STATE UNIVERSITY OF NEW YORK
AT STONY BROOK

COLLEGE OF
ENGINEERING

REPORT No. 117

SOME COMMENTS ON THE SOLUTION OF LINEAR EQUATIONS

by

R. P. Tewarson and B. Ramnath

AUGUST 7, 1968

REPORT NO. 117

SOME COMMENTS ON THE SOLUTION OF LINEAR EQUATIONS

by

R. P. Tewarson and B. Ramnath

AUGUST 7, 1968

This research was supported by the National Aeronautics and Space Administration, Washington, D. C., Grant No. NGR-33-015-013.

SOME COMMENTS ON THE SOLUTION OF LINEAR EQUATIONS*

R.P. Tewarson and B. Ramnath

Abstract. Homogeneous, ill-conditioned and singular linear equations are considered and some methods for their solution are described.

1. Introduction.

Let us consider the system of simultaneous linear equations

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

where A is an m x n matrix of rank r, x and b are column vectors with n and m elements respectively. The elements of A, x, b are all real. The techniques for the solution of (1.1) depend on many factors, for example, if A is sparse and m = n = r, then we can make use of the methods given in [1]. In case $m > n = r$, and if we want to find $\min_x \max_i |(b - Ax)_i|$, where $(b - Ax)_i$ denotes the i^{th} element of the residual vector b - Ax, then we can utilize [2]. Some of the other cases are considered in this paper. We will discuss homogeneous equations (1.1) with b = 0 in section 2, ill-conditioned equations with r = m = n in section 3 and rank deficient systems viz., $r < \min(m, n)$, in section 4.

2. Homogeneous Equations.

Taking b = 0 in (1.1), we have

$$Ax = 0 \ , \tag{2.1}$$

the solution of which is easy to find, if $\eta(A)$ - the null-space of A - is known. Let $\hat{S}$ be an orthogonal matrix of order n such that

---

*This research was supported by the National Aeronautics and Space Administration Grant No. NGR-33-015-013.

$$A\hat{S} = (B, 0) , \tag{2.2}$$

where B is m x r and $\hat{S}\hat{S}^T = \hat{S}^T\hat{S} = I_n$, the identity matrix of order n. Then it folows that

$$A = (B, 0)\hat{S}^T = B\hat{S}_r^T , \tag{2.3}$$

where $\hat{S}_r$ denotes the first r rows of $\hat{S}^T$. If the last (n - r) rows of $\hat{S}^T$ are denoted by $\hat{S}_{n-r}^T$, then

$$\hat{S}_r^T \hat{S}_{n-r} = 0 \text{ since } \hat{S}^T\hat{S} = I_n . \tag{2.4}$$

Let β denote an arbitrary column vector with (n - r) elements, then in view of (2.1), (2.3) and (2.4) we have

$$Ax = 0 \Rightarrow B\hat{S}_r^T x = 0 \Rightarrow x = \hat{S}_{n-r}\beta .$$

In fact, the columns of $\hat{S}_{n-r}$ form an orthonormal basis for $\eta(A)$. The construction of $\hat{S}$ in (2.2) as a product of matrices of the type $I - 2\omega\omega^T$ is described in [3]. Notice that in (2.2), A is post multiplied by elementary orthogonal matrices and column permutations (which are absorbed in $\hat{S}$) are usually required, as the r linearly independent columns of A are not necessarily the first r columns. If A is sparse, then the reader is referred to section 5 of [4]. In any case, the above-mentioned orthogonal triangularizations are highly stable with respect to the rounding errors [3]. For an alternative discussion of the solution of (2.2) see [5], where an operation similar to the Gauss Jordan Elimination is applied to the matrix $\begin{pmatrix} A \\ I \end{pmatrix}$. In other words, $\hat{S}$ is computed as a nonsingular but not necessarily orthogonal matrix. Unfortunately, the scheme is unstable with respect to the rounding errors.

3. Ill-Conditioned Equations.

: For any given A in (1.1), there exist orthogonal matrices Q and S such that

$$Q\ A\ S = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \tag{3.1}$$

where D is a diagonal matrix with elements $\mu_1 \geq \mu_2 \geq \ldots\ldots \geq \mu_r > 0$, which are called the singular values of A [6, 7]. If $r = m = n$, then the condition number of A is defined as

$$\text{cond}\,(A) = \frac{\mu_1}{\mu_r} \geq 1\ . \tag{3.2}$$

If A is normal, then Klinger [8] has proved the following theorem. We shall prove it for an arbitrary non-singular matrix A.

Theorem 3.1. If A is nonsingular, then for any $\epsilon > 0$, A is not better conditioned than $A + \epsilon(A^T)^{-1}$.

Proof: In view of (3.1), since $m = n = r$, we have

$$S^T A^T Q^T = D^T \Rightarrow Q(A^T)^{-1}S = D^{-1} \Rightarrow Q[A + \epsilon(A^T)^{-1}]S = D + \epsilon D^{-1}.$$

Therefore the singular values of $A + \epsilon(A^T)^{-1}$ are $\mu_i + \frac{\epsilon}{\mu_i}$, $i = 1, 2, \ldots, n$.

Now from (3.1) and (3.2), we have

$$\text{cond}\,[A + \epsilon\,(A^T)^{-1}] = \frac{\max\limits_i\ (\mu_i + \frac{\epsilon}{\mu_i})}{\min\limits_i\ (\mu_i + \frac{\epsilon}{\mu_i})} = \frac{\mu_p + \frac{\epsilon}{\mu_p}}{\mu_q + \frac{\epsilon}{\mu_q}}\ , \text{ (say)}$$

$$\leq \frac{\mu_p}{\mu_q}\ , \text{ if } \mu_p \geq \mu_q\ ,$$

$$\leq \frac{\mu_q}{\mu_p}\ , \text{ if } \mu_p < \mu_q\ .$$

But, $\frac{\mu_p}{\mu_q} \leq \frac{\mu_1}{\mu_r}$ and $\frac{\mu_q}{\mu_p} \leq \frac{\mu_1}{\mu_r}$, and therefore we conclude that

$$\text{Cond}\,[A + \epsilon(A^T)^{-1}] \leq \frac{\mu_1}{\mu_r} = \text{Cond}\,(A)\ .$$

which completes the proof of the theorem.

Thus instead of (1.1) we solve

$$[A + \epsilon(A^T)^{-1}]\, x = b, \tag{3.3}$$

ϵ being chosen reasonably small [8]. Now from (3.3) we have

$$x = [A + \epsilon(A^T)^{-1}]^{-1} b = [A^T A + \epsilon I]^{-1} A^T b\ . \tag{3.4}$$

At a first glance this appears attractive, as it avoids inverting twice. Unfortunately it has a serious disadvantage, as is indicated in the sequel. From (3.1) and the fact that $m = n = r$ we have $A = Q^T D S^T$ and clearly $AA^T + \epsilon I = Q^T(DD^T + \epsilon I)Q$. This implies that $\text{Cond}\,[AA^T + \epsilon I] = \dfrac{\mu_1^2 + \epsilon}{\mu_r^2 + \epsilon}$.

Hence it is easily seen that $\text{Cond}\,(AA^T + \epsilon I) \leq \text{Cond}\, A \Rightarrow \epsilon \geq \mu_1 \mu_r$. Evidently such a choice of ϵ is not possible without significantly changing the system (1.1). This technique can, however, be exploited as follows. If the rows of A can be partitioned into two sets of rows such that $P\,A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$, where P is an m x m permutation matrix and the rows in the submatrix A_2 constitute the ill-conditioned portion of A. Such ill-conditioned rows can be isolated, e.g. by the Householder triangularization when the norm of every one of the remaining transformed row divided by its initial norm is small (compare [9]). Now, instead of solving $Ax = b$, we solve

$$\begin{bmatrix} A_1 \\ A_2 + \epsilon(A_2 A_2^T)^{-1} A_2 \end{bmatrix} x = Pb\ . \tag{3.5}$$

In the next section we shall prove that (1.1) is not better conditioned than (3.5). Note that in solving (3.5), though two inversions are required, one of them does not require much computational effort because $A_2 A_2^T$ is usually a matrix of small order. If we were solving (1.1) by

Gaussian-Elimination (complete pivoting [3, p. 212]), then at some stage of the computation, we have

$$\begin{bmatrix} U & G \\ O & H \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} e \\ f \end{bmatrix} ,$$

where U is an upper triangular matrix, H is a square matrix and $x \equiv \begin{bmatrix} y \\ z \end{bmatrix}$. If H corresponds to the ill-conditioned rows of A, then at this stage we replace H by $H + \epsilon(H^T)^{-1}$ before continuing on. The solution of $Hz = f$ can be written as $z = (H^TH + \epsilon I)^{-1}H^Tf$ and if all the singular values of H are less than one, then the choice $\epsilon \geq \mu_p\mu_q$ (where μ_p and μ_q are the smallest and the largest singular values of H respectively) is small enough to give a reasonable solution. Faddeva [10] has suggested solving $(A + \epsilon I)\, x = b$ instead of (1.1), if A is ill-conditioned. The method is simple but no general analysis, analogous to theorem 3.1 seems possible. Replogle, Holcomb and Burrus [11] recommend adding constraints to smooth the solution and then using linear programming to find it; this requires a considerable amount of computational effort.

4. Rank Difficient or Singular Systems.

In the previous section we assumed that in (1.1) $r = m = n$. In this section we will consider the situation when $r < \min(m,n)$ and b may or may not lie in $\mathcal{R}(A)$, the range of A. It is well known [12] that in this case the solution of (1.1) is given by

$$x = A^{+}b + (I_n - A^{+}A)\,\gamma, \tag{4.1}$$

where A^{+} is the generalized inverse of A and γ is an arbitrary column vector of n elements and I_n is the identity matrix of order n. Any x given by (4.1) minimizes $\|b - Ax\|_2$ viz., the Euclidean length of the residual, and out of all such x's, $x = A^{+}b$ has the least Euclidean length viz., $\|x\|_2$ is minimum. In other words $x = A^{+}b$ is the unique minimum norm least square solution of (1.1). Various methods of computing generalized inverses are available e.g. [13], the explicit computation of A^{+} is not required when solving (1.1). In any case, if we define the condition number of the singular matrix A as in (3.2) when $r < \min(m,n)$, then we can state the following:

Theorem 4.1: If A is an mxn matrix of rank $r \leq \min(m,n)$, then for any $\epsilon > 0$, A is not better conditioned than $A + \epsilon(A^{T})^{-1}$.

Proof: From (3.1) we have

$$S^{T}A^{T}Q^{T} = \begin{bmatrix} D^{T} & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow Q(A^{T})^{+}S = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \text{ since } D^{T} = D$$

Therefore,

$$Q[\,A + \epsilon(A^{T})^{+}]S = \begin{bmatrix} D+\epsilon\, D^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

and by the same arguments as were used in the proof of Theorem 3.1, the result follows.

In view of the above Theorem, we can now prove that in (3.5) the matrix $A_2 + \epsilon(A_2 A_2^T)^{-1}A_2$ is better conditioned than A_2. Since A_2^T has full column rank (though ill conditioned), it follows that [13],

$$(A_2^T)^+ = (A_2 A_2^T)^{-1}A_2 \Rightarrow A_2 + \epsilon(A_2 A_2^T)^{-1}A_2 = A_2 + \epsilon(A_2^T)^+,$$

which is better conditioned (at worst it has the same condition number) than A_2 according to Theorem 4.1.

One of the problems in computing generalized inverses is the determination of rank r. The problem becomes difficult if there are singular values close to zero, viz., $\mu_r \approx 0$. In this case, the following corollary to Theorem 4.1 is useful.

Corollary 4.1: The minimum nonzero singular value of $[A + \epsilon(A^T)^+] > \mu_r$.

Proof: Let $\min_i (\mu_i + \frac{\epsilon}{\mu_i}) = \mu_q + \frac{\epsilon}{\mu_q}$, then $\mu_p \geq \mu_r \Rightarrow \mu_p - \mu_r \geq 0 \Rightarrow \mu_p - \mu_r + \frac{\epsilon}{\mu_p} > 0 \Rightarrow \mu_p + \frac{\epsilon}{\mu_p} > \mu_r$.

Thus we have seen that the perturbation $\epsilon(A^T)^+$ in A, even for small ϵ, moves the small singular values further away from zero.

Rosen [14] introduced the concept of basic solution of (1.1) viz., at most r elements of x in (4.1) are nonzero. We give here a method of computing a basic solution of (1.1). Analogous to (2.2), let us construct an Orthogonal matrix $\hat{Q}$ such that

$$\hat{Q}A = \begin{bmatrix} U & C \\ 0 & 0 \end{bmatrix} \qquad (4.2)$$

where U is an upper triangular rxr matrix and is nonsingular. Then, $A = \hat{Q}_r^T(U,C)$, where $\hat{Q}_r^T$ denotes the first r columns of $\hat{Q}^T$ and $A = (\hat{Q}_r^T U, \hat{Q}_r^T C)$. Then as proved by Rosen [14], the basic least square solution of (1.1) is given by

$$\hat{x} = \begin{bmatrix} (\hat{Q}_r^T U)^+ \\ 0 \end{bmatrix} b = \begin{pmatrix} U^{-1}\hat{Q}_r \\ 0 \end{pmatrix} b. \qquad (4.3)$$

But $\hat{Q}b = \begin{pmatrix} g \\ h \end{pmatrix}$ (say) $\Rightarrow \hat{Q}_r b = g$ and (4.3) gives $\hat{x} = \begin{pmatrix} U^{-1}g \\ 0 \end{pmatrix}$, in which the inversion of the upper triangular matrix U is easy. It is easy to see that the computation of a basic solution would be more accurate than that of (4.1) or even the computation of A^+b.

In passing, we show the equivalence of the following definition of A^+ due to Albert and Sittler [15] and the one implied by (3.1).

Definition: $A^+ = \lim_{\epsilon \to 0} (A^TA + \epsilon I)^{-1}A^T$.

From (3.1), we have

$$A = Q^T \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} S^T \Rightarrow A^TA + \epsilon I = S \begin{bmatrix} D^2+\epsilon I & 0 \\ 0 & \epsilon I \end{bmatrix} S^T$$

$$\Rightarrow (A^TA + \epsilon I)^{-1}A^T = S \begin{bmatrix} D^2+\epsilon I & 0 \\ 0 & \frac{1}{\epsilon} I \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} Q$$

$$= S \begin{bmatrix} (D^2+\epsilon I)^{-1}D & 0 \\ 0 & 0 \end{bmatrix} Q$$

But $\lim_{\epsilon \to 0} (D^2+\epsilon I)^{-1}D = \lim_{\epsilon \to 0} \begin{bmatrix} \frac{\mu_1}{\mu_1^2+\epsilon} & & 0 \\ & \ddots & \\ 0 & & \frac{\mu_r}{\mu_r^2+\epsilon} \end{bmatrix}$

$$= \begin{bmatrix} \frac{1}{\mu_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\mu_r} \end{bmatrix} = D^{-1}.$$

$$\text{Hence} \quad \lim_{\epsilon \to 0} (A^T A + \epsilon I)^{-1} A^T = S \begin{bmatrix} \lim_{\epsilon \to 0} (D^2 + \epsilon I)^{-1} D & 0 \\ 0 & 0 \end{bmatrix} Q$$

$$= S \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q = A^+.$$

5. Final Remarks.

We have attempted to give an idea of some of the problems one has to face in the solution of simultaneous linear equations. the "a priori" estimation of ϵ in Sections 3 and 4 is difficult and theoretical as well as computational work in this area is very much needed. Also, simple methods for computing the rank of A in the face of roundoff errors would be highly desirable.

August 7, 1968

State University of New York

Stony Brook, N. Y., U. S. A.

REFERENCES

1. R. P. Tewarson, Solution of a System of Simultaneous Linear Equations With a Sparse Coefficient Matrix by Elimination Methods, BIT 7 (1967), pp. 226-239.

2. R. P. Tewarson, On The Chebyshev Solution of Inconsistent Linear Equations, BIT 8 (1968). (to appear)

3. J. H. Wilkinson, The Algebraic Eigenvalue Problem, London, Oxford University Press (1965), p. 152, 245.

4. R. P. Tewarson, On The Orthonormalization of Sparse Vectors, Computing (1968), (to appear)

5. V. C. Kuznecov, Solution of a System of Linear Equations, Z. Vychist. Mat. Mat. Fiz. 7 (1967) pp. 157-160.

6. G. E. Forsythe and C. B. Moler, Computer Solution of Linear Algebraic Systems, Prentice Hall, Inc. Englewood Cliffs, N.J. (1967), p. 10.

7. J. B. Hawkins and A. Ben Israel, On Generalized Matrix Functions, System Research Memorandem, No. 193, The Technological Inst. Northwestern Univ. Jan. 1968.

8. A. Klinger, Approximate Pseudoinverse Solutions to Ill-Conditioned Linear Systems, J. Optimization Th. and Appl., 2(1968) pp. 117-124.

9. E. E. Osborne, Smallest Least Square Solutions of Linear Equatins, SIAM J. Num. Anal. 2(1965) pp. 300-307.

10. V. N. Faddeeva, Shift for Systems With Badly Posed Matrices, Zh. Vychist. Mat. Mat. Fiz. 5 (1965), pp. 907-911.

11. J. Reblogle, B. H. Holcomb and W. R. Burrus, The Use of Mathematical Programming for Solving Singular and Poorly Conditioned Systems of Equations, J. Math. Anal. Appl. 20 (1967), pp. 310-324.

12. R. Penrose, On Best Approximate Solutions of Linear Matrix Equations, Proc. Cambridge Phil. Soc., 52(1956), pp. 17-19.

13. R. P. Tewarson, A Direct Method for Generalized Matrix Inversion, SIAM J. Num. Anal., 4(1967), pp. 499-507.

14. J. B. Rosen, Minimum and Basic Solutions To Singular Linear Systems, SIAM J., 12(1964), pp. 156-162.

15. A. Albert and R. W. Sittler, A Method of Computing Least Square Estimators That Keep Up With The Data, SIAM J. On Control, 3(1965), pp. 1-31.