NPL Report DNACS 34/80
October 1980

ON LINEAR SYSTEMS ARISING FROM FINITE DIFFERENCE APPROXIMATIONS TO ELLIPTIC DIFFERENTIAL EQUATIONS

by
S Hammarling and J H Wilkinson

ISSN 0143 - 7348

National Physical Laboratory,
Teddington, Middlesex TW11 OLW UK

Approved on behalf of Director, NPL by Mr E L Albasiny,
Superintendent of Division of Numerical Analysis & Computer Science

October 1980

N A T I O N A L P H Y S I C A L L A B O R A T O R Y

ON LINEAR SYSTEMS ARISING FROM FINITE DIFFERENCE APPROXIMATIONS TO ELLIPTIC DIFFERENTIAL EQUATIONS

by

S Hammarling and J H Wilkinson

Division of Numerical Analysis and Computer Science

ABSTRACT

We discuss the accuracy of computed solutions of the linear systems that arise from finite difference approximations to elliptic differential equations.

The discussion shows that there is no reason to expect that the accuracy of the solutions will vary exceptionally with the nature of the boundary conditions. In general the accuracy of the computed solutions will fully reflect the condition number of the matrix of coefficients.

1 INTRODUCTION

If $Ax=b$ is a non-singular $n \times n$ system of linear equations, with $b \neq 0$, then we have

$$\|A\| \, \|x\| \geq \|b\|, \quad \|x\| \leq \|A^{-1}\| \, \|b\| \tag{1.1}$$

and hence

$$\|A\|^{-1} \leq \|x\| / \|b\| \leq \|A^{-1}\|. \tag{1.2}$$

When the matrix norm is subordinate to the vector norm then both extreme values can be attained and as b varies with a fixed value of $\|b\|$ we have

$$\max \|x\| \leq \|A\| \, \|A^{-1}\| \min \|x\| = K(A) \min \|x\|. \tag{1.3}$$

When A is ill-conditioned with respect to inversion K is large and there is a correspondingly large range of values attainable by $\|x\|$ for a given $\|b\|$. We shall refer to an x for which $\|x\|$ is near the maximum attainable value as a "large" solution and to one at the other end of the scale as a "small" solution.

If

$$A = U\Sigma V^H, \quad \Sigma = \operatorname{diag}(\sigma_i), \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0 \tag{1.4}$$

is the singular value decomposition (SVD) of A then

$$A \frac{v_i}{\sigma_i} = u_i \quad (i = 1, \ldots, n). \tag{1.5}$$

The u_i are unit vectors in the $\ell 2$ norm and the extreme values of $\|x\|$ are attained for $b=u_1$ and $b=u_n$, the corresponding solutions being v_1/σ_1 and v_n/σ_n with norms σ_1^{-1} and σ_n^{-1} respectively. When A is ill-conditioned $\sigma_1 >> \sigma_n$. If we take b to be a random unit vector then unless it happens to have an exceptionally small component in the direction of u_n, the corresponding x will be a large solution with $\|x\|_2$ of the order of magnitude of σ_n^{-1}. Indeed if some of the other σ_i^{-1} are of the same order of magnitude as σ_n^{-1} then $\|x\|$ will be of the order of magnitude of σ_n^{-1} unless b has exceptionally small components in the direction of <u>all</u> of the corresponding u_i. We may summarise this by saying that the probability that a random right-hand side will correspond to a "large" solution is very high. On the other hand if we take a random unit x and write

$$x = \sum \alpha_i v_i \quad (\sum \alpha_i^2 = 1), \tag{1.6}$$

then

$$Ax = \sum \alpha_i \sigma_i u_i \ , \tag{1.7}$$

and

$$\| Ax \| = (\sum \alpha_i^2 \sigma_i^2)^{\frac{1}{2}} \ . \tag{1.8}$$

From this it is clear that for such an x, Ax will usually have very small components in the direction of u_n and also of any other u_i that corresponds to very small σ_i. Hence the right-hand sides b generated from arbitrary x are usually such as will yield small solutions, in that $\| x \|$ will be of the order of magnitude of $\| b \| / \sigma_1$ ie $\| b \| / \| A \|$ and not $\| A^{-1} \| \ \| b \|$.

These considerations are very important in connexion with the linear systems arising from finite difference approximations to elliptic differential equations. For a very wide class of problems the exact solution of the exact finite difference approximations tends to the solution of the continuous problem as the mesh length tends to zero. Hence in theory one could obtain a solution of any required accuracy by using a fine enough mesh. However, in practice one is usually restricted to using numbers of a prescribed precision and this places limitations on the attainable accuracy for three reasons.

(i) The matrix A arising from the finite difference approximations may not be exactly representable in the computer.

(ii) The right-hand side b may not be exactly representable.

(iii) The solution of the system Ax = b will involve rounding errors and hence we shall not even obtain an exact solution of the "computer equations".

As far as (i) is concerned many important problems, such as those arising from the Laplacian operator, lead to matrices in which the elements are small integers and therefore are exactly representable. On the other hand an elliptic operator of the form

$$p(x,y) \frac{\partial^2 w}{\partial x^2} + q(x,y) \frac{\partial^2 w}{\partial y^2} \tag{1.9}$$

will usually not give exactly representable matrices and this places an inherent limitation on attainable accuracy if one is restricted to values of $p(x,y)$ and $q(x,y)$ of (say) single word length on the computer. The optimum mesh length in this case is that for which the error in the solution due to the truncation error resulting from the use of the finite difference approximation is of the same

order of magnitude as the error in the solution resulting from the errors of representation in A. Note that this difficulty cannot be overcome by using iterative refinement unless a more accuate representation of A is used when determining the residuals.

It will be much more common for the right-hand side to be of such a nature as to preclude exact representation. Indeed the right-hand side will usually depend on the boundary values and these will sometimes be known only to a limited accuracy quite apart from the problem of computer representation.

Although anxiety about (iii) was at one time the over-riding pre-occupation, that is not now so. In fact provided the matrix A is not almost singular to working precision, accurate solutions of the computer equations could be determined by iterative refinement so that there is a sense in which (iii) is the least important of the problems. However in saying this we are assuming the use of higher precision inner-products when determining the residuals in iterative refinement but not the use of higher precision representations of A and b in determining those residuals. The latter policy would make it possible to overcome the limitations in the representation of A and b without using higher precision throughout.

As far as (ii) is concerned the nature of the right-hand side, that is, whether or not it corresponds to one of the larger solutions could be of critical importance for the following reasons. Let us assume that A is exactly representable and that b is known exactly but is not exactly representable. Then the computer will be presented with the system

$$Ax = b + e \tag{1.10}$$

where e is a random vector of rounding errors satisfying

$$|e_i| \leq \tfrac{1}{2}\beta^{1-t}|b_i| . \tag{1.11}$$

Here we are assuming floating point arithmetic working in the base β with a mantissa of t digits. Now e being a random vector the solution $A^{-1}e$ will, with very high probability, be such that

$$\|A^{-1}e\| = O\left[\|A^{-1}\|\ \|e\|\right] \leq \tfrac{1}{2}\beta^{1-t}\|A^{-1}\|\ \|b\| . \tag{1.12}$$

If b is such that

$$\| A^{-1}b \| = 0 \left[\| A^{-1} \| \, \| b \| \right] \tag{1.13}$$

then the relative error in the exact solution of (1.10) will be $0(\beta^{1-t})$ and this will be perfectly satisfactory. If on the other hand b is such that

$$\| A^{-1}b \| = 0 \left[\| b \| / \| A \| \right] \tag{1.14}$$

then the relative error in the exact solution of (1.10) will be $0(K(A)\beta^{1-t})$ and for large K this may be unacceptable. Usually $K(A)$ will increase as the mesh length is decreased and although the exact solution of Ax=b will be tending to the correct solution, the exact solution of Ax=b+e will not. Again the optimum mesh length will be that for which $K(A)\beta^{1-t}$ is the same order of magnitude as the error due to the truncation error resulting from the finite difference approximation, as was in fact the case with (1.9).

With elliptic differential equations we shall see that there are important classes of problems for which the right-hand sides always correspond to large solutions and other important classes for which they always correspond to small solutions.

In sections 2 and 3 we consider, by means of simple examples, the exact solution of the finite difference approximation to an elliptic differential equation. In section 4 we additionally consider the effects of the rounding errors made in solving the finite difference equations, by applying standard results concerning the accuracy of computed solutions. Section 5 gives the results of some experiments designed to illustrate the points raised in sections 2–4 and section 6 comments upon these results.

2 A SIMPLE ORDINARY DIFFERENTIAL EQUATION

Consider first the equation

$$\frac{d^2w}{dx^2} = f(x) \, , \quad 0 \leq x \leq 1, \quad w(0) = w(1) = 0 \tag{2.1}$$

and the simple finite difference approximation

$$\frac{d^2w}{dx^2} = \frac{w(x+h)-2w(x)+w(x-h)}{h^2} \, . \tag{2.2}$$

The corresponding system of equations Au=b of order n-1 is such that

$$a_{ii} = 2,\ a_{i,i+1} = a_{i+1,i} = -1,\ a_{ij} = 0 \text{ otherwise,} \tag{2.3}$$

$$b_i = h^2 f(x_i) \tag{2.4}$$

where we assume a uniform mesh length $h = 1/n$. We observe that $\| A \|_\infty$ is independent of h but that $\| b \|_\infty$ is of order $1/n^2$ and $\| b \|_2$ is of order of $1/n^{3/2}$. Now the solution u is tending to the solution of the continuous problem and hence does not involve the factor $1/n^2$. This means that b must be such that $\| u \|_\infty$ is of the order of $n^2 \| b \|_\infty$. Since A is symmetric positive definite the singular values are the eigenvalues and $u_i = v_i$, each being an eigenvector, and we have

$$\sigma_{n-i} = 4 \sin^2 i\pi/2n \quad \text{and} \quad u_{n-i,j} = \sqrt{2/n} \sin j\theta_i \text{ where } \theta_i = i\pi/n. \tag{2.5}$$

For large n

$$\sigma_1 \doteq 4, \quad \sigma_{n-1} \doteq \pi^2/n^2 \tag{2.6}$$

and this means that the right-hand side <u>cannot</u> be pathologically defective in the eigenvectors corresponding to the smaller eigenvalues. It is easy to see that such pathological defectiveness is indeed ruled out. Although we are dealing with the discrete approximation, when n is large the components of b in the direction of the eigenvectors corresponding to the smaller eigenvalues will approximate to those in the Fourier expansion of f(x) in terms of sin πx, sin 2πx, etc. Notice that sin πx corresponds to the smallest eigenvalue σ_{n-1}, sin 2πx to the next smallest etc. If f(x) is sufficiently smooth for the differential problem to be of interest then it is the components in the direction of the graver modes which will be of normal size and the components in the direction of the higher modes (note that these correspond to σ_1, σ_2, ... in the discrete problem) which are tending to zero. Hence it is the components corresponding to the smaller eigenvalues which are well represented, giving the required amplification of roughly n^2 in the solution. Notice though that the right-hand sides are not particularly special. If we take a random right-hand side then as we remarked in section 1 the probability is high that $\| x \|_2 / \| b \|_2 = O(\| A^{-1} \|) = O(n^2/\pi^2)$. Hence smoothness of f(x) does not

play an over-ridingly important role, though it does guarantee that as $n \to \infty$ there will be substantial components of eigenvectors corresponding to the smallest eigenvalues whereas for a random vector we merely have a high probability that this is true.

Let us now consider the same problem but with $f(x) = 0$ and $w(0) = w_0$, $w(1) = w_1$. The finite difference equations have the same matrix A but the right-hand side is very special. We have

$$b_1 = w_0, \; b_{n-1} = w_1, \; b_i = 0 \quad \text{otherwise} \tag{2.7}$$

so that $\| b \|_\infty$ does not involve a power of n. The exact solution u of the discrete system again tends to that of the continuous problem and therefore does not involve the factor n^2. It must therefore be such that $\| u \|_\infty / \| b \|_\infty$ is of order unity, ie of order $\| A \|^{-1}$, and not $\| A^{-1} \|$. In this simple problem it is obvious that every component of the exact solution of the finite difference approximation lies between w_0 and w_1. As $n \to \infty$ the right-hand side must be almost completely deficient in the eigenvectors corresponding to the smallest eigenvalues(the gravest modes) since such components are amplified by a factor of order n^2. It is easy to verify that this is true. We have for example

$$b^T v_{n-1} = (w_0+w_n)(\sin \tfrac{\pi}{n}) \sqrt{\tfrac{2}{n}} = O(n^{-3/2}) , \tag{2.8}$$

$$b^T v_{n-2} = (w_0-w_n)(\sin \tfrac{2\pi}{n}) \sqrt{\tfrac{2}{n}} = O(n^{-3/2}) . \tag{2.9}$$

Note that we have the factor $n^{-3/2}$ rather than n^{-2} since the use of inner products related to the $\ell 2$ norm while the previous remarks were about the $\ell\infty$ norm.

We see then that although the right-hand sides corresponding to a forcing function f and homogeneous boundary conditions behave almost like random right-hand sides, those coming from a zero forcing function and non-homogeneous boundary conditions are very special indeed. An interesting feature of this second problem is that most of the elements of the right-hand side are zero and are therefore exactly representable in the computer. The only non-zero elements are b_1 and b_{n-1} and even if they are not exactly representable the error e in b will not only be of the order of magnitude $2^{-t} \| b \|$ (this is true in general) but will also be such that e_i is zero except for the two end values. Hence it, too, will be equally deficient in eigenvectors corresponding to the smaller

eigenvalues. The exact solution of Ax = b+e will therefore have only a small relative error in spite of the fact that b has negligible components in the direction of the eigenvectors corresponding to the smaller eigenvalues. This latter situation is usually a dangerous one, as we saw earlier following equation (1.14).

If we consider the general problem with both a forcing function f and non-zero boundary values w_0 and w_1 then the right-hand side is the sum of the two earlier ones. Here and later we assume that f and the boundary values are of similar orders of magnitude. The first and last components are $w_0 + f(\frac{1}{n})/n^2$ and $w_1 + f(1-\frac{1}{n})/n^2$ while $b_i = f(\frac{i}{n})/n^2$ at internal points. When n is large the part which comes from the forcing function is very small compared with that coming from the boundary conditions. In fact in the first and last terms $f(\frac{1}{n})/n^2$ and $f(1-\frac{1}{n})/n^2$ will usually be so small compared with w_0 and w_1 that in numbers of finite precision the information in $f(\frac{1}{n})/n^2$ and $f(1-\frac{1}{n})/n^2$ will be almost completely frozen out when the addition is performed. Further, during the elimination the multiples of b_1 added to the other b_i freeze out a good deal of information in those b_i. In fact if $n^2 > \beta^t$ then the information in the b_i will be frozen out completely. This looks (and indeed <u>is</u>) very serious, but when $n^2 > \beta^{-t}$ A is singular to working precision and one cannot expect to obtain any accuracy with any right-hand side without working to higher precision. When n^2 is large but not yet as large as β^t the loss of precision resulting from this freezing out of information is of comparable order of magnitude with that which results from the perturbations in A equivalent to the rounding errors made in the course of the solution.

3 PARTIAL DIFFERENTIAL EQUATIONS

Although it is a trivial problem the simple differential equation of section 2 exhibits some features which are common to many partial differential equations of elliptic type. If we consider for example Poisson's equation

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = f(x,y) \qquad (3.1)$$

over a square $0 \leq x \leq 1$, $0 \leq y \leq 1$ with $w = 0$ on the boundary then with a regular mesh of length $1/n$ and the usual 5 point approximation the right-hand side has elements $f(\frac{i}{n}, \frac{j}{n})/n^2$ while the exact solution u is tending to the solution of the continuous problem which is independent of n. The value of $\|u\|$ corresponding

to the solution of the discrete system must therefore be of the order of magnitude of n^2 times the right-hand side. The eigenvalues of the matrix of the linear system are

$$4 \sin^2\theta_i + 4 \sin^2\theta_j, \quad \theta_i = i\pi/2n \quad (i, j = 1, \ldots, n-1) \qquad (3.2)$$

so that

$$\lambda_{max} \doteq 8, \quad \lambda_{min} \doteq 2\pi^2/n^2 \qquad (3.3)$$

and hence, as before, the right-hand side must have substantial components in the direction of eigenvectors corresponding to the smaller eigenvalues. As $n \to \infty$, even a moderately smooth $f(x,y)$ guarantees that this is true but again we stress that there is already a very high probability that it is true for a random right-hand side.

However, if we go over to Laplace's equation, ie to $f(x,y) = 0$ and to non-zero values on the boundary a high percentage of elements of the right-hand side are zero; only those involving boundary values are non-zero. The right-hand side no longer involves n directly. The solution of the linear system tends to that of the continuous problem which is independent of n. For arbitrary boundary values, whether coming from a smooth function or not, the components of the right-hand side in the direction of the eigenvectors corresponding to the smallest eigenvalues must be very small since such components are amplified in the solution by a factor n^2. Again it is easy to verify this by examining the inner-products of the right-hand sides with these eigenvectors; we see that the non-zero components of the right-hand side align with negligible components in these eigenvectors. Since there are at most five non-zero elements in each row of A and these are -1, -1, 4, -1, -1 the elements of the exact solution of the discrete problem lie between the extreme values on the boundary, again showing that no amplifying factor n^2 occurs in this solution. Most elements of the right-hand side are zero and are exactly representable on the computer. Even if the non-zero elements are not exactly representable the error vector has the same distribution of non-zero elements as does the right-hand side itself and hence, as in the simpler problem, the potential danger of having a right-hand side of this special nature does not materialize.

For Poisson's equation with prescribed boundary conditions the right-hand side will be derived from both sources, the forcing function giving a contribution of order $1/n^2$ and the boundary a contribution which is not dependent on a power of n.

4 ERRORS IN THE COMPUTED SOLUTION

If we regard the equations Ax=b produced within the computer as exact for the moment, it is natural to ask whether the computed solutions derived by Gaussian elimination (before iterative refinement) will be of different accuracies according as the right-hand side is or is not deficient in eigenvectors corresponding to the smallest eigenvalues.

Now the rounding error analysis of Gaussian elimination (or of the Cholesky method, when A is positive definite) shows that the computed solution x_c of Ax=b is always the exact solution of some

$$(A+E_b)x_c = b \tag{4.1}$$

where E_b is indeed a function of b, but the analysis gives a uniform bound for it independent of b. The precise nature of the bound depends on the details of the computer arithmetic, whether partial or complete pivoting is used and whether growth takes place in the size of the elements during the course of elimination. When A is positive definite this last possibility is ruled out. (Since elliptic problems will mostly give positive definite systems this remark is of great relevance in our present discussion). If the statistical distribution of the rounding errors is taken into account then the error analysis gives very good reasons for expecting a bound of the form

$$\| E_b \| \leq \beta^{-t} f(n) \| A \| \tag{4.2}$$

with f(n) a very modest function of n. (Here n is the order of the matrix A and not the reciprocal of the mesh length). Finite difference approximations to elliptic differential equations usually lead to equations of band form with a band width which is small compared with the order of A and hence we may expect a particularly small E_b to be adequate.

Let us examine the consequences of (4.1) and (4.2) as far as the residuals are concerned. We have the standard result for linear systems

$$Ax_c = b-E_bx_c, \quad \| E_bx_c \| \leq \beta^{-t} f(n) \| A \| \| x_c \| . \tag{4.3}$$

Now in the boundary value problems without a forcing function we have seen that the solution is of the order of magnitude $\| b \| / \| A \|$ (and not the much larger $\| A^{-1} \| \| b \|$). Hence we have

$$\| E_bx_c \| \leq \beta^{-t} f(n) \| b \| \tag{4.4}$$

and assuming that f(n) is very modest indeed, Ax_c will be equal to b to within a very small relative error. One must not allow oneself to be deceived by this. It might be felt that since b comes entirely from the boundary conditions we have the exact solution corresponding to very slightly perturbed boundary values. This is not so. Most of the right-hand side consists of zeros and the components of Ax_c in these positions will be of the order of magnitude of $\beta^{-t}f(n)\|b\|$. As we shall see later this means that x_c is not nearly so accurate as it may appear from the closeness of Ax_c to the boundary values, though the latter is a reassuring feature of the computed solution.

Turning to the problem with zero boundary values but a non-zero f we now have an $\|x_c\|$ which is of the order of magnitude of $\|A^{-1}\|\|b\|$ and hence

$$\|E_b x_c\| \leq \beta^{-t}f(n)\|A\|\|A^{-1}\|\|b\| = \beta^{-t}f(n)K(A)\|b\| \qquad (4.5)$$

The residual relative to b is therefore likely to be affected by the factor K. (Remember that comments refer to probabilities; there is always the possibility that the rounding errors could by accident be particularly favourable).

When we have non-zero boundary values and a non-zero f then $\|x\|$ is again of the order of magnitude $\|b\|/\|A\|$. This may seem surprising but it should be remembered that the part of the right-hand side arising from f will be much smaller than that coming from the boundary values. For Poisson's equation, for example, it is smaller by the factor h^2. The norm of the right-hand side is therefore of the general order of magnitude of the boundary function and is largely independent of n. The residual is bounded by $f(n)\beta^{-t}\|b\|$ but this will nevertheless correspond to a high relative error for all those elements which are derived solely from f and do not involve boundary values; this, of course, will be most of them!

As far as the relative error in x_c itself is concerned we have

$$x_c = (A+E_b)^{-1}b = (I+A^{-1}E_b)^{-1}A^{-1}b = (I+A^{-1}E_b)^{-1}x \qquad (4.6)$$

and hence

$$x-x_c = A^{-1}E_b x_c \qquad (4.7)$$

$$\|x-x_c\| \leq \|A^{-1}E_b x_c\|. \qquad (4.8)$$

We have immediately that

$$\| x-x_c \| / \| x_c \| \leq \| A^{-1} \| \ \| E_b \| \leq \beta^{-t} f(n) \| A^{-1} \| \ \| A \| \qquad (4.9)$$

which is a standard result for linear systems and holds whatever the nature of the right-hand side. This bound contains the factor K and it is natural to ask whether this is always representative or whether for certain types of right-hand side

$$\| A^{-1} E_b x_c \| << \| A^{-1} \| \ \| E_b \| \ \| x_c \|. \qquad (4.10)$$

If this were to be true it would imply that for certain types of right-hand side E_b would be correlated with b in such a way as to make the inequality

$$\| A^{-1} E_b x_c \| \leq \| A^{-1} \| \ \| E_b \| \ \| x_c \| \qquad (4.11)$$

"exceptionally weak". The most promising right-hand sides would appear to be those for which x_c is 'large', ie of the order of magnitude of $\| A^{-1} \| \ \| b \|$ such as characterises those coming from zero-boundary conditions and non-zero forcing functions. We have been unable to identify any feature in the error analysis which would indicate that there is such a correlation. Indeed the most important contribution to E_b comes from the factorization of A in the Gaussian elimination, (or the related Cholesky factorization) and this is independent of b. It is not easy to see how this contribution could have a special relationship with any particular type of b. The further errors arising in the forward and backward substitutions are the ones which make E_b a function of b, but they do not usually contribute much to the error in x_c.

Thus the error analysis suggests that whatever the nature of the right-hand side b, the accuracy of the computed solution will fully reflect the condition of the matrix of coefficients. If this were not to be so there would need to be some special correlation in the errors arising in the solution process. That there is no such correlation in general would seem borne out by the experimental results which are described in the next two sections.

5 EXPERIMENTAL RESULTS

Numerical experiments were carried out for each of the following problems.

(I) $\frac{d^2w}{du^2} = f(x)$, $0 \leq x \leq 1$, $w(0) = w_0$, $w(1) = w_1$.

(a) With $f(x) = 0$. (b) With $w_0 = w_1 = 0$ and various functions $f(x)$.

(II) $\frac{d^4w}{dx^4} = f(x)$, $0 \leq x \leq 1$, $w(0) = w(1) = w''(0) = w''(1) = 0$ with various functions $f(x)$.

(III) $\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = f(x,y)$, $0 \leq x \leq 1$, $0 \leq y \leq 1$.

(a) With $f(x) = 0$ and w specified on the boundary.

(b) With $w = 0$ on the boundary and various functions $f(x,y)$.

The computations were performed on a PDP10 which has a 27 binary digit word and uses rounding. Hence $\frac{1}{2}\beta^{(1-t)} = 2^{-27} \doteq 7.45 \times 10^{-9}$. The method used was based on the LDL^T factorization of the matrix A. In all cases the computed solution x_1 was obtained using single-precision computation throughout with no double-precision accumulation of inner-products. One stage of iterative refinement was then performed using double-precision accumulation of inner-products in the computation of the residuals; this refined solution is denoted by x_2. In the tables below we list the following quantities.

(i) $$T = \|r\|_2/(\|x_1\|_2 \|A\|_2) = \|b-Ax_1\|_2/(\|x_1\|_2 \|A\|_2) . \quad (5.1)$$

The relation $r = b-Ax_1$ implies that

$$(A + rx_1^T/\|x_1\|_2^2)x_1 = b \quad (5.2)$$

and the computed solution x_1 is therefore the exact solution of $(A+E)x_1 = b$ with $E = rx_1^T/\|x_1\|_2^2$. Hence

$$\|E\|_2/\|A\|_2 = \|r\|_2 \|x_1\|_2/(\|x_1\|_2^2 \|A\|_2) = \|r\|_2/(\|x_1\|_2 \|A\|_2) = T \quad (5.3)$$

and the computed solution is exact for a perturbed matrix A with a relative perturbation T.

(ii) $$RE = \| x_1 - x_2 \|_2 / \| x_1 \|_2 . \tag{5.4}$$

The true relative error of the solution of the matrix equations is given by

$$\text{relative error} = \| x_1 - x_e \|_2 / \| x_e \|_2 , \tag{5.5}$$

where x_e denotes the exact solution. However, since in all of our examples x_1 had some correct figures and x_2 had additional correct figures, $\| x_1 \|_2 / \| x_e \|_2$ and $\| x_1 - x_2 \| / \| x_1 - x_e \|$ are both unity to at least a few binary digits, RE is essentially the relative error.

(iii) $$K(A) = \| A \|_2 \| A^{-1} \|_2 . \tag{5.6}$$

This is the ℓ_2 condition number of A with respect to inversion. For the simple problems of this exercise the singular values (which for these examples are the eigenvalues) are known explicitly and $K(A)$ was computed from the relation $K(A) = \sigma_1/\sigma_n$.

(iv) $$RT = RE/T . \tag{5.7}$$

Since $(A+E)x_1 = b$ we have

$$x_1 + A^{-1}Ex_1 = A^{-1}b = x_e \tag{5.8}$$

$$\| x_1 - x_e \|_2 \leq \| A^{-1} \|_2 \| E \|_2 \| x_1 \|_2 \tag{5.9}$$

$$\| x_1 - x_e \|_2 / \| x_1 \|_2 \leq \| A^{-1} \|_2 T \| A \|_2 = K(A)T \tag{5.10}$$

The left-hand side of (5.10) is essentially RE and we therefore have

$$RE \leq K(A)T . \tag{5.11}$$

From (5.7) this means that in all our examples we expect the relation $RT \leq K(A)$ to be satisfied. If the equivalent perturbations E in A are not in any way special then the inequalities on which the relation $RT \leq K(A)$ depends would not, in general, be weak and we would expect RT to be quite comparable with $K(A)$.

PROBLEM I

$\frac{d^2w}{dx^2} = 0.$ $w(0) = 0$, $w(1) = 1$

n	T	RE	K	RT
50	$(1.65)10^{-9}$	$(7.59)10^{-7}$	$(1.01)10^{3}$	$(4.59)10^{2}$
100	$(1.59)10^{-9}$	$(2.70)10^{-6}$	$(4.05)10^{3}$	$(1.70)10^{3}$
200	$(1.76)10^{-9}$	$(9.79)10^{-6}$	$(1.62)10^{4}$	$(5.57)10^{3}$
500	$(1.91)10^{-9}$	$(7.54)10^{-5}$	$(1.01)10^{5}$	$(3.94)10^{4}$
1,000	$(1.89)10^{-9}$	$(3.00)10^{-4}$	$(4.05)10^{5}$	$(1.59)10^{5}$
2,000	$(1.93)10^{-9}$	$(1.21)10^{-3}$	$(1.62)10^{6}$	$(6.26)10^{5}$
5,000	$(1.84)10^{-9}$	$(7.72)10^{-3}$	$(1.01)10^{7}$	$(4.19)10^{6}$
10,000	$(1.72)10^{-9}$	$(3.16)10^{-2}$	$(4.05)10^{7}$	$(1.84)10^{7}$

$\frac{d^2w}{dx^2} = 1$, $w(0) = w(1) = 0$

n	T	RE	K	RT
50	$(1.77)10^{-9}$	$(9.81)10^{-7}$	$(1.01)10^{3}$	$(5.56)10^{2}$
100	$(1.99)10^{-9}$	$(3.33)10^{-6}$	$(4.05)10^{3}$	$(1.67)10^{3}$
200	$(2.11)10^{-9}$	$(1.28)10^{-5}$	$(1.62)10^{4}$	$(6.08)10^{3}$
500	$(2.10)10^{-9}$	$(9.34)10^{-5}$	$(1.01)10^{5}$	$(4.45)10^{4}$
1,000	$(2.13)10^{-9}$	$(3.80)10^{-4}$	$(4.05)10^{5}$	$(1.78)10^{5}$
2,000	$(2.00)10^{-9}$	$(1.53)10^{-3}$	$(1.62)10^{6}$	$(7.64)10^{5}$
5,000	$(2.15)10^{-9}$	$(9.67)10^{-3}$	$(1.01)10^{7}$	$(4.50)10^{6}$
10,000	$(2.20)10^{-9}$	$(4.19)10^{-2}$	$(4.05)10^{7}$	$(1.90)10^{7}$

$\frac{d^2w}{dx^2} = x$, $w(0) = w(1) = 0$

n	T	RE	K	RT
50	$(2.20)10^{-9}$	$(9.79)10^{-7}$	$(1.01)10^{3}$	$(4.45)10^{2}$
100	$(2.44)10^{-9}$	$(3.32)10^{-6}$	$(4.05)10^{3}$	$(1.36)10^{3}$
200	$(2.15)10^{-9}$	$(1.26)10^{-5}$	$(1.62)10^{4}$	$(5.87)10^{3}$
500	$(2.38)10^{-9}$	$(9.36)10^{-5}$	$(1.01)10^{5}$	$(3.93)10^{4}$
1,000	$(2.39)10^{-9}$	$(3.78)10^{-4}$	$(4.05)10^{5}$	$(1.58)10^{5}$
2,000	$(2.34)10^{-9}$	$(1.52)10^{-3}$	$(1.62)10^{6}$	$(6.49)10^{5}$
5,000	$(2.30)10^{-9}$	$(9.64)10^{-3}$	$(1.01)10^{7}$	$(4.19)10^{6}$
10,000	$(2.19)10^{-9}$	$(4.20)10^{-2}$	$(4.05)10^{7}$	$(1.92)10^{7}$

$\frac{d^2w}{dx^2} = \sin \pi x$, $w(0) = w(1) = 0$

n	T	RE	K	RT
50	$(2.24)10^{-9}$	$(1.02)10^{-6}$	$(1.01)10^{3}$	$(4.57)10^{2}$
100	$(2.36)10^{-9}$	$(3.38)10^{-6}$	$(4.05)10^{3}$	$(1.43)10^{3}$
200	$(2.13)10^{-9}$	$(1.29)10^{-5}$	$(1.62)10^{4}$	$(6.04)10^{3}$
500	$(2.31)10^{-9}$	$(9.34)10^{-5}$	$(1.01)10^{5}$	$(4.03)10^{4}$
1,000	$(2.31)10^{-9}$	$(3.81)10^{-4}$	$(4.05)10^{5}$	$(1.65)10^{5}$
2,000	$(2.30)10^{-9}$	$(1.53)10^{-3}$	$(1.62)10^{6}$	$(6.68)10^{5}$
5,000	$(2.30)10^{-9}$	$(9.68)10^{-3}$	$(1.01)10^{7}$	$(4.22)10^{6}$
10,000	$(2.38)10^{-9}$	$(4.20)10^{-2}$	$(4.05)10^{7}$	$(1.77)10^{7}$

PROBLEM I

$\frac{d^2w}{dx^2} = \sin n\pi x,\ w(0) = w(1) = 0$ $\qquad$ $\frac{d^2w}{dx^2} = e^x,\ w(0) = w(1) = 0$

n	T	RE	K	RT	T	RE	K	RT
50	$(4.08)10^{-9}$	$(1.95)10^{-7}$	$(1.01)10^{3}$	$(4.78)10^{1}$	$(2.15)10^{-9}$	$(9.77)10^{-7}$	$(1.01)10^{3}$	$(4.55)10^{2}$
100	$(3.66)10^{-9}$	$(8.06)10^{-7}$	$(4.05)10^{3}$	$(2.20)10^{2}$	$(2.26)10^{-9}$	$(3.37)10^{-6}$	$(4.05)10^{3}$	$(1.49)10^{3}$
200	$(3.46)10^{-9}$	$(1.00)10^{-6}$	$(1.62)10^{4}$	$(2.90)10^{2}$	$(2.22)10^{-9}$	$(1.27)10^{-5}$	$(1.62)10^{4}$	$(5.73)10^{3}$
500	$(3.46)10^{-9}$	$(1.12)10^{-5}$	$(1.01)10^{5}$	$(3.24)10^{3}$	$(2.41)10^{-9}$	$(9.35)10^{-5}$	$(1.01)10^{5}$	$(3.89)10^{4}$
1,000	$(3.77)10^{-9}$	$(8.22)10^{-5}$	$(4.05)10^{5}$	$(2.18)10^{4}$	$(2.21)10^{-9}$	$(3.80)10^{-4}$	$(4.05)10^{5}$	$(1.72)10^{5}$
2,000	$(3.78)10^{-9}$	$(2.17)10^{-4}$	$(1.62)10^{6}$	$(5.74)10^{4}$	$(2.25)10^{-9}$	$(1.53)10^{-3}$	$(1.62)10^{6}$	$(6.81)10^{5}$
5,000	$(2.50)10^{-9}$	$(9.29)10^{-3}$	$(1.01)10^{7}$	$(3.72)10^{6}$	$(2.30)10^{-9}$	$(6.73)10^{-3}$	$(1.01)10^{7}$	$(2.93)10^{6}$
10,000	$(2.32)10^{-9}$	$(4.08)10^{-2}$	$(4.05)10^{7}$	$(1.76)10^{7}$	$(2.33)10^{-9}$	$(4.20)10^{-2}$	$(4.05)10^{7}$	$(1.81)10^{7}$

PROBLEM II

$\frac{d^4w}{dx^4} = 1, w(0) = w(1) = w''(0) = w''(1) = 0$ $\qquad$ $\frac{d^4w}{dx^4} = x, w(0) = w(1) = w''(0) = w''(1) = 0$

n	T	RE	K	RT	T	RE	K	RT
30	$(2.30)10^{-9}$	$(9.87)10^{-5}$	$(1.33)10^{5}$	$(4.29)10^{4}$	$(2.30)10^{-9}$	$(9.97)10^{-5}$	$(1.33)10^{5}$	$(4.34)10^{4}$
60	$(2.39)10^{-9}$	$(1.88)10^{-3}$	$(2.13)10^{6}$	$(7.88)10^{5}$	$(2.43)10^{-9}$	$(1.88)10^{-3}$	$(2.13)10^{6}$	$(7.75)10^{5}$
90	$(2.45)10^{-9}$	$(1.06)10^{-2}$	$(1.08)10^{7}$	$(4.33)10^{6}$	$(2.17)10^{-9}$	$(1.06)10^{-2}$	$(1.08)10^{7}$	$(4.90)10^{6}$
120	$(2.36)10^{-9}$	$(2.96)10^{-2}$	$(3.41)10^{7}$	$(1.25)10^{7}$	$(2.36)10^{-9}$	$(2.94)10^{-2}$	$(3.41)10^{7}$	$(1.24)10^{7}$
150	$(2.17)10^{-9}$	$(6.80)10^{-2}$	$(8.32)10^{7}$	$(3.13)10^{7}$	$(2.22)10^{-9}$	$(6.77)10^{-2}$	$(8.32)10^{7}$	$(3.04)10^{7}$

PROBLEM II

$\frac{d^4w}{dx^4} = \sin \pi x, w(0) = w(1) = w''(0) = w''(1) = 0$ $\qquad$ $\frac{d^4w}{dx^4} = e^x, \; w(0) = w(1) = w''(0) = w''(1) = 0$

n	T	RE	K	RT	T	RE	K	RT
30	$(2.67)10^{-9}$	$(9.84)10^{-5}$	$(1.33)10^{5}$	$(3.68)10^{4}$	$(2.52)10^{-9}$	$(9.92)10^{-5}$	$(1.33)10^{5}$	$(3.93)10^{4}$
60	$(2.64)10^{-9}$	$(1.88)10^{-3}$	$(2.13)10^{6}$	$(7.14)10^{5}$	$(2.30)10^{-9}$	$(1.88)10^{-3}$	$(2.13)10^{6}$	$(8.19)10^{5}$
90	$(2.46)10^{-9}$	$(1.06)10^{-2}$	$(1.08)10^{7}$	$(4.31)10^{6}$	$(2.36)10^{-9}$	$(1.06)10^{-2}$	$(1.08)10^{7}$	$(4.49)10^{6}$
120	$(2.35)10^{-9}$	$(2.96)10^{-2}$	$(3.41)10^{7}$	$(1.26)10^{7}$	$(2.47)10^{-9}$	$(2.95)10^{-2}$	$(3.41)10^{7}$	$(1.20)10^{7}$
150	$(2.23)10^{-9}$	$(6.80)10^{-2}$	$(8.32)10^{7}$	$(3.06)10^{7}$	$(2.44)10^{-9}$	$(6.79)10^{-2}$	$(8.32)10^{7}$	$(2.78)10^{7}$

PROBLEM III

$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 0$, w=1, 2, 3, 4 respectively on four sides of boundary

$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 0$, w=1 at just one point on boundary, w=0 at all other boundary points

N	T	RE	K	RT	T	RE	K	RT
100	$(5.34)10^{-9}$	$(6.58)10^{-8}$	$(4.84)10^{1}$	$(1.23)10^{1}$	$(5.37)10^{-9}$	$(3.01)10^{-8}$	$(4.84)10^{1}$	$(5.61)10^{0}$
225	$(6.66)10^{-9}$	$(3.10)10^{-8}$	$(1.03)10^{2}$	$(4.66)10^{0}$	$(5.62)10^{-9}$	$(3.77)10^{-8}$	$(1.03)10^{2}$	$(6.71)10^{0}$
400	$(7.85)10^{-9}$	$(5.56)10^{-8}$	$(1.78)10^{2}$	$(7.08)10^{0}$	$(5.12)10^{-9}$	$(5.01)10^{-8}$	$(1.78)10^{2}$	$(9.77)10^{0}$
625	$(8.68)10^{-9}$	$(2.15)10^{-7}$	$(2.73)10^{2}$	$(2.47)10^{1}$	$(6.27)10^{-9}$	$(1.13)10^{-7}$	$(2.73)10^{2}$	$(1.80)10^{1}$
900	$(8.78)10^{-9}$	$(2.52)10^{-7}$	$(3.89)10^{2}$	$(2.87)10^{1}$	$(5.38)10^{-9}$	$(1.66)10^{-7}$	$(3.89)10^{2}$	$(3.09)10^{1}$
1225	$(9.83)10^{-9}$	$(7.07)10^{-7}$	$(5.25)10^{2}$	$(7.19)10^{1}$	$(6.49)10^{-9}$	$(2.62)10^{-7}$	$(5.25)10^{2}$	$(4.04)10^{1}$

$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 1$, w = 0 on boundary

$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = x+y$, w = 0 on boundary

N	T	RE	K	RT	T	RE	K	RT
25	$(3.68)10^{-9}$	$(8.17)10^{-9}$	$(1.39)10^{1}$	$(2.22)10^{0}$	$(3.75)10^{-9}$	$(1.23)10^{-8}$	$(1.39)10^{1}$	$(3.27)10^{0}$
100	$(5.69)10^{-9}$	$(9.02)10^{-8}$	$(4.84)10^{1}$	$(1.58)10^{1}$	$(5.54)10^{-9}$	$(6.30)10^{-8}$	$(4.84)10^{1}$	$(1.14)10^{1}$
225	$(6.62)10^{-9}$	$(5.00)10^{-8}$	$(1.03)10^{2}$	$(7.55)10^{0}$	$(6.57)10^{-9}$	$(6.77)10^{-8}$	$(1.03)10^{2}$	$(1.03)10^{1}$
400	$(7.63)10^{-9}$	$(6.77)10^{-8}$	$(1.78)10^{2}$	$(8.88)10^{0}$	$(8.08)10^{-9}$	$(6.11)10^{-8}$	$(1.78)10^{2}$	$(7.56)10^{0}$
625	$(8.57)10^{-9}$	$(1.92)10^{-7}$	$(2.73)10^{2}$	$(2.25)10^{1}$	$(8.67)10^{-9}$	$(1.80)10^{-7}$	$(2.73)10^{2}$	$(2.08)10^{1}$
900	$(9.05)10^{-9}$	$(2.31)10^{-7}$	$(3.89)10^{2}$	$(2.55)10^{1}$	$(9.36)10^{-9}$	$(2.06)10^{-7}$	$(3.89)10^{2}$	$(2.20)10^{1}$

6 COMMENTS ON RESULTS

For problems I and II the largest value of T is $(4.08)10^{-9}$, this being achieved for problem I with $f(x) = \sin n\pi x$, $w(0) = w(1) = 0$ and $n = 50$, the smallest value of n appearing in the tables. The average value of T is approximately 2×10^{-9}. Notice that even the maximum value is less than $(7.45)10^{-9}$, the computer precision. The cumulative effect of all the rounding errors made in the course of the solution could be caused by remarkably small perturbations in the matrix A; indeed in every example this equivalent perturbation is actually considerably smaller than that which could arise merely from the digital representation of A within the computer if the differential equation were such that A were not exactly representable.

For problems I and II, RT is generally of the order of magnitude $0.4K$. We have remarked that the maximum value of RT is K; the fact that it is so 'close' to K indicates that the equivalent perturbations in A corresponding to the rounding errors are not in any way special.

Since the values of T are so small and are roughly constant for all the examples of problems I and II the relative error in the solution is primarily determined by K. For problem I, even with a value of n as large as 10,000 the computed solution still has some significance; for problem II the value of K is already roughly 10^8 when $n = 150$ and hence this is about the largest value of n for which we can expect the solution to have any significance on a PDP10. The matrices associated with problems I and II are positive definite band matrices of widths 3 and 5 respectively. The standard backward error analysis predicts quite accurately that T can be expected to be independent of n.

Turning now to problem III we give two sets of results for Laplace's equation and two for Poisson's equations with zero boundary conditions. In the tables N gives the number of internal mesh points in the square so that $N = 900$ here corresponds to $n = 30$ for problem I. The maximum value of T is now $(9.83)10^{-9}$ and in general T tends to increase with the number of mesh points; since the matrices are positive definite bands of width $2\sqrt{N}$, error analysis predicts that allowing for the statistical distribution of rounding errors T might be expected to increase roughly in line with $N^{1/4}$ and this is indeed true. As with the problems associated with ordinary differential equations the errors in the solution could be caused by very small perturbations in the elements, perturbations of 1 in the last digit being almost adequate.

The values of RT, which we know must be bounded by those of K, do in fact vary from $0.25K$ to $0.04K$, there being little significant difference between the four sets of examples; no marked trend with the variation in N is observable though for each of the examples in which N=25 (the smallest quoted value of N) RT/K takes one of its higher values. The right-hand sides associated with the first two sets of examples contain negligible components of the eigenvectors corresponding to the smaller eigenvalues while those associated with the other two sets are particularly rich in just these eigenvectors. In spite of this none of the parameters we list is at all sensitive to this feature, showing, as was expected, that there is no special correlation between the equivalent perturbations in A and the right-hand side.

ACKNOWLEDGEMENT

The authors wish to thank D W Martin for his careful reading of the manuscript and for his helpful suggestions on the presentation.

REFERENCES

1 CLINE, A K, MOLER, C B, STEWART, G W and WILKINSON, J H. An estimate of the condition number of a matrix. SIAM J. Numer. Anal., 1979, 16, 368-375.

2 WILKINSON, J H. Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach., 1961, 8, 281-330.

3 WILKINSON, J H. Rounding errors in algebraic processes. Notes on Applied Science No 32, Her Majesty's Stationery Office, London, 1963.

4 WILKINSON, J H. The algebraic eigenvalue problem. Oxford University Press, London, 1965.