} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} \mathbf{W} \end{bmatrix}^{-1}$$

then, since the transformation is orthogonal, $\|\mathbf{W}\|_2 = \|\mathbf{W}^{-1}\|_2$. Thus, for any norm the conditioning number is the square of the norm (Equation 5) and is approximated by $\ell^{2\beta x}$. The full tragedy becomes evident, however, if the matrix is rewritten as the sum of the rank -2 matrices

$$\frac{\ell^{\beta x}}{2}\begin{bmatrix} \cos\beta x & \frac{1}{2\beta}(\sin\beta x+\cos\beta x) & \frac{1}{2\beta^2}\sin\beta x & \frac{1}{4\beta^3}(\sin\beta x-\cos\beta x) \\ -\beta(\sin\beta x-c x\beta x) & \cos\beta x & \frac{1}{2\beta}(\sin\beta x+\cos\beta x) & \frac{1}{2\beta^2}\sin\beta x \\ -2\beta^2\sin\beta x & -\beta(\sin\beta x-\cos\beta x) & \cos\beta x & \frac{1}{2\beta}(\sin\beta x+\cos\beta x) \\ -2\beta^3(\sin\beta x+\cos\beta x) & -2\beta^2\sin\beta x & -\beta(\sin\beta x-\cos\beta x) & \cos\beta x \end{bmatrix}$$

$$+\frac{\ell^{-\beta x}}{2}\begin{bmatrix} \cos\beta x & \frac{1}{2\beta}(\sin\beta x-\cos\beta x) & -\frac{1}{2\beta^2}\sin\beta x & \frac{1}{4\beta^3}(\sin\beta x+\cos\beta x) \\ -\beta(\sin\beta x+\cos\beta x) & \cos\beta x & \frac{1}{2\beta}(\sin\beta x-\cos\beta x) & -\frac{1}{2\beta^2}\sin\beta x \\ 2\beta^2\sin\beta x & -\beta(\sin\beta x+\cos\beta x) & \cos\beta x & \frac{1}{2\beta}(\sin\beta x-\cos\beta x) \\ -2\beta^3(\sin\beta x-\cos\beta x) & 2\beta^2\sin\beta x & -\beta(\sin\beta x+\cos\beta x) & \cos\beta x \end{bmatrix}$$

Further probing will show that the eigenvalues are

$$\lambda = \ell^{\pm\beta(1\pm i)x}$$

Any local disturbances to a beam on elastic foundation are rapidly damped. Thus, physically sensible solutions belong to the second of the rank-2 matrices shown. Clearly as βx becomes large, any errors which project into the space of the first rank-2 matrix will expand rapidly just as the true solution is decaying.

Returning to the choice of basis functions, suppose a double coordinate system with a variable $z = \ell - x$ had been chosen, for example:

$$\phi_1 = \ell^{-\beta x}\cos\beta x, \qquad \phi_2 = \ell^{-\beta x}\sin\beta x, \qquad \phi_3 = \ell^{-\beta z}\cos\beta z, \qquad \phi_4 = \ell^{-\beta z}\sin\beta z$$

The matrix **W(x)** for this basis would be

$$\mathbf{W}(x) = \begin{bmatrix} \cos\beta x & \sin\beta x & \cos\beta z & \sin\beta z \\ -\cos(\beta x-\frac{\pi}{4}) & -\sin(\beta x-\frac{\pi}{4}) & \cos(\beta z-\frac{\pi}{4}) & \sin(\beta z-\frac{\pi}{4}) \\ \cos(\beta x-\frac{\pi}{2}) & \sin(\beta x-\frac{\pi}{2}) & \cos(\beta z-\frac{\pi}{2}) & \sin(\beta z-\frac{\pi}{2}) \\ -\cos(\beta x-\frac{3\pi}{4}) & -\sin(\beta x-\frac{3\pi}{4}) & \cos(\beta z-\frac{3\pi}{4}) & \sin(\beta z-\frac{3\pi}{4}) \end{bmatrix} \begin{bmatrix} \ell^{-dx} & 0 & 0 & 0 \\ 0 & \ell^{-dx} & 0 & 0 \\ 0 & 0 & \ell^{-dz} & 0 \\ 0 & 0 & 0 & \ell^{-dz} \end{bmatrix}$$

Now the improved conditioning is seen in the diagonal matrix of the matrix of exponentials. When one attempts to fit boundary conditions, the form of the equations indicates that, for very long domains, only two functions may be associated with each of the boundaries. The problem separates naturally. The matrix $\mathbf{W}(x)$ then partitions as (Reference 16):

$$\mathbf{W}(x) = \left[\begin{array}{c|c} U_{11} & U_{12} \\ \hline U_{21} & U_{22} \end{array}\right]$$

where U_{11} is a semi-infinite solution starting at $x = 0$, U_{22} is a semi-infinite solution starting at $x = \ell$ and the submatrices U_{12} and U_{21} represent coupling between the semi-infinite solutions. The coupling becomes small as ℓ becomes large.

SUMMARY, PRACTICAL CONSIDERATIONS AND CONCLUSIONS

In summary, it has been shown in the first Section, that susceptibility to numerical difficulties can be expressed quantitatively. Ill-conditioning is seen to be the absence of information about the lower modes of the matrix. In the second Section, it was seen that numerical instability does arise in physically stable problems. Further, examples were shown in which the source of the instability could be identified and avoided. Let us now turn to some practical considerations.

Much of the older literature implies that obtaining eigenvalues is too expensive for use in error analysis. This needs some clarification. For the common case of a positive, definite matrix, not too large for core, whose inverse is obtained as part of the calculation, the machine determination of the dominant eigenvalues of the matrix and its inverse, is a trivial problem. The power method, using Rayleighs Quotient, is cheap to use and easy to code. If the matrix is symmetrical, but not positive definite the PQ algorithm may be safer (Reference 18). If the matrix is not symmetric, the spectral norm may be used. If the inverse is not obtained as a part of the calculation, the method of extremes will work (References 19 and 20).

If the matrix is too large for core, the tape operations required to iterate through an eigenvalue problem may be too expensive. If, however, the inverse is obtained, the other norms, including the Euclidean norm, are easy to obtain. The smallest of these may then be used to bound the extreme eigenvalues. Finally if the matrix is too large for core and the inverse is not obtained as part of the calculation, error analysis may become expensive. One is then left to discover some particular pattern for the problem at hand.

The eigenvectors associated with the extreme eigenvalues may provide additional insight and should be plotted where applicable.

We conclude that:

1. The conditioning numbers of the first Section are adequate tools for matrix error analysis.
2. The measurement of conditioning is usually inexpensive and should be routine for most calculations.
3. Since the significance of ill conditioning is the absence of information, the future of matrix error analysis lies with avoidance of ill conditioning. Pre-conditioning and careful inversion techniques can be of only limited help.

REFERENCES

1. Wilkinson, J. H., Rounding Errors in Algebraic Processes, Prentice Hall Inc., Englewood Cliffs, N. J., 1963.

2. Rall, L. B., Ed., Error in Digital Computation Vol. I (Especially the Article, "Error in Digital Solution of Linear Problems," Albasiny, Ernest L.) John Wiley and Sons, Inc., New York, 1965.

3. Fadeeva, V. N., Computational Methods of Linear Algebra, Dover Publications Inc., New York, 1959.

4. Paige, Lowell J., and Swift, J. Dean, Elements of Linear Algebra, Ginn and Co., New York, 1961.

5. Comrie, L. J., Chambers Shorter Six Figure Mathematical Tables, Chemical Publishing Co., Inc., New York, 1954.

6. Bodewig, E., Matrix Calculus, North Holland Publishing Co., Amsterdam, 1959.

7. Wilf, Herbert, Mathematics for the Physical Sciences, John Wiley and Sons, Inc., New York, 1962.

8. Householder, Alston S., The Theory of Matrices in Numerical Analysis, Blaisdell Publishing Co., New York, 1964.

9. Savage, Richard, and Lucacks, Eugene, Tables of Inverses of Finite Segments of the Hilbert Matrix, National Bureau of Standards Applied Mathematics Series No. 39, 1954.

10. Forthergill, J. W., Jr., and Rosanoff, R. A., High Precision Values of the Latent Roots and Vectors of Some Hilbert Matrices, Thiokol Chemical Corp., Wasatch Division of Brigham City, Utah, 18 October 1963.

11. Hilderbrand, F.B., Methods of Applied Mathematics, Prentice Hall Inc., Englewood Cliffs, N. J., 1952.

12. Rosanoff, R. A., and Rosen, R., On the Choice of Finite Difference Approximations, NAA/S and ID SID 65-719.

13. Radkowski, P. P., Davis, R. M., and Bolduc, M. R., "Numerical Analysis of Equations of Thin Shells of Revolution," ARS Journal, pp. 36-41, January 1962.

14. Hurty, Walter C., and Rubinstein, Moshe F., Dynamics of Structures, Prentice Hall, Inc., New York, 1964.

15. Przemieniecki, J. S., "Matrix Structural Analysis of Substructures," AIAA Journal, pp. 138-147, January 1963.

16. Rosanoff, Richard A., and Mah, Gordon, Boundary Value Problem in Ordinary Differential Equations, NAA/S and ID SID 64-662.

17. Hetenyi, M., Beams on Elastic Foundation, Ann Arbor: The University of Michigan Press.

18. Lancoz, Cornelius. Applied Analysis, Prentice Hall Inc., Englewood Cliffs, N. J.

19. Allen, D. N., Relaxation Methods, McGraw-Hill Book Co., Inc., New York.

20. Fothergill, J. W., Jr., and Rosanoff, R. A., A Note on the Intensification Method and Rayleigh's Relaxation Method, Thiokol Chemical Corp., Wasatch Division, Brigham City, Utah, 18 October 1963.