BIT 7 (1967), 1–21

SOLVING LINEAR LEAST SQUARES PROBLEMS BY GRAM-SCHMIDT ORTHOGONALIZATION

ÅKE BJÖRCK

Abstract.

A general analysis of the condition of the linear least squares problem is given. The influence of rounding errors is studied in detail for a modified version of the Gram–Schmidt orthogonalization to obtain a factorization $\boldsymbol{A}=\boldsymbol{QR}$ of a given $m\times n$ matrix $\boldsymbol{A}$, where $\boldsymbol{R}$ is upper triangular and $\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{I}$. Let $\boldsymbol{x}$ be the vector which minimizes $\|\boldsymbol{b}-\boldsymbol{Ax}\|_2$ and $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{Ax}$. It is shown that if inner-products are accumulated in double precision then the errors in the computed $\boldsymbol{x}$ and $\boldsymbol{r}$ are less than the errors resulting from some simultaneous initial perturbation $\delta\boldsymbol{A}$, $\delta\boldsymbol{b}$ such that

$$\|\delta\boldsymbol{A}\|_E/\|\boldsymbol{A}\|_E \approx \|\delta\boldsymbol{b}\|_2/\|\boldsymbol{b}\|_2 \approx 2\cdot n^{3/2} \text{ machine units}.$$

No reorthogonalization is needed and the result is independent of the pivoting strategy used.

Key words: Least square, linear equations, orthogonalization.

1. Introduction.

Let $\boldsymbol{A}$ be a given $m\times n$ real matrix of rank n, $m\geqq n$, and $\boldsymbol{b}$ a given $m\times 1$ real vector. Then there exists a unique vector $\boldsymbol{x}$ which solves the least squares problem of minimizing

$$\|\boldsymbol{b}-\boldsymbol{Ax}\|_2 .$$

It is well known that the solution $\boldsymbol{x}$ satisfies the condition

$$\boldsymbol{A}^T(\boldsymbol{b}-\boldsymbol{Ax}) = \boldsymbol{0}\,, \tag{1.1}$$

i.e. the residual vector $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{Ax}$ is orthogonal to the columns of $\boldsymbol{A}$. It follows that we can compute $\boldsymbol{x}$ from the normal equations

$$\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{A}^T\boldsymbol{b}\,. \tag{1.2}$$

We define following Bauer [1] the (columnwise) condition of the rectangular matrix $\boldsymbol{A}$ to be

$$\operatorname{cond}(\boldsymbol{A}) = \max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{Ax}\|/\min_{\|\boldsymbol{x}\|=1}\|\boldsymbol{Ax}\|\,.$$

The condition number corresponding to the L_2-norm is denoted by $\varkappa(\boldsymbol{A})$. In section 7 we will show that under some restrictions, $\varkappa(\boldsymbol{A})$ can be con-

sidered as an approximate condition number for the problem of determining $\boldsymbol{x}$.[1] On the other hand

$$\varkappa(\boldsymbol{A}^T\boldsymbol{A}) = \varkappa^2(\boldsymbol{A}) .$$

This shows that in general using t-digit binary arithmetic, we will not be able to obtain even an approximate solution to (1.2) unless

$$\varkappa(\boldsymbol{A}) \leqq 2^{t/2} .$$

Now let $\boldsymbol{B}=\boldsymbol{A}\boldsymbol{S}$, where $\boldsymbol{S}$ is square and non-singular. Then from (1.1) follows $\boldsymbol{B}^T(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})=0$ and thus the equations

$$\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}^T\boldsymbol{b} \tag{1.3}$$

can be used instead of the normal equations. A natural question to ask is if we can chose $\boldsymbol{B}$ in such a way that

$$\varkappa(\boldsymbol{B}^T\boldsymbol{A}) = \varkappa(\boldsymbol{A}) . \tag{1.4}$$

This is indeed possible. Since the columns of $\boldsymbol{A}$ are linearly independent we have a factorization of $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \tag{1.5}$$

where $\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{I}$ and $\boldsymbol{R}$ is an upper triangle. Moreover, each row of $\boldsymbol{R}$ and each column of $\boldsymbol{Q}$ is uniquely determined up to a scalar factor of modulus one ([4] p. 7).

Choosing $\boldsymbol{B}=\boldsymbol{A}\boldsymbol{R}^{-1}=\boldsymbol{Q}$ the matrix $\boldsymbol{B}^T\boldsymbol{A}=\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}=\boldsymbol{R}$ is triangular. The equations (1.3) are thus easily solved by back-substitution. Further the condition (1.4) is satisfied since

$$\varkappa(\boldsymbol{R}) = \varkappa(\boldsymbol{Q}\boldsymbol{R}) = \varkappa(\boldsymbol{A}) . \tag{1.6}$$

This factorization can be obtained mainly in two different ways, and both have been proposed for solving least squares problems. One way is to annihilate the subdiagonal elements of $\boldsymbol{A}$ by a sequence of elementary Hermitian matrices

$$\boldsymbol{H}^{(k)} = (\boldsymbol{I}-2\boldsymbol{w}_k\boldsymbol{w}_k{}^T), \qquad k = 1,2,\ldots,n ,$$
$$\boldsymbol{H}^{(n)}\ldots\boldsymbol{H}^{(2)}\boldsymbol{H}^{(1)}\boldsymbol{A} = \boldsymbol{Q}^T\boldsymbol{A} = \boldsymbol{R} .$$

A detailed error estimate for this algorithm has been given by Wilkinson [7] pp. 152–161, and the application to least squares problems is treated by Golub in [2].

A second way to obtain the factorization is by applying the Gram–Schmidt orthogonalization procedure to the columns of $\boldsymbol{A}$. This is however known to have poor numerical stability.

[1] The reader may prefer to read this section at this stage.

In this paper we will use a slightly modified version of the Gram–Schmidt procedure, which is equivalent to the elimination with weighted row combinations in [1]. This will be applied to the solution of linear least squares problems. The propagation of rounding errors will be studied in detail and in particular an estimate of the deviation from orthogonality of the computed Q and an error bound for the computed solution will be derived.

2. Description of the algorithm.

In the actual computation it is preferable to use a different normalization of the factors in (1.5). We avoid computing square roots if we take instead

$$A = Q'R' \tag{2.1}$$

where R' is a *unit* upper triangle and $(Q')^T Q'$ diagonal. For completeness we give below the algorithms both for the classical and the modified Gram–Schmidt procedure. In both cases we compute a sequence of matrices

$$A = A^{(1)}, A^{(2)}, \ldots, A^{(n+1)} = Q' .$$

Classical Gram–Schmidt:

Here the elements of R' are computed one column at a time. We define

$$A^{(k)} = (q_1', \ldots, q_{k-1}', a_k, \ldots, a_n)$$

and assume

$$(q_p')^T q_r' = \delta_{pr} d_p, \qquad 1 \leqq p, r \leqq k-1 .$$

In step k we compute

$$r_{jk}' = (q_j')^T a_k / d_j, \qquad 1 \leqq j \leqq k-1 ,$$

$$q_k' = a_k - \sum_{j=1}^{k-1} r_{jk}' q_j', \qquad d_k = \|q_k'\|_2^2 .$$

After step n we treat the right hand side b in a similar way and compute

$$y_j' = (q_j')^T b / d_j, \quad 1 \leqq j \leqq n, \quad r = b - \sum_{j=1}^{n} y_j' q_j' .$$

Modified Gram–Schmidt:

Here the elements of R' are computed one row at a time. We define

$$A^{(k)} = (q_1', \ldots, q_{k-1}', a_k^{(k)}, \ldots, a_n^{(k)})$$

and assume

$$(\boldsymbol{q}_p')^T\boldsymbol{q}_r' = \delta_{pr}\boldsymbol{d}_p, \quad (\boldsymbol{q}_p')^T\boldsymbol{a}_j^{(k)} = 0, \quad 1 \leqq p,r \leqq k-1, \quad k \leqq j \leqq n .$$

In step k we take $\boldsymbol{q}_k' = \boldsymbol{a}_k^{(k)}$ and compute

$$\begin{aligned} d_k &= \|\boldsymbol{q}_k'\|_2^2, \quad r_{kj}' = (\boldsymbol{q}_k')^T\boldsymbol{a}_j^{(k)}/d_k , \\ \boldsymbol{a}_j^{(k+1)} &= \boldsymbol{a}_j^{(k)} - r_{kj}'\boldsymbol{q}_k', \quad k+1 \leqq j \leqq n . \end{aligned} \tag{2.2}$$

The right hand side is transformed in a similar way

$$\boldsymbol{b} = \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots, \boldsymbol{b}^{(n+1)} = \boldsymbol{r}$$

where in step k we compute

$$y_k' = (\boldsymbol{q}_k')^T\boldsymbol{b}^{(k)}/d_k, \quad \boldsymbol{b}^{(k+1)} = \boldsymbol{b}^{(k)} - y_k'\boldsymbol{q}_k' . \tag{2.3}$$

For both procedures we have taking $r_{kk}' = 1$

$$\boldsymbol{a}_k = \sum_{j=1}^{k} r_{jk}'\boldsymbol{q}_j', \quad \boldsymbol{r} = \boldsymbol{b} - \sum_{j=1}^{n} y_j'\boldsymbol{q}_j' .$$

Thus they both produce, if performed without rounding errors, the wanted unique factorizations

$$\boldsymbol{A} = \boldsymbol{Q}'\boldsymbol{R}', \qquad \boldsymbol{b} = \boldsymbol{Q}'\boldsymbol{y}' + \boldsymbol{r} ,$$

where

$$(\boldsymbol{Q}')^T\boldsymbol{Q}' = \boldsymbol{D} = \operatorname{diag}\{d_1, d_2, \ldots, d_n\} .$$

The two procedures, however, have completely different numerical properties when $n > 2$. If $\boldsymbol{A}$ is at all ill-conditioned, then using the classical procedure, the computed columns of $\boldsymbol{Q}'$ will soon completely lose their orthogonality. Consequently this procedure should never be used without reorthogonalization, which greatly increases the amount of computation. As we shall show, reorthogonalization is never needed when using the modified procedure for solving least squares problems.

The following example due to Läuchli illustrates well the different behavior of the two procedures. Suppose

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}$$

where ε is a small number such that due to rounding the result of the operation $1+\varepsilon^2$ is everywhere put equal to 1. It is easily verified, that if no other rounding errors are made, then the maximum deviation from orthogonality of the computed columns of $\boldsymbol{Q}'$ is given by

$$\text{classical: } \frac{|(\boldsymbol{q}_3')^T\boldsymbol{q}_2'|}{\|\boldsymbol{q}_3'\|_2\|\boldsymbol{q}_2'\|_2} = \tfrac{1}{2}; \quad \text{modified: } \frac{|(\boldsymbol{q}_3')^T\boldsymbol{q}_1'|}{\|\boldsymbol{q}_3'\|_2\|\boldsymbol{q}_1'\|_2} = \sqrt{\tfrac{2}{3}}\varepsilon\,.$$

Note that $\varkappa(\boldsymbol{A}) \approx \sqrt{3}\varepsilon^{-1}$ and that $\boldsymbol{A}^T\boldsymbol{A}$ (rounded) has rank 1, so that the normal equations are singular.

The modified procedure also has the advantage of easily allowing for instance one of the following pivoting strategies to be used. We can choose as $\boldsymbol{q}_k'$ the column of $\boldsymbol{A}^{(k)}$ for which $\|\boldsymbol{a}_j^{(k)}\|_2$, $k \leqq j \leqq n$, is maximized. Then as is seen from (2.2) the elements of $\boldsymbol{R}'$ will satisfy $|r_{kj}'| \leqq 1$. If we wish to express $\boldsymbol{b}$ in as few columns of $\boldsymbol{A}$ as possible, we can choose instead as $\boldsymbol{q}_k'$ the column for which $|p_{kj}|$ is maximized, where

$$p_{kj} = (\boldsymbol{a}_j^{(k)})^T\boldsymbol{b}^{(k)}/\|\boldsymbol{a}_j^{(k)}\|_2, \qquad k \leqq j \leqq n\,.$$

Both $\|\boldsymbol{a}_j^{(k)}\|_2$ and p_{kj} are easily calculated by recursion. This is probably one reason why the modified procedure often has been preferred in practice.

We will show that the superiority of the modified procedure is due to other factors and independent of the pivoting strategy. This is in agreement with experimental results reported by Rice [5].

3. Basic definitions for the error analysis.

In the error analysis we shall assume that floating point arithmetic is used, and follow the technique and the notations introduced by Wilkinson [6]. Thus let 'op' denote any of the four arithmetic operators $+ \; - \; * \; /$. Then an equation of the form

$$z = \mathrm{fl}(x\text{'op'}y)$$

will imply that x, y and z are floating point numbers and z obtained from x and y using the appropriate floating point operation. We assume that the rounding errors in these operations are such that

$$\mathrm{fl}(x\text{'op'}y) = (x\text{'op'}y)(1+\varepsilon), \qquad |\varepsilon| \leqq 2^{-t}\,. \tag{3.1}$$

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors of dimension n, where $n \cdot 2^{-t} \leqq 0.1$. Then the following error bounds for the computed innerproduct of $\boldsymbol{x}$ and $\boldsymbol{y}$ are valid ([7] pp. 114, 117):

$$|\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \leqq n \cdot 2^{-t_1}|\boldsymbol{x}^T||\boldsymbol{y}|\,, \tag{3.2}$$

$$|\mathrm{fl}_2(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \leqq 2^{-t}|\boldsymbol{x}^T\boldsymbol{y}| + \tfrac{3}{2} \cdot n2^{-2t_2}|\boldsymbol{x}^T||\boldsymbol{y}| \tag{3.3}$$

where

$$t_1 = t - \log_2(1.06), \quad 2t_2 = 2t - \log_2(1.06)\,. \tag{3.4}$$

Here $|x|$ denotes a vector with components $|x_i|$ and $\mathrm{fl}_2(\ )$ applies to the case when the inner-product is accumulated in double precision and then rounded.

Computed quantities will in the following be denoted by bars. Thus, we write

$$\bar{A}^{(k)} = (\bar{q}_1', \ldots, \bar{q}_{k-1}', \bar{a}_k^{(k)}, \ldots, \bar{a}_n^{(k)})$$

and the formulas (2.2) and (2.3) become

$$\begin{aligned} \bar{d}_k &= \mathrm{fl}(\|\bar{q}_k'\|_2^2), \quad \bar{r}_{kj}' = \mathrm{fl}((\bar{q}_k')^T \bar{a}_j^{(k)} / \bar{d}_k)\,, \\ \bar{a}_j^{(k+1)} &= \mathrm{fl}(\bar{a}_j^{(k)} - \bar{r}_{kj}' \bar{q}_k'), \quad k+1 \leqq j \leqq n\,, \end{aligned} \tag{3.5}$$

and

$$\bar{y}_k' = \mathrm{fl}((\bar{q}_k')^T \bar{b}^{(k)} / \bar{d}_k), \quad \bar{b}^{(k+1)} = \mathrm{fl}(\bar{b}^{(k)} - \bar{y}_k' \bar{q}_k')\,. \tag{3.6}$$

This algorithm will not break down unless $\bar{d}_k = 0$ for some k. This may happen due to rounding errors even when the rank of A equals n. According to (3.2) and (3.3) $\bar{d}_k = 0$ implies that $\bar{q}_k' = \mathbf{0}$. To account for this case we add the (trivial) singular rule that when $\bar{q}_k' = \mathbf{0}$ we take

$$\bar{r}_{kj}' = 0, \quad k+1 \leqq j \leqq n, \quad \bar{y}_k' = 0\,. \tag{3.7}$$

Then $\bar{R}'$ and $\bar{y}'$ are always uniquely defined.

For the convenience of notation we also introduce the normalized quantities

$$\bar{q}_k = d_k^{-\frac{1}{2}} \bar{q}_k', \quad \bar{r}_{kj} = d_k^{\frac{1}{2}} \bar{r}_{kj}', \quad \bar{y}_k = d_k^{\frac{1}{2}} \bar{y}_k' \tag{3.8}$$

where

$$d_k^{\frac{1}{2}} = \begin{cases} \|\bar{q}_k'\|_2, & \bar{q}_k' \neq \mathbf{0} \\ 1\ , & \bar{q}_k' = \mathbf{0} \end{cases}$$

Note that these quantities are never computed and thus (3.8) are exact relations.

4. Errors in an elementary projection.

If $\bar{q}_k' \neq 0$, then in the k:th step we compute the vectors $\bar{a}_j^{(k+1)}$, $j = k+1, \ldots, n$ as the projection of $\bar{a}_j^{(k)}$ on the subspace complementary to $\bar{q}_k'$. If this is performed without rounding errors the result is

$$(I - \bar{q}_k \bar{q}_k^T) \bar{a}_j^{(k)} = \bar{a}_j^{(k)} - r_{kj} \bar{q}_k$$

where r_{kj} is the *exact* multiplier corresponding to the *computed* $\bar{q}_k$ and $\bar{a}_j^{(k)}$. If using the *computed* multiplier $\bar{r}_{kj}'$ the subtraction is performed exactly, then the result is

$$\bar{a}_j^{(k)} - \bar{r}_{kj}' \bar{q}_k' = \bar{a}_j^{(k)} - \bar{r}_{kj} \bar{q}_k\,.$$

We define the related errors $\boldsymbol{\delta}_j^{(k)}$ and $\boldsymbol{\eta}_j^{(k)}$ by

$$\bar{\boldsymbol{a}}_j^{(k+1)} = \bar{\boldsymbol{a}}_j^{(k)} - \bar{r}_{kj}\bar{\boldsymbol{q}}_k + \boldsymbol{\delta}_j^{(k)} \tag{4.1}$$

$$\bar{\boldsymbol{a}}_j^{(k+1)} = (\boldsymbol{I} - \bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)\bar{\boldsymbol{a}}_j^{(k)} + \boldsymbol{\eta}_j^{(k)} . \tag{4.2}$$

In the singular case when $\bar{\boldsymbol{q}}_k' = \boldsymbol{0}$ these relations are satisfied with

$$\bar{\boldsymbol{a}}_j^{(k+1)} = \bar{\boldsymbol{a}}_j^{(k)}, \quad \boldsymbol{\delta}_j^{(k)} = \boldsymbol{\eta}_j^{(k)} = \boldsymbol{0} . \tag{4.3}$$

In the non-singular case we shall prove the following estimates which are basic for our analysis

$$\|\boldsymbol{\delta}_j^{(k)}\|_2 \leqq 1.45 \cdot 2^{-t}\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 \tag{4.4}$$

$$\|\boldsymbol{\eta}_j^{(k)}\|_2 \leqq \begin{cases} 3.23 \cdot 2^{-t}\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 \text{ d.p.acc.} \\ (2m+3) \cdot 2^{-t_1}\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 \text{ s.p.} \end{cases} \tag{4.5}$$

where (4.4) is valid also when single precision is used. For simplicity of notation we omit for a while the indices j and k and put

$$\bar{\boldsymbol{a}}_j^{(k+1)} = \boldsymbol{z}, \quad \bar{\boldsymbol{a}}_j^{(k)} = \boldsymbol{y}, \quad \bar{\boldsymbol{q}}_k = \boldsymbol{x} .$$

Then (3.5), (4.1) and (4.2) become

$$\bar{r}' = \mathrm{fl}\big((\boldsymbol{x}')^T\boldsymbol{y}/(\boldsymbol{x}')^T\boldsymbol{x}'\big), \quad \boldsymbol{z} = \mathrm{fl}(\boldsymbol{y} - \bar{r}'\boldsymbol{x}') , \tag{4.6}$$

$$r = \boldsymbol{x}^T\boldsymbol{y}, \quad \boldsymbol{z} = \boldsymbol{y} - \bar{r}\boldsymbol{x} + \boldsymbol{\delta}, \quad \boldsymbol{z} = \boldsymbol{y} - r\boldsymbol{x} + \boldsymbol{\eta} . \tag{4.7}$$

From (3.1) and (3.8) follows

$$z_i = \big(y_i - \bar{r}x_i(1+\varepsilon_1)\big)(1+\varepsilon_2), \quad |\varepsilon_1| \leqq 2^{-t}, \quad |\varepsilon_2| \leqq 2^{-t} .$$

Using this to eliminate y_i from the definition of $\boldsymbol{\delta}$ we get

$$\delta_i = \frac{\varepsilon_2}{1+\varepsilon_2} z_i - \varepsilon_1 \cdot \bar{r}x_i ,$$

and hence since $\|\boldsymbol{x}\|_2 = 1$

$$\|\boldsymbol{\delta}\|_2 \leqq \frac{2^{-t}}{1-2^{-t}}\|\boldsymbol{z}\|_2 + 2^{-t}|\bar{r}| . \tag{4.8}$$

Immediately from (4.7)

$$\|\boldsymbol{\eta}\|_2 \leqq \|\boldsymbol{\delta}\|_2 + |r - \bar{r}| . \tag{4.9}$$

To estimate the error in the multiplier $\bar{r}$ we use (3.1)–(3.3) and the identity

$$\frac{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y})}{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{x})} - r = \frac{\boldsymbol{x}^T\boldsymbol{x}}{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{x})}\left[\frac{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{x}} - r\frac{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{x}) - \boldsymbol{x}^T\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}\right].$$

If we compute $\bar{r}'$ as

$$\bar{d} = \mathrm{fl}_2((x')^T x'), \quad \bar{r}' = \mathrm{fl}_2\big((x')^T y/\bar{d}\big)$$

accumulating inner-products in double precision and divide $\bar{d}$ into the double precision mantissa before rounding, then

$$|\bar{r}-r| \leqq \frac{2^{-t}(2+\frac{3}{2}\cdot m\cdot 2^{-t_1})|r| + \frac{3}{2}\cdot m\cdot 2^{-t_2}\|y\|_2}{1-2^{-t}(1+\frac{3}{2}\cdot m\cdot 2^{-t_1})}.$$

Using instead single precision we have

$$|\bar{r}-r| \leqq \frac{(m+1)2^{-t_1}|r| + m(1+2^{-t})2^{-t_1}\|y\|_2}{1-m\cdot 2^{-t_1}}.$$

In order to simplify these and later inequalities we make in the following the reasonable assumptions

$$m \geqq 2, \quad 2(m+1)2^{-t_1} < 0.01\,. \tag{4.10}$$

Then we certainly have

$$|\bar{r}-r| < \begin{cases} (2.01\cdot|r| + 0.01\|y\|_2)2^{-t} & \text{d.p.acc.} \\ ((m+1)\cdot|r| + m\|y\|_2)2^{-t_1} & \text{s.p.} \end{cases} \tag{4.11}$$

where we have taken into account that because of (4.10) the factor 1.06 in 2^{-t_1} is now sufficiently generous to allow also for the factors $1-m\cdot 2^{-t_1}$ and $1+2^{-t}$.

Since $(z-\eta)$ is orthogonal to x it follows

$$\|z\|_2 \leqq (\|y\|_2^2 - r^2)^{\frac{1}{2}} + \|\eta\|_2\,. \tag{4.12}$$

Substituting this and (4.8) in (4.9) and solving for $\|\eta\|_2$ gives after some rearranging

$$(1-2\cdot 2^{-t})\|\eta\|_2 \leqq 2^{-t}\big((\|y\|_2^2 - r^2)^{\frac{1}{2}} + |r|\big) + |r-\bar{r}|\,.$$

Hence using (4.11)

$$(1-2\cdot 2^{-t})\|\eta\|_2 \leqq \begin{cases} ((\|y\|_2^2 - r^2)^{\frac{1}{2}} + 3|r| + 0.02\|y\|_2)2^{-t}, & \text{d.p.acc.} \\ ((\|y\|_2^2 - r^2)^{\frac{1}{2}} + (m+2)|r| + m\|y\|_2)2^{-t_1}, & \text{s.p.} \end{cases} \tag{4.13}$$

Maximizing over $|r|$, where $0 \leqq |r| \leqq \|y\|_2$ we have

$$(\|y\|_2^2 - r^2)^{\frac{1}{2}} + k|r| \leqq (1+k^2)^{\frac{1}{2}}\|y\|_2\,, \tag{4.14}$$

and finally (4.5) follows from (4.13) by invoking (4.10).

For the estimation of $\|\delta\|_2$ (4.8) and (4.12) yields

$$\|\delta\|_2 \leqq \frac{2^{-t}}{1-2^{-t}}\big((\|y\|_2^2 - r^2)^{\frac{1}{2}} + |r| + \|\eta\|_2 + |r-\bar{r}|\big)\,.$$

Using (4.14) for the first two terms and single precision bounds for the last two we get

$$\|\delta\|_2 \leqq \frac{2^{-t}}{1-2^{-t}}\left(2^{\frac{1}{2}}+4(m+1)\cdot 2^{-t_1}\right)\|y\|_2 .$$

and invoking (4.10) again this proves (4.4).

In exact computation we certainly have $\|\bar{a}_j^{(k+1)}\|_2 \leqq \|\bar{a}_j^{(k)}\|_2$. This however need not be true when rounding errors are taken into account. Estimating $\|\eta\|_2$ from (4.13) in (4.12) and again maximizing the resulting right hand side as a function of $|r|$, we get after some computations

$$\|\bar{a}_j^{(k+1)}\|_2 < \begin{cases} (1+1.05\cdot 2^{-t})\|\bar{a}_j^{(k)}\|_2, & \text{d.p.acc.} \\ (1+1.01(m+2)\cdot 2^{-t_1})\|\bar{a}_j^{(k)}\|_2, & \text{s.p.} \end{cases} \tag{4.15}$$

Note that these bounds can only be approached when $\bar{a}_j^{(k)}$ is almost orthogonal to $\bar{q}_k$.

5. Errors in the factorization.

Using the error analysis of the elementary projection we can now derive bounds for the errors related to the factorization of A and b. *While we will be able to show that the error* $(\bar{Q}\bar{R}-A)$ *is small independently of* $\varkappa(A)$, *this will not be true of the error* $(\bar{R}-\bar{Q}^T A)$. This leads us to introduce the matrices

$$\tilde{Q}=(\tilde{q}_1,\tilde{q}_2,\ldots,\tilde{q}_n), \quad \tilde{\tilde{Q}}=(\tilde{\tilde{q}}_1,\tilde{\tilde{q}}_2,\ldots,\tilde{\tilde{q}}_n)$$

where

$$\begin{aligned} \tilde{q}_k &= (I-\bar{q}_1\bar{q}_1{}^T)(I-\bar{q}_2\bar{q}_2{}^T)\ldots(I-\bar{q}_{k-1}\bar{q}_{k-1}^T)\bar{q}_k , \\ \tilde{\tilde{q}}_k &= (I-\bar{q}_n\bar{q}_n{}^T)(I-\bar{q}_{n-1}\bar{q}_{n-1}^T)\ldots(I-\bar{q}_{k+1}\bar{q}_{k+1}^T)\bar{q}_k . \end{aligned} \tag{5.1}$$

When $\bar{q}_j \neq 0$ the matrix

$$P^{(j)} = (I-\bar{q}_j\bar{q}_j{}^T)$$

is the projector for the subspace complementary to $\bar{q}_j$. Otherwise it is equal to the unit matrix. It follows that

$$\|P^{(j)}\|_2 = 1, \quad \|\tilde{q}_k\|_2 \leqq 1, \quad \|\tilde{\tilde{q}}_k\|_2 \leqq 1 . \tag{5.2}$$

We shall first prove a simple lemma concerning $\tilde{Q}$ and $\tilde{\tilde{Q}}$.

LEMMA 5.1. *Let* U *be the upper triangular matrix with elements*

$$u_{ij} = \bar{q}_i{}^T\bar{q}_j, \quad j > i, \quad u_{ij} = 0, \quad j \leqq i .$$

Then

$$\bar{Q}=\tilde{Q}(I+U), \quad \bar{Q}=\tilde{\tilde{Q}}(I+U^T) .$$

We first prove by induction in k that for $k=1,2,\ldots,n$

$$(\boldsymbol{I}-\bar{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T)\ldots(\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T) = \boldsymbol{I}-\tilde{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T-\ldots-\tilde{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T\,.$$

This is true for $k=1$ and we have

$$(\boldsymbol{I}-\bar{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T)\ldots(\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)(\boldsymbol{I}-\bar{\boldsymbol{q}}_{k+1}\bar{\boldsymbol{q}}_{k+1}^T) = -\tilde{\boldsymbol{q}}_{k+1}\bar{\boldsymbol{q}}_{k+1}^T+ +(\boldsymbol{I}-\bar{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T)\ldots(\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)\,,$$

which proves the hypothesis.

Using this result we get from the definition of $\tilde{\boldsymbol{q}}_k$

$$\tilde{\boldsymbol{q}}_k = (\boldsymbol{I}-\tilde{\boldsymbol{q}}_{k-1}\bar{\boldsymbol{q}}_{k-1}^T-\ldots-\tilde{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T)\bar{\boldsymbol{q}}_k\,,$$

and it follows that for $k=1,2,\ldots,n$

$$\bar{\boldsymbol{q}}_k = \tilde{\boldsymbol{q}}_k+(\bar{\boldsymbol{q}}_k^T\bar{\boldsymbol{q}}_{k-1})\tilde{\boldsymbol{q}}_{k-1}+\ldots+(\bar{\boldsymbol{q}}_k^T\bar{\boldsymbol{q}}_1)\tilde{\boldsymbol{q}}_1\,.$$

This is the k:th column of the equality $\bar{\boldsymbol{Q}}=\tilde{\boldsymbol{Q}}(\boldsymbol{I}+\boldsymbol{U})$ and the first part of the lemma is proved.

The second part is almost obvious from the symmetry of definition of $\tilde{\boldsymbol{Q}}$ and $\tilde{\tilde{\boldsymbol{Q}}}$, and is proved in the same way.

By taking the transpose of the induction hypothesis used in the proof of the lemma we obtain the

COROLLARY. *If* $\tilde{\boldsymbol{Q}}$ *is defined by* (5.1) *we have the identity*

$$\boldsymbol{I}-\bar{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^T = (\boldsymbol{I}-\bar{\boldsymbol{q}}_n\bar{\boldsymbol{q}}_n^T)\ldots(\boldsymbol{I}-\bar{\boldsymbol{q}}_2\bar{\boldsymbol{q}}_2^T)(\boldsymbol{I}-\bar{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1^T)\,.$$

We now define

$$\boldsymbol{E}_1 = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}-\boldsymbol{A}, \quad \boldsymbol{e}_1 = \bar{\boldsymbol{Q}}\bar{\boldsymbol{y}}+\bar{\boldsymbol{b}}^{(n+1)}-\boldsymbol{b}\,, \tag{5.3}$$

$$\boldsymbol{E}_2 = (\boldsymbol{I}-\bar{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^T)\boldsymbol{A}, \quad \boldsymbol{e}_2 = (\boldsymbol{I}-\bar{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^T)\boldsymbol{b}-\bar{\boldsymbol{b}}^{(n+1)}\,, \tag{5.4}$$

$$\boldsymbol{E}_3 = \bar{\boldsymbol{R}}-\tilde{\boldsymbol{Q}}^T\boldsymbol{A}, \quad \boldsymbol{e}_3 = \bar{\boldsymbol{y}}-\tilde{\boldsymbol{Q}}^T\boldsymbol{b}\,. \tag{5.5}$$

We shall prove the following estimates, which are valid if *inner-products are accumulated in double precision*[1]

$$\begin{aligned} \|\boldsymbol{E}_1\|_E &\leqq 1.5(n-1)\cdot 2^{-t}\|\boldsymbol{A}\|_E \\ \|\boldsymbol{e}_1\|_2 &\leqq 1.5n\cdot 2^{-t}\|\boldsymbol{b}\|_2 \end{aligned} \tag{5.6}$$

$$\begin{aligned} \|\boldsymbol{E}_2\|_E &\leqq 3.25(n-1)\cdot 2^{-t}\|\boldsymbol{A}\|_E \\ \|\boldsymbol{e}_2\|_2 &\leqq 3.25n\cdot 2^{-t}\|\boldsymbol{b}\|_2 \end{aligned} \tag{5.7}$$

$$\begin{aligned} \|\boldsymbol{E}_3\|_E &\leqq 1.9(n-1)^{\frac{1}{2}}n\cdot 2^{-t}\|\boldsymbol{A}\|_E \\ \|\boldsymbol{e}_3\|_2 &\leqq 1.9n^{\frac{1}{2}}(n+1)\cdot 2^{-t}\|\boldsymbol{b}\|_2 \end{aligned} \tag{5.8}$$

[1] Here the suffix E denotes the Frobenius norm i.e. $\|A\|_E=(\sum\sum|a_{ij}|^2)^{\frac{1}{2}}$.

Summing (4.1) for $k=1,2,\ldots,j-1$ and using

$$\bar{\boldsymbol{a}}_j^{(j)} = \bar{r}_{jj}\bar{\boldsymbol{q}}_j, \quad \bar{\boldsymbol{a}}_j^{(1)} = \boldsymbol{a}_j$$

we get

$$\sum_{k=1}^{j} \bar{r}_{kj}\cdot\bar{\boldsymbol{q}}_k - \boldsymbol{a}_j = \sum_{k=1}^{j-1} \boldsymbol{\delta}_j^{(k)} = \boldsymbol{\delta}_j .$$

From (4.4) follows

$$\|\boldsymbol{\delta}_j\|_2 \leqq 1.45\cdot 2^{-t}\sum_{k=1}^{j-1}\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 .$$

Using the $\mathrm{fl}_2(\ )$ bound in (4.15)

$$\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 < (1+1.05\cdot 2^{-t})^{k-1}\|\boldsymbol{a}_j\|_2$$

and thus invoking (4.10)

$$\|\boldsymbol{a}_j^{(k)}\|_2 < (1+1.05n\ 2^{-t_1})\|\boldsymbol{a}_j\|_2 < 1.006\|\boldsymbol{a}_j\|_2 . \tag{5.9}$$

This inequality holds even when using single precision if (4.10) is replaced by the stronger condition

$$2n(m+2)2^{-t_1} < 0.01 .$$

We then certainly have

$$\|\boldsymbol{\delta}_j\|_2 < 1.5\cdot(j-1)2^{-t}\|\boldsymbol{a}_j\|_2 \tag{5.10}$$

and

$$\|\boldsymbol{E}_1\|_E = \left(\sum_{j=1}^{n}\|\boldsymbol{\delta}_j\|_2^2\right)^{\frac{1}{2}} < 1.5\cdot(n-1)2^{-t}\left(\sum_{j=1}^{n}\|\boldsymbol{a}_j\|_2^2\right)^{\frac{1}{2}}$$

which proves the first part of (5.6). Since the right hand side $\boldsymbol{b}$ is treated in the same way as the columns of $\boldsymbol{A}$, the second part follows immediately.

Next we solve the difference equation (4.2) for $\bar{\boldsymbol{a}}_j^{(k)}$ and get for $k \leqq j$

$$\begin{aligned}\bar{\boldsymbol{a}}_j^{(k)} = \boldsymbol{a}_j^{(k)} + \boldsymbol{P}^{(k-1)}\ldots\boldsymbol{P}^{(3)}\boldsymbol{P}^{(2)}\boldsymbol{\eta}_j^{(1)} + \ldots \\ + \boldsymbol{P}^{(k-1)}\boldsymbol{\eta}_j^{(k-2)} + \boldsymbol{\eta}_j^{(k-1)}\end{aligned} \tag{5.11}$$

where

$$\boldsymbol{a}_j^{(k)} = \boldsymbol{P}^{(k-1)}\ldots\boldsymbol{P}^{(2)}\boldsymbol{P}^{(1)}\boldsymbol{a}_j .$$

Hence $\boldsymbol{a}_j^{(k)}$ is the vector we should obtain if we performed the computation *exactly* with the *computed* $\bar{\boldsymbol{q}}_k$.

Since $\bar{\boldsymbol{q}}_j' = \bar{\boldsymbol{a}}_j^{(j)}$ we have $(\boldsymbol{I}-\bar{\boldsymbol{q}}_j\bar{\boldsymbol{q}}_j^T)\bar{\boldsymbol{a}}_j^{(j)} = \boldsymbol{0}$. Consequently (5.11) holds also for $j+1 \leqq k \leqq n+1$ with

$$\bar{\boldsymbol{a}}_j^{(k)} = \boldsymbol{0}, \quad \boldsymbol{\eta}_j^{(j)} = \ldots = \boldsymbol{\eta}_j^{(k-1)} = \boldsymbol{0} .$$

Taking norms in (5.11) and using (5.2) we get

$$\|\bar{\boldsymbol{a}}_j^{(k)} - \boldsymbol{a}_j^{(k)}\|_2 \leqq \sum_{i=1}^{s-1} \|\eta_j^{(i)}\|_2, \quad s = \min(j,k)\,.$$

From this follows, using $\mathrm{fl}_2(\ )$ bounds from (4.5)

$$\|\bar{\boldsymbol{a}}_j^{(k)} - \boldsymbol{a}_j^{(k)}\|_2 \leqq 3.23\, 2^{-t} \sum_{i=1}^{s-1} \|\bar{\boldsymbol{a}}_j^{(i)}\|_2 \leqq 3.25(s-1)2^{-t}\|\boldsymbol{a}_j\|_2\,. \tag{5.12}$$

Using the corollary of lemma 5.1 we can write this for $k = n+1$ as

$$\|(\boldsymbol{I} - \bar{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^T)\boldsymbol{a}_j\|_2 \leqq 3.25(j-1)\cdot 2^{-t}\|\boldsymbol{a}_j\|_2\,.$$

This and a similar relation for $\boldsymbol{b}$ proves (5.7).

Using the identity

$$\tilde{\boldsymbol{q}}_k{}^T\boldsymbol{a}_j = \bar{\boldsymbol{q}}_k{}^T(\boldsymbol{I} - \bar{\boldsymbol{q}}_{k-1}\bar{\boldsymbol{q}}_{k-1}^T)\ldots(\boldsymbol{I} - \bar{\boldsymbol{q}}_1\bar{\boldsymbol{q}}_1{}^T)\boldsymbol{a}_j = \bar{\boldsymbol{q}}_k{}^T\boldsymbol{a}_j^{(k)}$$

we can write the component (k,j) of $\bar{\boldsymbol{R}} - \tilde{\boldsymbol{Q}}^T\boldsymbol{A}$

$$\bar{r}_{kj} - \tilde{\boldsymbol{q}}_k{}^T\boldsymbol{a}_j = (\bar{r}_{kj} - r_{kj}) + \bar{\boldsymbol{q}}_k{}^T(\bar{\boldsymbol{a}}_j^{(k)} - \boldsymbol{a}_j^{(k)}), \quad k < j\,.$$

It follows that

$$|\bar{r}_{kj} - \tilde{\boldsymbol{q}}_k{}^T\boldsymbol{a}_j| \leqq |\bar{r}_{kj} - r_{kj}| + \|\bar{\boldsymbol{a}}_j^{(k)} - \boldsymbol{a}_j^{(k)}\|_2\,.$$

Now (5.12) and (4.11) gives

$$|\bar{r}_{kj} - \tilde{\boldsymbol{q}}_k{}^T\boldsymbol{a}_j| = \big(2.02 + 3.25(k-1)\big)2^{-t}\|\boldsymbol{a}_j\|_2 < 3.25k\cdot 2^{-t}\|\boldsymbol{a}_j\|_2\,.$$

The L_2-norm of the j:th column in $\boldsymbol{E}_3$ is therefore bounded by

$$3.25\cdot 2^{-t}\left(\sum_{k=1}^{j-1} k^2\right)^{\frac{1}{2}}\|\boldsymbol{a}_j\|_2 < 1.9(j-1)^{\frac{1}{2}}j\cdot 2^{-t}\|\boldsymbol{a}_j\|_2\,,$$

which finally proves (5.8).

Note that *when $\boldsymbol{A}$ is ill-conditioned cancellation will occur, so that $\|\bar{\boldsymbol{a}}_j^{(k)}\|_2 \ll \|\boldsymbol{a}_j\|_2$ often for a rather small value of k.* Then the estimate (5.9) which we have used in (5.10) and (5.12) is very crude, and (5.6)–(5.8) *will considerably overestimate the error.*

If we use single precision (5.6) holds unchanged, but in (5.7) and (5.8) the bounds must be increased by a factor of $(\frac{2}{3}m+1)$.

6. Orthogonality of the computed vectors.

All the error bounds given in section 5 apply also in the singular case when $\bar{\boldsymbol{q}}_k{}' = 0$ for some k, i.e. when the rank of $\bar{\boldsymbol{A}} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}$ is less than n. We now derive a sufficient condition for $\bar{\boldsymbol{A}}$ to have rank n. Let the exact factorization of $\boldsymbol{A}$ be $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$. Then using (5.3) we can write

$$\bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}} = (\boldsymbol{A}+\boldsymbol{E}_1)^T(\boldsymbol{A}+\boldsymbol{E}_1) = \boldsymbol{R}^T(\boldsymbol{I}+\boldsymbol{F}_1)\boldsymbol{R} \tag{6.1}$$

where

$$\boldsymbol{F}_1 = (\boldsymbol{Q}^T\boldsymbol{E}_1\boldsymbol{R}^{-1})^T + \boldsymbol{Q}^T\boldsymbol{E}_1\boldsymbol{R}^{-1} + (\boldsymbol{E}_1\boldsymbol{R}^{-1})^T\boldsymbol{E}_1\boldsymbol{R}^{-1} . \tag{6.2}$$

Taking norms in (6.2) we get

$$\|\boldsymbol{F}_1\|_2 \leqq 2\|\boldsymbol{E}_1\|_E\|\boldsymbol{R}^{-1}\|_2 + \|\boldsymbol{E}_1\|_E^2\|\boldsymbol{R}^{-1}\|_2^2 \tag{6.3}$$

and thus $\|\boldsymbol{F}_1\|_2 < 1$ if

$$\|\boldsymbol{E}_1\|_E\|\boldsymbol{R}^{-1}\|_2 < \sqrt{2} - 1 .$$

Using (5.6) it follows that $\bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}}$ is non-singular and $\bar{\boldsymbol{A}}$ has rank n if

$$1.5(n-1)2^{-t}\|\boldsymbol{A}\|_E\|\boldsymbol{R}^{-1}\|_2 < \sqrt{2} - 1 . \tag{6.4}$$

We assume in the following that (6.4) is satisfied.

The orthogonality of the computed vectors $\bar{\boldsymbol{q}}_1, \bar{\boldsymbol{q}}_2, \ldots, \bar{\boldsymbol{q}}_n$ can be measured by the norm of the matrix $(\boldsymbol{I} - \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}})$. Now

$$\boldsymbol{I} - \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}} = -(\boldsymbol{U} + \boldsymbol{U}^T) \tag{6.5}$$

where $\boldsymbol{U}$ is the matrix introduced in lemma 5.1.

Summing (4.1) for $k = i+1, i+2, \ldots, j-1$ we get

$$\bar{\boldsymbol{a}}_j^{(i+1)} = \sum_{k=i+1}^{j} \bar{r}_{kj}\bar{\boldsymbol{q}}_k - \sum_{k=i+1}^{j-1} \boldsymbol{\delta}_j^{(k)} . \tag{6.6}$$

From (4.2) follows

$$\bar{\boldsymbol{q}}_i^T\bar{\boldsymbol{a}}_j^{(i+1)} = \bar{\boldsymbol{q}}_i^T\boldsymbol{\eta}_j^{(i)}$$

and hence multiplying (6.6) by $\bar{\boldsymbol{q}}_i^T$ we get

$$s_{ij} = \sum_{k=i+1}^{j} \bar{r}_{kj}(\bar{\boldsymbol{q}}_i^T\bar{\boldsymbol{q}}_k) = \bar{\boldsymbol{q}}_i^T\left(\boldsymbol{\eta}_j^{(i)} + \sum_{k=i+1}^{j-1} \boldsymbol{\delta}_j^{(k)}\right), \tag{6.7}$$

where s_{ij} is the component (i,j) of the matrix $\boldsymbol{S} = \boldsymbol{U}\bar{\boldsymbol{R}}$.
Using the $\mathrm{fl}_2(\,)$ bound from (4.10), (4.11) and (5.9) we get

$$|s_{ij}| \leqq \left(3.25 + 1.5(j-i-1)\right)\cdot 2^{-t}\|\boldsymbol{a}_j\|_2$$

and since

$$\sum_{i=1}^{j-1}(j-i+\tfrac{7}{6})^2 \leqq \sum_{i=1}^{n-1}(i+\tfrac{7}{6})^2 = \tfrac{1}{3}(n-1)\left((n+\tfrac{3}{2})^2 + \tfrac{11}{6}\right) < \frac{1.0001}{3}n(n+1)^2$$

we can bound the L_2-norm of each column of S by

$$0.87\cdot n^{\frac{1}{2}}(n+1)2^{-t}\|\boldsymbol{a}_j\|_2 .$$

After a similar estimation using the bounds for single precision we obtain

$$\|U\bar{R}\|_E < \begin{cases} 0.87 \cdot n^{\frac{1}{2}}(n+1)2^{-t}\|A\|_E & \text{d.p.acc.} \\ 0.87 \cdot n^{\frac{1}{2}}(n+1+2.5m)2^{-t}\|A\|_E & \text{s.p.} \end{cases}$$

From (6.5) follows that when accumulating inner-products in double precision we have the a posteriori estimation

$$\|I - \bar{Q}^T\bar{Q}\|_2 \leqq 2\|U\|_2 \leqq 1.74n^{\frac{1}{2}}(n+1)2^{-t}\|A\|_E\|\bar{R}^{-1}\|_2 . \tag{6.8}$$

To estimate $\|\bar{R}^{-1}\|_2$ in (6.8) we use the identity

$$\bar{R}^T\bar{R} = (A+E_1)^T(A+E_1) - \bar{R}^T(U+U^T)\bar{R}$$

which can be derived from (5.3) and (6.5). We write this

$$\bar{R}^T\bar{R} = R^T(I+F_2)R \tag{6.9}$$

where

$$F_2 = F_1 - R^{-T}(U\bar{R})^T\bar{R}R^{-1} - R^{-T}\bar{R}^T(U\bar{R})R^{-1} .^1$$

By (6.9) follows

$$\|\bar{R}R^{-1}\|_2^2 \leqq 1 + \|F_2\|_2 \tag{6.10}$$

and hence using (6.3) we have the inequality

$$\|F_2\|_2 \leqq 2\|E_1\|_E\|R^{-1}\|_2 + \|E_1\|_E^2\|R^{-1}\|_2^2 + 2\|U\bar{R}\|_E\|R^{-1}\|_2(1+\|F_2\|_2)^{\frac{1}{2}} .$$

Using the derived estimates for $\|E_1\|_E$ and $\|U\bar{R}\|_E$ and the inequality $(n-1) \leqq 0.3n^{\frac{1}{2}}(n+1)$ we get, assuming d.p.acc.

$$\|F_2\|_2 \leqq 2 \cdot 0.3 \cdot 1.5c + (0.3 \cdot 1.5 \cdot c)^2 + 2 \cdot 0.87c(1+\|F_2\|_2)^{\frac{1}{2}}$$

where

$$c = n^{\frac{1}{2}}(n+1) \cdot 2^{-t}\|A\|_E\|R^{-1}\|_2 .$$

It follows that $\|F_2\|_2 \leqq b$ where b is the positive root to the equation

$$b = 0.9 \cdot c + (0.45 \cdot c)^2 + 1.74 \cdot c(1+b)^{\frac{1}{2}} .$$

Since $b = 1$ corresponds to $c = 1/3.419\ldots$ it is readily verified that if we require

$$\beta = 3.42c = 3.42n^{\frac{1}{2}}(n+1)2^{-t}\|A\|_E\|R^{-1}\|_2 < 1 , \tag{6.12}$$

which is a stronger condition than (6.4), then $\|F_2\|_2 < \beta$ and from (6.9)

$$\|\bar{R}^{-1}\|_2^2 \leqq \frac{1}{1-\beta}\|R^{-1}\|_2^2 . \tag{6.13}$$

This inequality holds for single precision if we define β by

$$\beta = 3.42n^{\frac{1}{2}}(n+1+2.5m)2^{-t}\|A\|_E\|R^{-1}\|_2 .$$

Finally from (6.8) and (6.13) we get the a priori estimate

[1] The notation R^{-T} is used for $(R^{-1})^T$.

$$\|\boldsymbol{I}-\overline{\boldsymbol{Q}}^T\overline{\boldsymbol{Q}}\|_2 \leqq \frac{1.74}{(1-\beta)^{\frac{1}{2}}}\, n^{\frac{1}{2}}(n+1)\cdot 2^{-t}\|\boldsymbol{A}\|_E\|\boldsymbol{R}^{-1}\|_2\,. \tag{6.14}$$

In [5] Rice, after extensive computational experiments, reports that when using the modified Gram–Schmidt method "pivoting results in a perceptible, but small (even neglible) improvement" in the orthogonality of the computed vectors. Our analysis gives a theoretical explanation of this. *The bound in* (6.14) *is independent of pivoting, and since*

$$\|\boldsymbol{A}\|_E\|\boldsymbol{R}^{-1}\|_2 \leqq n^{\frac{1}{2}}\varkappa(\boldsymbol{A})$$

the deviation from orthogonality of $\overline{\boldsymbol{Q}}$ *can be expected to depend essentially only on* $\varkappa(\boldsymbol{A})$.

7. Condition of the least squares problem.

Before estimating the errors in the computed solution, we shall study the sensitivity to perturbations in $\boldsymbol{A}$ and $\boldsymbol{b}$ of the least squares problem, which, following a suggestion by Golub, we write

$$\begin{pmatrix} \boldsymbol{r}+\boldsymbol{A}\boldsymbol{x} \\ \boldsymbol{A}^T\boldsymbol{r} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} \tag{7.1}$$

Let the perturbed system be

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{A}+\delta\boldsymbol{A} \\ \boldsymbol{A}^T+\delta\boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{r}+\delta\boldsymbol{r} \\ \boldsymbol{x}+\delta\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}+\delta\boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix}$$

and let

$$\tilde{\boldsymbol{A}} = \boldsymbol{A}+\delta\boldsymbol{A} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{R}}, \quad \tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}} = \boldsymbol{I} \tag{7.2}$$

be the factorization of the perturbed matrix. Then

$$\begin{pmatrix} \delta\boldsymbol{r} \\ \delta\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{A}}^T & 0 \end{pmatrix}^{-1} \left[-\begin{pmatrix} \boldsymbol{0} & \delta\boldsymbol{A} \\ \delta\boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{pmatrix} + \begin{pmatrix} \delta\boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} \right].$$

Using (7.2) the inverse can readily be calculated ([4] p. 17) and we get

$$\begin{pmatrix} \delta\boldsymbol{r} \\ \tilde{\boldsymbol{R}}\delta\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}-\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^T & \tilde{\boldsymbol{V}} \\ \tilde{\boldsymbol{V}}^T & -\boldsymbol{I} \end{pmatrix} \begin{pmatrix} -\delta\boldsymbol{A}\boldsymbol{x}+\delta\boldsymbol{b} \\ -\tilde{\boldsymbol{R}}^{-T}\delta\boldsymbol{A}^T\boldsymbol{r} \end{pmatrix} \tag{7.3}$$

The eigenvalues of the symmetric matrix $(\boldsymbol{I}-\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^T)$ are all equal to 0 or 1. Hence $\|\boldsymbol{I}-\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^T\|_2 \leqq 1$ and (7.3) yields

$$\begin{pmatrix} \|\delta\boldsymbol{r}\|_2 \\ \|\tilde{\boldsymbol{R}}\delta\boldsymbol{x}\|_2 \end{pmatrix} \leqq \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \|\tilde{\boldsymbol{R}}^{-1}\|_2\|\delta\boldsymbol{A}\|_2\|\boldsymbol{r}\|_2 \\ \|\delta\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2+\|\delta\boldsymbol{b}\|_2 \end{pmatrix} \tag{7.4}$$

To estimate $\|\tilde{\boldsymbol{R}}^{-1}\|_2$, let the factorization of $\boldsymbol{A}$ be $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R}$. Then we can write

$$\tilde{\boldsymbol{R}}^T\tilde{\boldsymbol{R}} = \tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}} = (\boldsymbol{A}+\delta\boldsymbol{A})^T(\boldsymbol{A}+\delta\boldsymbol{A}) = \boldsymbol{R}^T(\boldsymbol{I}+\boldsymbol{K})\boldsymbol{R}$$

where

$$\boldsymbol{K} = (\boldsymbol{Q}^T\delta\boldsymbol{A}\boldsymbol{R}^{-1})^T + \boldsymbol{Q}^T\delta\boldsymbol{A}\boldsymbol{R}^{-1} + (\delta\boldsymbol{A}\boldsymbol{R}^{-1})^T\delta\boldsymbol{A}\boldsymbol{R}^{-1}\,.$$

Put

$$\alpha = (\sqrt{2}+1)\|\boldsymbol{R}^{-1}\|_2\|\delta\boldsymbol{A}\|_2\,, \tag{7.5}$$

and *assume that* $\alpha<1$. Then, it is easily shown that

$$\|\boldsymbol{K}\|_2 \leqq 2\|\boldsymbol{R}^{-1}\|_2\|\delta\boldsymbol{A}\|_2 + \|\boldsymbol{R}^{-1}\|_2^2\|\delta\boldsymbol{A}\|_2^2 < \alpha < 1\,,$$

and hence

$$\|\tilde{\boldsymbol{R}}^{-1}\|_2^2 \leqq \|\boldsymbol{R}^{-1}\|_2\|(\boldsymbol{I}+\boldsymbol{K})^{-1}\|_2\|\boldsymbol{R}^{-T}\|_2 < \frac{1}{1-\alpha}\|\boldsymbol{R}^{-1}\|_2^2\,.$$

Hence by (1.6)

$$\|\boldsymbol{A}\|_2\|\tilde{\boldsymbol{R}}^{-1}\|_2 \leqq (1-\alpha)^{-\frac{1}{2}}\|\boldsymbol{R}\|_2\|\boldsymbol{R}^{-1}\|_2 = (1-\alpha)^{-\frac{1}{2}}\varkappa(\boldsymbol{A})\,.$$

Since $\|\delta\boldsymbol{x}\|_2 \leqq \|\tilde{\boldsymbol{R}}^{-1}\|_2\|\tilde{\boldsymbol{R}}\delta\boldsymbol{x}\|_2$ it now follows from (7.4)

$$\begin{pmatrix} \|\delta\boldsymbol{r}\|_2 \\ \|\boldsymbol{A}\|_2\|\delta\boldsymbol{x}\|_2 \end{pmatrix} \leqq \begin{pmatrix} \dfrac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}} & 1 \\ \dfrac{\varkappa^2(\boldsymbol{A})}{1-\alpha} & \dfrac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}} \end{pmatrix} \begin{pmatrix} \|\boldsymbol{r}\|_2\,\dfrac{\|\delta\boldsymbol{A}\|_2}{\|\boldsymbol{A}\|_2} \\ \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2\,\dfrac{\|\delta\boldsymbol{A}\|_2}{\|\boldsymbol{A}\|_2} + \|\delta\boldsymbol{b}\|_2 \end{pmatrix} \tag{7.6}$$

It is possible to construct examples where these bounds are nearly obtained.

The result (7.6) has the following interesting interpretations. Notice that

$$\boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{x}$$

and thus

$$\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2 + \varkappa(\boldsymbol{A})\|\boldsymbol{r}\|_2 \leqq \varkappa(\boldsymbol{A})(\|\boldsymbol{A}\boldsymbol{x}\|_2 + \|\boldsymbol{r}\|_2) \leqq \sqrt{2}\varkappa(\boldsymbol{A})\|\boldsymbol{b}\|_2\,.$$

Then from the first row of (7.6) follows

$$\frac{\|\delta\boldsymbol{r}\|_2}{\|\boldsymbol{b}\|_2} \leqq \sqrt{2}\,\frac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}}\,\frac{\|\delta\boldsymbol{A}\|_2}{\|\boldsymbol{A}\|_2} + \frac{\|\delta\boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2}. \tag{7.7}$$

Hence as *condition number for the decomposition of* $\boldsymbol{b}$ *into orthogonal components* $\boldsymbol{A}\boldsymbol{x}$ *and* $\boldsymbol{r}$ *we can take* $\sqrt{2}(1-\alpha)^{-\frac{1}{2}}\varkappa(\boldsymbol{A})$.

If we assume that $\|\boldsymbol{x}\|_2 \neq 0$, then we can also derive a condition number for the determination of $\boldsymbol{x}$. From the second row (7.6) follows

$$\frac{\|\delta \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leqq \frac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}}\left(1+\frac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}}\,\frac{\|\boldsymbol{r}\|_2}{\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2}\right)\frac{\|\delta \boldsymbol{A}\|_2}{\|\boldsymbol{A}\|_2}+\frac{\varkappa(\boldsymbol{A})}{\sqrt{1-\alpha}}\,\frac{\|\delta \boldsymbol{b}\|_2}{\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2}.$$

Notice in particular the importance of the ratio $\|\boldsymbol{r}\|_2/\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$. *If the equations are nearly compatible* in the sense that $\|\boldsymbol{r}\|_2 \ll \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$, then the situation is much the same when $m>n$ as when $m=n$, and $\varkappa(\boldsymbol{A})$ *is an approximate condition number for the determination of* $\boldsymbol{x}$. However, if $\|\boldsymbol{r}\|_2$ is of the same magnitude as $\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$ then the condition number (in an ill-conditioned case) is more like $\varkappa^2(\boldsymbol{A})$.

8. Errors in the computed solution.

As the computed solution we take

$$\overline{\boldsymbol{r}}=\overline{\boldsymbol{b}}^{(n+1)}, \quad \overline{\boldsymbol{x}} = fl((\overline{\boldsymbol{R}}')^{-1}\overline{\boldsymbol{y}}') \,. \tag{8.1}$$

If we define $\tilde{\boldsymbol{R}}$ so that $\overline{\boldsymbol{x}}$ is the exact solution of

$$\tilde{\boldsymbol{R}}\overline{\boldsymbol{x}} = (\overline{\boldsymbol{R}}+\delta\overline{\boldsymbol{R}})\overline{\boldsymbol{x}} = \overline{\boldsymbol{y}} \,,$$

then using definitions (5.4) and (5.5) we can write (8.1)

$$\overline{\boldsymbol{r}} = (\boldsymbol{I}-\overline{\boldsymbol{Q}}\,\tilde{\boldsymbol{Q}}^T)\boldsymbol{b}-\boldsymbol{e}_2 \,, \qquad \overline{\boldsymbol{x}} = \tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{b}+\boldsymbol{e}_3) \,. \tag{8.2}$$

Substituting $\boldsymbol{b}=\boldsymbol{r}+\boldsymbol{A}\boldsymbol{x}$ in (8.2) we get

$$\begin{pmatrix}\overline{\boldsymbol{r}}\\ \overline{\boldsymbol{x}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}-\overline{\boldsymbol{Q}}\,\tilde{\boldsymbol{Q}}^T & (\boldsymbol{I}-\overline{\boldsymbol{Q}}\,\tilde{\boldsymbol{Q}}^T)\boldsymbol{A}\\ \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{Q}}^T & \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{Q}}^T\boldsymbol{A}\end{pmatrix}\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{x}\end{pmatrix}+\begin{pmatrix}-\boldsymbol{e}_2\\ \tilde{\boldsymbol{R}}^{-1}\boldsymbol{e}_3\end{pmatrix} \tag{8.3}$$

Now since $\boldsymbol{A}^T\boldsymbol{r}=0$ we have from (5.3)

$$\overline{\boldsymbol{Q}}^T\boldsymbol{r} = \overline{\boldsymbol{R}}^{-T}(\boldsymbol{A}+\boldsymbol{E}_1)^T\boldsymbol{r} = \overline{\boldsymbol{R}}^{-T}\boldsymbol{E}_1{}^T\boldsymbol{r}$$

and from lemma 5.1 it follows

$$\tilde{\boldsymbol{Q}}^T = (\boldsymbol{I}+\boldsymbol{U}^T)^{-1}\overline{\boldsymbol{Q}}^T, \quad \tilde{\tilde{\boldsymbol{Q}}} = \overline{\boldsymbol{Q}}(\boldsymbol{I}+\boldsymbol{U}^T)^{-1} \,.$$

Thus the errors in the computed solution can be written

$$\begin{pmatrix}\overline{\boldsymbol{r}}-\boldsymbol{r}\\ \overline{\boldsymbol{x}}-\boldsymbol{x}\end{pmatrix} = \begin{pmatrix}-\tilde{\tilde{\boldsymbol{Q}}}\,\overline{\boldsymbol{R}}^{-T}\boldsymbol{E}_1{}^T & (\boldsymbol{I}-\overline{\boldsymbol{Q}}\,\tilde{\boldsymbol{Q}}^T)\boldsymbol{A}\\ (\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\tilde{\boldsymbol{R}})^{-1}\boldsymbol{E}_1{}^T & \tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{A}-\tilde{\boldsymbol{R}})\end{pmatrix}\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{x}\end{pmatrix}+\begin{pmatrix}-\boldsymbol{e}_2\\ \tilde{\boldsymbol{R}}^{-1}\boldsymbol{e}_3\end{pmatrix} \tag{8.4}$$

We shall now estimate the L_2-norms of the submatrices appearing in (8.4). We assume for shortness that the $\mathrm{fl}_2(\ )$ mode of computation is used both in the decomposition and in the backsubstitution. The estimates relevant when using single precision are not very different and can be derived in a similar way.

Since according to (5.2) $\|\tilde{\tilde{\boldsymbol{Q}}}\|_E \leqq n^{\frac{1}{2}}$ we have immediately from (5.6) and (5.7)

$$\|-\tilde{\tilde{\boldsymbol{Q}}}\,\overline{\boldsymbol{R}}^{-T}\boldsymbol{E}_1{}^T\|_2 \leqq 1.5\cdot n^{\frac{1}{2}}(n-1)\cdot 2^{-t}\|\overline{\boldsymbol{R}}^{-1}\|_2\|\boldsymbol{A}\|_E\,, \tag{8.5}$$

$$\|(\boldsymbol{I}-\overline{\boldsymbol{Q}}\,\tilde{\boldsymbol{Q}}^T)\boldsymbol{A}\|_2 \leqq 3.25\cdot(n-1)\cdot 2^{-t}\|\boldsymbol{A}\|_E\,. \tag{8.6}$$

We rewrite the remaining two submatrices in (8.4) as

$$(\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\tilde{\boldsymbol{R}})^{-1}\boldsymbol{E}_1{}^T = \tilde{\boldsymbol{R}}^{-1}\overline{\boldsymbol{R}}(\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\overline{\boldsymbol{R}})^{-1}\boldsymbol{E}_1{}^T\,, \tag{8.7}$$

$$\tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{A}-\tilde{\boldsymbol{R}}) = \tilde{\boldsymbol{R}}^{-1}\overline{\boldsymbol{R}}(\overline{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{A}-\overline{\boldsymbol{R}})-\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}})\,. \tag{8.8}$$

Now from (6.1) and (6.5) it follows

$$\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\overline{\boldsymbol{R}} = \bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}}-\overline{\boldsymbol{R}}^T\boldsymbol{U}\overline{\boldsymbol{R}} = \boldsymbol{R}^T(\boldsymbol{I}+\boldsymbol{F}_3)\boldsymbol{R}$$

where

$$\boldsymbol{F}_3 = \boldsymbol{F}_1-\boldsymbol{R}^{-T}\overline{\boldsymbol{R}}^T(\boldsymbol{U}\overline{\boldsymbol{R}})\boldsymbol{R}^{-1}\,.$$

Comparing this with (6.9) and (6.13) we observe that certainly $\|\boldsymbol{F}_3\|_2 \leqq \beta$ and

$$\|(\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\overline{\boldsymbol{R}})^{-1}\|_2 \leqq \frac{1}{1-\beta}\,\|\boldsymbol{R}^{-1}\|_2^2$$

If we put

$$\|\tilde{\boldsymbol{R}}^{-1}\overline{\boldsymbol{R}}\|_2 = \|(\boldsymbol{I}+\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}})^{-1}\|_2 = 1+\eta\,, \tag{8.9}$$

then using (5.6) we get from (8.7)

$$\|(\overline{\boldsymbol{R}}^T(\boldsymbol{I}+\boldsymbol{U}^T)\tilde{\boldsymbol{R}})^{-1}\boldsymbol{E}_1{}^T\|_2 \leqq \frac{1+\eta}{1-\beta}\,1.5\cdot(n-1)\cdot 2^{-t}\|\boldsymbol{R}^{-1}\|_2^2\|\boldsymbol{A}\|_E\,, \tag{8.10}$$

and from (5.5) and (8.8)

$$\|\tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{A}-\tilde{\boldsymbol{R}})\|_2 \leqq (1+\eta)(\|\boldsymbol{E}_3\|_E\|\overline{\boldsymbol{R}}^{-1}\|_2+\|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2)\,. \tag{8.11}$$

Wilkinson shows in [6] pp. 99–104 that if double precision accumulation is used, then

$$\|\delta\overline{\boldsymbol{R}}'\|_E \leqq (1.001\cdot 2^{-t}+\tfrac{3}{2}(n+1)2^{-2t_2})\|\overline{\boldsymbol{R}}'\|_E \leqq 2^{-t_1}\|\overline{\boldsymbol{R}}'\|_E\,. \tag{8.12}$$

Since $\|\overline{\boldsymbol{R}}'\|_E \leqq n^{\frac{1}{2}}\|\overline{\boldsymbol{R}}'\|_2$ we have

$$\|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2 = \|(\overline{\boldsymbol{R}}')^{-1}\delta\overline{\boldsymbol{R}}'\|_2 \leqq n^{\frac{1}{2}}\cdot 2^{-t_1}\varkappa(\overline{\boldsymbol{R}}')\,. \tag{8.13}$$

But (8.12) is also valid without primes and as

$$\|\overline{\boldsymbol{R}}\|_2 \leqq (1+\beta)^{\frac{1}{2}}\|\boldsymbol{R}\|_2$$

when $\beta<1$, it follows

$$\|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2 \leqq n^{\frac{1}{2}}2^{-t_1}\varkappa(\overline{\boldsymbol{R}}) \leqq (1+\beta)^{\frac{1}{2}}n^{\frac{1}{2}}2^{-t_1}\|\overline{\boldsymbol{R}}^{-1}\|_2\|\boldsymbol{A}\|_E\,. \tag{8.14}$$

Substituting in (8.11) from (5.8) and (8.14) we can show that

$$\|\tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{Q}}^T\boldsymbol{A}-\tilde{\boldsymbol{R}})\|_2 \leqq (1+\eta)\cdot 1.9n^{\frac{1}{2}}(n+1)\cdot 2^{-t}\|\overline{\boldsymbol{R}}^{-1}\|_2\|\boldsymbol{A}\|_E\,. \tag{8.15}$$

For the estimation of η (8.9) yields

$$\eta \leqq \|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2/(1-\|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2)\,. \tag{8.16}$$

From (8.14), (6.13) and (6.12) we get

$$\|\overline{\boldsymbol{R}}^{-1}\delta\boldsymbol{R}\|_2 \leqq 1.06\left(\frac{1+\beta}{1-\beta}\right)^{\frac{1}{2}}\cdot n^{\frac{1}{2}}2^{-t}\|\boldsymbol{R}^{-1}\|_2\|\boldsymbol{A}\|_E \leqq \frac{1.06\beta}{3.42(n+1)}\left(\frac{1+\beta}{1-\beta}\right),$$

and if we make the additional assumption $\beta \leqq 0.9$ then

$$\|\overline{\boldsymbol{R}}^{-1}\delta\overline{\boldsymbol{R}}\|_2 \leqq 1.22/(n+1) \leqq 0.41\,.$$

Using this to estimate the denominator in (8.16) we can show that

$$\eta \leqq 1.8n^{\frac{1}{2}}2^{-t}\varkappa(\overline{\boldsymbol{R}}')\,, \tag{8.17}$$

$$\eta \leqq \frac{0.53}{n+1}\left(\frac{1+\beta}{1-\beta}\right)^{\frac{1}{2}}\beta\,. \tag{8.18}$$

Although it is not generally true that $\varkappa(\overline{\boldsymbol{R}}') \leqq \varkappa(\boldsymbol{A})$, often $\varkappa(\overline{\boldsymbol{R}}') \ll \varkappa(\overline{\boldsymbol{R}}) \approx \varkappa(\boldsymbol{A})$. In fact for many very ill-conditioned systems $\varkappa(\overline{\boldsymbol{R}}')$ will be of order unity. Then (8.14) and consequently (8.18) will considerably overestimate the error from the backsubstitution, which nearly always is of little importance.

We now take norms in (8.4) and use (8.5), (8.6), (8.10) and (8.15) together with the estimates for $\|\boldsymbol{e}_2\|_2$ and $\|\boldsymbol{e}_3\|_2$ in (5.7) and (5.8). Eliminating $\|\overline{\boldsymbol{R}}^{-1}\|_2$ by means of the inequality

$$\|\overline{\boldsymbol{R}}^{-1}\|_2\|\boldsymbol{A}\|_2 \leqq (1-\beta)^{-\frac{1}{2}}\cdot\varkappa(\boldsymbol{A})$$

and invoking the trivial inequalities

$$n-1 \leqq 0.3n^{\frac{1}{2}}(n+1), \quad n \leqq \tfrac{2}{3}n^{\frac{1}{2}}(n+1)$$

we arrive at the result

$$\begin{pmatrix} \|\overline{\boldsymbol{r}}-\boldsymbol{r}\|_2 \\ \|\boldsymbol{A}\|_2\|\overline{\boldsymbol{x}}-\boldsymbol{x}\|_2 \end{pmatrix} \leqq \begin{pmatrix} \dfrac{0.79\varkappa(\boldsymbol{A})}{\sqrt{1-\beta}} & 0.81 \\ \dfrac{0.24\varkappa^2(\boldsymbol{A})}{1-\beta} & \dfrac{\varkappa(\boldsymbol{A})}{\sqrt{1-\beta}} \end{pmatrix} \begin{pmatrix} \|\boldsymbol{r}\|_2\dfrac{\|\boldsymbol{A}\|_E}{\|\boldsymbol{A}\|_2} \\ \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2\dfrac{\|\boldsymbol{A}\|_E}{\|\boldsymbol{A}\|_2}+\|\boldsymbol{b}\|_2 \end{pmatrix}\cdot f(n)2^{-t} \tag{8.19}$$

where

$$f(n) = (1+\eta)1.9n^{\frac{1}{2}}(n+1)\,.$$

We observe that (8.19) gives a smaller bound for the error than that obtained from (7.6) when

$$\|\delta \boldsymbol{A}\|_2 = f(n)\cdot 2^{-t}\|\boldsymbol{A}\|_E, \quad \|\delta \boldsymbol{b}\|_2 = f(n)2^{-t}\|\boldsymbol{b}\|_2 ,$$

as this according to (6.12) and (7.5) implies

$$\alpha = (1+\eta)(\sqrt{2}+1)\cdot 1.9n^{\frac{1}{2}}(n+1)\cdot 2^{-t}\|\boldsymbol{R}^{-1}\|_2\|\boldsymbol{A}\|_E > \beta .$$

As the factor $(1+\eta)$ usually is of no importance, we can say that *the deviation of the computed solution from the true solution is roughly less than the deviation resulting from some perturbation $\delta \boldsymbol{A}$, $\delta \boldsymbol{b}$ such that*

$$\|\delta \boldsymbol{A}\|_E/\|\boldsymbol{A}\|_E \approx \|\delta \boldsymbol{b}\|_2/\|\boldsymbol{b}\|_2 \approx 2\cdot n^{3/2}\cdot 2^{-t} .$$

Moreover when $\boldsymbol{A}$ is ill-conditioned we can, as remarked in the end of section 5, *expect a much better result than this.*

We finally observe that when the modified procedure is used for computing $\boldsymbol{R}$, then the classical procedure should on no account be used for calculating $\boldsymbol{y}$. If this was done we would have

$$\boldsymbol{y} \approx \bar{\boldsymbol{Q}}^T\boldsymbol{b} = (\bar{\boldsymbol{Q}}-\tilde{\boldsymbol{Q}})^T\boldsymbol{b}+\tilde{\boldsymbol{Q}}^T\boldsymbol{b}$$

and consequently

$$\|\boldsymbol{y}-\tilde{\boldsymbol{Q}}^T\boldsymbol{b}\|_2 \approx \|\bar{\boldsymbol{Q}}-\tilde{\boldsymbol{Q}}\|_2\|\boldsymbol{b}\|_2 = \|\boldsymbol{U}\tilde{\boldsymbol{Q}}\|_2\|\boldsymbol{b}\|_2 = \text{const.}\,\varkappa(\boldsymbol{A})2^{-t}\|\boldsymbol{b}\|_2 .$$

It can be verified that if $\|\boldsymbol{r}\|_2 \ll \|\boldsymbol{b}\|_2$ this means that the errors in the computed solution usually would be multiplied by $\varkappa(\boldsymbol{A})$!

9. Conclusions.

The modified Gram–Schmidt procedure has been successfully used since 1960 as part of a data processing system for optical spectra [8]. It has about the same high numerical stability as other comparable algorithms based on nearly orthogonal transformations, [7] ch. 3, and generally should be preferred to the classical Gram–Schmidt procedure. Reorthogonalization is then never necessary, and the solution is obtained in approximately mn^2 multiplications. Pivoting can easily be done, but is not required for stability. The storage requirement is approximately $mn+n^2/2$. This should be compared to $2mn^2+4n^3/3$ s.p. multiplications and n^2 storage locations needed for forming and solving the normal equations in double precision.

More accurate solutions to linear least squares problems can be obtained by using iterative refinement. This was first proposed by Golub [2] and used also in [1]. Different schemes for the refinement are discussed

in [3]. The results obtained so far, however, are only satisfactory when the equations are almost compatible, i.e. when

$$\varkappa^2(\boldsymbol{A})\|\boldsymbol{r}\|_2 < \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2 .$$

In a forthcoming paper, where also numerical examples will be provided, the author will apply iterative refinement to linear least squares problems by calculating the residuals to the system (7.1) in double precision, and use the inverse

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix}^{-1} \approx \begin{pmatrix} \boldsymbol{I}-\bar{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^T & \tilde{\boldsymbol{Q}}\bar{\boldsymbol{R}}^{-T} \\ \bar{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{Q}}^T & -\bar{\boldsymbol{R}}^{-1}(\boldsymbol{I}+\boldsymbol{U})^{-T}\bar{\boldsymbol{R}}^{-T} \end{pmatrix}$$

to solve for the increments in $\boldsymbol{r}$ and $\boldsymbol{x}$. This inverse results from a slight generalization of the modified Gram–Schmidt procedure. It will be shown that the iterates generally will converge to the exact solution correctly rounded even when $\varkappa^2(\boldsymbol{A})\|\boldsymbol{r}\|_2$ is of the same order of magnitude as $2^t\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$.

REFERENCES

1. Bauer, F. L., *Elimination with Weighted Row Combinations for Solving Linear Equations and Least Squares Problems*, Num. Math. 7 (1965), 338–352.
2. Golub, G. H., *Numerical Methods for Solving Linear Least Squares Problems*, Num. Math. 7 (1965), 206–216.
3. Golub, G. H. and Wilkinson, J. H., *Note on the Iterative Refinement of Least Squares Solution*, Num. Math. 9 (1966), 139–148.
4. Householder, A. S., *The Theory of Matrices in Numerical Analysis*, New York: Blaisdell 1964.
5. Rice, J. R., *Experiments on Gram–Schmidt Orthogonalization*, Math. Comp. 20 (1966), 325–328.
6. Wilkinson, J. H., *Rounding Errors in Algebraic Processes*, London: H.M.S.O. 1963.
7. Wilkinson, J. H.: *The Algebraic Eigenvalue Problem*, Oxford: Clarendon Press 1965.
8. Åslund, N., *A Data Processing System for Spectra of Diatomic Molecules*, Arkiv Fysik 30 (1965), 377–396.

INSTITUTE FOR INFORMATION PROCESSING
ROYAL INSTITUTE OF TECHNOLOGY
STOCKHOLM, SWEDEN