BIT 7 (1967), 257–278

ITERATIVE REFINEMENT OF LINEAR LEAST SQUARES SOLUTIONS I

ÅKE BJÖRCK

Abstract.

An iterative procedure is developed for reducing the rounding errors in the computed least squares solution to an overdetermined system of equations $\boldsymbol{Ax}=\boldsymbol{b}$, where $\boldsymbol{A}$ is an $m \times n$ matrix ($m \geqq n$) of rank n. The method relies on computing accurate residuals to a certain augmented system of linear equations, by using double precision accumulation of inner products. To determine the corrections, two methods are given, based on a matrix decomposition of $\boldsymbol{A}$ obtained either by orthogonal Householder transformations or by a modified Gram–Schmidt orthogonalization. It is shown that the rate of convergence in the iteration is independent of the right hand side, $\boldsymbol{b}$, and depends linearly on the condition number, $\varkappa(\boldsymbol{A})$, of the rectangular matrix $\boldsymbol{A}$. The limiting accuracy achieved will be approximately the same as that obtained by a double precision factorization.

In a second part of this paper the case when $\boldsymbol{x}$ is subject to linear constraints and/or $\boldsymbol{A}$ has rank less than n is covered. Here also ALGOL-programs embodying the derived algorithms will be given.

1. Introduction.

Let $\boldsymbol{Ax}=\boldsymbol{b}$ be a given overdetermined system of linear equations where $\boldsymbol{A}$ is an $m \times n$ matrix ($m \geqq n$) and $\boldsymbol{b}$ is a vector. A vector $\boldsymbol{x}$ which minimizes $\|\boldsymbol{b}-\boldsymbol{Ax}\|_2$ is called a least squares solution to the system.

Least squares problems are often ill-conditioned. Rounding errors may then seriously contaminate the solution. For the linear equation case ($m=n$) Wilkinson [8] has proposed the following process of iterative refinement for reducing the rounding errors: Compute the sequence of vectors $\boldsymbol{x}^{(s)}$, $s=0,1,2,\ldots$ defined by

$$\begin{aligned} \boldsymbol{x}^{(0)} &= 0,\ \boldsymbol{r}^{(s)} = \boldsymbol{b}-\boldsymbol{Ax}^{(s)}\,, \\ \delta\boldsymbol{x}^{(s)} &= \boldsymbol{A}^{-1}\boldsymbol{r}^{(s)},\ \boldsymbol{x}^{(s+1)} = \boldsymbol{x}^{(s)}+\delta\boldsymbol{x}^{(s)}\,. \end{aligned} \tag{1.1}$$

Here the residual vector $\boldsymbol{r}^{(s)}$ is computed using double precision accumulation of inner products. Single precision is used in all other steps. In particular, the corrections $\delta\boldsymbol{x}^{(s)}$ are computed using a suitable single precision factorization of $\boldsymbol{A}$.

This work was sponsored by the Swedish Natural Science Research Council.

The performance of this process in floating point arithmetic has been analysed by Moler [7]. It has also been embodied in an ALGOL procedure by Martin, Peters and Wilkinson [6] for the special case when $\boldsymbol{A}$ is positive definite. Now assume that $m > n$ and that $\boldsymbol{A}$ has rank n. Then there exists a decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{R}$ is upper triangular and $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$, and it is well known that the least squares solution is given by

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

Thus it is natural to use the refinement procedure (1.1) with

$$\delta\boldsymbol{x}^{(s)} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{r}^{(s)} \tag{1.2}$$

for refining least squares solutions. This was first pointed out by Golub [4] and used also by Bauer in [1]. However it has been shown by Golub and Wilkinson in [5] that this process works satisfactorily only when the overdetermined system is nearly compatible.

In part I of this paper a procedure for the iterative refinement of least squares solutions without this restriction will be developed and analysed. In chapter 2 we describe the procedure as a special case of (1.1) and show, why the analysis in [7] is too general to be of any use here. In chapter 3 we formulate the assumptions on the arithmetic underlying our analysis. Any method for solving least squares problems can, after modification, be used in our procedure to solve for the corrections. In chapter 4 we analyse the errors which are independent of the particular method chosen. In recent papers it has been pointed out that methods related to an orthogonal triangularisation of the matrix $\boldsymbol{A}$ either by Householder transformations [4] or by a modified Gram–Schmidt procedure [2], have several advantages over the classical method of solving the normal equations. In chapter 5 an algorithm is derived for the Householder method and a detailed error analysis for a single step is carried out. In chapter 6 these results are used to derive estimates for the rate of convergence and the limiting accuracy. Finally in chapter 7 the corresponding algorithm for the Gram–Schmidt method is analysed.

In part II the case when $\boldsymbol{x}$ is subject to linear constraints and/or $\boldsymbol{A}$ has rank less than n is covered. Here ALGOL-programs embodying the derived algorithms will also be given.

2. The refinement procedure.

It is well known that a least squares solution is characterized by the property that the residual vector $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ is orthogonal to the columns of $\boldsymbol{A}$. Thus for the unknowns $\boldsymbol{r}$ and $\boldsymbol{x}$ we have the system of $(m+n)$ equations

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} \tag{2.1}$$

If we assume that the rank of A equals n, then this system is non-singular and determines $\boldsymbol{r}$ and $\boldsymbol{x}$ uniquely.

We now propose to use the iterative procedure (1.1) for the refinement of the solution to the *augmented system* (2.1). As initial approximation we take

$$\boldsymbol{r}^{(0)} = \boldsymbol{0}, \quad \boldsymbol{x}^{(0)} = \boldsymbol{0} .$$

The sth iteration consists of the three steps:

(i) Compute the residuals

$$\begin{pmatrix} \boldsymbol{f}^{(s)} \\ \boldsymbol{g}^{(s)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{r}^{(s)} \\ \boldsymbol{x}^{(s)} \end{pmatrix} \tag{2.2}$$

Here inner products are accumulated in double precision.

(ii) Solve for the corrections $\delta\boldsymbol{r}^{(s)}$ and $\delta\boldsymbol{x}^{(s)}$ from

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \delta\boldsymbol{r}^{(s)} \\ \delta\boldsymbol{x}^{(s)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}^{(s)} \\ \boldsymbol{g}^{(s)} \end{pmatrix} \tag{2.3}$$

Note that when $s \neq 0$ we generally have $\boldsymbol{g}^{(s)} \neq 0$. Thus, modifications to the usual methods for solving linear least squares problems are necessary in order to solve (2.3).

(iii) Add the corrections

$$\begin{pmatrix} \boldsymbol{r}^{(s+1)} \\ \boldsymbol{x}^{(s+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}^{(s)} \\ \boldsymbol{x}^{(s)} \end{pmatrix} + \begin{pmatrix} \delta\boldsymbol{r}^{(s)} \\ \delta\boldsymbol{x}^{(s)} \end{pmatrix} \tag{2.4}$$

If we put $\boldsymbol{r}^{(s)} = \boldsymbol{0}$ for all s, then this procedure degenerates into

$$\boldsymbol{f}^{(s)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(s)}, \quad \delta\boldsymbol{x}^{(s)} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{f}^{(s)} ,$$

which is precisely the scheme (1.2) proposed by Golub. This indicates that when the overdetermined system is compatible, the final performance of the two schemes should be the same.

Since the proposed procedure (2.2)–(2.4) is a special case of the general scheme (1.1), the analysis in [7] applies. Because of the special structure of the matrix

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix} \tag{2.5}$$

this analysis, however, does not give a true assessment of the performance. According to this general analysis the condition number $\varkappa(\boldsymbol{B})$ should play an essential role. We now make the following observation. If we scale the matrix $\boldsymbol{A}$ so that

$$\boldsymbol{A} := \alpha^{-1}\boldsymbol{A}, \quad \alpha = 2^{-q} \ (q \text{ integer})$$

then the fractional parts of the floating point numbers in our algorithm remain the same. The iterations are now, however, associated with the matrix

$$\boldsymbol{B}_\alpha = \begin{pmatrix} \alpha \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix}.$$

We will show that the condition number of $\boldsymbol{B}_\alpha$ *varies considerably with* α. Let λ be an eigenvalue of $\boldsymbol{B}_\alpha$ and $(\boldsymbol{x},\boldsymbol{y})^T$ the corresponding eigenvector. Then

$$\begin{aligned} \alpha\boldsymbol{x} + \boldsymbol{A}\boldsymbol{y} &= \lambda\boldsymbol{x} \\ \boldsymbol{A}^T\boldsymbol{x} \qquad &= \lambda\boldsymbol{y} \end{aligned}$$

and it follows that

$$\alpha\lambda\boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{y} = \lambda^2\boldsymbol{y}\,.$$

Now $\boldsymbol{y} \neq 0$ implies that $\boldsymbol{y}$ is an eigenvector and $(\lambda^2 - \alpha\lambda)$ is an eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$. On the other hand $\boldsymbol{y} = 0$ implies that

$$\boldsymbol{A}^T\boldsymbol{x} = 0, \quad \alpha\boldsymbol{x} = \lambda\boldsymbol{x}, \quad \boldsymbol{x} \neq 0\,.$$

Thus the eigenvalues of $\boldsymbol{B}_\alpha$ are

$$\lambda(\boldsymbol{B}_\alpha) = \begin{cases} \dfrac{\alpha}{2} \pm \left(\dfrac{\alpha^2}{4} + \sigma_i^2\right)^{\frac{1}{2}} \\ \alpha \end{cases}$$

where σ_i^2, $i = 1, 2, \ldots, n$ are the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ and the eigenvalue α has multiplicity $(m-n)$. From this it can be deduced that

$$\min_\alpha \varkappa(\boldsymbol{B}_\alpha) = \tfrac{1}{2} + \left(\tfrac{1}{4} + 2\,\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\right)^{\frac{1}{2}} \leqq 2\varkappa(\boldsymbol{A})\,, \tag{2.6}$$

where the minimum is attained for $\alpha = 2^{-\frac{1}{2}}\sigma_{\min}$. Here $\varkappa(\boldsymbol{A})$ is the Euclidian condition number for the rectangular matrix $\boldsymbol{A}$,

$$\varkappa(\boldsymbol{A}) = \sigma_{\max}/\sigma_{\min} \geqq 1\,.$$

Furthermore, if we take $\alpha = 2^{-\frac{1}{2}}\sigma_{\max}$, then

$$\varkappa(\boldsymbol{B}_\alpha) > \varkappa^2(\boldsymbol{A}) \left(\frac{1}{2\sqrt{2}} + \frac{3}{2\sqrt{2}}\right)\left(\frac{1}{2\sqrt{2}} + \frac{1}{2\sqrt{2}}\right) = \varkappa^2(\boldsymbol{A})\,. \tag{2.7}$$

This shows that a special analysis is needed.

3. Preliminaries for the error analysis.

We assume in the following that normalized floating point arithmetic with a single (double) precision mantissa of t_1 (t_2) digits of base β is used. By accumulating an inner product in double precision we mean that multiplications which produce a t_2-digit product from t_1-digit numbers and double precision addition is used, the result finally being rounded to single precision.

More precisely, we assume that if x and y are single precision numbers and z a double precision number then

$$\begin{aligned} \mathrm{fl}(x'op'y) &= (x'op'y)(1+\delta_1), \quad 'op' = +,-,\times,/\,, \\ \mathrm{fl}_2(x\times y+z) &= (x\times y+z)(1+\delta_2) \end{aligned} \tag{3.1}$$

where

$$|\delta_i| \leqq \varepsilon_i, \quad \varepsilon_i = \beta^{1-t_i}, \quad i = 1,2\,. \tag{3.2}$$

If the single precision machine operations are rounded rather than chopped, then ε_1 can usually be halved. For a detailed discussion, see Wilkinson [8]. We furthermore assume that all quantities remain within the permitted range of the computer.

It has been pointed out that the accumulation of inner products is essential only in the calculation of the residuals. For convenience we will, however, assume that this is done also when computing the decomposition of A, and when solving the systems (2.3). This is not an essential restriction. If single precision inner products are used in these steps, most derived error bounds will only increase by a factor less than m.

4. Rounding errors in the residuals.

We consider here the rounding errors introduced in the calculation of the residuals (2.2). Here and in the following we distinguish computed quantities by using a bar. To make for easier accumulation of errors we assume that $m\varepsilon_2 \leqq 0.1$ and define

$$\varepsilon_2' = 1.06(1+\varepsilon_1)\varepsilon_2\,. \tag{4.1}$$

Following Moler [7] (p. 318) we have for the computed ith component of $\boldsymbol{f}^{(s)} = \boldsymbol{b} - \bar{\boldsymbol{r}}^{(s)} - \boldsymbol{A}\bar{\boldsymbol{x}}^{(s)}$

$$\bar{f}_i^{(s)} = (1+\delta)[-\bar{r}_i^{(s)}(1+\eta_1)+\big(b_i(1+\eta_2)-t(1+\eta_3)\big)(1+\eta_4)]$$

$$t = \sum_{j=1}^{n} a_{ij}\bar{x}_j^{(s)}(1+\gamma_j)\,,$$

where

$$|\delta| \leqq \varepsilon_1, \ |\eta_i| \leqq \varepsilon_2, \ i = 1,2,3,4, \ |\gamma_j| \leqq 1.06(n-j+2)\varepsilon_2\,.$$

If we denote the errors in the computed residuals by

$$\delta \boldsymbol{f}^{(s)} = \bar{\boldsymbol{f}}^{(s)} - \boldsymbol{f}^{(s)} ,$$

then combining these results

$$\delta f_i^{(s)} = (b_i - \bar{r}_i^{(s)} - \sum_{j=1}^{n} a_{ij}\bar{x}_j^{(s)})\delta + (1+\delta)$$
$$[-\bar{r}_i^{(s)}\eta_1 + b_i((1+\eta_2)(1+\eta_4)-1) - \sum_{j=1}^{n} a_{ij}\bar{x}_j^{(s)}((1+\gamma_j)(1+\eta_3)(1+\eta_4)-1)] ,$$

and thus

$$\|\delta \boldsymbol{f}^{(s)}\|_2 \leqq \varepsilon_1 \|\boldsymbol{b} - \bar{\boldsymbol{r}}^{(s)} - \boldsymbol{A}\bar{\boldsymbol{x}}^{(s)}\|_2 +$$
$$(1+\varepsilon_1)\varepsilon_2(\|\bar{\boldsymbol{r}}^{(s)}\|_2 + 2\cdot 1.06\|\boldsymbol{b}\|_2 + 1.06(n+3)\|\boldsymbol{A}\|_2\|\bar{\boldsymbol{x}}^{(s)}\|_2) .$$

Using $\boldsymbol{b} = \boldsymbol{r} + \boldsymbol{A}\boldsymbol{x}$ and the estimate

$$\|\boldsymbol{b}\|_2 \leqq \|\boldsymbol{r}\|_2 + \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$$

we obtain

$$\begin{aligned}\|\delta \boldsymbol{f}^{(s)}\|_2 \leqq\ & 3\varepsilon_2'\|\boldsymbol{r}\|_2 + (n+5)\varepsilon_2'\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2 + \\ & (\varepsilon_1 + \varepsilon_2')\|\boldsymbol{r} - \bar{\boldsymbol{r}}^{(s)}\|_2 + (\varepsilon_1 + (n+3)\varepsilon_2')\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \bar{\boldsymbol{x}}^{(s)}\|_2 .\end{aligned} \tag{4.2}$$

In the same way we can bound the rounding error in the computed residual vector $\boldsymbol{g}^{(s)} = -\boldsymbol{A}^T\bar{\boldsymbol{r}}^{(s)}$. We obtain

$$\delta g_i^{(s)} = -\delta \sum_{i=1}^{m} a_{ij}\bar{r}_i^{(s)} - (1+\delta)\sum_{i=1}^{m} a_{ij}\bar{r}_i^{(s)}\gamma_i ,$$

where

$$|\gamma_i| \leqq 1.06(m-i+2)\varepsilon_2, \quad |\delta| \leqq \varepsilon_1 .$$

From this it follows that

$$\|\delta \boldsymbol{g}^{(s)}\|_2 \leqq \|\boldsymbol{A}\|_2[(m+1)\varepsilon_2'\|\boldsymbol{r}\|_2 + (\varepsilon_1 + (m+1)\varepsilon_2')\|\boldsymbol{r} - \bar{\boldsymbol{r}}^{(s)}\|_2] . \tag{4.3}$$

5. Error analysis of Householder's method.

In [4] Golub has described a method for solving linear least squares problems using a sequence of elementary orthogonal transformations of Householder type:

$$\boldsymbol{P}^{(r)} = \boldsymbol{I} - 2\boldsymbol{w}^{(r)}\boldsymbol{w}^{(r)T} .$$

Here $\boldsymbol{w}^{(r)} = (0,\ldots,0,w_{r+1}^{(r)},\ldots,w_m^{(r)})^T$, $r = 0,1,\ldots,n-1$ is chosen so that

$$\boldsymbol{P}^{(n-1)}\ldots\boldsymbol{P}^{(1)}\boldsymbol{P}^{(0)}\boldsymbol{A} = \begin{pmatrix}\boldsymbol{U}\\ \hline \boldsymbol{0}\end{pmatrix}\begin{matrix}\}n\\ \}m-n\end{matrix} \tag{5.1}$$

where $\boldsymbol{U}$ is an upper triangular matrix.

We now show how to adopt this method to solve the system (2.3) which we write, for the moment,

$$\begin{aligned} \boldsymbol{r}+\boldsymbol{A}\boldsymbol{x} &= \boldsymbol{f} \\ \boldsymbol{A}^T\boldsymbol{r} \quad &= \boldsymbol{g} \end{aligned} \tag{5.2}$$

Multiplying the first set of these equations from the left with $\boldsymbol{Q}=\boldsymbol{P}^{(n-1)}\ldots\boldsymbol{P}^{(1)}\boldsymbol{P}^{(0)}$ and using (5.1) we get

$$\begin{aligned} \boldsymbol{Q}\boldsymbol{r}+\left(\frac{\boldsymbol{U}}{0}\right)\boldsymbol{x} &= \boldsymbol{Q}\boldsymbol{f} \\ (\boldsymbol{U}^T|0)\boldsymbol{Q}\boldsymbol{r} &= \boldsymbol{g} \end{aligned} \tag{5.3}$$

Thus, it is easily seen that $\boldsymbol{r}$ and $\boldsymbol{x}$ can be computed by the following algorithm:

$$\begin{aligned} \boldsymbol{h} &= \boldsymbol{U}^{-T}\boldsymbol{g}, \quad \boldsymbol{d} = \boldsymbol{Q}\boldsymbol{f} = \begin{pmatrix} \boldsymbol{d}_1 \\ \hline \boldsymbol{d}_2 \end{pmatrix}\begin{matrix} \}n \\ \}m-n \end{matrix} \\ \boldsymbol{r} &= \boldsymbol{Q}^T\left(\frac{\boldsymbol{h}}{\boldsymbol{d}_2}\right), \quad \boldsymbol{x} = \boldsymbol{U}^{-1}(\boldsymbol{d}_1-\boldsymbol{h}) \end{aligned} \tag{5.4}$$

Wilkinson [9] has, under the assumption that inner products are accumulated with $t_2=2t_1$ digits, given an error analysis of orthogonal transformations of the type used here. We state below in (5.5)–(5.8) those of his results needed here.

Consider the computed sequence of transformed matrices

$$\boldsymbol{A} = \bar{\boldsymbol{A}}^{(0)}, \ \bar{\boldsymbol{A}}^{(r+1)} = \mathrm{fl}_2(\bar{\boldsymbol{P}}^{(r)}\bar{\boldsymbol{A}}^{(r)}), \quad r = 0,1,\ldots,n-1 .$$

For a certain prescribed method of computation there exists a sequence of elementary orthogonal transformations $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\ldots,\boldsymbol{P}^{(n-1)}$ (*not* the matrices corresponding to exact computation throughout) such that

$$\boldsymbol{P}^{(n-1)}\ldots\boldsymbol{P}^{(1)}\boldsymbol{P}^{(0)}(\boldsymbol{A}+\boldsymbol{E}) = \bar{\boldsymbol{A}}^{(n)} = \left(\frac{\bar{\boldsymbol{U}}}{0}\right) \tag{5.5}$$

where[1]

$$\|\boldsymbol{E}\|_E \leqq n\beta(1+\beta)^{n-1}\|\boldsymbol{A}\|_E, \quad \beta = 12.36\,\varepsilon_1 \tag{5.6}$$

and $\bar{\boldsymbol{U}}$ is the computed upper triangular matrix.

If the computed transformation $\bar{\boldsymbol{Q}}=\bar{\boldsymbol{P}}^{(n-1)}\ldots\bar{\boldsymbol{P}}^{(1)}\bar{\boldsymbol{P}}^{(0)}$ is applied to a vector $\boldsymbol{b}$,

$$\boldsymbol{b} = \bar{\boldsymbol{b}}^{(0)}, \ \bar{\boldsymbol{b}}^{(r+1)} = \mathrm{fl}_2(\bar{\boldsymbol{P}}^{(r)}\bar{\boldsymbol{b}}^{(r)}), \quad r = 0,1,\ldots,n-1 .$$

then

$$\bar{\boldsymbol{b}}^{(r+1)} = \boldsymbol{P}^{(r)}\bar{\boldsymbol{b}}^{(r)}+\boldsymbol{f}^{(r)}, \ \|\boldsymbol{f}^{(r)}\|_2 \leqq \beta\|\bar{\boldsymbol{b}}^{(r)}\|_2 . \tag{5.7}$$

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

From (5.7) it follows by induction that

$$\|\bar{\boldsymbol{b}}^{(r)}\|_2 \leqq (1+\beta)\|\bar{\boldsymbol{b}}^{(r-1)}\|_2 \leqq (1+\beta)^r\|\boldsymbol{b}\|_2$$

and

$$\boldsymbol{d}^{(r+1)} = \boldsymbol{P}^{(r)}\boldsymbol{d}^{(r)}+\boldsymbol{f}^{(r)}, \quad \boldsymbol{d}^{(r)} = \bar{\boldsymbol{b}}^{(r)}-\boldsymbol{P}^{(r-1)}\ldots\boldsymbol{P}^{(0)}\boldsymbol{b}\,.$$

Since $\boldsymbol{d}^{(0)}=\boldsymbol{0}$ and $\boldsymbol{d}^{(n)}=\bar{\boldsymbol{b}}^{(n)}-\boldsymbol{Q}\boldsymbol{b}$, it follows that

$$\|\bar{\boldsymbol{b}}^{(n)}-\boldsymbol{Q}\boldsymbol{b}\|_2 \leqq \sum_{r=0}^{n-1}\beta(1+\beta)^r\|\boldsymbol{b}\|_2 < n\beta(1+\beta)^{n-1}\|\boldsymbol{b}\|_2\,. \tag{5.8}$$

We now derive an only slightly different result for the error when $\bar{\boldsymbol{Q}}^T=\bar{\boldsymbol{P}}^{(0)}\bar{\boldsymbol{P}}^{(1)}\ldots\bar{\boldsymbol{P}}^{(n-1)}$ is applied to a vector $\boldsymbol{c}$;

$$\boldsymbol{c} = \bar{\boldsymbol{c}}^{(0)}, \quad \bar{\boldsymbol{c}}^{(r+1)} = \mathrm{fl}_2(\bar{\boldsymbol{P}}^{(n-r-1)}\bar{\boldsymbol{c}}^{(r)}), \qquad r = 0,1,\ldots,n-1\,.$$

In analogy to (5.7) we obviously have

$$\bar{\boldsymbol{c}}^{(r+1)} = \boldsymbol{P}^{(n-r-1)}\bar{\boldsymbol{c}}^{(r)}+\boldsymbol{g}^{(r)}, \quad \|\boldsymbol{g}^{(r)}\|_2 \leqq \beta\|\bar{\boldsymbol{c}}^{(r)}\|_2\,. \tag{5.9}$$

As $\boldsymbol{P}^{(n-r-1)}$ is orthogonal and symmetric we get

$$\bar{\boldsymbol{c}}^{(r)} = \boldsymbol{P}^{(n-r-1)}(\bar{\boldsymbol{c}}^{(r+1)}+\boldsymbol{g}^{(r)})$$

and thus

$$\|\bar{\boldsymbol{c}}^{(r)}\|_2 \leqq (1-\beta)^{-1}\|\bar{\boldsymbol{c}}^{(r+1)}\|_2 \leqq (1-\beta)^{-(n-r)}\|\bar{\boldsymbol{c}}^{(n)}\|_2\,.$$

By induction from (5.9)

$$\boldsymbol{e}^{(r+1)} = \boldsymbol{P}^{(n-r-1)}\boldsymbol{e}^{(r)}+\boldsymbol{g}^{(r)}, \quad \boldsymbol{e}^{(r)} = \bar{\boldsymbol{c}}^{(r)}-\boldsymbol{P}^{(n-r)}\ldots\boldsymbol{P}^{(n-1)}\boldsymbol{c}\,,$$

where $\boldsymbol{e}^{(0)}=\boldsymbol{0}$ and $\boldsymbol{e}^{(n)}=\bar{\boldsymbol{c}}^{(n)}-\boldsymbol{Q}^T\boldsymbol{c}$. Hence

$$\|\bar{\boldsymbol{c}}^{(n)}-\boldsymbol{Q}^T\boldsymbol{c}\|_2 \leqq \sum_{r=0}^{n-1}\beta(1-\beta)^{-(n-r)}\|\bar{\boldsymbol{c}}^{(n)}\|_2 < n\beta(1-\beta)^{-n}\|\bar{\boldsymbol{c}}^{(n)}\|_2 \tag{5.10}$$

For convenience we assume in the following that

$$12.5\cdot n\cdot\varepsilon_1 \leqq 0.01\,, \tag{5.11}$$

which, as is easily shown, implies that

$$n\beta(1+\beta)^{n-1} < n\beta(1-\beta)^{-n} \leqq 12.485\cdot n\cdot\varepsilon_1\,. \tag{5.12}$$

From (5.8), (5.10) and (5.12) it now follows that the computed quantities $\bar{\boldsymbol{d}}$ and $\bar{\boldsymbol{r}}$, in the algorithm (5.4), satisfy

$$\bar{\boldsymbol{d}} = \boldsymbol{Q}(\boldsymbol{f}+\boldsymbol{e}_1), \quad \bar{\boldsymbol{r}} = \boldsymbol{Q}^T\begin{pmatrix}\bar{\boldsymbol{h}}\\ \bar{\boldsymbol{d}}_2\end{pmatrix}+\boldsymbol{e}_2 \tag{5.13}$$

where

$$\|\boldsymbol{e}_1\|_2 \leqq 12.5n\varepsilon_1\|\boldsymbol{f}\|_2, \quad \|\boldsymbol{e}_2\|_2 \leqq 12.5n\varepsilon_1\|\bar{\boldsymbol{r}}\|_2\,. \tag{5.14}$$

Further errors are made in the two back-substitutions

$$\bar{\boldsymbol{U}}^T\boldsymbol{h} = \boldsymbol{g}, \quad \bar{\boldsymbol{U}}\boldsymbol{x} = \bar{\boldsymbol{d}}_1 - \bar{\boldsymbol{h}} .$$

Wilkinson [8], pp. 99–104, has shown that the computed quantity $\bar{\boldsymbol{h}}$ satisfies exactly

$$(\bar{\boldsymbol{U}}+\boldsymbol{F}_2)^T\bar{\boldsymbol{h}} = \boldsymbol{g} \tag{5.15}$$

where

$$\|\boldsymbol{F}_2\|_E \leqq (\varepsilon_1+n\varepsilon_2)\|\bar{\boldsymbol{U}}\|_E . \tag{5.16}$$

A similar result holds for $\bar{\boldsymbol{x}}$, which satisfies

$$(\bar{\boldsymbol{U}}+\boldsymbol{F}_1)\bar{\boldsymbol{x}} = \bar{\boldsymbol{d}}_1 - \bar{\boldsymbol{h}} , \tag{5.17}$$

where

$$\|\boldsymbol{F}_1\|_E \leqq (1-\varepsilon_1)^{-1}(2\varepsilon_1+n\varepsilon_2)\|\bar{\boldsymbol{U}}\|_E . \tag{5.18}$$

The slightly greater bound for $\boldsymbol{F}_1$ accounts for the rounding of the difference $(\bar{\boldsymbol{d}}_1-\bar{\boldsymbol{h}})$. From (5.5), (5.6) and (5.12) we have

$$\|\bar{\boldsymbol{U}}\|_E = \|\boldsymbol{A}+\boldsymbol{E}\|_E \leqq \|\boldsymbol{A}\|_E+\|\boldsymbol{E}\|_E \leqq 1.01\|\boldsymbol{A}\|_E .$$

When $t_2 \geqq 2t_1$ we certainly have $\varepsilon_2 \leqq 2\varepsilon_1{}^2$. Hence if we assume $n \geqq 2$, from (5.16) and (5.18) we have

$$\|\boldsymbol{F}_i\|_E \leqq (1-\varepsilon_1)^{-1}(1+n\varepsilon_1)1.01n\varepsilon_1\|\boldsymbol{A}\|_E, \quad i = 1,2 ,$$

or, after using (5.12) twice,

$$\|\boldsymbol{F}_i\|_E \leqq 1.012n\varepsilon_1\|\boldsymbol{A}\|_E, \quad i = 1,2 . \tag{5.19}$$

Now define

$$\boldsymbol{H}_i = \boldsymbol{E}+\boldsymbol{Q}^T\boldsymbol{F}_i, \quad i = 1,2 \tag{5.20}$$

where $\boldsymbol{E}$ is the matrix in (5.5); then the relation

$$\boldsymbol{Q}(\boldsymbol{A}+\boldsymbol{H}_i) = \left(\frac{\bar{\boldsymbol{U}}+\boldsymbol{F}_i}{\boldsymbol{0}}\right), \quad i = 1,2 \tag{5.21}$$

holds exactly. From (5.6), (5.12), (5.19) and (5.20) it follows that

$$\|\boldsymbol{H}_i\|_E \leqq 13.5n\varepsilon_1\|\boldsymbol{A}\|_E, \quad i = 1,2 . \tag{5.22}$$

We summarize the results obtained in this section:

Assume that the solutions $\boldsymbol{r}$ *and* $\boldsymbol{x}$ *of the system of equations* (5.2) *are computed by the algorithm* (5.4), *using a certain method of computation described in* [9]. *Then, provided* $12.5n\varepsilon_1 \leqq 0.01$ *and* $n \geqq 2$, *the computed solution* $\bar{\boldsymbol{r}}$ *and* $\bar{\boldsymbol{x}}$ *is the exact solution to the perturbed system*

$$\left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A}+\boldsymbol{H}_1 \\ \hline (\boldsymbol{A}+\boldsymbol{H}_2)^T & \boldsymbol{0} \end{array}\right)\left(\begin{array}{c} \bar{\boldsymbol{r}}-\boldsymbol{e}_2 \\ \hline \bar{\boldsymbol{x}} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{f}-\boldsymbol{e}_1 \\ \hline \boldsymbol{g} \end{array}\right), \tag{5.23}$$

where $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{H}_1$ *and* $\boldsymbol{H}_2$ *satisfies the bounds given in* (5.14) *and* (5.22).

Perturbed systems of type (5.23) with symmetric perturbations $\boldsymbol{H}_1 = \boldsymbol{H}_2$ have been studied by Björck in [2]. To obtain from (5.23), an estimate of the errors in $\bar{\boldsymbol{r}}$ and $\bar{\boldsymbol{x}}$ we need a slightly more general result, which we state in the following theorem:

THEOREM 1. *Let* $\bar{\boldsymbol{r}}$ *and* $\bar{\boldsymbol{x}}$ *satisfy a perturbed system of equations*

$$\left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A}+\boldsymbol{H}_1 \\ \hline (\boldsymbol{A}+\boldsymbol{H}_2)^T & \boldsymbol{0} \end{array}\right)\left(\begin{array}{c} \bar{\boldsymbol{r}} \\ \hline \bar{\boldsymbol{x}} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{f}+\boldsymbol{d}_1 \\ \hline \boldsymbol{g}+\boldsymbol{d}_2 \end{array}\right)$$

where $\boldsymbol{A}$ *is a given* $m \times n$ *matrix of rank* n, *and let* $\boldsymbol{r}$ *and* $\boldsymbol{x}$ *denote the solutions to the corresponding unperturbed system.*

Assume that there exists an orthogonal matrix $\boldsymbol{Q}$ *such that the matrices* $\boldsymbol{Q}(\boldsymbol{A}+\boldsymbol{H}_i)$ $i = 1,2$ *are upper triangular, and let the perturbations satisfy the bounds*

$$\|\boldsymbol{H}_i\|_2 \leqq \eta \|\boldsymbol{A}\|_2, \quad i = 1,2 ,$$
$$\|\boldsymbol{d}_1\|_2 \leqq \tau_2, \quad \|\boldsymbol{d}_2\|_2 \leqq \tau_1 \|\boldsymbol{A}\|_2 .$$

Then, provided

$$\alpha = (\sqrt{2}+1)\varkappa(\boldsymbol{A})\cdot\eta < 1 ,$$

the following estimate holds

$$\begin{pmatrix} \|\bar{\boldsymbol{r}}-\boldsymbol{r}\|_2 \\ \|\boldsymbol{A}\|_2 \|\bar{\boldsymbol{x}}-\boldsymbol{x}\|_2 \end{pmatrix} \leqq \begin{pmatrix} \varkappa' & 1 \\ (\varkappa')^2 & \varkappa' \end{pmatrix} \left\{ \eta \begin{pmatrix} \|\boldsymbol{r}\|_2 \\ \|\boldsymbol{A}\|_2 \|\boldsymbol{x}\|_2 \end{pmatrix} + \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} \right\} \tag{5.24}$$

where

$$\varkappa' = (1-\alpha)^{-\frac{1}{2}} \varkappa(\boldsymbol{A}) . \tag{5.25}$$

PROOF. The theorem is proved by the same technique as used in [2] pp. 15–16, only trivial changes being necessary.

6. Convergence and limiting accuracy.

Denote by $\mathscr{M}(\varkappa')$ the set of 2×2 non-negative matrices of the type

$$\boldsymbol{M} = \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (a,b) . \tag{6.1}$$

Then, obviously, $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathscr{M}(\varkappa')$ implies that $\boldsymbol{M}_1 + \boldsymbol{M}_2 \in \mathscr{M}(\varkappa')$ and $\boldsymbol{M}_1 \boldsymbol{M}_2 \in \mathscr{M}(\varkappa')$. More generally, if $\boldsymbol{M} \in \mathscr{M}(\varkappa')$ and $\boldsymbol{B}$ is an arbitrary non-negative 2×2 matrix then $\boldsymbol{M}\boldsymbol{B} \in \mathscr{M}(\varkappa')$.

We note that the matrix

$$M_0 = \begin{pmatrix} \varkappa' & 1 \\ (\varkappa')^2 & \varkappa' \end{pmatrix} = \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (\varkappa', 1)$$

in theorem 1 (5.24), belongs to $\mathscr{M}(\varkappa')$. This set of matrices will play an essential part in the analysis of the refinement procedure. We now (in lemmas 1 and 2) prove some simple properties of these matrices.

Lemma 1. *Let $M \in \mathscr{M}(\varkappa')$ be defined by* (6.1). *Then the spectral radius of M is given by*

$$\varrho(M) = a + \varkappa' b \, .$$

Proof. M has by definition rank one and thus only one eigenvalue different from zero. From the non-negativity it follows that

$$\varrho(M) = \operatorname{trace}(M) = a + \varkappa' b \, .$$

Lemma 2. *Let $M_1, M_2 \in \mathscr{M}(\varkappa')$. Then the following multiplication rule holds*

$$M_1 M_2 = \varrho(M_1) M_2 \, . \tag{6.2}$$

Proof. We have

$$M_1 M_2 = \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (a_1, b_1) \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (a_2, b_2) = (a_1 + \varkappa' b_1) \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (a_2, b_2) = \varrho(M_1) M_2 \, .$$

Corollary. $$M^n = \varrho(M)^{n-1} M, \quad n \geqq 1 \, . \tag{6.3}$$

The estimate (5.24), moreover, suggests that we define a pseudo-norm in the space R^{m+n} by

$$\|z\|_A = \begin{pmatrix} \|r\|_2 \\ \|A\|_2 \|x\|_2 \end{pmatrix}, \quad z = \begin{pmatrix} r \\ - \\ x \end{pmatrix} \begin{matrix} \}m \\ \\ \}n \end{matrix} \in R^{m+n} \, . \tag{6.4}$$

Note that with the pseudo-distance in R^{m+n}

$$d_A(z_1, z_2) = \begin{pmatrix} \|r_1 - r_2\|_2 \\ \|A\|_2 \|x_1 - x_2\|_2 \end{pmatrix}, \quad z_1, z_2 \in R^{m+n} \, ,$$

R^{m+n} becomes a pseudo-metric space cf. Collatz [3] (p. 40).

With the notation introduced in (6.1) and (6.4) we can write (5.24) in the simple form

$$\|\bar{z} - z\|_A \leqq M_0 (\eta \|z\|_A + \tau) \, . \tag{6.5}$$

We will now analyse the iterative refinement procedure defined by (2.2)–(2.4) assuming that the Householder method is used for the calculation of the corrections. Let $f^{(s)}$, $g^{(s)}$ and $\delta r^{(s)}$, $\delta x^{(s)}$ denote the exact residuals and corrections corresponding to the computed approximations $\bar{\boldsymbol{r}}^{(s)}$, $\bar{\boldsymbol{x}}^{(s)}$. Then we have

$$\boldsymbol{r} = \bar{\boldsymbol{r}}^{(s)} + \delta \boldsymbol{r}^{(s)}, \quad \boldsymbol{x} = \bar{\boldsymbol{x}}^{(s)} + \delta \boldsymbol{x}^{(s)} .$$

From (5.23) it follows that the computed corrections satisfy

$$\left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A}+\boldsymbol{H}_1^{(s)} \\ \hline (\boldsymbol{A}+\boldsymbol{H}_2^{(s)})^T & \boldsymbol{0} \end{array}\right) \begin{pmatrix} \delta\bar{\boldsymbol{r}}^{(s)} - \boldsymbol{e}_2^{(s)} \\ \delta\bar{\boldsymbol{x}}^{(s)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}^{(s)} + \boldsymbol{e}_1^{(s)} \\ \boldsymbol{g}^{(s)} \end{pmatrix} + \begin{pmatrix} \delta \boldsymbol{f}^{(s)} \\ \delta \boldsymbol{g}^{(s)} \end{pmatrix} \tag{6.6}$$

where

$$\|\boldsymbol{e}_1^{(s)}\|_2 \leqq 12.5 n \varepsilon_1 \|\bar{\boldsymbol{f}}^{(s)}\|_2, \quad \|\boldsymbol{e}_2^{(s)}\|_2 \leqq 12.5 n \varepsilon_1 \|\delta\bar{\boldsymbol{r}}^{(s)}\|_2 .$$

Let $\delta\tilde{\boldsymbol{r}}^{(s+1)}$ and $\delta\tilde{\boldsymbol{x}}^{(s+1)}$ denote the errors in the exact sums

$$\tilde{\boldsymbol{r}}^{(s+1)} = \bar{\boldsymbol{r}}^{(s)} + \delta\bar{\boldsymbol{r}}^{(s)}, \quad \tilde{\boldsymbol{x}}^{(s+1)} = \bar{\boldsymbol{x}}^{(s)} + \delta\bar{\boldsymbol{x}}^{(s)} .$$

Then it follows that

$$\delta\tilde{\boldsymbol{r}}^{(s+1)} = \delta\bar{\boldsymbol{r}}^{(s)} - \delta\boldsymbol{r}^{(s)}, \quad \delta\tilde{\boldsymbol{x}}^{(s+1)} = \delta\bar{\boldsymbol{x}}^{(s)} - \delta\boldsymbol{x}^{(s)} . \tag{6.7}$$

Using (6.7) and the identity $\boldsymbol{f}^{(s)} = \delta\boldsymbol{r}^{(s)} + \boldsymbol{A}\delta\boldsymbol{x}^{(s)}$ we have

$$\|\boldsymbol{e}_1^{(s)}\|_2 \leqq 12.5 n\varepsilon_1 (\|\delta\boldsymbol{r}^{(s)}\|_2 + \|\boldsymbol{A}\|_2 \|\delta\boldsymbol{x}^{(s)}\|_2 + \|\delta\boldsymbol{f}^{(s)}\|_2) ,$$
$$\|\boldsymbol{e}_2^{(s)}\|_2 \leqq 12.5 n\varepsilon_1 (\|\delta\boldsymbol{r}^{(s)}\|_2 + \|\delta\tilde{\boldsymbol{r}}^{(s+1)}\|_2) .$$

From (5.22) and (the often very weak) inequality $\|\boldsymbol{A}\|_E \leqq n^{\frac{1}{2}} \|\boldsymbol{A}\|_2$ it follows that we can apply theorem 1 with

$$\alpha = (\sqrt{2}+1) 13.5 n^{3/2} \varepsilon_1 \varkappa(\boldsymbol{A}) . \tag{6.8}$$

Assuming $n \geqq 2$ it certainly follows that

$$12.5 n < \tfrac{2}{3} 13.5 n^{3/2} .$$

Taking (5.11) into account we obtain

$$\begin{pmatrix} 0.99\|\delta\tilde{\boldsymbol{r}}^{(s+1)}\|_2 \\ \|\boldsymbol{A}\|_2 \|\delta\tilde{\boldsymbol{x}}^{(s+1)}\|_2 \end{pmatrix} \leqq 13.5 n^{3/2} \varepsilon_1 \left\{ \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} [(\varkappa' \;\; 1) + \tfrac{2}{3}(1 \;\; 1)] + \right.$$
$$\left. \tfrac{2}{3} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right\} \begin{pmatrix} \|\delta\boldsymbol{r}^{(s)}\|_2 \\ \|\boldsymbol{A}\|_2 \|\delta\boldsymbol{x}^{(s)}\|_2 \end{pmatrix} + \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (\varkappa', 1) \begin{pmatrix} 1.01\|\delta\boldsymbol{f}^{(s)}\|_2 \\ \|\delta\boldsymbol{g}^{(s)}\|_2 / \|\boldsymbol{A}\|_2 \end{pmatrix} .$$

From this follows, using a more compact notation,

$$0.99\|\delta\tilde{\boldsymbol{z}}^{(s+1)}\|_A \leqq 13.5 n^{3/2} \varepsilon_1 \boldsymbol{M}' \|\delta\boldsymbol{z}^{(s)}\|_A + 1.01 \boldsymbol{M}_0 \boldsymbol{\tau} \tag{6.9}$$

where

$$\boldsymbol{M}' = \begin{pmatrix} \varkappa' + \frac{4}{3} & \frac{5}{3} \\ \varkappa'(\varkappa' + \frac{2}{3}) & \varkappa' \frac{5}{3} \end{pmatrix} < \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} ((\varkappa' + \tfrac{4}{3}), \tfrac{5}{3}) \tag{6.10}$$

and from (4.2) and (4.3) we have

$$\boldsymbol{r} \leqq \left\{\varepsilon_1 \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix} + \varepsilon_2' \begin{pmatrix} m+1 & 0\\ 1 & n+3\end{pmatrix}\right\} \|\delta \boldsymbol{z}^{(s)}\|_A + \varepsilon_2' \begin{pmatrix} m+1 & 0\\ 3 & n+5\end{pmatrix} \|\boldsymbol{z}\|_A \,. \tag{6.11}$$

Assuming $n \geqq 2$ we have

$$\begin{aligned} \boldsymbol{M}_0 \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix} &= \begin{pmatrix}1\\ \varkappa'\end{pmatrix} ((\varkappa'+1),1) < \boldsymbol{M}' \,,\\ \boldsymbol{M}_0 \begin{pmatrix} m+1 & 0\\ 1 & n+3\end{pmatrix} &= \begin{pmatrix}1\\ \varkappa'\end{pmatrix} \left(\left(\varkappa' + \frac{1}{m+1}\right), \frac{n+3}{m+1}\right) < \boldsymbol{M}' \end{aligned} \tag{6.12}$$

Now consider the errors made when adding the corrections. For the error in the ith component of $\bar{\boldsymbol{r}}^{(s+1)}$

$$\bar r_i^{(s+1)} - r_i = (1+\delta)\tilde r_i^{(s+1)} - r_i = (1+\delta)(\tilde r_i^{(s+1)} - r_i) + r_i\delta \,,$$

where $|\delta| \leqq \varepsilon_1$. Hence

$$\|\bar{\boldsymbol{r}}^{(s+1)} - \boldsymbol{r}\|_2 \leqq (1+\varepsilon_1)\|\tilde{\boldsymbol{r}}^{(s+1)} - \boldsymbol{r}\|_2 + \varepsilon_1\|\boldsymbol{r}\|_2 \,.$$

Using a similar result for $\boldsymbol{x}^{(s+1)}$ we get

$$\|\delta \boldsymbol{z}^{(s+1)}\|_A \leqq (1+\varepsilon_1)\|\delta \tilde{\boldsymbol{z}}^{(s+1)}\|_A + \varepsilon_1\|\boldsymbol{z}\|_A \,. \tag{6.13}$$

We now summarize the results obtained in (6.9)–(6.13):
Define the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ by

$$\boldsymbol{C} = c\boldsymbol{M}_1, \quad \boldsymbol{D} = d\boldsymbol{M}_2 \tag{6.14}$$

where

$$\begin{aligned} c &= [(13.64n^{3/2} + 1.021)\varepsilon_1 + 1.021(m+1)\varepsilon_2'](1+\varepsilon_1) \,,\\ d &= 1.021(m+1)\varepsilon_2'(1+\varepsilon_1) \,, \end{aligned} \tag{6.15}$$

and

$$\begin{aligned} \boldsymbol{M}_1 &= \begin{pmatrix}1\\ \varkappa'\end{pmatrix} ((\varkappa' + \tfrac{4}{3}), \tfrac{5}{3}) \,,\\ \boldsymbol{M}_2 &= \begin{pmatrix}1\\ \varkappa'\end{pmatrix} \left(\left(\varkappa' + \frac{3}{m+1}\right), \frac{n+5}{m+1}\right). \end{aligned} \tag{6.16}$$

Then the errors in $\tilde{\boldsymbol{z}}^{(s+1)}$ and $\bar{\boldsymbol{z}}^{(s+1)}$ satisfy the recurrence relations

$$\begin{aligned} \|\delta \tilde{\boldsymbol{z}}^{(s+1)}\|_A &\leqq \boldsymbol{C}\|\delta \boldsymbol{z}^{(s)}\|_A + \boldsymbol{D}\|\boldsymbol{z}\|_A\\ \|\delta \boldsymbol{z}^{(s+1)}\|_A &\leqq \boldsymbol{C}\|\delta \boldsymbol{z}^{(s)}\|_A + (\boldsymbol{D} + \varepsilon_1\boldsymbol{I})\|\boldsymbol{z}\|_A \,. \end{aligned} \tag{6.17}$$

Thus, by induction,

$$\begin{aligned} \|\delta \boldsymbol{z}^{(s)}\|_A &\leqq \boldsymbol{C}^s\|\boldsymbol{z}^{(0)}\|_A + (\boldsymbol{I} + \boldsymbol{C} + \ldots + \boldsymbol{C}^{s-1})(\boldsymbol{D} + \varepsilon_1\boldsymbol{I})\|\boldsymbol{z}\|_A\\ \|\delta \tilde{\boldsymbol{z}}^{(s)}\|_A &\leqq \boldsymbol{C}^s\|\boldsymbol{z}^{(0)}\|_A + [(\boldsymbol{C} + \boldsymbol{C}^2 + \ldots + \boldsymbol{C}^{s-1})(\boldsymbol{D} + \varepsilon_1\boldsymbol{I}) + \boldsymbol{D}]\|\boldsymbol{z}\|_A \,. \end{aligned} \tag{6.18}$$

Since $\boldsymbol{C}$, $\boldsymbol{D} \in \mathscr{M}(\varkappa')$, we can use the multiplication rule (6.2) from lemma 2. Hence, if we put

$$\varrho = \varrho(\boldsymbol{C}) = \tfrac{8}{3}c(\varkappa' + \tfrac{1}{2}) \tag{6.19}$$

and assume $\varrho < 1$, then, for $s \geqq 1$, $\boldsymbol{C}^s = \varrho^{s-1}\boldsymbol{C}$ and

$$\begin{aligned} (\boldsymbol{I}+\boldsymbol{C}+\ldots+\boldsymbol{C}^{s-1})\boldsymbol{D} &= (1+\varrho+\ldots+\varrho^{s-1})\boldsymbol{D} < (1-\varrho)^{-1}\boldsymbol{D} \\ (\boldsymbol{C}+\boldsymbol{C}^2+\ldots+\boldsymbol{C}^{s-1}) &= (1+\varrho+\ldots+\varrho^{s-2})\boldsymbol{C} < (1-\varrho)^{-1}\boldsymbol{C}\,. \end{aligned}$$

Substituting this into (6.18) and noting that $\boldsymbol{z}^{(0)} = \boldsymbol{z}$ we finally get

$$\begin{aligned} \|\delta\boldsymbol{z}^{(s)}\|_A &< \left(\varrho^{s-1}\boldsymbol{C} + (1-\varrho)^{-1}(\boldsymbol{D}+\varepsilon_1\boldsymbol{C}) + \varepsilon_1\boldsymbol{I}\right)\|\boldsymbol{z}\|_A \\ \|\delta\tilde{\boldsymbol{z}}^{(s)}\|_A &< \left(\varrho^{s-1}\boldsymbol{C} + (1-\varrho)^{-1}(\boldsymbol{D}+\varepsilon_1\boldsymbol{C})\right)\|\boldsymbol{z}\|_A\,. \end{aligned} \tag{6.20}$$

If we make the reasonable assumption

$$(m+1)\varepsilon_2' < \varepsilon_1$$

then, as a consequence of (5.11) and the assumption $n \geqq 2$

$$c < 1.0004(13.64 + 2\cdot 2^{-3/2}1.021)n^{3/2}\varepsilon_1\,,$$

or

$$c < 14.4n^{3/2}\varepsilon_1\,. \tag{6.21}$$

In the initial stages the term $\varrho^{s-1}\boldsymbol{C}$ in (6.20) dominates and we have approximately

$$\|\delta\boldsymbol{z}^{(s)}\|_A \leqq \varrho^{s-1}\boldsymbol{C}\|\boldsymbol{z}\|_A\,.$$

This justifies calling ϱ the initial rate of convergence. From (6.19) *and* (6.21) *we have the estimate*

$$\varrho < 38.4n^{3/2}(\varkappa' + \tfrac{1}{2})\varepsilon_1\,. \tag{6.22}$$

We note that this bound for ϱ is independent of the right hand side and roughly proportional to $\varkappa(\boldsymbol{A})$. Thus, $\varkappa^2(\boldsymbol{A})$ does not enter, which, remembering (2.7), might have been conjectured from the general analysis.

When $\varrho < 1$, the term $\varrho^{s-1}\boldsymbol{C}\|\boldsymbol{z}\|_A$ in (6.20) *approaches zero as $s \to \infty$ and the limiting accuracy in $\tilde{\boldsymbol{z}}^{(s)}$ is*

$$\lim_{s\to\infty} \|\delta\tilde{\boldsymbol{z}}^{(s)}\|_A \leqq (1-\varrho)^{-1}(\boldsymbol{D}+\varepsilon_1\boldsymbol{C})\|\boldsymbol{z}\|_A = \begin{pmatrix} 1 \\ \varkappa' \end{pmatrix} (1-\varrho)^{-1}K \tag{6.23}$$

where

$$\begin{aligned} K = {} & 1.022\varepsilon_2'\left(\varkappa'(m+4)\|\boldsymbol{r}\|_2 + (n+5)\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2\right) + \\ & 14.4n^{3/2}\varepsilon_1{}^2\left((\varkappa' + \tfrac{4}{3})\|\boldsymbol{r}\|_2 + \tfrac{5}{3}\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2\right)\,. \end{aligned} \tag{6.24}$$

The first term in K is proportional to ε_2 and comes from the errors made when computing the residuals. The second term, which is proportional to ε_1^2, comes from the errors introduced when solving the system of equations defining the corrections. *To get full benefit from the refinement it is obviously necessary to have $t_2 \geqq 2t_1$.* The first term in K will then usually be negligible, and the limiting accuracy will approximately be $(1-\varrho)^{-1}\varepsilon_1 C\|z\|_A$. This can be compared to the bound for the error $\|\bar{z}^{(1)}-z\|_A$ derived from (6.18), which is $C\|z\|_A$. *Hence, we can expect to gain t_1 figures during the refinement, and to achieve almost the same accuracy as if, without any refinement, $2t_1$ digit precision had been used throughout the computation.*

Since either $\boldsymbol{r}$ or $\boldsymbol{x}$ can be equal to zero, we obviously can not always expect to achieve a small relative error in either $\bar{\boldsymbol{r}}^{(s)}$ or $\bar{\boldsymbol{x}}^{(s)}$. We now derive simple sufficient conditions for the relations

$$\lim_{s\to\infty} \|\bar{\boldsymbol{r}}^{(s)}-\boldsymbol{r}\|_2 \leqq 2\varepsilon_1\cdot\|\boldsymbol{r}\|_2\,, \tag{6.25}$$

$$\lim_{s\to\infty} \|\bar{\boldsymbol{x}}^{(s)}-\boldsymbol{x}\|_2 \leqq 2\varepsilon_1\cdot\|\boldsymbol{x}\|_2 \tag{6.26}$$

to be satisfied. We assume that $\varrho \leqq \frac{1}{4}$ and that the second term in K can be neglected, since the most important case in practice is $t_2 \geqq 2t_1$. From (6.14), (6.20) and (6.23) it follows that (6.25) holds provided

$$\tfrac{4}{3}c(\varkappa'+\tfrac{4}{3})(1+\tfrac{5}{3}\gamma^{-1}) \leqq 1$$

where we have put

$$\gamma = \frac{(\varkappa'+\frac{4}{3})\|\boldsymbol{r}\|_2}{\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2}\,. \tag{6.27}$$

Substituting for c from (6.19) we obtain

$$1+\tfrac{5}{3}\gamma^{-1} \leqq \frac{2}{\varrho}\,\frac{\varkappa'+\frac{1}{2}}{\varkappa'+\frac{4}{3}}\,.$$

Since $\varkappa' \geqq 1$ this relation is satisfied if

$$\gamma^{-1} \leqq \tfrac{3}{5}\left(\frac{2\cdot 9}{\varrho\cdot 14}-1\right).$$

Remembering that $\varrho \leqq \frac{1}{4}$ it follows that (6.25) certainly holds if

$$\gamma^{-1} \leqq \tfrac{87}{140}\varrho^{-1}\,. \tag{6.28}$$

Similarly (6.26) holds if

$$\gamma+\tfrac{5}{3} \leqq \frac{2}{\varrho}\,\frac{\varkappa'+\frac{1}{2}}{\varkappa'}$$

or if

$$\gamma \leqq \frac{1}{\varrho}(2-\tfrac{5}{12}) = \tfrac{19}{12}\varrho^{-1}\,. \tag{6.29}$$

We conclude that if γ satisfies

$$1.61\varrho \leqq \gamma \leqq 1.58\frac{1}{\varrho}$$

then, for sufficiently large s, both $\bar{\boldsymbol{r}}^{(s)}$ and $\bar{\boldsymbol{x}}^{(s)}$ will have a relative accuracy better than $2\varepsilon_1$. It is obvious that, for a given matrix $\boldsymbol{A}$, there will always exist right hand sides for which this is not true either for $\bar{\boldsymbol{r}}^{(s)}$ or for $\bar{\boldsymbol{x}}^{(s)}$. However, since $\varrho \leqq \frac{1}{4}$ implies $1.61\varrho < 1.58\varrho^{-1}$, *at least one of the relations* (6.25) *and* (6.26) *is always satisfied.*

7. Error analysis of the modified Gram-Schmidt method.

This method is based on a decomposition of $\boldsymbol{A}$ obtained in the following way: Let $\boldsymbol{A}^{(1)} = \boldsymbol{A}$ and $\boldsymbol{A}^{(k)}$, $k = 2,3,\ldots,n+1$ be defined by

$$\begin{aligned} \boldsymbol{r}_k^T &= d_k^{-1}\boldsymbol{q}_k^T\boldsymbol{A}^{(k)}, \quad d_k = \|\boldsymbol{q}_k\|_2^2\,, \\ \boldsymbol{A}^{(k+1)} &= \boldsymbol{A}^{(k)} - \boldsymbol{q}_k\boldsymbol{r}_k^T \end{aligned} \tag{7.1}$$

where $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ is a suitable sequence of linearly independent vectors. By induction it follows that

$$\boldsymbol{A}^{(n+1)} = \boldsymbol{A} - \boldsymbol{Q}\boldsymbol{R} \tag{7.2}$$

where $\boldsymbol{R}^T = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n)$, and for $k = 1,2,\ldots,n$, that

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

Thus, if we define the lower triangular matrix $\boldsymbol{L}$ by

$$\boldsymbol{L} = \{l_{kj}\}, \qquad l_{kj} = \begin{cases} \boldsymbol{q}_k^T\boldsymbol{q}_j, & k \geqq j \\ 0\quad, & k < j \end{cases}, \tag{7.3}$$

then

$$\boldsymbol{L}\boldsymbol{R} = \boldsymbol{Q}^T\boldsymbol{A}\,. \tag{7.4}$$

We now choose $\boldsymbol{q}_k$ so that in step k the kth column of $\boldsymbol{A}^{(k)}$ is annihilated, i.e. so that

$$\boldsymbol{A}^{(k)} = (0,\ldots,0,\boldsymbol{a}_k^{(k)},\ldots,\boldsymbol{a}_n^{(k)})\,.$$

This is obviously achieved if we take $\boldsymbol{q}_k = \boldsymbol{a}_k^{(k)}$. Then, by (7.2),

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \tag{7.5}$$

and $\boldsymbol{R}$ becomes unit upper triangular. If the calculations are performed *exactly* the columns of $\boldsymbol{Q}$ will be mutually orthogonal and thus $\boldsymbol{L}$ will be a diagonal matrix.

The method described can be interpreted as an elimination with weighted row combinations (Bauer [1]) or as a modification of the classical Gram–Schmidt orthogonalization method (Björck [2]).

We now adopt this method to solve the system (5.2). This must be done carefully, since, as is well known, [2], the computed columns of $\boldsymbol{Q}$ may deviate considerably from orthogonality. In the derivation we, therefore, do not assume $\boldsymbol{L}$ to be diagonal. If we define

$$\boldsymbol{y} = (y_1,y_2,\ldots,y_n)^T, \quad \boldsymbol{h} = (h_1,h_2,\ldots,h_n)^T$$

by

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x}, \quad \boldsymbol{h} = \boldsymbol{R}^{-T}\boldsymbol{g} \tag{7.6}$$

and use (7.5), then we can write (5.2) as

$$\begin{aligned} \boldsymbol{r}+\boldsymbol{Q}\boldsymbol{y} &= \boldsymbol{f} \\ \boldsymbol{Q}^T\boldsymbol{r} \quad &= \boldsymbol{h}\,. \end{aligned}$$

Multiplying the first set of equations in (5.2) from the left by $\boldsymbol{Q}^T$ and using (7.4) we get

$$\boldsymbol{Q}^T\boldsymbol{r}+\boldsymbol{L}\boldsymbol{y} = \boldsymbol{Q}^T\boldsymbol{f}\,.$$

Thus $\boldsymbol{r}$ and $\boldsymbol{y}$ are determined by

$$\boldsymbol{L}\boldsymbol{y} = \boldsymbol{Q}^T\boldsymbol{f}-\boldsymbol{h}, \quad \boldsymbol{r} = \boldsymbol{f}-\boldsymbol{Q}\boldsymbol{y}\,. \tag{7.7}$$

Now let $\boldsymbol{f}^{(1)}=\boldsymbol{f}$ and define $\boldsymbol{f}^{(k)}$, $k=2,3,\ldots,n+1$, and y by

$$y_k = (\boldsymbol{q}_k{}^T\boldsymbol{f}^{(k)}-h_k)/d_k, \quad \boldsymbol{f}^{(k+1)} = \boldsymbol{f}^{(k)}-\boldsymbol{q}_k y_k\,. \tag{7.8}$$

By induction it follows that

$$\begin{aligned} \boldsymbol{f}^{(k)} &= \boldsymbol{f}-(\boldsymbol{q}_1 y_1+\ldots+\boldsymbol{q}_{k-1}y_{k-1})\,, \\ \boldsymbol{q}_k{}^T(\boldsymbol{q}_1 y_1+\ldots+\boldsymbol{q}_{k-1}y_{k-1}+\boldsymbol{q}_k y_k) &= \boldsymbol{q}_k{}^T\boldsymbol{f}-h_k\,, \end{aligned}$$

and thus $\boldsymbol{y}$ and $\boldsymbol{r}=\boldsymbol{f}^{(n+1)}$ satisfy (7.7). Hence, (7.6) and (7.8) is the desired algorithm for solving (5.2).

In the error analysis below, we again assume that (5.11) holds and that $t_2 \geqq 2t_1$. Let the computed factors in the decomposition (7.5) of $\boldsymbol{A}$ be $\overline{\boldsymbol{R}}$ and $\overline{\boldsymbol{Q}}=(\overline{\boldsymbol{q}}_1,\overline{\boldsymbol{q}}_2,\ldots,\overline{\boldsymbol{q}}_n)$. In the back-substitutions (7.6) the computed quantities $\boldsymbol{h}$ and $\boldsymbol{x}$ satisfy, cf. (5.16),

$$\overline{\boldsymbol{R}}_1\overline{\boldsymbol{x}} = \overline{\boldsymbol{y}}, \quad \overline{\boldsymbol{R}}_2{}^T\overline{\boldsymbol{h}} = \boldsymbol{g}\,,$$

where $\overline{\boldsymbol{R}}_i=\overline{\boldsymbol{R}}+\boldsymbol{G}_i$, and

$$\|\boldsymbol{G}_i\|_E \leqq (\varepsilon_1+n\varepsilon_2)\|\overline{\boldsymbol{R}}\|_E \leqq 1.002\varepsilon_1\|\overline{\boldsymbol{R}}\|_E, \quad i = 1,2\,. \tag{7.9}$$

Let $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{f}}^{(1)},\tilde{\boldsymbol{f}}^{(2)},\ldots,\tilde{\boldsymbol{f}}^{(n+1)}=\tilde{\boldsymbol{r}}$ denote the *exact* results when performing (7.8) using the *computed* quantities $\overline{\boldsymbol{q}}_k$ and $\overline{h}_k$. If we put

$$e_1 = \tilde{\boldsymbol{r}} - \bar{\boldsymbol{r}}, \quad e_2 = \tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}\,,$$

then the following expression for the errors in $\bar{\boldsymbol{r}}$ and $\bar{\boldsymbol{x}}$ holds:

$$\begin{pmatrix} \boldsymbol{r}-\bar{\boldsymbol{r}} \\ \boldsymbol{x}-\bar{\boldsymbol{x}} \end{pmatrix} = (\boldsymbol{I}-\bar{\boldsymbol{B}}^{-1}\boldsymbol{B}) \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{pmatrix} + \begin{pmatrix} e_1 \\ \bar{\boldsymbol{R}}_1^{-1} e_2 \end{pmatrix} \tag{7.10}$$

where

$$\boldsymbol{B} = \left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A} \\ \hline \boldsymbol{A}^T & 0 \end{array}\right), \quad \bar{\boldsymbol{B}}^{-1} = \left(\begin{array}{c|c} \boldsymbol{I} & -\bar{\boldsymbol{Q}} \\ \hline 0 & \bar{\boldsymbol{R}}_1^{-1} \end{array}\right) \left(\begin{array}{c|c} \boldsymbol{I} & 0 \\ \hline 0 & \bar{\boldsymbol{L}}^{-1} \end{array}\right) \left(\begin{array}{c|c} \boldsymbol{I} & 0 \\ \hline \bar{\boldsymbol{Q}}^T & -\bar{\boldsymbol{R}}_2^{-T} \end{array}\right).$$

Note that, in order to make the errors small, $\bar{\boldsymbol{B}}^{-1}$, which is not symmetric, has to be a good *left-hand* inverse. It is readily verified that

$$\boldsymbol{I}-\bar{\boldsymbol{B}}^{-1}\boldsymbol{B} = \left(\begin{array}{c|c} \bar{\boldsymbol{Q}}\boldsymbol{F}_1 & -\boldsymbol{E}_2 \\ \hline -\bar{\boldsymbol{R}}_1^{-1}\boldsymbol{F}_1 & \bar{\boldsymbol{R}}_1^{-1}(\boldsymbol{E}_3+\boldsymbol{G}_1) \end{array}\right) \tag{7.11}$$

where

$$\begin{aligned} \boldsymbol{F}_1 &= \bar{\boldsymbol{L}}^{-1}\bar{\boldsymbol{R}}_2^{-T}(\boldsymbol{E}_1+\bar{\boldsymbol{Q}}\boldsymbol{G}_2)^T, & \boldsymbol{E}_1 &= \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}-\boldsymbol{A}\,, \\ \boldsymbol{E}_2 &= (\boldsymbol{I}-\bar{\boldsymbol{Q}}\bar{\boldsymbol{L}}^{-1}\bar{\boldsymbol{Q}}^T)\boldsymbol{A}, & \boldsymbol{E}_3 &= \bar{\boldsymbol{R}}-\bar{\boldsymbol{L}}^{-1}\bar{\boldsymbol{Q}}^T\boldsymbol{A}\,, \end{aligned} \tag{7.12}$$

and from [2] p. 10 we have the following results:

$$\begin{aligned} &\|\boldsymbol{E}_1\|_E \leqq 1.5(n-1)\varepsilon_1\|\boldsymbol{A}\|_E, \quad \|\boldsymbol{E}_2\|_E \leqq 3.25(n-1)\varepsilon_1\|\boldsymbol{A}\|_E \\ &\|\boldsymbol{E}_3\|_E \leqq 1.9(n-1)^{\frac{1}{2}} n\varepsilon_1\|\boldsymbol{A}\|_E\,. \end{aligned} \tag{7.13}$$

We now estimate the rounding errors made in (7.8). To simplify the analysis, we assume that

$$\|\bar{\boldsymbol{q}}_k\|_2 = 1, \quad k = 1,2,\ldots,n\,.$$

Since this can be achieved by a proper scaling of the columns of $\boldsymbol{A}$, this assumption will not influence the derived floating point error bounds. Then we have

$$\tilde{\boldsymbol{f}}^{(k+1)} = (\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)\tilde{\boldsymbol{f}}^{(k)}+\bar{\boldsymbol{q}}_k\bar{h}_k\,.$$

If we define the error vector $\eta^{(k)}$ by

$$\bar{\boldsymbol{f}}^{(k+1)} = (\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)\bar{\boldsymbol{f}}^{(k)}+\bar{\boldsymbol{q}}_k\bar{h}_k+\eta^{(k)}\,, \tag{7.14}$$

then, subtracting,

$$\bar{\boldsymbol{f}}^{(k+1)}-\tilde{\boldsymbol{f}}^{(k+1)} = (\boldsymbol{I}-\bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^T)(\bar{\boldsymbol{f}}^{(k)}-\tilde{\boldsymbol{f}}^{(k)})+\boldsymbol{\eta}^{(k)}\,.$$

Since $\bar{\boldsymbol{f}}^{(1)}=\tilde{\boldsymbol{f}}^{(1)}=\boldsymbol{f}$ it follows that

$$\|\bar{\boldsymbol{f}}^{(k+1)}-\tilde{\boldsymbol{f}}^{(k+1)}\|_2 \leqq \sum_{\nu=1}^{k} \|\eta^{(\nu)}\|_2, \qquad k = 1,2,\ldots,n\,. \tag{7.15}$$

For the error in $\bar{y}_k$ we write

$$\bar{y}_k-\tilde{y}_k = \bar{y}_k-y_k'+\bar{\boldsymbol{q}}_k^T(\bar{\boldsymbol{f}}^{(k)}-\tilde{\boldsymbol{f}}^{(k)})$$

where y_k' is the exact multiplier corresponding to the computed vector $\bar{\boldsymbol{f}}^{(k)}$. Then

$$|\bar{y}_k-\tilde{y}_k| \leqq |\bar{y}_k-y_k'|+\|\bar{\boldsymbol{f}}^{(k)}-\tilde{\boldsymbol{f}}^{(k)}\|_2 \tag{7.16}$$

and

$$|y_k'| \leqq \|\bar{\boldsymbol{f}}^{(k)}\|_2+|\bar{h}_k| \,.$$

Now, if we make the reasonable assumption that $2(m+1)\varepsilon_1 \leqq 0.01$, then we can use results from [2] p. 7–8, duly modified, to obtain

$$|\bar{y}_k-y_k'| < 2.02\varepsilon_1(\|\bar{\boldsymbol{f}}^{(k)}\|_2+|\bar{h}_k|) \tag{7.17}$$

and

$$\|\boldsymbol{\eta}^{(k)}\|_2 < (1-\varepsilon_1)^{-1}\varepsilon_1\|\bar{\boldsymbol{f}}^{(k+1)}\|_2+\varepsilon_1|\bar{y}_k|+|\bar{y}_k-y_k'| \,.$$

Consequently

$$(1-\varepsilon_1)\|\boldsymbol{\eta}^{(k)}\|_2 \leqq \varepsilon_1\|\bar{\boldsymbol{f}}^{(k+1)}\|_2+3.02\varepsilon_1(\|\bar{\boldsymbol{f}}^{(k)}\|_2+|\bar{h}_k|) \tag{7.18}$$

From (7.14) it follows that

$$\|\bar{\boldsymbol{f}}^{(k+1)}\|_2 \leqq \|\bar{\boldsymbol{f}}^{(k)}\|_2+|\bar{h}_k|+\|\boldsymbol{\eta}^{(k)}\|_2$$

and using (7.18) we have

$$(1-2\varepsilon_1)\|\bar{\boldsymbol{f}}^{(k+1)}\|_2 \leqq (1+2.02\varepsilon_1)(\|\bar{\boldsymbol{f}}^{(k)}\|_2+|\bar{h}_k|) \,.$$

The assumption (5.11) certainly implies that

$$(1+2.02\varepsilon_1)/(1-2\varepsilon_1) < 1.002^{1/n}$$

and thus, for $k=1,2,\ldots,n$,

$$\|\bar{\boldsymbol{f}}^{(k+1)}\|_2 < 1.002(\|\boldsymbol{f}\|_2+s_k), \quad s_k = \sum_{j=1}^{k}|\bar{h}_j| \,.$$

Using this bound for the growth of the computed vectors $\bar{f}^{(k)}$ in (7.17) and (7.18) we obtain

$$\begin{aligned} |\bar{y}_k-y_k'| &< 4.05\tfrac{1}{2}\varepsilon_1(\|\boldsymbol{f}\|_2+s_k)\,, \\ \|\boldsymbol{\eta}^{(k)}\|_2 &< 4.05\varepsilon_1(\|\boldsymbol{f}\|_2+s_k)\,. \end{aligned} \tag{7.19}$$

Using Schwarz's inequality we get

$$\sum_{\nu=1}^{k} s_\nu \leqq \left(\sum_{\nu=1}^{k}\nu^2\right)^{\frac{1}{2}}\|\bar{\boldsymbol{h}}\|_2 < \frac{1}{\sqrt{3}}k^{\frac{1}{2}}(k+1)\|\bar{\boldsymbol{h}}\|_2$$

and

$$\tfrac{1}{2}s_k+\sum_{\nu=1}^{k-1}s_\nu < \frac{1}{\sqrt{3}}k^{3/2}\|\bar{\boldsymbol{h}}\|_2 \,.$$

Hence, from (7.15), (7.16) and (7.19) we have

$$\|\boldsymbol{e}_1\|_2 = \|\bar{\boldsymbol{f}}^{(n+1)} - \tilde{\boldsymbol{f}}^{(n+1)}\|_2 < 4.05(n+1)\varepsilon_1\varphi\,, \qquad (7.20)$$
$$|\bar{y}_k - \tilde{y}_k| < 4.05k\varepsilon_1\varphi\,,$$

where

$$\varphi = \|\boldsymbol{f}\|_2 + \frac{1}{\sqrt{3}} n^{\frac{1}{2}} \|\bar{\boldsymbol{h}}\|_2\,. \qquad (7.21)$$

Using Schwarz's inequality again we obtain

$$\|\boldsymbol{e}_2\|_2 = \|\bar{\boldsymbol{y}} - \tilde{\boldsymbol{y}}\|_2 < \frac{1}{\sqrt{3}} 4.05 n^{\frac{1}{2}}(n+1)\varepsilon_1\varphi\,. \qquad (7.22)$$

From (5.2) we have $\boldsymbol{h} = \boldsymbol{R}^{-T}\boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{Q}^T\boldsymbol{r}$, and thus

$$\|\boldsymbol{f}\|_2 \leqq \|\boldsymbol{r}\|_2 + \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2, \quad \|\boldsymbol{h}\|_2 \leqq \|\boldsymbol{r}\|_2\,. \qquad (7.33)$$

Since the computed matrix $\bar{\boldsymbol{Q}}$ is not exactly orthogonal, a further strict analysis becomes very cumbersome. In [2] pp. 12–15 we have shown, however, that

$$\|\boldsymbol{I} - \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}}\|_2 \leqq \text{const.}\,\delta + 0(\delta^2)\,, \qquad (7.24)$$

where

$$\delta = n^2 \cdot \varepsilon_1 \varkappa(A)\,.$$

In the following we therefore assume that δ is a small quantity. From (7.24) it follows

$$\|\bar{\boldsymbol{Q}}\|_2 = 1 + 0(\delta), \quad \|\bar{\boldsymbol{L}}^{-1}\|_2 = 1 + 0(\delta)\,.$$

Let $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q}$ is normalized so that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Then we also have ([2] p. 14)

$$\|\bar{\boldsymbol{R}}\|_E = \|\boldsymbol{R}\|_E(1 + 0(\delta)) = \|\boldsymbol{A}\|_E(1 + 0(\delta))\,.$$

Using these relations, we get from (7.9) and (7.12)

$$\begin{aligned} \|\boldsymbol{E}_1 + \bar{\boldsymbol{Q}}\boldsymbol{G}_2\|_E &\leqq 1.5n\varepsilon_1\|\boldsymbol{A}\|_E(1 + 0(\delta))\,, \\ \|\boldsymbol{E}_3 + \boldsymbol{G}_1\|_E &\leqq 1.9 \cdot n^{3/2}\varepsilon_1\|\boldsymbol{A}\|_E(1 + 0(\delta)) \end{aligned} \qquad (7.25)$$

Neglecting powers of δ we derive from the inequality

$$\|\boldsymbol{A}\|_E\|\bar{\boldsymbol{R}}_i^{-1}\|_2 \leqq n^{\frac{1}{2}}\|\boldsymbol{A}\|_2\|\bar{\boldsymbol{R}}_i^{-1}\|_2 = n^{\frac{1}{2}}\varkappa(\boldsymbol{A})(1 + 0(\delta))$$

$i = 1, 2$, (7.10)–(7.12) and (7.25) the estimate

$$\|z - \bar{z}\|_A \leqq 1.5n^{\frac{1}{2}}(n+1)\varepsilon_1 \begin{pmatrix} \varkappa & 1.27\sqrt{3} \\ \varkappa^2 & \varkappa \cdot 1.27 \cdot n^{\frac{1}{2}} \end{pmatrix} \|z\|_A + \begin{pmatrix} \|\boldsymbol{e}_1\|_2 \\ \varkappa\|\boldsymbol{e}_2\|_2 \end{pmatrix}.$$

Further, from $\|\bar{\boldsymbol{h}}\|_2 = \|\boldsymbol{h}\|_2(1+0(\delta))$, (7.21) and (7.23) it follows that

$$\varphi \leqq \|\boldsymbol{r}\|_2\left(1+\frac{n^{\frac{1}{2}}}{\sqrt{3}}\right)+\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2+0(\delta)\,.$$

Hence, we obtain from (7.20) and (7.22), neglecting powers of δ

$$\begin{pmatrix}\|\boldsymbol{e}_1\|_2\\ \varkappa\|\boldsymbol{e}_2\|_2\end{pmatrix} \leqq 4.05(n+1)\varepsilon_1\begin{pmatrix}1\\ \varkappa\dfrac{n^{\frac{1}{2}}}{\sqrt{3}}\end{pmatrix}\left(1+\frac{n^{\frac{1}{2}}}{\sqrt{3}},\ 1\right)\|\boldsymbol{z}\|_A\,.$$

After some manipulation we have finally

$$\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|_A \leqq 1.5n^{\frac{1}{2}}(n+1)\varepsilon_1\begin{pmatrix}\varkappa+0.9(\sqrt{3}+3n^{-\frac{1}{2}}) & 1.27\sqrt{3}(1+1.25n^{-\frac{1}{2}})\\ \varkappa(\varkappa+0.9(n^{\frac{1}{2}}+\sqrt{3})) & \varkappa\cdot 1.27(n^{\frac{1}{2}}+1.25)\end{pmatrix}\|\boldsymbol{z}\|_A\,.$$

If $n \geqq 3$, which we therefore assume, we can obviously write this relation in the form

$$\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|_A \leqq 1.5n^{\frac{1}{2}}(n+1)\varepsilon_1\boldsymbol{M}_{GS}\|\boldsymbol{z}\|_A+0(\delta^2) \tag{7.26}$$

where

$$\boldsymbol{M}_{GS} = \begin{pmatrix}1\\ \varkappa\end{pmatrix}\left(\varkappa+0.9(n^{\frac{1}{2}}+\sqrt{3}),\ 1.27(n^{\frac{1}{2}}+1.25)\right) \in \mathscr{M}(\varkappa)\,.$$

Using this as a starting point, the iterative refinement with the Gram–Schmidt method can be analysed in a way similar to that used for the Householder method. Obviously, the effects of the errors in the residuals will, in the first approximation, be the same.

It will again be true, that if $t_2 \geqq 2t_1$, we can expect to gain t_1 figures during the refinement. To compare the limiting accuracy of the methods it is therefore sufficient to look at the error bounds for the solution of the system of equations (5.2).

A result corresponding to (7.26) for the Householder method follows from (6.9), if we put $s=0$ and $\tau=0$. Then, we get

$$\|\boldsymbol{z}-\boldsymbol{z}\|_A \leqq 13.65n^{3/2}\varepsilon_1\boldsymbol{M}_H\|\boldsymbol{z}\|_A+0(\delta^2) \tag{7.27}$$

where

$$\boldsymbol{M}_H = \begin{pmatrix}1\\ \varkappa\end{pmatrix}\left(\varkappa+\tfrac{4}{3},\tfrac{5}{3}\right)\,.$$

Comparing (7.26) and (7.27) it is seen that when $n \leqq 110$ the bound for the Gram–Schmidt method is always smaller. In particular for large values of $\varkappa$ and γ (defined in (6.26)), the ratio is approximately 1:9.

By lemma 1 and 2 the rate of convergence for the Gram–Schmidt method is approximately equal to

$$\varrho_{GS} = 1.5n^{\frac{1}{2}}(n+1)\varepsilon_1\varrho(\boldsymbol{M}_{GS}) \tag{7.28}$$

where

$$\begin{aligned}\varrho(\boldsymbol{M}_{GS}) &= \varkappa+0.9(n^{\frac{1}{2}}+\sqrt{3})+1.27(n^{\frac{1}{2}}+1.25)\\ &< 1.27(n^{\frac{1}{2}}+2)(\varkappa+0.71)\,.\end{aligned}$$

This can be compared to the corresponding expression for the Householder method which is

$$\varrho_H = 13.65n^{3/2}\varepsilon_1\tfrac{8}{3}(\varkappa+\tfrac{1}{2})\,. \tag{7.29}$$

For reasonable values of n, the Gram–Schmidt method is again seen to be more favourable. The advantage when $\varkappa \gg 1$ and $\gamma \gg 1$ is however less marked.

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. Björck, Å., *Solving Linear Least Squares Problems by Gram–Schmidt Orthogonalization*, BIT 7 (1967), 1–21.
3. Collatz, L., *Funktionalanalysis und Numerische Mathematik*, Berlin: Springer-Verlag (1964).
4. Golub, G. H., *Numerical Methods for Solving Linear Least Squares Problems*, Num. Math. 7 (1965), 206–216.
5. Golub, G. H. and Wilkinson, J. H., *Note on the Iterative Refinement of Least Squares Solution*, Num. Math. 9 (1966), 139–148.
6. Martin, R. S., Peters, G. and Wilkinson, J. H., *Iterative Refinement of the Solution of a Positive Definite System of Equations*, Num. Math. 8, (1966), 203–216.
7. Moler, C. B., *Iterative Refinement in Floating Point*, J. Assoc. Comput. Mach. 14, (1967), 316–321.
8. Wilkinson, J. H., *Rounding Errors in Algebraic Processes*, London: Her Majesty's Stationary Office; Englewood Cliffs, N.J.: Prentice-Hall 1963.
9. Wilkinson, J. H., *Error Analysis of Transformations Based on the Use of Matrices of the Form $\boldsymbol{I}-2\boldsymbol{w}\boldsymbol{w}^T$*, Error in Digital Computations. Volume II. Rall, L. B. ed. New York: John Wiley (1965), 77–101.