Numerische Mathematik 9, 139—148 (1966)

Note on the Iterative Refinement of Least Squares Solution

G. H. GOLUB and J. H. WILKINSON

Received June 12, 1966

1. Introduction

In the standard least squares problem the vector x is required for which $\|b-Ax\|_2$ is a minimum. Here A is an $m\times n$ matrix ($m\geqq n$) and b and x are vectors. In a recent paper GOLUB [4] described a method for solving this problem using orthogonal transformations of the type $P^{(r)}=I-2w^{(r)}(w^{(r)})^T$ where

$$(w^{(r)})^T=(0,\ldots,0,w^{(r)}_{r+1},\ldots,w^{(r)}_n),\qquad \|w^{(r)}\|_2=1. \tag{1}$$

Using pre-multiplications by $P^{(0)}, P^{(1)},\ldots,P^{(n-1)}$ (the last matrix is not required if $m=n$) the original problem is reduced to that of minimising $\|c-Bx\|_2$ where

$$B=P^{(n-1)}\ldots P^{(1)}P^{(0)}A=\begin{bmatrix}U\\ \cdots\\ 0\end{bmatrix}\begin{matrix}\}n\\ \}m-n\end{matrix} \tag{2}$$

$$c=P^{(n-1)}\ldots P^{(1)}P^{(0)}b=\begin{bmatrix}p\\ \cdots\\ q\end{bmatrix}\begin{matrix}\}n\\ \}m-n\end{matrix} \tag{3}$$

Obviously the solution is given by

$$Ux=p \tag{4}$$

since writing $P=P^{(n-1)}\ldots P^{(1)}P^{(0)}$ we have

$$\|b-Ax\|_2=\|P(b-Ax)\|_2=\left\|\begin{matrix}p-Ux\\ \cdots\cdots\\ q\end{matrix}\right\|_2. \tag{5}$$

When $m=n$ the method reduces to HOUSEHOLDER'S triangularization [5] for solving linear equations; the vector q then has no components and provided A (and therefore U) is non-singular, x is the unique solution.

In practice, whether or not $m=n$, rounding errors contaminate the solution. If $x^{(1)}$ is the computed solution and $\delta^{(1)}$ is determined so that $\|(b-Ax^{(1)})-A\delta^{(1)}\|_2$ is a minimum, then since this implies that $\|b-A(x^{(1)}+\delta^{(1)})\|_2$ is a minimum, $x^{(1)}+\delta^{(1)}$ is the least squares solution of the original problem. Hence if $r^{(1)}$ is the residual vector defined by

$$r^{(1)}=b-Ax^{(1)} \tag{6}$$

the required correction is the least squares solution of the system with matrix A and right-hand side $r^{(1)}$. The computed $P^{(r)}$ and U may be used to derive an approximation to this correction and the process may be continued to give suc-

cessive refinement of the solution. BUSINGER and GOLUB [2] described this process in detail and have published an ALGOL procedure embodying it.

Although the process is formally the same whether or not $m=n$, when $m>n$ the behaviour of the refinement procedure exhibits some features which are not present when $m=n$. These features are the main topic of this note.

2. The Basic Error Analysis

The error analysis depends to some extent on the details of the arithmetic operations used. Since it is *essential* that inner-products are accumulated to double-precision in the calculation of the residual vectors (otherwise no refinement is achieved) we shall assume that this is done throughout so that only one inner-product procedure is used.

WILKINSON [8] has given a fairly general error analysis of orthogonal transformations of this type based on the assumptions that floating-point arithmetic is used with a mantissa of t binary digits and that inner-products are accumulated wherever possible. This analysis applies directly to the above algorithm and although the bounds apply to a specific rounding procedure the differences arising from alternative procedures are of minor importance. The results may be summarised as follows:

The computed matrices $\bar{B}$ and $\bar{b}$, which are of the forms

$$\bar{B}=\begin{bmatrix}\bar{U}\\ \cdots\\ 0\end{bmatrix}, \qquad \bar{b}=\begin{bmatrix}\bar{p}\\ \cdots\\ \bar{q}\end{bmatrix} \tag{7}$$

are such that there exists an exactly orthogonal matrix Q (*not* the matrix corresponding to exact computation throughout) such that

$$\bar{B}=Q(A+E), \qquad \bar{b}=Q(b+e). \tag{8}$$

Here E and e satisfy the bounds

$$\|E\|_F \leqq 12.5\, n\, 2^{-t}\|A\|_F, \qquad \|e\|_2 \leqq 12.5\, n\, 2^{-t}\|b\|_2, \tag{9}$$

where the suffix F denotes the Frobenius norm (i.e. $\|A\|_F=(\Sigma\Sigma|a_{ij}|^2)^{\frac{1}{2}}$).

Further errors are made in the back-substitution $\bar{U}x=\bar{p}$ and it has been shown by WILKINSON [7, 8] that the computed solution $x^{(1)}$ satisfies the equation

$$(\bar{U}+F^{(1)})\,x^{(1)}=b \tag{10}$$

where

$$\|F^{(1)}\|_F \leqq (2^{-t}+\tfrac{3}{2}\, n\, 2^{-2t})\,\|U\|_F. \tag{11}$$

If we define $G^{(1)}$ by the relation $G^{(1)}=Q^T F^{(1)}$ then

$$\|G^{(1)}\|_F = \|F^{(1)}\|_F \tag{12}$$

and the computed $x^{(1)}$ is therefore the exact least squares solution of the system

$$(A+E+G^{(1)}\,|\,b+e)=(A+H^{(1)}\,|\,b+e) \quad \text{(say)} \tag{13}$$

where

$$\|H^{(1)}\|_F = \|E+G^{(1)}\|_F \leqq \|E\|_F + \|F^{(1)}\|_F. \tag{14}$$

Having performed the orthogonal reduction of A we can treat any right-hand side. The matrix E is independent of the right-hand side, but this is not true of $G^{(1)}$ which is why we have included the superscript. However, the important point for the subsequent analysis is that we have the uniform bound (11) for $\|F^{(1)}\|_F$ and therefore for $\|H^{(1)}\|_F$.

3. The Linear Equation Case

When $m=n$ the analysis of the previous section implies that the computed solution satisfies the equation

$$(A+H^{(1)})\,x^{(1)}=b+e. \tag{15}$$

Obviously if A is too ill-conditioned $A+H^{(1)}$ could even be singular. We show that if

$$27n^{\frac{3}{2}}2^{-t}\|A\|_2\,\|A^{-1}\|_2<2^{-p} \quad (p\geqq 0) \tag{16}$$

then

$$\|x-x^{(1)}\|_2/\|x\|_2<2^{-p}/(1-2^{-p-1}) \tag{17}$$

where $x=A^{-1}b$ is the true solution.

From Eq. (15)

$$x^{(1)}=(A+H^{(1)})^{-1}(b+e) \tag{18}$$

providing $A+H^{(1)}$ is non-singular, giving

$$\|x^{(1)}-x\|_2\leqq(\|A^{-1}\|_2\,\|H^{(1)}\|_2\,\|x\|_2+\|A^{-1}\|\,\|e\|)/(1-\|A^{-1}\|_2\,\|H^{(1)}\|_2). \tag{19}$$

Writing $\alpha=12.5n2^{-t}$ and remembering that $\|\overline{B}\|_F=\|\overline{U}\|_F$ we have from (14), (9), (11), (8) and (7)

$$\|H^{(1)}\|_F\leqq\alpha\|A\|_F+(2^{-t}+\tfrac{3}{2}n2^{-2t})(1+\alpha)\|A\|_F \tag{20}$$

and since $\|A\|_2\,\|A^{-1}\|_2\geqq 1$, it follows from (16) that $H^{(1)}$ certainly satisfies the bound

$$\|H^{(1)}\|_F<13.5n2^{-t}\|A\|_F. \tag{21}$$

Using the relations $\|A\|_F\leqq n^{\frac{1}{2}}\|A\|_2$ (a very weak inequality for most A), $\|H^{(1)}\|_2\leqq\|H^{(1)}\|_F$ and $\|e\|_2\leqq 12.5n2^{-t}\|b\|\leqq 12.5n2^{-t}\|A\|_2\|x\|_2$ we have finally

$$\frac{\|x^{(1)}-x\|}{\|x\|_2}\leqq\frac{13.5n^{\frac{3}{2}}2^{-t}\|A^{-1}\|_2\,\|A\|_2+12.5n2^{-t}\|A^{-1}\|_2\,\|A\|_2}{1-13.5n^{\frac{3}{2}}2^{-t}\|A^{-1}\|_2\,\|A\|_2}\leqq\frac{26n^{\frac{3}{2}}2^{-t}\varkappa(A)}{1-13.5n^{\frac{3}{2}}2^{-t}\varkappa(A)} \tag{22}$$

where $\varkappa(A)$ is the usual spectral condition number. The condition (16) now gives the required result.

Since $H^{(1)}$ is uniformly bounded for all right-hand sides this shows that provided A is "not too ill-conditioned" one can be certain that the computed solution will have some correct figures. The bound (16) is extremely conservative. Experience suggest that (17) is true when p is defined by some such relation as

$$n^{\frac{1}{2}}2^{-t}\|A\|_2\|A^{-1}\|_2=2^{-p}. \tag{23}$$

Unless A is too ill-conditioned the iterative refinement procedure is certain to work with any right-hand side *provided the errors made in computing the residual are unimportant.* Indeed if the residual were computed exactly the analysis

guarantees that the s-th computed correction $\delta^{(s)}$ satisfies the relation

$$\|\delta^{(s)} - (x - x^{(s)})\|_2 / \|x - x^{(s)}\| \leqq 2^{-p}/(1 - 2^{-p-1}). \tag{24}$$

If, further, no error were made in adding $\delta^{(s)}$ to $x^{(s)}$ we have

$$\|x^{(s+1)} - x\|_2 = \|x - x^{(s+1)}\|_2 \leqq \|x - x^{(s)}\|_2 2^{-p}/(1 - 2^{-p-1}). \tag{25}$$

On the average $x^{(s)}$ gains roughly the same number of figures in each iteration until it is correct to working accuracy. Indeed by a slight modification of the procedure it is usually possible to obtain an $x^{(s)}$ of any accuracy using the original factorization of A. In practice it is adequate to compute $r^{(s)}$, the s-th residual, using accumulation of inner-products, since if $\bar{r}^{(s)}$ is the computed residual and $r^{(s)}$ the true residual then $\bar{r}^{(s)} = r^{(s)} + f^{(s)}$ where

$$\|f^{(s)}\|_2 \leqq 2^{-t}\|r^{(s)}\|_2 + \tfrac{3}{2} n^{\frac{3}{2}} 2^{-2t} \|A\|_2 \|x^{(s)}\|_2. \tag{26}$$

(This result is an immediate consequence of the error analysis given by WILKINSON [7], pp. 82—85.)

The accumulation of inner-products is not vital at any other stage of the procedure. Failure to use it elsewhere merely leads to a strengthening of condition (16) thereby reducing the range of condition numbers for which the iterative refinement process will succeed and also slows down the rate of improvement to some extent. Provided $\varkappa(A)$ satisfies the requisite bound an $x^{(s)}$ is ultimately attained for which

$$\|x - x^{(s)}\| / \|x\| \leqq 2^{-t} \tag{27}$$

at which stage x is "correct to working accuracy".

4. The Least Squares Case with $m > n$

A moment's reflection will make it obvious that we cannot obtain a result which is exactly analogous to (27) in the least squares case, because if $A^T b = 0$ the correct solution is $x = 0$. Nevertheless, we might expect that $\|x - x^{(s)}\|_2$ will show a progressive decrease in magnitude provided only that A satisfies some bound of the form shown in (16).

However, a detailed study of the linear equation case shows that even here the behaviour is not quite as simple as one might expect. As far as the computed $x^{(s)}$ are concerned the behaviour is analogous to that of $x^{(s)}$ defined exactly by the relations

$$r^{(s)} = b - A x^{(s)}, \qquad \delta^{(s)} = (A + H^{(s)})^{-1} r^{(s)}, \qquad x^{(s+1)} = x^{(s)} + \delta^{(s)} \tag{28}$$

where the $H^{(s)}$ are uniformly bounded. But the $\|r^{(s)}\|_2$ defined by Eqs. (28) diminish at much the same rate as the $\|x - x^{(s)}\|_2$ whereas in the practical iterative refinement procedure $\|r^{(s)}\|_2$ remains roughly constant. This behaviour is predicted by the detailed error analysis (see, for example, WILKINSON [7], pp. 121—126).

5. Sensitivity of the Solution

We consider first the inherent sensitivity of the solution of the least squares problem. For this purpose it is convenient to introduce the condition number $\varkappa(A)$ of a non-square matrix A. This is defined by

$$\varkappa(A) = \sigma_1/\sigma_n, \qquad \sigma_1 = \max_{x \neq 0} \|A x\|_2 / \|x\|_2, \qquad \sigma_n = \min_{x \neq 0} \|A x\|_2 / \|x\|_2 \tag{29}$$

so that σ_1^2 and σ_n^2 are the greatest and the least eigenvalues of A^TA. From its definition it is clear that $\varkappa(A)$ is invariant with respect to unitary transformations. If U is defined as in Eq. (2) then

$$\sigma_1(U)=\sigma_1(A), \qquad \sigma_n(U)=\sigma_n(A), \qquad \varkappa(U)=\varkappa(A), \tag{30}$$

while

$$\sigma_1(U)=\|U\|_2 \quad \text{and} \quad \sigma_n(U)=1/\|U^{-1}\|_2. \tag{31}$$

The commonest method of solving least squares problems is via the normal equation

$$A^TA\,x=A^Tb. \tag{32}$$

The matrix A^TA is square and we have

$$\varkappa(A^TA)=\varkappa^2(A). \tag{33}$$

This means that if A has a condition number of the order of $2^{\frac{1}{2}t}$ then A^TA has a condition number of order 2^t and it will not be possible using t-digit arithmetic to solve the Eqs. (32). Now from Eq. (4) the method of orthogonal transformations replaces the least squares problem by the solution of the equations $Ux=p$ and $\varkappa(U)=\varkappa(A)$. It would therefore seem to have substantial advantages since we avoid working with a matrix with condition number $\varkappa^2(A)$.

We now show that this last remark is an oversimplification. To this end, we compare the solution of the original system $(A\,|\,b)$ with that of a perturbed system. It is convenient to assume that

$$\sigma_1=\|A\|_2=\|b\|_2=1; \tag{34}$$

this is not in any sense a restriction since we can make $\|A\|_2$ and $\|b\|_2$ of order unity merely by scaling by an appropriate power of two. We now have

$$\varkappa(A)=\varkappa(U)=\|U^{-1}\|_2=1/\sigma_n. \tag{35}$$

Consider the perturbed system

$$(A+\varepsilon E\,|\,b+\varepsilon e), \qquad \|E\|_2=\|e\|_2=1, \tag{36}$$

where ε is to be arbitrarily small. The solution $\bar{x}$ of the perturbed system satisfies the equation

$$(A+\varepsilon E)^T(A+\varepsilon E)\,\bar{x}=(A+\varepsilon E)^T(b+\varepsilon e). \tag{37}$$

If x is the exact solution of the original system and P is the exact orthogonal transformation corresponding to A we have

$$PA=\begin{bmatrix} U \\ \cdots \\ 0 \end{bmatrix}, \qquad P(A+\varepsilon E)=\begin{bmatrix} U+\varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix}, \qquad Pe=\begin{bmatrix} f \\ \cdots \\ g \end{bmatrix} \tag{38}$$

and

$$r=b-Ax, \qquad A^Tr=0. \tag{39}$$

Eq. (37) therefore becomes

$$(A+\varepsilon E)^T(A+\varepsilon E)=(A^T+\varepsilon E^T)(Ax+r+\varepsilon e) \tag{40}$$

giving

$$\left[\begin{array}{c} U+\varepsilon F \\ \hdashline \varepsilon G \end{array}\right]^T \left[\begin{array}{c} U+\varepsilon F \\ \hdashline \varepsilon G \end{array}\right] \bar{x} = \left[\begin{array}{c} U+\varepsilon F \\ \hdashline \varepsilon G \end{array}\right]^T \left(\left[\begin{array}{c} U \\ \hdashline 0 \end{array}\right] x + \varepsilon \left[\begin{array}{c} f \\ \hdashline g \end{array}\right] \right) + \varepsilon E^T r. \quad (41)$$

Neglecting ε^2 where advantageous

$$\begin{aligned} (U+\varepsilon F)^T (U+\varepsilon F)\,\bar{x} &= (U+\varepsilon F)^T U x + \varepsilon (U+\varepsilon F)^T f + \varepsilon E^T r + O(\varepsilon^2) \\ \bar{x} &= (U+\varepsilon F)^{-1} U x + \varepsilon (U+\varepsilon F)^{-1} f + \\ &\quad + \varepsilon (U^T U)^{-1} E^T r + O(\varepsilon^2) \\ &= x - \varepsilon U^{-1} F x + \varepsilon U^{-1} f + \varepsilon (U^T U)^{-1} E^T r + O(\varepsilon^2) \end{aligned} \quad (42)$$

giving

$$\begin{aligned} \|\bar{x} - x\|_2 &\leqq \varepsilon \|U^{-1}\|_2 \|F\|_2 \|x\|_2 + \varepsilon \|U^{-1}\|_2 \|f\|_2 + \varepsilon \|U^{-1}\|_2^2 \|E\|_2 \|r\|_2 + O(\varepsilon^2) \\ &\leqq \varepsilon \kappa(A) \|x\|_2 + \varepsilon \kappa(A) + \varepsilon \kappa^2(A) \|r\|_2 + O(\varepsilon^2). \end{aligned} \quad (43)$$

We observe that the bounds include a term $\varepsilon \kappa^2(A) \|r\|_2$. It is easy to verify by means of a 3×2 matrix A that this bound is realistic and that an error of this order of magnitude does indeed result from almost any such perturbation E of A. *We conclude that although the use of the orthogonal transformation avoids some of the ill effects inherent in the use of the normal equations the value of $\kappa^2(A)$ is still relevant to some extent.*

When the equations are compatible $\|r\| = 0$ and the term in $\kappa^2(A)$ disappears. In the non-singular linear equation case r is always null and hence it is always $\kappa(A)$ rather than $\kappa^2(A)$ which is relevant.

6. Correction Derived from the Correct Solution

An indication of the fundamental limitations of the algorithm may be obtained by considering the "correction" which corresponds to the correct solution. This correction will be the computed solution of the system $(A \mid r)$. We know that this computed solution y is the exact solution of some system $(A+H \mid r+e)$ where

$$\|H\|_F \leqq 13.5\, n\, 2^{-t} \|A\|_F, \qquad \|e\|_2 \leqq 12.5\, n\, 2^{-t} \|r\|_2, \quad (44)$$

and we have as in section 5

$$\begin{aligned} \|y\| &\leqq n\, 2^{-t} [(12.5)/\sigma_n + (13.5)\, \|A\|_F / \sigma_n^2] \|r\|_2 \\ &\leqq 26 n^{\frac{1}{2}} 2^{-t} \kappa^2(A) \|r\|_2, \end{aligned} \quad (45)$$

the last inequality following trivially because $\kappa(A) \leqq \kappa^2(A)$ and $1 < n^{\frac{1}{2}}$. Hence even when we start with the correct solution the procedure cannot recognize this. In the case when $x \neq 0$ the relative error in the corrected solution is bounded by $26 n^{\frac{1}{2}} 2^{-t} \kappa^2(A) \|r\|_2 / \|x\|_2$. We can quarantee that the correction will leave x unaffected to working accuracy only if

$$26 n^{\frac{1}{2}} \kappa^2(A) \|r\|_2 / \|x\|_2 < 1. \quad (46)$$

The factor $26 n^{\frac{1}{2}}$ need not be taken too seriously since it will, in general, be a severe overestimate, but even if it is replaced by unity we still require

$$\kappa^2(A) \|r\|_2 / \|x\|_2 < 1. \quad (47)$$

Hence whatever precision of computation is used there will be right-hand sides for which iterative refinement will never give solutions which are correct to working accuracy. This is in striking contrast with the linear equation case.

7. Modified Iterative Procedure

A further indication of the limitations of the iterative refinement procedure is provided by the following considerations. Suppose we were able to perform the iterative refinement procedure *exactly* using the *exact* matrices $P^{(0)}, \ldots, P^{(n-1)}$ defined by the algorithm for the *computed* $A^{(r)}$ obtained in the reduction of A to B. Writing $P^{(n-1)} \ldots P^{(0)} = P$ this hypothetical procedure would give

$$P r^{(s)} = \begin{bmatrix} p^{(s)} \\ \cdots \\ q^{(s)} \end{bmatrix}, \qquad \bar{U}\, \delta^{(s)} = p^{(s)} \tag{48}$$

from which

$$\begin{bmatrix} p^{(s)} \\ \cdots \\ q^{(s)} \end{bmatrix} = P r^{(s)} = P(b - A\, x^{(s)}) \tag{49}$$

giving

$$\begin{bmatrix} \bar{U} \\ \cdots \\ 0 \end{bmatrix}^T \begin{bmatrix} p^{(s)} \\ \cdots \\ q^{(s)} \end{bmatrix} = \begin{bmatrix} \bar{U} \\ \cdots \\ 0 \end{bmatrix}^T P(b - A\, x^{(s)}). \tag{50}$$

Now $\delta^{(s)} = 0$ when $p^{(s)} = 0$, that is from Eq. (50) when

$$\begin{bmatrix} \bar{U} \\ \cdots \\ 0 \end{bmatrix}^T P(b - A\, x^{(s)}) = 0$$

or

$$(A + E)^T (b - A\, x^{(s)}) = 0. \tag{51}$$

This implies that even this hypothetical procedure would ultimately give an $x^{(s)}$ such that the residual is orthogonal to the columns of $A + E$ instead of A. In the linear equation case $(A + E)^T$ is a square non-singular matrix and Eq. (51) implies $b - A\, x^{(s)} = 0$. In other words in the linear equation case the hypothetical procedure gives a zero correction if and only if $x^{(s)}$ is the exact solution.

This has led KAHAN (oral remark) to suggest an alternative procedure which has some advantages. Ignoring rounding errors we have

$$A^T A = A^T P^T P A = \begin{bmatrix} U \\ \cdots \\ 0 \end{bmatrix}^T \begin{bmatrix} U \\ \cdots \\ 0 \end{bmatrix} = U^T U \tag{52}$$

and the usual normal equations are equivalent to

$$U^T U x = A^T b. \tag{53}$$

Hence using the computed $\bar{U}$ we may carry out the iterative refinement procedure defined by

$$\begin{aligned} x^{(0)} &= 0, \qquad r^{(s)} = b - A\, x^{(s)} \\ \bar{U}^T \bar{U}\, \delta^{(s)} &= A^T r^{(s)}, \qquad x^{(s+1)} = x^{(s)} + \delta^{(s)}. \end{aligned} \tag{54}$$

At each stage we multiply $r^{(s)}$ by A^T and perform a forward and a backward substitution to obtain $\delta^{(s)}$. This certainly has the advantage that when b is exactly

orthogonal to the columns of A then $A^T r^{(0)} = A^T b = 0$, so that the computed correction is the null vector and the process "converges" to the correct answer. (This is not true of the procedure of BUSINGER and GOLUB.)

Unless we are prepared to go further than the mere accumulation of inner-products the limiting accuracy attainable with this algorithm is generally much the same as with that of BUSINGER and GOLUB. For if $x^{(s)}$ is, in fact, the correctly rounded solution then even if we accumulate inner-products when computing $r^{(s)}$, each of its components will have to be rounded, in general, before pre-multiplying with A^T. Instead of obtaining the true residual r, the best we can hope to obtain is $r+h$ where $\|h\| \leqq 2^{-t}\|r\|_2$. (Even the correctly rounded r will have an error of this nature.) Hence Eq. (54) gives

$$\bar{U}^T \bar{U} \delta^{(s)} = A^T h \tag{55}$$

since $A^T r$ is null, and if $\|A\| = 1$ we have effectively

$$\|\delta^{(s)}\|_2 \leqq 2^{-t} \varkappa^2(A) \|r\|_2. \tag{56}$$

However, if we multiply the accumulated $r^{(s)}$ by A^T without previous rounding to form $A^T r^{(s)}$ (this involves the multiplication of a *double-precision* vector by a *single-precision* matrix, accumulating each element in double-precision and rounding only on completion) then, if $x^{(s)}$ is the correct solution, $\delta^{(s)}$ will be given by

$$\bar{U}^T \bar{U} \delta^{(s)} = A^T h \quad \text{where} \quad \|h\| \leqq 2^{-2t}\|r\|_2. \tag{57}$$

With this modified arithmetic procedure we can expect $\delta^{(s)}$ to satisfy an inequality of the form

$$\|\delta^{(s)}\|_2 \leqq 2^{-2t} \varkappa^2(A) \|r\|_2. \tag{58}$$

It is therefore much more satisfactory but a substantial amount of extra work is involved.

8. Numerical Example

The predictions of the error analysis are well illustrated by the numerical examples given by BUSINGER and GOLUB [2]. The matrix A consists of the first five columns of the inverse of the 6×6 Hilbert matrix. A matrix of 39 binary digits was used and all inner-products were accumulated.

The right-hand side b_1 was chosen so that $Ax = b_1$ is compatible. The third iteration gave the correctly rounded solution to working accuracy as is to be expected.

The second right-hand side b_2 is such that $A^T b_2 = 0$ so that the correct solution is the null vector and $r = b_2$. The first computed solution had a component which was of the order of magnitude 10^{-3} and subsequent iterations produced random fluctuations in the sixth significant figure of the first iterate, the norm remaining roughly constant.

The successive residuals also remain roughly constant and differ only slightly from the original right-hand side which is, in any case, the exact residual corresponding to the correct solution. At first sight it may seem surprising that the original right-hand side gives a computed solution of the order of magnitude 10^{-3}, yet the subsequent corrections are of order 10^{-8} although they too are derived from a right-hand side which is much the same as the original. The application

of two stages of our error analysis shows that this is precisely what is to be expected. The first solution contains a term $(A^TA)^{-1}(H^{(1)})^Tr$. This produces a contribution to the first residual which, though small, is just adequate to cancel out the term of order $(A^TA)^{-1}(H^{(2)})^Tr$ which would otherwise arise in the first correction.

Example

A (exact)

$3.60000_{10}+1$	$-6.30000_{10}+2$	$3.36000_{10}+3$	$-7.56000_{10}+3$	$7.56000_{10}+3$
$-6.30000_{10}+2$	$1.47000_{10}+4$	$-8.82000_{10}+4$	$2.11680_{10}+5$	$-2.20500_{10}+5$
$3.36000_{10}+3$	$-8.82000_{10}+4$	$5.64480_{10}+5$	$-1.41120_{10}+6$	$1.51200_{10}+6$
$-7.56000_{10}+3$	$2.11680_{10}+5$	$-1.41120_{10}+6$	$3.62880_{10}+6$	$-3.96900_{10}+6$
$7.56000_{10}+3$	$-2.20500_{10}+5$	$1.51200_{10}+6$	$-3.96900_{10}+6$	$4.41000_{10}+6$
$-2.77200_{10}+3$	$8.31600_{10}+5$	$-5.8212_{10}+5$	$1.55232_{10}+6$	$-1.74636_{10}+6$

b_1 *(exact)*	$x^{(3)}$ *(Third iteration)*
$4.63000_{10}+2$	$1.00000\ 00000_{10}+0$
$-1.38600_{10}+4$	$5.00000\ 00000_{10}-1$
$9.70200_{10}+4$	$3.33333\ 33333_{10}-1$
$-2.58720_{10}+5$	$2.50000\ 00000_{10}-1$
$2.91060_{10}+5$	$2.00000\ 00000_{10}-1$
$-1.16424_{10}+5$	

First iteration

b_2 *(exact)*	$x^{(1)}$	$r^{(1)}$	$\delta^{(1)}$
$4.62000_{10}+3$	$-1.34714\ 39142_{10}-3$	$4.62000\ 19967_{10}+3$	$1.11227\ 27028_{10}-8$
$3.96000_{10}+3$	$-4.48614\ 75412_{10}-4$	$3.96000\ 02989_{10}+3$	$3.43958\ 93111_{10}-9$
$3.46500_{10}+3$	$-1.91979\ 16562_{10}-4$	$3.46499\ 95056_{10}+3$	$1.37004\ 51340_{10}-9$
$3.08000_{10}+3$	$-8.39004\ 73880_{10}-5$	$3.07999\ 91136_{10}+3$	$5.62977\ 18071_{10}-10$
$2.77200_{10}+3$	$-2.98101\ 00501_{10}-5$	$2.77199\ 89175_{10}+3$	$1.89861\ 20958_{10}-10$
$2.52000_{10}+3$		$2.51999\ 88239_{10}+3$	

Second iteration

$x^{(2)}$	$r^{(2)}$	$\delta^{(2)}$
$-1.34713\ 27915_{10}-3$	$4.62000\ 19966_{10}+3$	$-1.11227\ 27027_{10}-8$
$-4.48611\ 31453_{10}-4$	$3.96000\ 02988_{10}+3$	$-5.05313\ 76285_{10}-9$
$-1.91977\ 29557_{10}-4$	$3.46499\ 95056_{10}+3$	$-2.42938\ 00078_{10}-9$
$-8.38999\ 10403_{10}-5$	$3.07999\ 91136_{10}+3$	$-1.12757\ 48695_{10}-9$
$-2.98099\ 10639_{10}-5$	$2.77199\ 89175_{10}+3$	$-4.15470\ 31492_{10}-10$
	$2.51999\ 88238_{10}+3$	

9. General Comments

For the accurate solution of linear equations iterative refinement procedures are extremely attractive whether used in connexion with the Householder triangularization or the Crout triangularization with pivoting [*1, 3, 6*]. For large systems a single-precision factorization plus iterative refinement has the advantage

over double-precision factorization without refinement that it requires less storage (only about half), is faster if only a few right-hand sides are involved and gives useful indication about the condition of A. It has the disadvantage that if A is too ill-conditioned iterative refinement will not work whereas double-precision computation might well give an answer of acceptable (but unknown) accuracy.

Turning to the least squares problem the choice is effectively between forming the normal equations $A^TAx=A^Tb$ and using an LL^T decomposition of A^TA, all in double-precision or using the orthogonal reduction in single-precision with iterative refinement.

If, as is true in some fields of operation $m \gg n$, the matrix A^TA is small compared with A. The storage requirement then ceases to be a valid argument against the use of double-precision. The matrix A^TA has the condition number $\varkappa^2(A)$ but we have shown that the performance of the iterative refinement process is dependent to some extent on $\varkappa^2(A)$ (unless the system is compatible), in spite of the fact that PA and U have the same condition number $\varkappa(A)$ as A itself. In some fields of operation overdetermined systems are obtained which are almost compatible so that $\|r\|_2$ is far smaller than x. In such cases, particularly if m is not excessively large compared with n, the iterative refinement algorithm is very attractive.

In this paper we have discussed only the case when the matrix A has independent columns. Clearly the least squares problem merits further analysis, both theoretical and experimental.

Acknowledgements. The work of J. H. WILKINSON included here is part of the Research Programme of the National Physical Laboratory and is published by permission of the Director of the Laboratory.

The work of G. H. GOLUB was supported in part by the National Science Foundation.

References

[1] BOWDLER, HILARY, J., R. S. MARTIN, G. PETERS, and J. H. WILKINSON: Solution of real and complex systems of linear equations. Numerische Mathematik **8**, 217–234 (1966).

[2] BUSINGER, P., and G. H. GOLUB: Linear least squares solutions by Householder transformations. Numerische Mathematik **7**, 269–276 (1965).

[3] FORSYTHE, G. E.: Crout with pivoting. Commun. Ass. Comp. Mach. **3**, 507–508 (1960).

[4] GOLUB, G. H.: Numerical methods for solving linear least squares problems. Numerische Mathematik **7**, 206–216 (1965).

[5] HOUSEHOLDER, A. S.: Unitary triangularization of a non-symmetric matrix. J. Assoc. Comput. Mach. **5**, 339–342 (1958).

[6] MCKEEMAN, W. M.: Crout with equilibration and iteration. Commun. Ass. Comp. Mach. **5**, 552–555 (1962).

[7] WILKINSON, J. H.: Rounding errors in algebraic processes. Her Majesty's Stationery Office, London. New Jersey: Prentice-Hall 1963.

[8] — The algebraic eigenvalue problem. London: Oxford University Press 1965.

Computation Center
Stanford University
Stanford, Calif. 94305 (USA)

National Physical Lab.
Teddington, Middlesex
Great Britain