Error Analysis of Algorithms for Computing the Projection of a Point onto a Linear Manifold

M. Arioli and A. Laratta
Istituto di Elaborazione dell'Informazione
Via S. Maria 46
Pisa, Italy

Submitted by Hans Schneider

ABSTRACT

The accumulation of rounding errors in methods used to compute the projection of a point onto a linear manifold is studied. The methods are based on modified Gram-Schmidt, Householder, and Givens orthogonal factorizations.

1. INTRODUCTION

Problems in applied mathematics often require the computation of the solution of the linear system

$$A^T x = b \tag{1.1}$$

at the least distance from an assigned point p, i.e., the solution to the problem

$$\min_{A^T x = b} \tfrac{1}{2}\|x - p\|^2, \tag{1.2}$$

where the $n \times m$ real matrix A and the real m-vector b are known, and the n-vector x is unknown. In this paper, the Euclidean norm for the vectors and the corresponding induced norm for the matrices are denoted by $\|\cdot\|$, and the Frobenius norm by $\|\cdot\|_F$.

The system (1.1) is assumed consistent even if the rank of A is not full. In any case, a linear dependence test and a consistency test of the system (1.1) are both useful in solving (1.2).

For the case $p = 0$ Huang gives in [1] a method, based on the modified Gram-Schmidt decomposition of the matrix A, which includes these two tests.

52 Vanderbilt Ave., New York, NY 10017 0024-3795/86/$3.50

In [2] we have generalized this method by taking into account other orthogonal decompositions of the matrix A; in particular we have experimentally evaluated the numerical stability of the methods described. Furthermore, in [2] these methods are experimentally compared with a very good method for which in [3] a backward error analysis is given and the tests of consistency and linear dependence are studied.

Here we give a roundoff error analysis of the methods examined in [2]. In Sections 2 and 3 we will give some properties of the solution of (1.2) and will describe the methods used to compute the solution. The error analysis will be made in Sections 4 and 5. Finally, in Section 6, the tests of consistency and linear dependence will be discussed.

2. DATA PERTURBATION

In Sections 2, 3, 4, and 5, unless otherwise explicitly stated, we assume that $m \leqslant n$ and A is of full rank. The solution x^* of (1.2) can be written as

$$x^* = \left[I - A(A^TA)^{-1}A^T\right]p + A(A^TA)^{-1}b. \tag{2.1}$$

If

$$P(A) = I - A(A^TA)^{-1}A^T$$

denotes the orthogonal projection matrix onto the manifold defined by the system $A^Tx = 0$ and

$$A^+ = (A^TA)^{-1}A^T$$

denotes the Moore-Penrose pseudoinverse of A, from (2.1) we have

$$x^* = P(A)p + A^{+T}b.$$

Hence x^* is the sum of the minimum length solution of (1.1) and the projection of p onto the manifold $\{x: A^Tx = 0\}$. Moreover, the R.H.S. of (2.1) with an arbitrary u replacing p gives the general solution of (1.1).

Let us now describe some results of the perturbation theory. The matrix $\tilde{A}$ and vectors $\tilde{b}$ and $\tilde{p}$ are defined as

$$\begin{aligned} \tilde{A} &= A + G, \\ \tilde{b} &= b + \Delta b, \\ \tilde{p} &= p + \Delta p, \end{aligned}$$

and constants α, β, γ, δ, and $\hat{K}(A)$ are defined as follows:

$$\alpha = \frac{\|G\|}{\|A\|}, \qquad \beta = \frac{\|\Delta b\|}{\|b\|}, \qquad \gamma = \frac{\|b\|}{\|A\|\,\|x^*\|},$$

$$\delta = \frac{\|\Delta p\|}{\|p\|}, \qquad \hat{K}(A) = \frac{K(A)}{1 - \alpha K(A)}.$$

where $K(A)$ is the condition number of A: $K(A) = \|A\|\,\|A^+\|$. If $\alpha K(A) < 1$, then $\operatorname{rank}(\tilde{A}) = \operatorname{rank}(A)$, and in [4] it is shown that the solution $\tilde{x}$ of the problem

$$\min_{\tilde{A}^T x = \tilde{b}} \tfrac{1}{2}\|x - \tilde{p}\|^2 \tag{2.2}$$

is related to x^* by the inequality

$$\frac{\|x^* - \tilde{x}\|}{\|x^*\|} \leqslant \hat{K}(A)\left(\alpha + \gamma\beta + \alpha\frac{\|p - x^*\|}{\|x^*\|}\right) + \left(\frac{\|p - x^*\|}{\|x^*\|} + 1\right)\delta. \tag{2.3}$$

The bound (2.3) has already been used in [3] [formula (31)] together with a similar bound [formula (5)], which also has been derived from the results of [4]. These two bounds (and the corresponding roundoff error of the algorithm studied in [3]) are quantitatively slightly different. In this paper we prefer (2.3) because it makes some minor simplifications in the computations possible.

If $m = n$, (2.3) becomes the well-known formula

$$\frac{\|x^* - \tilde{x}\|}{\|x^*\|} \leqslant \hat{K}(A)(\alpha + \gamma\beta). \tag{2.4}$$

If $m = n$, $\operatorname{rank}(A) = \mu$ ($\mu < m$), we can extract from (1.1) a system

$$A'^T x = b'$$

of μ linearly independent equations. The solution x^* and a perturbation formula can be obtained from (2.1) and (2.3) replacing A with A' and b with b'.

3. THE ALGORITHMS

The algorithms described here for the solution of (1.2) are based on the factorization

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

where R is an $m \times m$ upper triangular matrix of elements r_{ij} and Q is an $n \times m$ matrix which has pairwise orthogonal columns q_i.

The Gram-Schmidt and modified Gram-Schmidt methods give (3.1) directly, where $r_{ii} = 1$ and q_i, $i = 1, \cdots, m$, are not normalized. However the Householder and Givens methods compute an $n \times n$ orthonormal matrix H, an $m \times m$ upper triangular matrix R, and an $(n - m) \times m$ null matrix 0 such that

$$HA = \begin{bmatrix} R \\ 0 \end{bmatrix}. \tag{3.2}$$

The factorization (3.1) is obtained from (3.2), partitioning H as follows:

$$H = \begin{bmatrix} Q^T \\ S^T \end{bmatrix}.$$

The columns of Q are the first m rows of the matrix H and now have unit length. Furthermore, the elements r_{ii} are generally not equal to one.

From (3.1) the system (1.1) can be written as

$$Q^T x = d, \qquad d = (R^T)^{-1} b;$$

the problem (1.2) as

$$\min_{Q^T x = d} \tfrac{1}{2} \|x - p\|^2;$$

and the solution x^* as

$$x^* = \left(I - Q(Q^TQ)^{-1}Q^T\right)p + Q(Q^TQ)^{-1}d$$

or

$$x^* = p - \sum_{j=1}^{m} \frac{q_j^T p}{q_j^T q_j} q_j + \sum_{j=1}^{m} \frac{d_j}{q_j^T q_j} q_j. \tag{3.3}$$

Let us now describe an algorithm which successively computes the solutions $x^{(i)}$, $i = 1, \cdots, m$, ($x^{(m)} = x^*$) of the problems

$$\min_{A_i^T x = b^{(i)}} \tfrac{1}{2}\|x - p\|^2,$$

where

$$A_i^T x = b^{(i)}$$

is the system of the first i equations of (1.1).

Because of (3.3), the solution $x^{(i)}$ can be written as

$$x^{(i)} = p - \sum_{j=1}^{i} \frac{q_j^T p}{q_j^T q_j} q_j + \sum_{j=1}^{i} \frac{d_j}{q_j^T q_j} q_j,$$

and we thus easily obtain the following algorithm:

$$\left.\begin{aligned} x^{(0)} &= p, \\ r^i(p) &= q_i^T p - d_i \\ x^{(i)} &= x^{(i-1)} - \frac{r^i(p)}{q_i^T q_i} q_i \end{aligned}\right\} \quad i = 1, \cdots, m. \tag{3.4}$$

Since $q_i^T x^{(i)} = d_i$, it follows from (3.4) that

$$q_i^T p = q_i^T x^{(i-1)}. \tag{3.5}$$

Then $x^{(i)}$ can be computed by the algorithm

$$\left.\begin{aligned} x^{(0)} &= p, \\ r^i(x^{(i-1)}) &= q_i^T x^{(i-1)} - d_i \\ x^{(i)} &= x^{(i-1)} - \frac{r^i(x^{(i-1)})}{q_i^T q_i} q_i \end{aligned}\right\} \quad i = 1, \cdots, m. \tag{3.6}$$

If in (3.6) q_i are computed by the modified Gram-Schmidt method, an algorithm is obtained which was studied in [1] for the computation of the minimum length solution. In [2] a numerical evaluation of (3.6) is given when q_i are computed by means of Householder or modified Gram-Schmidt methods.

In the next section we will study the effects of rounding errors in the algorithms (3.4) and (3.6). We will now attempt to show that it is possible to compute q_i with (3.4) and (3.6).

Denoting the columns of A by a_j, $j = 1, \cdots, m$, from (3.1), it easily follows that

$$r_{ii} q_i = P_{i-1} a_i, \qquad i = 1, \cdots, m, \tag{3.7}$$

where

$$P_0 = I,$$

and

$$P_{i-1} = I - \sum_{k=1}^{i-1} \frac{q_k q_k^T}{q_k^T q_k}$$

is the projection matrix onto the manifold $\{x: A_{i-1}^T x = 0\}$. If we assume $r_{ii} = 1$, $i = 1, \cdots, m$, we have that $q_1 = a_1$ and q_j, $j = 2, \cdots, m$, are the solutions of the problems

$$\min_{A_{j-1}^T x = 0} \tfrac{1}{2}\|x - a_j\|^2, \qquad j = 2, \cdots, m. \tag{3.8}$$

If $y_j^{(i)}$, $j = 2, \cdots, m$, denote the solutions of the problems

$$\min_{A_i^T x = 0} \tfrac{1}{2}\|x - a_j\|^2 \qquad i = 1, \cdots, j-1, \quad j = 2, \cdots, m,$$

then using (3.4) we obtain

$$\left.\begin{aligned} q_1 &= a_1, \\ y_j^{(0)} &= a_j, \\ y_j^{(i)} &= y_j^{(i-1)} - \frac{q_i^T a_j}{q_i^T q_i} q_i, \quad i = 1, \cdots, j-1 \\ q_j &= y_j^{(j-1)}, \end{aligned}\right\} \qquad j = 2, \cdots, m, \tag{3.9}$$

and using (3.6) we obtain

$$\left.\begin{aligned} q_1 &= a_1 \\ y_j^{(0)} &= a_j \\ y_j^{(i)} &= y_j^{(i-1)} - \frac{q_i^T y_j^{(i-1)}}{q_i^T q_i} q_i, \quad i = 1, \cdots, j-1 \\ q_j &= y_j^{(j-1)} \end{aligned}\right\} \quad j = 2, \cdots, m. \tag{3.10}$$

The method (3.9) is the Gram-Schmidt orthogonalization method, and it is easy to prove that the computation of q_i by (3.10) is numerically equivalent to the modified Gram-Schmidt method.

4. BASIC RESULTS FOR THE ERROR ANALYSIS

Let us produce some perturbation theory results which will be used in the next section, where the computational errors of algorithms (3.4) and (3.6) will be studied.

Let $\tilde{Q}$ and $\tilde{R}$ be approximations to the matrices Q and R. Let $\tilde{Q}$ be of full rank, and let $\tilde{R}$ be an upper triangular nonsingular matrix. Considering the problem

$$\min_{\tilde{Q}^T x = \tilde{d}} \tfrac{1}{2}\|x - p\|^2, \qquad \tilde{d} = (\tilde{R}^T)^{-1} b, \tag{4.1}$$

let D be the $m \times m$ diagonal matrix of diagonal elements d_{ii} defined by

$$d_{ii} = \begin{cases} 1 & \text{if } Q^T Q = I, \\ \|\tilde{q}_i\| & \text{if } Q^T Q \neq I. \end{cases}$$

Because $\|\tilde{q}_i\| \neq 0$ we can define

$$Q' = \tilde{Q} D^{-1}, \qquad R' = D\tilde{R}, \qquad d' = D^{-1}\tilde{d}.$$

The problem (4.1) can be written as follows:

$$\min_{Q'^T x = d'} \tfrac{1}{2}\|x - p\|^2, \qquad d' = (R'^T)^{-1} b, \tag{4.2}$$

and its solution $\hat{x}$ as

$$\hat{x} = p - Q'\left(Q'^T Q'\right)^{-1}\left(Q'^T p - d'\right). \tag{4.3}$$

Let us denote by $\hat{x}^{(i)}$, $i = 1, \cdots, m$, ($\hat{x} = \hat{x}^{(m)}$) the vectors defined by

$$\hat{x}^{(i)} = p - Q_i'\left(Q_i'^T Q_i'\right)^{-1}\left[Q_i'^T p - d'^{(i)}\right], \tag{4.4}$$

where Q_i' is the matrix of the first i columns of Q' and $d'^{(i)}$ is the vector of the first i elements of d'.

The algorithms (3.4) and (3.6), where the vectors q_j are replaced by the columns of q_j' of Q', and the elements r_{ij} of R by the elements r_{ij}' of R', compute the vectors $x'^{(i)}$ defined by

$$\left.\begin{aligned} x'^{(i)} &= p - Q_i'\left(Q_i'^T p - d'^{(i)}\right), \\ x' &= x'^{(m)}. \end{aligned}\right\} \tag{4.5}$$

Denoting

$$G = \tilde{Q}\tilde{R} - A = Q'R' - A \tag{4.6}$$

and

$$E = Q'^T Q' - I, \tag{4.7}$$

we will compute upper bounds, depending on $\|G\|$ and $\|E\|$, for the errors $\|x^* - \hat{x}\|$ and $\|\hat{x} - x'\|$, and therefore also for

$$\|x^* - x'\| \leqslant \|x^* - \hat{x}\| + \|\hat{x} - x'\|.$$

THEOREM 4.1. *If*

$$\frac{\|G\|}{\|A\|} K(A) < 1,$$

then if $m < n$ *it follows that*

$$\frac{\|x^* - \hat{x}\|}{\|x^*\|} \leqslant \hat{K}(A) \frac{\|G\|}{\|A\|} \left\{ 1 + \frac{\|x^* - p\|}{\|x^*\|} \right\}, \tag{4.8}$$

and if $m = n$ it follows that

$$\frac{\|x^* - \hat{x}\|}{\|x^*\|} \leqslant \hat{K}(A)\frac{\|G\|}{\|A\|}. \tag{4.9}$$

Proof. The system $\tilde{Q}^T x = \tilde{d}$ can be written as $(A + G)^T x = b$. Formulas (4.8) and (4.9) then follow from (2.3) and (2.4). ∎

THEOREM 4.2. *If $\|E\| < 1$ then*

$$\|\hat{x} - x'\| \leqslant \|Q'\|^2 \frac{\|E\|}{1 - \|E\|}\|\hat{x} - p\|. \tag{4.10}$$

Proof. From (4.3) and (4.5), because $d' = Q'^T\hat{x}$, we obtain

$$\hat{x} - x' = Q'(Q'^TQ')^{-1}(I - Q'^TQ')Q'^T(\hat{x} - p).$$

Using

$$\|(Q'^TQ')^{-1}\| = \|(I + E)^{-1}\| \leqslant \frac{1}{1 - \|E\|},$$

(4.10) easily follows. ∎

Using (4.8), (4.9), and (4.10), we have:

COROLLARY 4.1. *Under the hypotheses of the Theorems* 4.1 *and* 4.2 *we have, if $m < n$,*

$$\frac{\|\hat{x} - x'\|}{\|x^*\|} \leqslant \|Q'\|^2\hat{K}(A)\frac{\|G\|}{\|A\|}\frac{\|E\|}{1 - \|E\|}\left\{1 + \frac{\|x^* - p\|}{\|x^*\|}\right\} + \|Q'\|^2\frac{\|E\|}{1 - \|E\|}\frac{\|x^* - p\|}{\|x^*\|}, \tag{4.11}$$

and if $m = n$,

$$\frac{\|\hat{x} - x'\|}{\|x^*\|} \leqslant \|Q'\|^2\hat{K}(A)\frac{\|G\|}{\|A\|}\frac{\|E\|}{1 - \|E\|} + \|Q'\|^2\frac{\|E\|}{1 - \|E\|}\frac{\|x^* - p\|}{\|x^*\|}. \tag{4.12}$$

Finally, using Theorems 4.1, 4.2 and Corollary 4.1, we have the following theorem.

THEOREM 4.3. *Under the hypotheses of Theorems* 4.1 *and* 4.2 *it follows that if* $m < n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \hat{K}(A)\frac{\|G\|}{\|A\|}\left\{1 + \|Q'\|^2\frac{\|E\|}{1-\|E\|}\right\}\left\{1 + \frac{\|x^* - p\|}{\|x^*\|}\right\}$$

$$+ \|Q'\|^2\frac{\|E\|}{1-\|E\|}\frac{\|x^* - p\|}{\|x^*\|}, \tag{4.13}$$

and if $m = n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \hat{K}(A)\frac{\|G\|}{\|A\|}\left\{1 + \|Q'\|^2\frac{\|E\|}{1-\|E\|}\right\}$$

$$+ \|Q'\|^2\frac{\|E\|}{1-\|E\|}\frac{\|x^* - p\|}{\|x^*\|}. \tag{4.14}$$

5. ROUNDOFF ERROR ANALYSIS

Let us now study the computational errors when the problem (1.2) is solved by the algorithm (3.4) or (3.6) on a computer with mixed precision arithmetic of relative precisions eps and eps^2 (the scalar products are computed by arithmetic with relative precision eps^2).

Because of rounding errors in the computation of the decomposition (3.1), the matrices $\bar{Q}$ and $\bar{R}$ (upper triangular) are computed instead of Q and R. Let $\bar{G}$ be the matrix such that

$$A + \bar{G} = \bar{Q}\,\bar{R}. \tag{5.1}$$

If the factorization (3.1) is executed by the modified Gram-Schmidt, Householder, or Givens method, then the matrix $\bar{G}$ will satisfy

$$\|\bar{G}\|_F \leqslant c(n, m)\|A\|_F\,\text{eps} + O(\text{eps}^2), \tag{5.2}$$

where

$$c(n,m)=\begin{cases}\dfrac{3}{2}(m-1), & \text{modified Gram-Schmidt method [5],}\\ 29\ m, & \text{Householder method [6,7],}\\ t(n+m-2), & \\ \quad (t=6,7.5,11,85) & \text{Givens method [6,8].}\end{cases}\tag{5.3}$$

The computed solution $\bar{d}$ of the triangular system $\bar{R}^T y=b$ (using the substitution method in mixed precision arithmetic) satisfies

$$(\bar{R}+U)^T\bar{d}=b,\tag{5.4}$$

where U is an upper triangular matrix such that

$$|U|\leqslant|\bar{R}|\text{eps}+O(\text{eps}^2)\tag{5.5}$$

($|U|$ and $|\bar{R}|$ denote the matrices of the absolute values of the elements of U and $\bar{R}$). Moreover, if eps is sufficiently small, the nonsingularity of $\bar{R}$ will involve the nonsingularity of $\bar{R}+U$. In the following, we assume that $\bar{R}+U$ is nonsingular.

If

$$\tilde{Q}=\bar{Q}$$

and

$$\tilde{R}=\bar{R}+U,$$

from (5.1) we have

$$A+G=\tilde{Q}\tilde{R},$$

$$G=\bar{G}+\tilde{Q}U.$$

If D, Q', R' are defined as in the previous section and, moreover, if we use U' to denote the matrix

$$U'=DU,$$

if follows that

$$A+G=Q'R'\tag{5.6}$$

and

$$G = \bar{G} + Q'U', \tag{5.7}$$

and from (5.5) we have

$$\|U'\|_F \leqslant \|D\bar{R}\|_F \,\mathrm{eps} + O(\mathrm{eps}^2). \tag{5.8}$$

Furthermore we denote

$$E = Q'^T Q' - I.$$

In conclusion, if the computation of the decomposition (3.1) and the solution of the triangular system $\bar{R}^T y = b$, using the substitution method, is performed by mixed precision arithmetic of relative precisions eps and eps^2, then the methods (3.4) and (3.6), applied in exact arithmetic, will give a vector x' instead of x^*. The vector x' is expressed by (4.5) with the Q' and R' just defined.

Let us now compute some upper bounds for $\|Q'\|$, $\|G\|/\|A\|$, and $\|E\|$ for the modified Gram-Schmidt, Householder, and Givens methods. From (4.13) and (4.14) it is possible to obtain the corresponding upper bounds for the error $\|x^* - x'\|/\|x^*\|$. The errors due to the floating point computation of the arithmetic expressions in (3.4) and (3.6) will be examined successively.

We will use the results in [5] for the modified Gram-Schmidt method, those in [7] for the Householder method, the results in [8] for the Givens method.

Furthermore it is useful to remember that if B is an $n \times m$ matrix, it follows that

$$\|B\| \leqslant \|B\|_F \leqslant m^{1/2}\|B\|.$$

(a) Modified Gram-Schmidt Method

We have

$$\|Q'\| \leqslant \|Q'\|_F = m^{1/2}.$$

Under the hypothesis

$$\rho = 3.42m(m+1)K(A)\,\mathrm{eps} < 1,$$

in [5] it is proved that

$$\|D\bar{R}\| \leqslant (1+\rho)^{1/2}\|A\| \leqslant 2^{1/2}\|A\|, \tag{5.9}$$

$$\|E\| \leqslant \frac{1.74}{(1-\rho)^{1/2}} m(m+1)K(A)\,\text{eps}. \tag{5.10}$$

From (5.7) and taking into account (5.2), (5.3), (5.8), and (5.9), it follows that

$$\frac{\|G\|}{\|A\|} \leqslant \left[2^{1/2}m + \tfrac{3}{2}m^{1/2}(m-1)\right]\text{eps} + O(\text{eps}^2). \tag{5.11}$$

From (4.13) and (4.14) some upper bounds on $\|x^* - x'\|/\|x^*\|$ can be found. Simple formulas are obtained if we assume

$$\rho \leqslant \tfrac{1}{2}.$$

In this case

$$\|E\| < \rho \leqslant \tfrac{1}{2},$$

$$\hat{K}(A)\frac{\|G\|}{\|A\|} < \frac{\rho}{1-\rho} \leqslant 1,$$

and from (4.13) and (4.14) it follows that if $m < n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \frac{\rho}{1-\rho}\left\{(2m+1)\frac{\|x^* - p\|}{\|x^*\|} + (m+1)\right\}, \tag{5.12}$$

and if $m = n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \frac{\rho}{1-\rho}\left\{m\frac{\|x^* - p\|}{\|x^*\|} + (m+1)\right\}. \tag{5.13}$$

(b) Householder and Givens Methods

In [6] it is proved that an $n \times m$ matrix Q^* with pairwise orthonormal columns exists such that

$$\bar{Q} = Q^* + F,$$

where F is an $n \times m$ matrix which satisfies the relation

$$\|F\|_F \leqslant c(n,m)\,\mathrm{eps} + O(\mathrm{eps}^2).$$

The quantity $c(n,m)$ is defined by (5.3). Moreover $\|\bar{R}\|_F \leqslant \|A\|_F + O(\mathrm{eps})$. Because $Q' = \bar{Q}$ and $U' = U$, it follows that

$$\|Q'\| = 1 + O(\mathrm{eps}),$$
$$\frac{\|G\|}{\|A\|} \leqslant \alpha' \mathrm{eps} + O(\mathrm{eps}^2),$$
$$\|E\| \leqslant \alpha'' \mathrm{eps} + O(\mathrm{eps}^2),$$

with

$$\alpha' = \begin{cases} (29m+1)m^{1/2}, & \text{Householder method,} \\ [t(n+m-2)+1]\,m^{1/2}, & \text{Givens method,} \end{cases}$$

and

$$\alpha'' = \begin{cases} 58m, & \text{Householder method,} \\ 2t(n+m-2), & \text{Givens method.} \end{cases}$$

From (4.13) and (4.14) is possible to find some upper bounds on $\|x^* - x'\|/\|x^*\|$. Simple formulas are obtained if we assume

$$\alpha' K(A)\,\mathrm{eps} \leqslant \tfrac{1}{2}.$$

In this case

$$\hat{K}(A)\frac{\|G\|}{\|A\|} \leqslant \frac{\alpha' K(A)\,\mathrm{eps}}{1 - \alpha' K(A)\,\mathrm{eps}} + O(\mathrm{eps}^2) \leqslant 1 + O(\mathrm{eps}^2),$$
$$\|E\| \leqslant \alpha'' \mathrm{eps} + O(\mathrm{eps}^2) \leqslant 1 + O(\mathrm{eps}^2),$$

and from (4.13) and (4.14) it follows that if $m < n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \frac{\alpha' K(A)\,\mathrm{eps}}{1 - \alpha' K(A)\,\mathrm{eps}}\left(1 + \frac{\alpha''\,\mathrm{eps}}{1 - \alpha''\,\mathrm{eps}}\right)\left(1 + \frac{\|x^* - p\|}{\|x^*\|}\right) + \frac{\alpha''\,\mathrm{eps}}{1 - \alpha''\,\mathrm{eps}}\frac{\|x^* - p\|}{\|x^*\|} + O(\mathrm{eps}^2), \tag{5.14}$$

and if $m = n$,

$$\frac{\|x^* - x'\|}{\|x^*\|} \leqslant \frac{\alpha' K(A)\,\mathrm{eps}}{1 - \alpha' K(A)\,\mathrm{eps}}\left(1 + \frac{\alpha''\,\mathrm{eps}}{1 - \alpha''\,\mathrm{eps}}\right) + \frac{\alpha''\,\mathrm{eps}}{1 - \alpha''\,\mathrm{eps}}\frac{\|x^* - p\|}{\|x^*\|} + O(\mathrm{eps}^2). \tag{5.15}$$

If mixed precision arithmetic is not used (i.e., scalar products are computed by arithmetic of relative precision eps), upper bounds on $\|U\|$, $\|G\|/\|A\|$, $\|E\|$ and then on $\|x^* - x'\|/\|x^*\|$ can be computed using the results of [5], [6], [7], [8].

We have thus examined the effects of the numerical computation of the decomposition (3.1) and of the numerical solution of the triangular system $\bar{R}^T y = b$. It must be remembered that due to rounding errors, using (3.4) and (3.6), $\bar{x}^{(i)}$ ($\bar{x} = \bar{x}^{(m)}$) will be computed instead of $x'^{(i)}$ ($x' = x'^{(m)}$). From the algorithm (3.4) we have

$$\bar{x}^{(i)} = \begin{cases} \mathrm{fl}\left\{\bar{x}^{(i-1)} - \bar{q}_i\left[\dfrac{\bar{q}_i^T p - \bar{d}_i}{\bar{q}_i^T \bar{q}}\right]\right\} & \text{if} \quad Q^T Q \neq I, \\ \mathrm{fl}\left\{\bar{x}^{(i-1)} - \bar{q}_i\left(\bar{q}_i^T p - \bar{d}_i\right)\right\} & \text{if} \quad Q^T Q = I, \end{cases} \tag{5.16}$$

and from the algorithm (3.6) we have

$$\bar{x}^{(i)} = \begin{cases} \mathrm{fl}\left\{\bar{x}^{(i-1)} - \bar{q}_i\left[\dfrac{\bar{q}_i^T \bar{x}^{(i-1)} - \bar{d}_i}{\bar{q}_i^T \bar{q}}\right]\right\} & \text{if} \quad Q^T Q \neq I, \\ \mathrm{fl}\left\{\bar{x}^{(i-1)} - \bar{q}_i\left(\bar{q}_i^T \bar{x}^{(i-1)} - \bar{d}_i\right)\right\} & \text{if} \quad Q^T Q = I. \end{cases} \tag{5.17}$$

Here we have used the $\mathrm{fl}(\cdot)$ operator to signify computed quantities. The total error will be given by

$$\|x^* - \bar{x}\| \leqslant \|x^* - x'\| + \|x' - \bar{x}\|.$$

$\|x' - \bar{x}\|$ will then be examined. For this reason let us now prove a lemma which is useful in proving theorems which give bounds for $\|x' - x\|$. In the following we will assume that

$$\mathrm{eps} \leqslant 0.01.$$

LEMMA 5.1. *If* $Q^TQ \neq I$ *then*

$$\|Q_i'^T p - d'^{(i)}\| \leqslant i^{1/2}\|\hat{x}^{(i)} - p\| \tag{5.18}$$

and

$$\|x'^{(i)}\| \leqslant \|p\| + i\|\hat{x}^{(i)} - p\|. \tag{5.19}$$

If $Q^TQ = I$ *and* $\|Q'\| = 1 + O(eps)$ *then*

$$\|Q_i'^T p - d'^{(i)}\| \leqslant \|\hat{x}^{(i)} - p\| + O(\text{eps}) \tag{5.20}$$

and

$$\|x'^{(i)}\| \leqslant \|p\| + \|\hat{x}^{(i)} - p\| + O(\text{eps}) \tag{5.21}$$

Proof. Because

$$\|Q_j'\| \leqslant \|Q_j'\|_F = j^{1/2} \qquad \text{if} \quad Q^TQ \neq I,$$

$$\|Q_j'\| = 1 + O(\text{eps}) \qquad \text{if} \quad Q^TQ = I,$$

from (4.4) we obtain (5.18) and (5.20), and from (4.5) we obtain (5.19) and (5.21).

THEOREM 5.1. *If* $Q^TQ \neq I$ *and* $\bar{x}^{(i)}$ $(\bar{x} = \bar{x}^{(m)})$ *is the computed value of* $x^{(i)}$ $(x^* = x^{(m)})$ *using the algorithm* (3.4), *we have*

$$\bar{x}^{(i)} = \bar{x}^{(i-1)} - \bar{q}_i \frac{\bar{q}_i^T p - \bar{d}_i}{\bar{q}_i^T q_i} + \sigma^{(i)} \tag{5.22}$$

with

$$\|\sigma^{(i)}\| \leqslant \left[\, 5\|\bar{x}^{(i)}\| + 6\|\bar{x}^{(i-1)}\| + \|p\|\right] \text{eps} + O(\text{eps}^2), \tag{5.23}$$

and furthermore

$$\bar{x}^{(i)} = x'^{(i)} + v^{(i)} \tag{5.24}$$

with

$$v^{(i)} = \sum_{j=1}^{i} \sigma^{(j)} \tag{5.25}$$

and

$$\|v^{(i)}\| \leqslant (12i+6)\left[\|p\| + \frac{i+1}{2}\|\hat{x}^{(i)} - p\|\right]\text{eps} + O(\text{eps}^2). \tag{5.26}$$

Proof. Computing in mixed precision arithmetic $\bar{x}^{(i)}$, expressed by the first row of (5.16), we obtain

$$\|\sigma^{(i)}\| \leqslant \text{eps}\left\{\frac{5|\bar{q}_i^T p - \bar{d}_i|}{\|\bar{q}_i\|} + \|\bar{x}^{(i-1)}\| + \|p\|\right\} + O(\text{eps}^2).$$

From this and because

$$\frac{|\bar{q}_i^T p - \bar{d}_i|}{\|\bar{q}_i\|} \leqslant \|\bar{x}^{(i)} - \bar{x}^{(i-1)}\| + \|\sigma^{(i)}\|$$

$$\leqslant \|\bar{x}^{(i)}\| + \|\bar{x}^{(i-1)}\| + \|\sigma^{(i)}\|,$$

(5.23) will follow. Formula (5.24) is obtained recursively from (5.22). We can now prove (5.26). From (5.23), (5.24), and (5.25) we obtain

$$\|v^{(i)}\| \leqslant 11\,\text{eps}\left\{\sum_{j=1}^{i}\left[\|v^{(j)}\| + \|x'^{(j)}\|\right]\right\}$$

$$+ (i+6)\|p\|\text{eps} + O(\text{eps}^2),$$

from which we have

$$\|v^{(i)}\| - \frac{11\,\text{eps}}{1-11\,\text{eps}}\sum_{j=1}^{i-1}\|v^{(j)}\| \leqslant \frac{11\,\text{eps}}{1-11\,\text{eps}}\left\{\sum_{j=1}^{i}\|x'^{(j)}\| + \frac{i+6}{11}\|p\|\right\}. \tag{5.27}$$

Denoting by L the lower triangular matrix of elements

$$l_{kk} = 1,$$

$$l_{kj} = -\frac{11\,\text{eps}}{1-11\,\text{eps}} < 0,\ k > j,$$

and by z and t the vectors of elements

$$z_k = \|v^{(k)}\|,$$

$$t_k = \sum_{j=1}^{k} \|x'^{(j)}\| + \frac{k+6}{11}\|p\|,$$

we can write (5.27) as

$$Lz \leqslant \frac{11\,\text{eps}}{1-11\,\text{eps}}\, t.$$

Because the nondiagonal elements of L are nonpositive and the easily computable matrix L^{-1} is nonnegative, the matrix L will be monotone. Therefore we have

$$z \leqslant \frac{11\,\text{eps}}{1-11\,\text{eps}}\, L^{-1} t.$$

From this it is easy to obtain

$$\|v^{(i)}\| \leqslant \frac{11\,\text{eps}}{1-11\,\text{eps}} \left[\sum_{j=1}^{i} \|x'^{(j)}\| + \frac{i+6}{11}\|p\| \right]$$

$$\times \left[1 + \frac{11\,\text{eps}}{1-11\,\text{eps}} \sum_{k=0}^{i-2} \left(\frac{1}{1-11\,\text{eps}} \right)^k \right],$$

and (5.26) thus follows. ■

THEOREM 5.2. *If* $Q^TQ \neq I$ *and* $\bar{x}^{(i)}$ $(\bar{x} = \bar{x}^{(m)})$ *is the value of* $x^{(i)}$ $(x^* = x^{(m)})$ *computed by the algorithm* (3.6), *then*

$$\bar{x}^{(i)} = \bar{x}^{(i-1)} - \bar{q}_i \frac{\bar{q}_i^T \bar{x}^{(i-1)} - \bar{d}_i}{\bar{q}_i^T \bar{q}_i} + \tau^{(i)} \tag{5.28}$$

with

$$\begin{aligned} &\|\tau^{(1)}\| \leqslant \operatorname{eps}\left[5\|x'^{(1)}\| + 7\|p\|\right] + O(\operatorname{eps}^2) \\ &\|\tau^{(i)}\| \leqslant \operatorname{eps}\left[5\|\bar{x}^{(i)}\| + 7\|\bar{x}^{(i-1)}\|\right] + O(\operatorname{eps}^2), \qquad i = 2, \cdots, m. \end{aligned} \tag{5.29}$$

Moreover

$$\bar{x}^{(i)} = x'^{(i)} + w^{(i)} \tag{5.30}$$

with

$$\begin{aligned} w^{(1)} &= \tau^{(1)} \\ w^{(i)} &= \frac{1}{\bar{q}_i^T \bar{q}_i} \bar{q}_i \bar{q}_i^T \bar{Q}_{i-1} D_{i-1}^{-2} \left[\bar{Q}_{i-1}^T p - \bar{d}^{(i-1)}\right] \\ &\quad + \left[I - \frac{\bar{q}_i \bar{q}_i^T}{(\bar{q}_i^T \bar{q}_i)}\right] w^{(i-1)} + \tau^{(i)}, \qquad i = 2, \cdots, m. \end{aligned} \tag{5.31}$$

and

$$\begin{aligned} \|w^{(1)}\| &\leqslant \operatorname{eps}\left\{5\|x'^{(1)}\| + 7\|p\|\right\} + O(\operatorname{eps}^2), \\ \|w^{(i)}\| &\leqslant (i-1) i^{1/2} \|\hat{x}^{(i)} - p\| \, \|E\| (1 + 6i \operatorname{eps}) \\ &\quad + 12\left[i\|p\| + (i^2+1)\|\hat{x}^{(i)} - p\|\right] \operatorname{eps} \\ &\quad + O(\operatorname{eps}^2), \qquad i = 2, \cdots, m. \end{aligned} \tag{5.32}$$

Proof. The proof of (5.28) and (5.29) is similar to the proof given in Theorem 5.1 for the formulas (5.22) and (5.23).

Let us now prove (5.30) and (5.31) by induction. These formulas are true for $i = 1$. We assume that they are true for $i = 2, \cdots, j-1$. We will now

prove that they are true for $i = j$. From (5.28) we have successively

$$\begin{aligned}
\bar{x}^{(j)} &= p - \bar{Q}_{j-1}D_{j-1}^{-2}\left(\bar{Q}_{j-1}^{T}p - \bar{d}^{(j-1)}\right) + w^{(j-1)} \\
&\quad - \frac{1}{\bar{q}_j^{T}\bar{q}_j}\bar{q}_j\left\{\bar{q}_j^{T}\left[p - \bar{Q}_{j-1}D_{j-1}^{-2}\left(\bar{Q}_{j-1}^{T}p - \bar{d}^{(j-1)}\right) + w^{(j-1)}\right] - \bar{d}_j\right\} \\
&\quad + \tau^{(j)} \\
&= p - \bar{Q}_{j-1}D_{j-1}^{-2}\left[\bar{Q}_{j-1}^{T}p - \bar{d}^{(j-1)}\right] - \frac{1}{\bar{q}_j^{T}\bar{q}_j}\bar{q}_j\left(\bar{q}_j^{T}p - \bar{d}_j\right) \\
&\quad + \frac{1}{\bar{q}_j^{T}\bar{q}_j}\bar{q}_j\bar{q}_j^{T}\bar{Q}_{j-1}D_{j-1}^{-2}\left(\bar{Q}_{j-1}^{T}p - \bar{d}^{(j-1)}\right) \\
&\quad + \left\{I - \frac{\bar{q}_j\bar{q}_j^{T}}{\bar{q}_j^{T}\bar{q}_j}\right\}w^{(j-1)} + \tau^{(j)} \\
&= p - \bar{Q}_jD_j^{-2}\left[\bar{Q}_j^{T}p - \bar{d}^{(j)}\right] + w^{(j)}.
\end{aligned}$$

Let us now prove (5.32). Because $\|q_i'\| = 1$, $\|q_i'^{T}Q_{i-1}'\| \leqslant \|E\|$, $\|I - q_i'q_i'^{T}\| = 1$, by (5.18), (5.29), and (5.30) we have

$$\begin{aligned}
\|w^{(i)}\| &\leqslant i^{1/2}\|E\|\,\|\hat{x}^{(i)} - p\| + \|w^{(i-1)}\| \\
&\quad + \operatorname{eps}\left\{5\|w^{(i)}\| + 7\|w^{(i-1)}\|\right. \\
&\quad\quad \left. + 5\|x'^{(i)}\| + 7\|x'^{(i-1)}\|\right\} + O(\operatorname{eps}^2).
\end{aligned}$$

From this, by (5.19), we obtain

$$\begin{aligned}
\|w^{(i)}\| &\leqslant \frac{1 + 7\operatorname{eps}}{1 - 5\operatorname{eps}}\|w^{(i-1)}\| \\
&\quad + \frac{1}{1 - 5\operatorname{eps}}\left\{i^{1/2}\|\hat{x}^{(i)} - p\|\,\|E\| + \operatorname{eps}\left[12\|p\| + (12i - 7)\|\hat{x}^{(i)} - p\|\right]\right\} \\
&\quad + O(\operatorname{eps}^2),
\end{aligned}$$

and recursively,

$$\|w^{(i)}\| \leqslant \left(\frac{1+7\,\mathrm{eps}}{1-5\,\mathrm{eps}}\right)^{i-1}\|r^{(1)}\| + \left[\left(\frac{1+7\,\mathrm{eps}}{1-5\,\mathrm{eps}}\right)^{i-1}-1\right]\left(\frac{1}{12\,\mathrm{eps}}\right) \times \left\{i^{1/2}\|\hat{x}^{(i)}-p\|\,\|E\|+\mathrm{eps}\left[12\|p\|+(12i-7)\|\hat{x}^{(i)}-p\|\right]\right\} + O(\mathrm{eps}^2).$$

From this last formula and taking into account (5.29), Formulas (5.31) and (5.32) are proved.

THEOREM 5.3. *If* $Q^TQ = I$, *if* $\|Q'\| = 1 + O(\mathrm{eps})$, *and if* $\bar{x}^{(i)}$ $(\bar{x} = \bar{x}^{(m)})$ *is the value of* $x^{(i)}$ $(x^* = x^{(m)})$ *computed by the algorithm* (3.4), *then we have*

$$\bar{x}^{(i)} = \bar{x}^{(i-1)} - \bar{q}_i\left(\bar{q}_i^T p - \bar{d}_i\right) + \sigma'^{(i)} \tag{5.33}$$

with

$$\|\sigma'^{(i)}\| \leqslant \left[3\|\bar{x}^{(i)}\| + 4\|\bar{x}^{(i-1)}\| + \|p\|\right]\mathrm{eps} + O(\mathrm{eps}^2). \tag{5.34}$$

Moreover

$$\bar{x}^{(i)} = x'^{(i)} + v'^{(i)} \tag{5.35}$$

with

$$v'^{(i)} = \sum_{j=1}^{i} \sigma'^{(j)} \tag{5.36}$$

and

$$\|v'^{(i)}\| \leqslant (8i+4)\left[\|p\| + \|\hat{x}^{(i)} - p\|\right]\mathrm{eps} + O(\mathrm{eps}^2). \tag{5.37}$$

Proof. This proof is similar to the proof of Theorem 5.1. ∎

THEOREM 5.4. *If* $Q^TQ = I$, *if* $\|Q'\| = 1 + O(\text{eps})$, *and if* $\bar{x}^{(i)}$ ($\bar{x} = \bar{x}^{(m)}$) *is the value of* $x^{(i)}$ ($x^* = x^{(m)}$) *computed by the algorithm* (3.6), *then*

$$\bar{x}^{(i)} = \bar{x}^{(i-1)} - \bar{q}_i\left(\bar{q}_i^T\bar{x}^{(i-1)} - \bar{d}_i\right) + \tau'^{(i)} \tag{5.38}$$

with

$$\begin{aligned}
&\|\tau'^{(1)}\| \leqslant \text{eps}\left[3\|x'^{(1)}\| + 5\|p\|\right] + O(\text{eps}^2)\\
&\|\tau'^{(i)}\| \leqslant \text{eps}\left[3\|\bar{x}^{(i)}\| + 5\|\bar{x}^{(i-1)}\|\right] + O(\text{eps}^2), \qquad i = 2, \cdots, m.
\end{aligned} \tag{5.39}$$

Moreover

$$\bar{x}^{(i)} = x'^{(i)} + w'^{(i)} \tag{5.40}$$

with

$$\begin{aligned}
w'^{(1)} &= \tau'^{(1)}\\
w'^{(i)} &= \bar{q}_i\bar{q}_i^T\bar{Q}_{i-1}\left[\bar{Q}_{i-1}^T p - \bar{d}^{(i-1)}\right]\\
&\quad + \left[I - \bar{q}_i\bar{q}_i^T\right]w'^{(i-1)} + \tau'^{(i)}, \qquad i = 2, \cdots, m,
\end{aligned} \tag{5.41}$$

and

$$\begin{aligned}
\|w'^{(1)}\| &\leqslant \text{eps}\left\{3\|x'^{(1)}\| + 5\|p\|\right\} + O(\text{eps}^2)\\
\|w'^{(i)}\| &\leqslant (i-1)\|\hat{x}^{(i)} - p\|\,\|E\|(1 + 4i\,\text{eps})\\
&\quad + 8i\,\text{eps}\left[\|p\| + \|\hat{x}^{(i)} - p\|\right]\\
&\quad + O(\text{eps}^2), \qquad i = 2, \cdots, m.
\end{aligned} \tag{5.42}$$

Proof. This proof is similar to the proof of Theorem 5.2. ∎

We can now compute bounds on $\|x' - \bar{x}\|/\|x^*\|$ for the algorithms (3.4) and (3.6) when the orthogonal decomposition of A is performed by the modified Gram-Schmidt, Householder, or Givens method.

(a) Modified Gram-Schmidt Method

(a1) Algorithm (3.4). From Theorem 5.1, with $\rho \leqslant 0.5$, we have, if $m < n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (12m + 6)\left[\frac{\|p\|}{\|x^*\|} + \frac{m+1}{2} + (m+1)\frac{\|x^* - p\|}{\|x^*\|}\right]\text{eps}, \tag{5.43}$$

and if $m = n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (12m + 6) \times \left[\frac{\|p\|}{\|x^*\|} + \frac{m+1}{2} + \frac{m+1}{2} \frac{\|x^* - p\|}{\|x^*\|} \right] \text{eps}. \tag{5.44}$$

(a2) Algorithm (3.6). Taking into account Theorem 5.2, with $\rho \leqslant 0.5$, we have, if $m < n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (m-1)m^{1/2}\rho \left[1 + 2\frac{\|x^* - p\|}{\|x^*\|} \right](1 + 6m \text{ eps}) + 12\text{eps} \left[m\frac{\|p\|}{\|x^*\|} + (m^2+1)\left(1 + 2\frac{\|x^* - p\|}{\|x^*\|} \right) \right], \tag{5.45}$$

and if $m = n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (m-1)m^{1/2}\rho \left[1 + \frac{\|x^* - p\|}{\|x^*\|} \right](1 + 6m \text{ eps}) + 12\text{eps} \left[m\frac{\|p\|}{\|x^*\|} + (m^2+1)\left(1 + \frac{\|x^* - p\|}{\|x^*\|} \right) \right]. \tag{5.46}$$

(b) Householder and Givens Methods

(b1) Algorithm (3.4). From Theorem 5.3, with $\alpha' K(A)\text{eps} \leqslant 0.5$, we have, if $m < n$

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (8m+4) \left[\frac{\|p\|}{\|x^*\|} + 2\frac{\|x^* - p\|}{\|x^*\|} + 1 \right] \text{eps} + O(\text{eps}^2), \tag{5.47}$$

and if $m = n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant (8m+4) \left[\frac{\|p\|}{\|x^*\|} + \frac{\|x^* - p\|}{\|x^*\|} + 1 \right] \text{eps} + O(\text{eps}^2). \tag{5.48}$$

(b2) Algorithm (3.6). Taking into account Theorem 5.4, with $\alpha' K(A)\text{eps} \leqslant 0.5$, we have, if $m < n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant \alpha''(m-1)\left[1 + 2\frac{\|x^* - p\|}{\|x^*\|}\right]\text{eps} + 8m\left[\frac{\|p\|}{\|x^*\|} + 2\frac{\|x^* - p\|}{\|x^*\|} + 1\right]\text{eps} + O(\text{eps}^2), \quad (5.49)$$

and if $m = n$,

$$\frac{\|x' - \bar{x}\|}{\|x^*\|} \leqslant \alpha''(m-1)\left[1 + \frac{\|x^* - p\|}{\|x^*\|}\right]\text{eps} + 8m\left[\frac{\|p\|}{\|x^*\|} + \frac{\|x^* - p\|}{\|x^*\|} + 1\right]\text{eps} + O(\text{eps}^2). \quad (5.50)$$

Let $\bar{Q}$ now be the computed value of Q using the modified Gram-Schmidt method. Let us give an upper bound on $\|q_j - \bar{q}_j\|$ which will be used in the following section.

From the results of Section 3, the algorithm (3.10) will compute the columns $\bar{q}_j$ of $\bar{Q}$. From (5.12) and (5.45) we will have then

$$\|q_j - \bar{q}_j\| \leqslant \|a_j\|\frac{\rho_{j-1}}{1 - \rho_{j-1}}(5j + 6) + 3\|a_j\|(12j^2 + 7j + 12)\text{eps}, \quad (5.51)$$

where ρ_{j-1} is

$$\rho_{j-1} = 3.42j(j-1)K(A_{j-1})\text{eps} < 1$$

6. LINEAR DEPENDENCE, CONSISTENCY, PIVOTING, AND SCALING

From the theoretical point of view, the above algorithms can also be used to check whether an equation of (1.1) is linearly dependent on the previous equations and, moreover, if the system (1.1) is consistent.

From (3.7), if $r_{ii}q_i = 0$, then the ith column of A will be linearly dependent on the previous columns. Consequently, if $q_i = 0$ occurs during the modified Gram-Schmidt orthogonal factorization or if $r_{ii} = 0$ occurs during the Householder or Givens factorization, then we must also check whether $a_i^T x^{(i-1)} - b_i = 0$. If so, the ith equation will be a linear combination of the previous equations; otherwise the system (1.1) will not be consistent.

Numerical checks are not so easy. In [3] a bound on the error $||r_{ii}| - |\bar{r}_{ii}||$ is computed, where $\bar{r}_{ii}$ is the computed value of r_{ii} with the Householder or Givens method. In this paper the second member of (5.51) is an upper bound on $|\,||q_i|| - ||\bar{q}_i||\,|$, where $\bar{q}_i$ is an approximation, computed using the modified Gram-Schmidt method, to the vector q_i. Both upper bounds are functions of $||a_i||$ and of the condition number of A_{i-1}. This suggests that in the algorithms for computing the decomposition (3.1), a scaling procedure to impose $||a_i|| = 1$, $i = 1, \cdots, m$, and a pivoting procedure, analogous to that performed in algorithms for least squares problems [9], should both be applied.

However, in ill-conditioned problems, it is very difficult and sometimes impossible to check the linear dependence of columns of the matrix A. Analogous troubles occur when performing consistency checks on the system (1.1) [3].

7. CONCLUSIONS

Let us now examine the bounds on the errors $||x^* - x'||/||x^*||$ and $||x' - \bar{x}||/||x^*||$ obtained in the previous sections. We particularly want to point out their dependence on $K(A)$ and $||x^* - p||/||x^*||$. This also explains the numerical results on the algorithm (3.6), given in [2].

The upper bounds obtained for $||x' - \bar{x}||/||x^*||$ are independent of the product $K(A)||x^* - p||/||x^*||$ except for the algorithm (3.6), in which the q_i are computed by the modified Gram-Schmidt method.

The upper bounds on $||x^* - x'||/||x^*||$ depend on the product $K(A)||x^* - p||/||x^*||$ except when $m = n$ and the decomposition (3.1) is computed by the Householder or Givens methods.

From the computational error point of view, there is no significant difference between the algorithms (3.4) and (3.6), if the matrix Q is computed using the same method in both. Moreover, the algorithms (3.4) and (3.6) do not substantially differ if the matrix Q is computed by the Householder or by the Givens method.

The numerical results obtained in [2] for algorithm (3.6), can be now explained: if $m < n$ there will be no substantial differences in the results when

Q is computed by the modified Gram-Schmidt or the Householder method; if $m = n$ better results are obtained if Q is computed by the Householder method.

REFERENCES

1 H. Y. Huang, A direct method for the general solution of a system of linear equations, *J. Optim. Theory Appl.* 16:429–445 (1975).
2 M. Arioli and A. Laratta, Metodi diretti per la risoluzione di sistemi lineari, *Calcolo* XXI:229–252 (1984).
3 M. Arioli and A. Laratta, Error analysis of an algorithm for solving an underdetermined linear system, *Numer. Math.* 46:255–268 (1985).
4 M. Arioli, A. Laratta, and O. Menchi, Numerical computation of the projection of a point onto a polyhedron, *J. Optim. Theory Appl.* 43:495–525 (1984).
5 A. Bjorck, Solving linear least squares problems by Gram-Schmidt orthogonalization, *BIT* 7:1–21 (1967).
6 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.
7 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J., 1974.
8 W. M. Gentleman, Error analysis of QR decompositions by Givens transformations, *Linear Algebra Appl.* 10:189–197 (1975).
9 J. J. Dongarra, J. R. Brunch, G. B. Moler, and G. W. Stewart, LINPACK *User's Guide*, SIAM, Philadelphia, 1979.

Received 13 August 1985; revised 3 October 1985