SIAM J. NUMER. ANAL.
Vol. 16, No. 2, April 1979

0036-1429/79/1602-0017 $01.00/0

AN ESTIMATE FOR THE CONDITION NUMBER OF A MATRIX*

A. K. CLINE,† C. B. MOLER,‡ G. W. STEWART¶ AND J. H. WILKINSON§

Abstract. It is important in practice when solving linear systems to have an economical method for estimating the condition number $\kappa(A)$ of the matrix of coefficients. An algorithm involving $O(n^2)$ arithmetic operations is described; it gives a reliable indication of the order of magnitude of $\kappa(A)$.

1. Introduction. The sensitivity of the solution of a nonsingular system of linear equations

$$Ax = b \tag{1.1}$$

with respect to perturbation E in A is directly related to the *condition number* $\kappa(A)$ of A with respect to inversion. If

$$(A+E)y = b \tag{1.2}$$

then

$$\|y-x\|/\|y\| \leqq \|A^{-1}E\|. \tag{1.3}$$

Let

$$\varepsilon = \|E\|/\|A\| \tag{1.4}$$

be the relative error in A. Then the error in x, measured relative to the perturbed solution y, satisfies

$$\|y-x\|/\|y\| \leqq \varepsilon \|A\| \, \|A^{-1}\|. \tag{1.5}$$

The condition number of A with respect to inversion is defined by

$$\kappa(A) = \|A\| \, \|A^{-1}\|. \tag{1.6}$$

Obviously $\kappa(A)$ is a function of the norm which is used. We shall be interested in l_1, l_2 and l_∞ norms and the variation of κ with the norm will be its least interesting aspect.

We observe that in deriving (1.3) we have majorized $\|A^{-1}Ey\|$ by $\|A^{-1}E\| \, \|y\|$ and for some y and E this will be an overestimate. When $Ey = 0$, for example, we have $y = x$. Further, in deriving (1.5) from (1.3), we have majorized $\|A^{-1}E\|$ by $\|A^{-1}\| \, \|E\|$. When $\kappa(A)$ is large this will be a severe overestimate for some E; for example, if $E = \varepsilon A$ we have $\|A^{-1}E\| = \varepsilon$ which is independent of κ.

In spite of these remarks the inequality (1.3) will be quite realistic for almost all perturbations E.

The sensitivity to perturbations e in b is also related to $\kappa(A)$. If

$$Ay = b + e, \qquad \|e\|/\|b\| \leqq \varepsilon, \tag{1.7}$$

we have

$$\|y-x\| \leqq \|A^{-1}e\| \leqq \varepsilon \|A^{-1}\| \, \|b\|, \tag{1.8}$$

* Received by the editors December 5, 1977. This work was performed under the auspices of the U.S. Energy Research and Development Administration and the U.S. National Science Foundation.

† Computer Science Department, University of Texas at Austin, Austin, Texas 78712.

‡ Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131.

¶ Computer Science Department, University of Maryland, College Park, Maryland 20742.

§ National Physical Laboratory, Teddington, England.

while from (1.1)

$$\|b\| \leqq \|A\| \, \|x\|. \tag{1.9}$$

Hence

$$\|y - x\|/\|x\| \leqq \varepsilon \kappa(A), \tag{1.10}$$

which is essentially the same bound as in (1.5). However, this last result is very deceptive. Although the bound is attainable the probability that (1.10) is realistic is rather low. Its derivation is dependent on (1.9) and this relation is very weak for almost all b when $\kappa(A)$ is large. Such probability considerations are of vital importance in what follows.

2. Analysis in terms of SVD of A. Deeper insight into the effect of perturbations may be gained via the singular value decomposition (SVD) of A. We write

$$A = U \Sigma V^T \tag{2.1}$$

where U and V are orthogonal and $\Sigma = \text{diag}\,(\sigma_i)$, the σ_i being the singular values of A. From (2.1)

$$A v_i = \sigma_i u_i, \qquad A^T u_i = \sigma_i v_i \tag{2.2}$$

and since $\|A\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = \sigma_n^{-1}$ we have $\|A\|_2 \|A^{-1}\|_2 = \sigma_1/\sigma_n$.

We may expand both b and e in terms of the orthogonal vectors u_i in the form

$$b = \|b\| \Sigma \alpha_i u_i \qquad (\Sigma \alpha_i^2 = 1), \tag{2.3}$$

$$e = \varepsilon \|b\| \Sigma \beta_i u_i \qquad (\Sigma \beta_i^2 = 1), \tag{2.4}$$

and from (2.2)

$$x = \|b\| \Sigma (\alpha_i/\sigma_i) v_i, \qquad y - x = \varepsilon \|b\| \Sigma (\beta_i/\sigma_i) v_i. \tag{2.5}$$

It is clear that

$$\|y - x\|/\|x\| = \varepsilon \sigma_1/\sigma_n = \varepsilon \kappa(A) \tag{2.6}$$

only in the extreme case when

$$b = \|b\| u_1, \qquad e = \varepsilon \|b\| u_n. \tag{2.7}$$

When σ_1/σ_n is large $\|y - x\|/\varepsilon \|x\|$ will be of the order of unity for any e satisfying (1.7) unless α_n, which gives the exponent of b in the direction of u_n, is exceptionally small. Also since

$$\|x\|/\|b\| = [\Sigma (\alpha_i/\sigma_i)^2]^{1/2} \tag{2.8}$$

this ratio will be of order σ_n^{-1} (i.e. $\|A^{-1}\|_2$) unless α_n is exceptionally small.

3. Estimates for $\kappa(A)$. It is important in practice when solving linear systems to have some estimate of $\kappa(A)$ which will give at least a reliable indication of its order of magnitude. When a linear system $Ax = b$ has been solved by a direct method one has some factorization of A and it is natural to make use of this in determining the estimate of $\kappa(A)$. For the moment we assume that we have exact factors of A and comment later on the effect of the rounding errors made when deriving the factorization.

The problem is perhaps simplest when we have a factorization of the form

$$A = QR \tag{3.1}$$

where Q is orthogonal and R is upper triangular. In this case

$$\|A\|_2 = \|R\|_2, \quad \|A^{-1}\|_2 = \|R^{-1}\|_2, \quad \kappa_2(A) = \kappa_2(R), \tag{3.2}$$

where $\kappa_2(\cdot)$ denotes the condition number corresponding to the l_2 norm. However, we concentrate on an estimate of $\kappa_1(R)$, since the l_1 norm can be computed with less expense. Certainly $\|R\|_1$ is immediately available, while $\|R^{-1}\|_1$ may be obtained by computing the columns of R^{-1} one by one. There is no need to store R^{-1} since $\|R^{-1}\|_1$ is merely the maximum of the l_1 norms of its columns. However, the computation of R^{-1} involves $n^3/6$ multiplication and one would like to derive an estimate of $\|R^{-1}\|$ by a process which involves only $O(n^2)$ operations and at most $O(n)$ additional storage locations.

Now we know that if

$$Rx = b \tag{3.3}$$

then $\max(\|x\|_2/\|b\|_2) = \sigma_n^{-1} = \|R^{-1}\|_2$, this bound being attained when $b = u_n$ where $U\Sigma V^T$ is now the SVD of R. Note that

$$A = QR = (QU)\Sigma V^T \tag{3.4}$$

so that the SVD's of A and R are closely related. Is it possible to choose a right-hand side b in some ad hoc manner so that $\|x\|_2$ has a near-maximum norm? From equation (2.8) we see that when $\kappa(A)$ is large the probability that $\|x\|_2/\|b\|_2$ will give a good estimate of σ_n^{-1} is quite high. Our object is to choose b in such a way as to reinforce this natural probability.

It is tempting to suggest that one takes a random b, computes the corresponding x and then uses this x as a new right-hand side. Indeed this is effectively what is done when one wishes to find the eigenvector of R of smallest modulus by inverse iteration. However, it is well known that a second step of inverse iteration is often extremely unsatisfactory. The problem is that in order to obtain a 'large' solution to the system $Rx = b$ one requires a right-hand side in which u_n is substantially represented. The x derived at the end of the first step will usually have v_n as its dominant component; if v_n is almost orthogonal to u_n the latter will be poorly represented in x. This is a surprisingly common situation.

Discussion in terms of the SVD, however, suggests the use of the two-step process

$$R^T x = b, \qquad Ry = x. \tag{3.5}$$

If then

$$b = \Sigma \alpha_i v_i \tag{3.6}$$

we have

$$x = \Sigma(\alpha_i/\sigma_i)u_i, \qquad y = \Sigma(\alpha_i/\sigma_i^2)v_i. \tag{3.7}$$

Provided b has a component of v_n which is not unduly small the vector y is likely to be completely dominated by its component of v_n. Unlike the situation with two steps of inverse iteration we now have the full benefit of the factor σ_i^{-2}. From (3.7)

$$\|y\|_2/\|x\|_2 = [\Sigma(\alpha_i/\sigma_i^2)^2 / \Sigma(\alpha_i/\sigma_i)^2]^{1/2} \tag{3.8}$$

and we shall have a good estimate of σ_n^{-1} provided α_n/σ_n is not very small relative to the other α_i/σ_i. This is a much healthier situation, the more so in the very important case when $\sigma_n \ll \sigma_1$, i.e. when A is ill-conditioned.

If we start then from a *random* vector b the probability of obtaining a good estimate for $\|R^{-1}\|$ and hence for $\|A^{-1}\|$ is high. We have the possibility of enhancing this probability by a judicious choice of b. Our object is to choose b so that the solution x of $R^T x = b$ is such that $\|x\|/\|b\|$ is as large as possible. The following simple strategy suggests itself.

Let us take $b_s = \pm 1$, the sign being determined at the stage when x_s is computed via the relation

$$r_{ss}x_s = b_s - (r_{1s}x_1 + \cdots + r_{s-1,s}x_{s-1}). \tag{3.9}$$

The sign of b_s is chosen to be the opposite of that of the inner-product in parentheses on the right-hand side of (3.9). This gives $|x_s|$ the larger of the two possible values.

This strategy was very successful in practice, but the following matrix of order 4 revealed its weakness. Let

$$R^T = \left[\begin{array}{c|c} I & 0 \\ \hline kE & I \end{array}\right], \qquad E = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{3.10}$$

Then

$$R^{-T} = \left[\begin{array}{c|c} I & 0 \\ \hline -kE & I \end{array}\right], \qquad \|R^T\|_1 = \|R^{-T}\|_1 = 1 + 2k. \tag{3.11}$$

If k is large, both $\|R^T\|_1$ and $\|R^{-T}\|_1$ are large. Without loss of generality we can assume $b_1 = +1$, giving $x_1 = 1$. Either choice of b_2 gives $|x_2| = 1$. If we take $b_2 = +1$ then both $|x_3|$ and $|x_4|$ are unity and we shall not have a 'large' x. Note that if the ambiguity in choice is resolved by taking $b_2 = -1$ we have a similar failure with

$$E = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{3.12}$$

Returning to the E of (3.10) and assuming that we take $b_s = +1$ on each occasion we have

$$R^T b = b, \qquad Rb = b \tag{3.13}$$

and the process fails to indicate the ill-condition of R.

Clearly the weakness in the strategy is that the decision on the sign of b_s takes place on a purely local criterion. The value of such x_s affects all later values x_i and we need some 'look-ahead' feature in our strategy. This leads naturally to the following improved strategy.

Each x_i $(i > s)$ is determined by the relation

$$r_{ii}x_i = (-r_{1i}x_1 - \cdots - r_{s-1,i}x_{s-1}) + (-r_{si}x_s - \cdots - r_{i-1,i}x_{i-1} + b_i) \tag{3.14}$$

where the right-hand side has been split into two parts. The first part is determined irrevocably once $x_1, \cdots, x_{s-1}$ have been assigned. We denote this part by $p_i^{(s-1)}$. At the stage when we are about to assign b_s and compute x_s let us assume that we have already computed and stored the quantities $p_i^{(s-1)}$ $(i = s, \cdots, n)$. Now the two possible values of x_s are given by

$$r_{ss}x_s = -p_s^{(s-1)} + b_s = -p_s^{(s-1)} \pm 1. \tag{3.15}$$

We denote these two possible values of x_s by x_s^+ and x_s^-. If we compute both of these for the moment we can use these to give two different sets of updates of the

$p_i^{(s-1)}$ to $p_i^{(s)}$. We may write

$$p_i^{(s)+} = p_i^{(s-1)} + r_{si}x_s^+, \qquad p_i^{(s)-} = p_i^{(s-1)} + r_{si}x_s^- \tag{3.16}$$

and our decision should depend on the size of these $|p_i^{(s)}|$ as well as on the size of x_s itself. As a reasonable criterion we could take $b_s = \pm 1$ according as

$$|-p_s^{(s-1)}+1| + \sum_{i=s+1}^{n} |p_i^{(s)+}| \geqq \text{ or } < |-p_s^{(s-1)}-1| + \sum_{i=s+1}^{n} |p_i^{(s)-}|. \tag{3.17}$$

There is about twice as much work in the solution of $R^Tx = b$ with this improved strategy as with the original but since we have subsequently to solve $Ry = x$ this reduces the overall factor to 1.5. It will be seen that this modification immediately deals with the example of (3.10). Indeed if

$$E = \begin{bmatrix} \pm 1 & \pm 1 \\ \pm 1 & \pm 1 \end{bmatrix} \tag{3.18}$$

then with all possible combinations of signs an x is produced such that $\|x\|_1 > k$.

Two comments may be made on this second strategy.

(i) Since it is the size of x which really interests us it might be advisable to replaced (3.17) by the criterion

$$\frac{|-p_s^{(s-1)}+1|}{|r_{ss}|} + \sum_{i=s+1}^{n} \frac{|p_i^{(s)+}|}{|r_{ii}|} \geqq \text{ or } < \frac{|-p_s^{(s-1)}-1|}{|r_{ss}|} + \sum_{i=s+1}^{n} \frac{|p_i^{(s)-}|}{|r_{ii}|}. \tag{3.19}$$

However, this modification increases the volume of computation appreciably.

(ii) Matrices which arise in practice can in no sense be said to have random elements. Many of them will turn out to be very special and choosing a vector b with elements ± 1 must increase the probability of an accidental failure. If one takes $b_i = \pm\theta_i$ with the sign chosen as in our second strategy but with each θ_i a random number between $\frac{1}{2}$ and 1 the chances of accidental failure must be diminished. Indeed if b_i is chosen in this way even the *first* strategy is virtually certain to succeed with the example of (3.10).

4. Estimate from LU decomposition. In practice dense systems of linear equations are most commonly solved by Gaussian elimination with some form of pivoting. This provides permutation matrices P and Q, a unit lower-triangular matrix L and an upper-triangular matrix U such that

$$PAQ = LU. \tag{4.1}$$

The matrices P and Q cover interchanges resulting from pivoting. With partial pivoting $Q = I$ and with either partial pivoting or complete pivoting

$$|l_{ij}| \leqq 1. \tag{4.2}$$

For simplicity we shall write A in place of PAQ from now on. Our previous discussion suggests that we now use the two-step process

$$(LU)^Tx = b, \qquad LUy = x \tag{4.3}$$

and use $\|y\|/\|x\|$ as our estimate of $\|A^{-1}\|$. The first step then is the solution of $U^TL^Tx = b$ which is done in the two stages

$$U^Tz = b, \qquad L^Tx = z. \tag{4.4}$$

We can use either of the strategies discussed in the previous section to attempt to maximize z.

That the first strategy described in § 3 may fail is shown by the example

$$U^T = \left[\begin{array}{c|c} I & 0 \\ \hline kE & I \end{array}\right], \qquad L^T = \left[\begin{array}{c|c} I & E \\ \hline U & I \end{array}\right] \tag{4.5}$$

with E as defined in (3.10). This leads to the choice $b_s = +1$ $(s = 1, \cdots, 4)$ and

$$U^T b = b, \qquad L^T b = b, \tag{4.6}$$

$$Lb = b, \qquad Ub = b. \tag{4.7}$$

No indication is given of the ill-condition of A. The second strategy immediately overcomes this difficulty.

There remains the potential weakness that any advantage we may have gained by a good choice of b when computing z may be vitiated when we compute x. (It should be appreciated that when solving the pair $R^T x = b$, $Ry = x$ this danger does not exist; moderate success with $R^T x = b$ ensures almost complete success when solving $Ry = x$.) However, the danger of this would appear to be slight. Pivoting usually ensures that any ill-condition in A is reflected in a corresponding ill-condition in U; L is usually quite well-conditioned. The ill-condition of U is to be expected since when A is exactly singular some u_{ii}, usually u_{nn}, must be zero. The most common situation with an ill-conditioned A is that u_{nn} is small. It must be emphasized though that A *can* be almost singular without *any* u_{ii} being unduly small.

That L can be ill-conditioned in spite of pivoting is illustrated by the example

$$l_{ii} = 1, \qquad l_{ij} = -1 \qquad (i > j), \tag{4.8}$$

for which the maximum element of L^{-1} is 2^{n-2}; indeed L is "almost singular" even for quite modest values of n. If such an L is associated with a well-conditioned U then decisions should be made when working with L and not with U. The danger is exemplified by the case when U is diagonal with $u_{nn} = 1$, $u_{ii} = -1$ $(i \neq n)$. Then the solution of $U^T z = b$ is $(-1, -1, \cdots, -1, 1)$ with either of our two strategies and the solution of $L^T x = z$ is $(0, 0, \cdots, 0, 1)$. In solving $LUy = x$ there is no further amplification. Examples of this kind are extremely special and the danger in practice is likely to be greatly diminished by taking $b_i = \pm\theta_i$ as discussed under (ii) in § 3.

5. Rounding errors in the factors. In practice the factors will not be exact but if the factorization algorithm is stable the computed factors will be exact for some matrix $(A + E)$ where $\|E\|/\|A\|$ is some fairly modest multiple of the machine precision. The errors made in solving the triangular systems are unlikely to be of much importance in practice particularly as we are trying to use right-hand sides such that the solution truly reflects the ill-condition, if any. However, it must be faced that at best we can obtain only $\kappa(A + E)$. Since $A + E$ may be singular when A is not and vice-versa, there is naturally a limitation on the information that can be derived from the computed factors.

This limitation becomes serious only when A is very close to singularity. For if

$$\|E\|/\|A\| = \varepsilon \tag{5.1}$$

we have

$$(1 - \varepsilon)\|A\| \leqq \|A + E\| \leqq (1 + \varepsilon)\|A\| \tag{5.2}$$

while

$$\frac{\|A^{-1}\|(1-2\varepsilon\kappa)}{1-\varepsilon\kappa} \leqq \|(A+E)^{-1}\| \leqq \frac{\|A^{-1}\|}{1-\varepsilon\kappa}. \tag{5.3}$$

Combining (5.2) and (5.3), we obtain

$$\frac{(1-\varepsilon)(1-2\varepsilon\kappa)}{1-\varepsilon\kappa} \leqq \frac{\kappa(A+E)}{\kappa(A)} \leqq \frac{(1+\varepsilon)}{1-\varepsilon\kappa}. \tag{5.4}$$

Results are not likely to be of interest unless, say, $\varepsilon\kappa(A) \leqq 0.1$ and we have then

$$\frac{8}{9}(1-\varepsilon) \leqq \frac{\kappa(A+E)}{\kappa(A)} \leqq \frac{10}{9}(1+\varepsilon).$$

Since we are primarily concerned with the order of magnitude of $\kappa(A)$, the implicit use of $A+E$ in place of A is of little importance.

6. Implementation. This technique is implemented in LINPACK, a collection of FORTRAN subroutines for solving various forms of linear equations that is being developed by Argonne National Laboratory and three universities. The actual details of the implementation are described in the LINPACK documentation [1].

Since the technique is designed to cause growth in the size of the elements of various vectors, there is a definite possibility of floating point arithmetic overflow. In the Gaussian elimination subroutines, for example, the most crucial point is the division by the diagonal elements of U in the solution of $U^T z = b$. The elements of b are in principle ± 1, but since only the direction of z is important, b can be rescaled as necessary to avoid any dangerous divisions. Consider the 3×3 example

$$\begin{pmatrix} u_{11} & 0 & 0 \\ u_{12} & \eta & 0 \\ u_{13} & u_{23} & u_{33} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \pm 1 \\ \pm 1 \\ \pm 1 \end{pmatrix}$$

where η is very small. The first equation gives $z_1 = 1/u_{11}$ with no difficulty. Then the η in the second equation is noted. The right-hand side and the solution developed so far can both be scaled by η, giving

$$z_1 = \eta/u_{11}, \qquad z_2 = \pm 1 - u_{12}/u_{11}.$$

Finally,

$$z_3 = (\pm\eta - u_{13}z_1 - u_{23}z_2)/u_{33}.$$

No division by η is required. In the extreme case when $\eta = 0$, the right-hand side becomes the zero vector and z becomes a null vector of U^T. The subsequent steps produce a null vector of A, which is exactly what is desired.

The LINPACK subroutines are designed with the assumption that overflows are fatal errors, but that underflows are quietly replaced by zeros. Consequently, a quantity RCOND which estimates $1/\kappa(A)$ is computed. If $\kappa(A)$ is very large, this may underflow to zero. If an exact zero diagonal element occurs in the triangular factors, the rescaling strategy automatically results in a zero RCOND. Consequently, there is no necessity to make exact singularity a special case.

The programs used in testing LINPACK evaluate the condition estimator by also inverting the matrices and computing the actual $\kappa_1(A)$. The ratio of the estimated condition to the actual condition is printed out as one of the testing diagnostics. This

ratio must be less than 1.0 if the programs are working correctly. The ratio is regarded as "suspicious" if it is less than the arbitrary value 0.1. We have no justification for picking 0.1, except that the estimator very rarely produces a suspicious estimate.

In one test, J. T. Goodman of the University of New Mexico generated 1350 matrices with elements chosen randomly from various distributions. The orders of the matrices ranged from $n = 10$ to $n = 50$.

Of 550 matrices generated with elements normally distributed with mean zero and variance one, only one matrix was noted in which the ratio was less than 0.1. The matrix was of order 30 and the ratio was 0.077.

Of 300 matrices generated with elements uniformly distributed on the interval $(-1, 1)$, all ratios were greater than 0.1 with the majority of the ratios between 0.55 and 0.65. No ratios higher than 0.8 were observed.

Of 400 matrices generated with elements -1, 0, or 1 with equal probability, two matrices exhibited ratios less than 0.1. One matrix was of order 20 with a ratio of 0.062, the other a matrix of order 30 with a ratio of 0.092. The majority of the ratios were between 0.45 and 0.55. No ratio higher than 0.8 was observed.

Of 100 matrices of order 10 generated from Householder reflections, all ratios were greater than 0.1. Indeed, the condition estimator performed better on these matrices of order 10 than on any of the other three types of matrices of order 10. Four ratios were observed in the 0.95 to 1.0 range. The majority of the ratios were in the 0.50 to 0.55 range.

None of these matrices were particularly badly conditioned and so ratios near 1.0 are not expected.

In a second test, several dozen matrices were generated which had condition numbers in the range from 10^3 to 10^{12}. For these matrices, the estimate is quite accurate and ratios between 0.99 and 1.00 were fairly common, although this is of no great importance.

REFERENCE

[1] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.