Extensions and Applications of the Householder Algorithm for Solving Linear Least Squares Problems*

By Richard J. Hanson and Charles L. Lawson

Abstract. The mathematical and numerical least squares solution of a general linear system of equations is discussed.

Perturbation and differentiability theorems for pseudoinverses are given.

Computational procedures for calculating least squares solutions using orthonormal transformations, multiplying matrices by a matrix of orthonormal basis vectors for the null-space of a given matrix, sequential processing of data, and processing of block diagonal matrices form a partial list of numerical topics presented.

Key Words: Linear Least Squares, Linear Equations, Pseudoinverse Solution, Numerical Methods for Linear Equations, Orthogonal Methods for Linear Systems.

Table of Contents

1. **Introduction**

2. **Preliminary Notation and Theorems**

   2.1 *Notation*
   2.2 *The Singular Value Decomposition and the Pseudoinverse*
   2.3 *General Perturbation Theorems*
   2.4 *Perturbation and Differentiability Theorems for Matrices of Full Rank*

3. **Two Basic Householder Orthonormal Transformation Algorithms**

   3.1 *Constructing* $I_m + \beta^{-1}uu^T$
   3.2 *Multiplication by* $I_m + \beta^{-1}uu^T$

4. **Extension of the Businger-Golub Algorithm to Permit Rank Deficiency**

   4.1 *The Businger-Golub Algorithm*
   4.2 *Extension for Rank Deficiency*
   4.3 *The Complete Solution*
   4.4 *Matrix Replacement, Pseudorank, and Column Interchanges*

5. **Other Linear Algebraic Computations**

   5.1 *Multiple Right Sides*
   5.2 *Computation of the Pseudoinverse,* A^+
   5.3 *Triangularization as a Preliminary Step*
   5.4 *Triangularization of a Block Diagonal Matrix*
   5.5 *The Case of* $m < n$
   5.6 *Linear Least Squares with Linear Equality Constraints*

1. Introduction. The central problem under consideration in this paper is the numerical solution of the following linear least squares problem:

Received August 27, 1968, revised February 19, 1969.

* This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. NAS 7-100, sponsored by the National Aeronautics and Space Administration.

Problem LS. *Given a real $m \times n$ matrix A of* rank k, *and a real m-vector b, find a real n-vector x_0 minimizing the euclidean length of $Ax - b$.*

It is known that this problem always has a solution. The solution is unique if $k = n$. If $k < n$ there is an $(n - k)$-dimensional linear flat of solutions. In this solution flat there is a unique vector of minimum euclidean length and we shall adopt this additional condition to impose uniqueness on the solution x_0 of Problem LS.

In the case of $k < n$ the general solution of Problem LS can be expressed in the parametric form $x_0 + Hy$, where H is an $n \times (n - k)$-matrix of rank $n - k$ satisfying $AH = 0$ and y is an arbitrary $(n - k)$-vector. We will treat the problem of computing H subject to the additional condition that the column vectors of H be orthonormal.

In Section 2 we recall certain relevant properties of the singular value decomposition of a matrix and the pseudoinverse of a matrix. Perturbation theorems are stated for pseudoinverses. In particular for the full rank case ($k = \min(m, n)$), Theorem 2.4.2 and the consequent Theorems 2.4.3, 2.4.4, and 2.4.5 are more precise and complete than similar previously published theorems of which we are aware. Closely related results are given in [8], [10], [16], and [17].

An algorithm using Householder orthonormal transformations for the solution of Problem LS when $k = n$ was given by Businger and Golub [1]. This algorithm has favorable numerical properties [14] due to the use of orthonormal transformations and the avoidance of the formation of the matrix $N = A^TA$. Many algorithmic variations of A^TA methods have been developed to meet special requirements, or to take advantage of special properties. Examples include sequential availability of data (rows of the augmented matrix $[A, b]$) in real time, linear equality constraints, computer storage limitations, rank deficient problems, block diagonal matrices, etc. We have investigated the possibility of adapting the Householder transformation technique to some of these special cases and have found that it is frequently possible to do so. Gram-Schmidt orthogonalization can also be organized to meet certain special requirements (see [11], [13], and [16] for example).

Specifically we propose to identify two basic algorithms, one to construct a Householder transformation matrix in the usual compactly stored form and the other to multiply a vector by such a matrix. We describe these basic algorithms in Section 3.

In Section 4 we use these two basic algorithms as an aid in describing an extension of the Businger-Golub algorithm to treat the case of $k < n$, including the computation of the matrix H needed to define the complete solution of Problem LS.

In Section 5 we illustrate the use of these two basic algorithms as components in the definition of algorithms for a variety of computational problems in linear algebra.

2. Preliminary Notation and Theorems.

2.1. *Notation.* A real matrix $A = \{a_{ij}\}$ having m rows and n columns will be called an $m \times n$ matrix. The designation $A_{m \times n}$ will also be used. The transpose of A will be denoted by A^T.

The symbols I_p and 0_q will denote respectively the $p \times p$ identity matrix and the $q \times q$ zero matrix.

If S is the $k \times k$ diagonal matrix

$$\begin{bmatrix} s_1 & & & & 0 \\ & s_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & s_k \end{bmatrix},$$

then we will write $S = \operatorname{diag}(s_1, \cdots, s_k)$.

An $m \times n$ matrix P, $(m \geqq n)$, is *orthonormal* if $P^TP = I_n$.

For any integer $q > 0$, we will denote the q-dimensional linear coordinate vector space by E^q.

For two q-dimensional vectors x_1 and x_2, $(x_1, x_2) = x_1{}^Tx_2 = x_2{}^Tx_1$, and if y is a vector in E^q with consecutive components $y_1, \cdots, y_q$, then

$$\|y\| = \left(\sum_{i=1}^{q} y_i^2\right)^{1/2} = (y^Ty)^{1/2} = (y, y)^{1/2}$$

is *the euclidean norm* of y. We will avoid referring to the dimension of the space to which a vector belongs except when failure to do so might lead to confusion.

The corresponding matrix norm, $\|A\|$, for an $m \times n$ matrix A, is the spectral norm, i.e. the square root of the maximal eigenvalue of A^TA (or, equivalently, AA^T).

If A is an $m \times n$ matrix, then the range of A is denoted by

$$\mathcal{R}(A) = \{w \in E^m | w = Ax, x \in E^n\} .$$

For an arbitrary real number a, define

$$\begin{aligned} \operatorname{sgn}(a) &= 1\,, \quad a \geqq 0 \\ &= -1\,, \quad a < 0\,. \end{aligned}$$

2.2. *The Singular Value Decomposition and the Pseudoinverse.*

THEOREM 2.2.1. *Let A be an arbitrary $m \times n$ real nonzero matrix of* rank $k \leqq$ min (m, n). *Then there exist matrices U, S, and V of respective dimensions $m \times k$, $k \times k$, and $n \times k$ such that*

$$A = USV^T \,. \tag{2.2.1}$$

The matrices U and V are both orthonormal. Thus

$$U^TU = I_k \tag{2.2.2}$$

and

$$V^TV = I_k \,. \tag{2.2.3}$$

The matrix

$$S = \operatorname{diag}(s_1, \cdots, s_k) \tag{2.2.4}$$

has diagonal terms $s_1, \cdots, s_k$ for which $s_1 \geqq s_2 \geqq \cdots \geqq s_k > 0$. These scalars are the nonzero singular values of the matrix A. The singular values are unique.

See [2] or [3], pages 9–10 for a proof of Theorem 2.2.1, and [4] for an efficient procedure for computation of this singular value composition.

Alternative statements of Theorem 2.2.1 are useful in certain circumstances. If $q = \min(m, n)$ one may write $A = U_{m\times q} S_{q\times q} (V_{n\times q})^T$ or $A = U_{m\times m} S_{m\times n} (V_{n\times n})^T$. In these alternative statements the matrix S is augmented by the inclusion of zero elements and the matrices U and V are augmented by additional orthonormal columns. In any case one would say that the matrix A has q singular values of which k are positive and $q - k$ are zero.

The following characterizations are useful for the largest singular value, s_1, the smallest positive singular value, s_k, and the smallest singular value, s_q.

(2.2.5) $$s_1 = \max\{\|Ax\| : \|x\| = 1\},$$

(2.2.6) $$s_k = \min\{\|Ax\| : \|x\| = 1 \text{ and } x \in \mathcal{R}(A^T)\},$$

and

(2.2.7) $$s_q = \min\{\|Ax\| : \|x\| = 1\}.$$

THEOREM 2.2.2. *Suppose that A is an m × n real matrix. Then there exists exactly one real n × m matrix X such that*

(2.2.8)
$$\begin{aligned} &\text{(a)} & AXA &= A, \\ &\text{(b)} & XAX &= X, \\ &\text{(c)} & (AX)^T &= AX, \\ &\text{(d)} & (XA)^T &= XA. \end{aligned}$$

The matrix $A^+ = X$ is called the pseudoinverse of A.

The proof of Theorem 2.2.2 can be found in [5].

THEOREM 2.2.3. *Let $A_{m\times n}$, $U_{m\times k}$, $B_{k\times k}$, and $V_{m\times k}$ be matrices such that B is nonsingular $A = UBV^T$, $U^TU = I_k$, and $V^TV = I_k$. Then*

(2.2.9) $$A^+ = VB^{-1}U^T.$$

This theorem is easily established by verifying that $VB^{-1}U^T$ satisfies the conditions for X of (a) through (d) of Eq. (2.2.8).

Useful special cases of Theorem 2.2.3 arise when some of the matrices U, B, or V are identity matrices. Furthermore Theorem 2.2.3 can be used to prove a similar theorem where B may now be singular and Eq. (2.2.9) becomes $A^+ = VB^+U^T$.

We list a number of useful identities involving the singular value decomposition, $A_{m\times n} = U_{m\times k} S_{k\times k} (V_{n\times k})^T$, and the pseudoinverse A^+. For convenience we assume $k \equiv \text{rank}(A) > 0$ so that $S_{k\times k}^{-1}$ exists.

(2.2.10) $$A = USV^T, \qquad A^T = VSU^T.$$

(2.2.11) $$\|A\| = \|A^T\| = \|S\| = s_1.$$

(2.2.12)
$$\begin{aligned} A^+ &= VS^{-1}U^T, & (A^+)^+ &= A, \\ (A^T)^+ &= (A^+)^T, & \|A^+\| &= \|S^{-1}\| = s_k^{-1}. \end{aligned}$$

(2.2.13) $$A^TA = VS^2V^T, \qquad AA^T = US^2U^T.$$

$$\|A^TA\| = \|AA^T\| = \|S^2\| = s_1^2 . \tag{2.2.14}$$

$$(A^TA)^+ = A^+A^{T+} = VS^{-2}V^T . \tag{2.2.15}$$

$$(AA^T)^+ = A^{T+}A^+ = US^{-2}U^T . \tag{2.2.16}$$

$$\|(A^TA)^+\| = \|(AA^T)^+\| = \|S^{-2}\| = s_k^{-2} . \tag{2.2.17}$$

$$A^+A = VV^T , \qquad AA^+ = UU^T . \tag{2.2.18}$$

$$\text{If rank } (A) = n, A^+ = (A^TA)^{-1}A^T \text{ and } A^+A = I_n . \tag{2.2.19}$$

$$\text{If rank } (A) = m, A^+ = A^T(AA^T)^{-1} \text{ and } AA^+ = I_m . \tag{2.2.20}$$

(2.2.21) *The vector $x_0 = A^+b$ is the unique minimum length solution of Problem* LS.

If $W_{m\times n}$ and $R_{n\times n}$ are each of rank n then

$$(WR)^+ = R^{-1}W^+ , \tag{2.2.22}$$

$$A^+A^{T+}A^T = A^+ , \tag{2.2.23}$$

$$A^T = A^TAA^+ . \tag{2.2.24}$$

The orthogonal projection of an m-vector b onto the range space of an $m \times n$ matrix A will be denoted by b_A and is defined by

$$b_A = AA^+b . \tag{2.2.25}$$

Note that $b_A = Ax_0$ where $x_0 = A^+b$.

If rank $(A) = k > 0$ define the spectral condition number of the matrix A, by

$$\kappa \equiv \kappa_A \equiv \text{Cond } (A) = s_1/s_k \tag{2.2.26}$$

where s_1 and s_k are respectively the largest and smallest nonzero singular values of A. If $A = 0$, so that $k = 0$, define Cond $(A) = 1$.

Note that [6] if $A \neq 0$,

$$\text{Cond } (A) = \|A\| \cdot \|A^+\| = \text{Cond } (A^T) = \text{Cond } (A^+) , \tag{2.2.27}$$

and that for any matrix A

$$\text{Cond } (A^TA) = \text{Cond } (AA^T) = [\text{Cond } (A)]^2 . \tag{2.2.28}$$

We now prove the following

THEOREM 2.2.4. *Suppose A is an $m \times n$ matrix of* rank k *and that H is an $n \times q$ matrix such that $H^TH = I_q$. If $\mathcal{R}(H) \subseteq \mathcal{R}(A^T)$, then*

$$\text{Cond } (AH) \leqq \text{Cond } (A) . \tag{2.2.29}$$

Proof. If $AH = 0$, the inequality of Eq. (2.2.29) is obvious. Suppose, then, that $AH \neq 0$. Let $\bar{s}_1 \geqq \cdots \geqq \bar{s}_q$ denote the q singular values of AH.

Then, on the one hand, $\bar{s}_1 = \|AH\| \leqq \|A\| \, \|H\| = \|A\| = s_1$, so that

$$\bar{s}_1 \leqq s_1 . \tag{2.2.30}$$

On the other hand, the smallest nonzero value of A (see Eq. (2.2.6)) is characterized by the equation

$$
\begin{aligned}
s_k &= \min_{w \in \mathcal{R}(A^T);\ w \neq 0} \frac{\|Aw\|}{\|w\|} \leqq \min_{w \in \mathcal{R}(H);\ w \neq 0} \frac{\|Aw\|}{\|w\|} \\
&= \min_{Hy;\ Hy \neq 0} \frac{\|AHy\|}{\|Hy\|} = \min_{y \neq 0} \frac{\|AHy\|}{\|y\|} = \bar{s}_q \,,
\end{aligned}
\tag{2.2.31}
$$

so that

$$s_k \leqq \bar{s}_q \,. \tag{2.2.32}$$

From Eq. (2.2.26) and the inequalities (2.2.30) and (2.2.32) the proof of Eq. (2.2.29) and Theorem 2.2.4 follow.

Note that the requirement that $\mathcal{R}(H) \subseteq \mathcal{R}(A^T)$ is certainly satisfied if rank $(A) = n$ since in that case $\mathcal{R}(A^T) = E^n$.

Also note that Theorem 2.2.4 implies that for any arbitrary $m \times n$ matrix A $(m \geqq n)$, of rank n, the $m \times (n - k)$ matrix A' formed by deleting k columns from A is such that

$$\text{Cond}\,(A') \leqq \text{Cond}\,(A) \,. \tag{2.2.33}$$

To verify this put

$$H = S \begin{bmatrix} I_{n-k} \\ 0 \end{bmatrix}, \tag{2.2.34}$$

where S is an $n \times n$ permutation matrix such that the last k columns of AS are to be deleted. Since $\mathcal{R}(H) \subseteq \mathcal{R}(A^T)$, it follows from Theorem 2.2.1, Eq. (2.2.34) and Theorem 2.2.4 that

$$\text{Cond}\,(A') = \text{Cond}\,(AH) \leqq \text{Cond}\,(AS) = \text{Cond}\,(A) \,.$$

In an analogous manner we can prove that for any arbitrary $m \times n$ matrix A, $(m \leqq n)$, of rank m, the $(m - k') \times n$ matrix A'' formed by deleting k' rows from A satisfies the inequality

$$\text{Cond}\,(A'') \leqq \text{Cond}\,(A) \,. \tag{2.2.35}$$

We summarize these remarks with

THEOREM 2.2.5. *If more independent parameters are introduced into an overdetermined linear least squares problem of full rank, then the resulting coefficient matrix has the same, or a greater, condition number.*

More precisely, and more generally, let A *be an* $m \times n$ *matrix,* $(m > n)$, *of* rank n. *Let* B *denote an* $m \times k$ $(k \leqq m - n)$, *matrix such that the augmented* $m \times (n + k)$ *matrix* $[A, B]$ *is of* rank $n + k$. Then

$$\text{Cond}\,(A) \leqq \text{Cond}\,([A, B]) \,. \tag{2.2.36}$$

2.3. *General Perturbation Theorems.* In the following discussion of perturbations we will use the notation

$$
\begin{aligned}
A_\delta &= A + dA \,, \qquad & \rho_A &= \|dA\|/\|A\| \,, \\
b_\delta &= b + db \,, \qquad & \rho_b &= \|db\|/\|b\| \,.
\end{aligned}
$$

THEOREM 2.3.1.

(a) *Suppose that* $b_A \neq 0$. *Then*

$$\|A^+b - A^+b_\delta\| \leqq \operatorname{Cond}(A) \frac{\|b - b_\delta\|}{\|b\|} \frac{\|b\|}{\|b_A\|} \|A^+b\| . \tag{2.3.1}$$

For every matrix A the inequality of Eq. (2.3.1) *may be an equality for certain vectors b and b_δ.*

(b) *Suppose that* $A \neq 0$ *and* $b_A \neq 0$. *Then*

$$\begin{aligned} \|A_\delta^+b - A^+b\| \leqq \operatorname{Cond}(A)&\left[\frac{\|A - A_\delta\|}{\|A\|} \|A_\delta^+b\| + \frac{\|b_A - b_{A\delta}\|}{\|b_A\|} \|A^+b\|\right] \\ &+ \|A^+A - A_\delta^+A_\delta\| \, \|A_\delta^+b\| . \end{aligned} \tag{2.3.2}$$

Proof. (a) We will first establish the inequality (2.3.1). The inequalities

$$\|A^+b - A^+b_\delta\| \leqq \|A^+\| \, \|b - b_\delta\| \tag{2.3.3}$$

and

$$\|b_A\| \leqq \|A\| \, \|A^+b\| \tag{2.3.4}$$

imply that

$$\begin{aligned} \|A^+b - A^+b_\delta\| &\leqq \|A\| \, \|A^+\| \frac{\|b - b_\delta\|}{\|b_A\|} \|A^+b\| \\ &= \operatorname{Cond}(A) \frac{\|b - b_\delta\|}{\|b\|} \frac{\|b\|}{\|b_A\|} \|A^+b\| , \end{aligned}$$

which proves Eq. (2.3.1).

For any matrix A this bound may be attained for certain vectors b and b_δ.

Let the columns of the matrices U and V of Eq. (1.8) be denoted by $u_1, \cdots, u_k$ and $v_1, \cdots, v_k$ respectively.

If $k < m$ let $\tilde{u}$ denote a unit vector orthogonal to all u_i, $(i = 1, \cdots, k)$. Then for $\delta > 0$ define $b = u_1 + \tilde{u}$ and $b_\delta = b + \sqrt{2}\delta u_k$. If $k = m$ define $b = u_1$ and $b_\delta = b + \delta u_k$ for $\delta > 0$.

In either case the equality condition in Eq. (2.3.1) is attained since the terms on both sides of Eq. (2.3.1) may be evaluated with:

$$\begin{aligned} \|A^+b - A^+b_\delta\| &= \sqrt{2}\delta s_k^{-1} \quad \text{if } k < m \\ &= \delta s_k^{-1} \qquad \text{if } k = m , \\ \operatorname{Cond}(A) &= s_1/s_k , \\ \|b - b_\delta\|/\|b\| &= \delta , \\ \|b\|/\|b_A\| &= \sqrt{2} \quad \text{if } k < m \\ &= 1 \qquad \text{if } k = m , \end{aligned}$$

and

$$\|A^+b\| = s_1^{-1} .$$

Part (a) of Theorem 2.3.1 is now completed. We now prove part (b).

From $b_{A_\delta} = A_\delta A_\delta^+ b$ and Eq. (2.2.8) we have the identity

$$A_\delta^+b - A^+b = A^+(A - A_\delta)A_\delta^+b - A^+(b_A - b_{A\delta}) - (A^+A - A_\delta^+A_\delta)A_\delta^+b . \tag{2.3.5}$$

With Eq. (2.3.5) and part (a) of this theorem we see that

$$\begin{aligned}\|A_\delta^+ b - A^+ b\| &\leqq \frac{\|A^+\|\,\|A\|\,\|A - A_\delta\|\,\|A_\delta^+ b\|}{\|A\|}\\ &\quad + \text{Cond }(A)\,\frac{\|b_A - b_{A\delta}\|}{\|b_A\|}\,\|A^+ b_A\|\\ &\quad + \|A^+A - A_\delta^+ A_\delta\|\,\|A_\delta^+ b\|\\ &= \text{Cond }(A)\left[\frac{\|A - A_\delta\|}{\|A\|}\,\|A_\delta^+ b\| + \frac{\|b_A - b_{A\delta}\|}{\|b_A\|}\,\|A^+ b\|\right]\\ &\quad + \|A^+A - A_\delta^+ A_\delta\|\,\|A_\delta^+ b\|\,,\end{aligned}$$

which proves Eq. (2.3.2) and completes the proof of Theorem 2.3.1.

The inequality of Eq. (2.3.1) shows that relative error in the solution can be bounded by a product of the relative error in the vector $b + db$, together with Cond (A) and the ratio $\|b\|/\|b_A\| \geqq 1$. Thus $\|b\|/\|b_A\|$ (the secant of the angle b makes with $\mathcal{R}(A)$) can be interpreted as a condition number of b relative to a given matrix A.

The inequality of Eq. (2.3.2) displays the relative error in the solution in terms of Cond (A), the relative error in the coefficient matrix, the resulting relative change in the projection of b onto $\mathcal{R}(A)$ and $\mathcal{R}(A_\delta)$, and the norm of the matrix $\|A^+A - A_\delta^+ A_\delta\|$.

THEOREM 2.3.2. *If* $\kappa\rho_A < 1$, *then* rank $(A + dA) \geqq$ rank (A).

Proof. Let $k =$ rank (A). If $k = 0$ the theorem is obviously true.

If $k > 0$ let $A = U_{m\times k}\, S_{k\times k}\, (V_{n\times k})^T$ be a singular value decomposition of A. We may write

$$U^T(A + dA)V = S + U^T(dA)V = S(I_k + P) \tag{2.3.6}$$

where $P = S^{-1}U^T(dA)V$. Then $\|P\| \leqq \|S^{-1}\|\cdot\|dA\| = \kappa\rho_A < 1$. It follows that $(I_k + P)$ is nonsingular, i.e. of rank k. Thus the rank of each of the three matrix factors on the left side of Eq. (2.3.6), one of which is $A + dA$, must be at least k, and the proof of Theorem 2.3.2 is completed.

THEOREM 2.3.3. *Let* $B = A + dA$. *If* rank $(B) >$ rank (A) *and* $\kappa\rho_A < 1$, *then*

$$\kappa_B \geqq (1/\rho_A) - 1\,.$$

Proof. Applying the contrapositive of Theorem 2.3.2 to B the condition rank (B) $>$ rank $(B - dA)$ implies that $\kappa_B(\|dA\|/\|B\|) \geqq 1$. Thus

$$\kappa_B \geqq \frac{\|B\|}{\|dA\|} \geqq \frac{\|A\| - \|dA\|}{\|dA\|} = (1/\rho_A) - 1$$

which completes the proof of Theorem 2.3.2.

One interpretation of the above theorem is that if a matrix B is very close to a matrix of lower rank, then B has a very large condition number. This provides some motivation for preferring the lowest possible rank when data errors lead to uncertainty of the rank of a matrix.

2.4. *Perturbation and Differentiability Theorems for Matrices of Full Rank.* A matrix $A_{m\times n}$ is of *full rank* if rank $(A) =$ min (m, n). For these matrices perturbation theorems can be obtained which are more explicit than the general theorems given in the previous section.

Theorem 2.4.1 extends to full rank rectangular matrices a theorem well known [3, p. 93] for square nonsingular matrices.

THEOREM 2.4.1. *Let* $k = \operatorname{rank}(A_{m\times n}) = \min\{m, n\} \geqq 1$. *If* $\kappa\rho_A < 1$ *then* $\operatorname{rank}(A + dA) = k$ *and* $\|(A + dA)^+\| \leqq \|A^+\|/(1 - \kappa\rho_A)$.

Proof. Let the singular values of A be denoted by $s_1 \geqq \cdots \geqq s_k > 0$ and those of $A + dA$ by $\tilde{s}_1 \geqq \cdots \geqq \tilde{s}_k \geqq 0$. The conclusions of the theorem are equivalent to assertions that $\tilde{s}_k > 0$ and $\tilde{s}_k^{-1} \leqq s_k^{-1}/(1 - s_k^{-1}\|dA\|)$. Thus it suffices to prove that $\tilde{s}_k \geqq s_k - \|dA\| > 0$.

The latter inequality is a direct consequence of the hypothesis $\kappa\rho_A < 1$ since $\kappa\rho_A = s_k^{-1}\|dA\|$.

Using Eq. (2.2.7) we write

$$\begin{aligned}\tilde{s}_k &= \min\{\|(A + dA)x\| : \|x\| = 1\}\\ &= \min\{\|Ax + dAx\| : \|x\| = 1\}\,.\end{aligned}$$

To evaluate this minimum note that if $\|x\| = 1$ then $\|dAx\| \leqq \|dA\| < s_k \leqq \|Ax\|$, and thus $\|Ax + dAx\| \geqq \|Ax\| - \|dAx\| \geqq s_k - \|dA\|$. Therefore $\tilde{s}_k \geqq s_k - \|dA\|$ which completes the proof of Theorem 2.4.1.

In the case of square nonsingular matrices there is a companion theorem to Theorem 2.4.1 which states that if $\kappa\rho_A < 1$, then

$$(2.4.1) \qquad \|(A + dA)^{-1} - A^{-1}\| \leqq \|A^{-1}\|\kappa\rho_A/(1 - \kappa\rho_A)\,.$$

The following theorem establishes the analogous, but somewhat more complicated, result for the pseudoinverse of a full rank rectangular matrix. The theorem is stated for the case $m \geqq n$.

THEOREM 2.4.2. *Assume* $m \geqq n \geqq 1$, $\operatorname{rank}(A) = n$ *and* $\kappa\rho_A < 1$. *Let* $P = (A + dA)^+ - A^+$. *Then*

$$(2.4.2) \qquad (a) \qquad P = A^+(BA^+ + CQ_2)\,,$$

where B, C, *and* Q_2 *are respectively* $m \times n$, $m \times (m - n)$, *and* $(m - n) \times m$ *matrices satisfying*

$$\|B\| \leqq \|dA\|/(1 - \kappa\rho_A)\,, \quad \|C\| \leqq \kappa\rho_A/(1 - \kappa\rho_A)^2\,,$$
$$Q_2Q_2^T = I_{m-n}\,, \quad Q_2A = 0$$

and furthermore

$$(2.4.3) \qquad (b) \qquad \|P\| \leqq \sqrt{2}\|A^+\|\kappa\rho_A/(1 - \kappa\rho_A)^2\,.$$

Remark. If $m = n$ then the matrix Q_2 (of dimension $(m - n) \times m$) is not present and the conclusions of Theorem 2.4.2 become $P = A^+BA^+$ and $\|P\| \leqq \|A^+\|\kappa\rho_A/(1 - \kappa\rho_A)$ in agreement with Eq. (2.4.1).

Proof. By either Theorem 2.3.2 or 2.4.1 the rank of $A + dA$ is n. Thus using Eq. (2.2.19) the pseudoinverse $(A + dA)^+$ is uniquely determined by the equation

$$(2.4.4) \qquad (A + dA)^T(A + dA)(A + dA)^+ = (A + dA)^T\,.$$

Let $P = (A + dA)^+ - A^+$. Replacing $(A + dA)^+$ by $A^+ + P$ in Eq. (2.4.4) leads to

$$(A + dA)^T(A + dA)P = (A + dA)^T - (A + dA)^T(A + dA)A^+$$

or since $A^T - A^TAA^+ = 0$

$$(2.4.5) \quad (A + dA)^T(A + dA)P = -(A + dA)^T(dA)A^+ + dA^T(I_m - AA^+) .$$

Using Eq. (2.2.23) we obtain

$$(2.4.6) \quad P = (A + dA)^+\{-(dA)A^+ + (A + dA)^{T+}(dA)^T(I_m - AA^+)\} .$$

The bound of Eq. (2.4.3) for $\|P\|$ can be obtained from Eq. (2.4.6) using Theorem 2.4.1. However, for some purposes (e.g. Theorem 2.4.5) it is advantageous to have a representation of P in which the leading factor is A^+ rather than $(A + dA)^+$. To this end let Q be an $m \times m$ orthonormal matrix, partitioned into submatrices Q_1 and Q_2 of dimensions $n \times m$ and $(m - n) \times m$, satisfying

$$QA = \begin{bmatrix} Q_1A \\ Q_2A \end{bmatrix} = \begin{bmatrix} R_{n \times n} \\ 0_{(m-n) \times n} \end{bmatrix}$$

where R is nonsingular. The existence of such matrices, Q and R, is assured by the singular value decomposition theorem. It is easily verified that

$$A = Q_1^TR , \qquad A^+ = R^{-1}Q_1 ,$$
$$R^{-1} = A^+Q_1^T , \qquad \|A\| = \|R\| ,$$

and

$$\|A^+\| = \|R^{-1}\| .$$

Let $E = (dA)R^{-1}$, then

$$A + dA = (Q_1^T + E)R .$$

Since all singular values of Q_1^T are unity and $\|E\| \leqq \kappa\rho_A < 1$, Theorem 2.4.1 implies that rank $(Q_1^T + E) = n$ and

$$\|(Q_1^T + E)^+\| \leqq 1/(1 - \kappa\rho_A) .$$

Using Eq. (2.2.22) we obtain

$$(2.4.7) \qquad (A + dA)^+ = R^{-1}(Q_1^T + E)^+ = A^+Q_1^T(Q_1^T + E)^+ .$$

Substituting Eq. (2.4.7) and $I_m - AA^+ = Q_2^TQ_2$ into Eq. (2.4.6) gives

$$P = A^+Q_1^T(Q_1^T + E)^+\{-(dA)A^+ + (A + dA)^{T+}(dA^T)Q_2^TQ_2\} .$$

Therefore

$$(2.4.8) \qquad P = A^+(BA^+ + CQ_2)$$

where

$$B = -Q_1^T(Q_1^T + E)^+(dA)$$
$$C = Q_1^T(Q_1^T + E)^+(A^T + dA^T)^+(dA^T)Q_2^T$$
$$\|B\| \leqq \|dA\|/(1 - \kappa\rho_A)$$
$$\|C\| \leqq \left[\frac{1}{1 - \kappa\rho_A}\right]\left[\frac{\|A^+\|}{1 - \kappa\rho_A}\right]\|dA\| = \kappa\rho_A/(1 - \kappa\rho_A)^2$$

which establishes assertion (a).

It follows immediately that

$$\|P\| \leqq 2\|A^+\|\kappa\rho_A/(1-\kappa\rho_A)^2$$

which is a slightly weaker result than assertion (b). To establish assertion (b) we must use the orthonormality of the matrix Q.

Let $\mathcal{U}$ denote the set of all m-vectors having unit euclidean length. Then $\|Q_1b\|^2 + \|Q_2b\|^2 = 1$ for all $b \in \mathcal{U}$, and therefore

$$\max_{b\in\mathcal{U}} (\|Q_1b\| + \|Q_2b\|) = \sqrt{2}\,.$$

Using $A^+ = R^{-1}Q_1$ in Eq. (2.4.8) we obtain

$$\begin{aligned}\|P\| &= \max_{b\in\mathcal{U}} \|Pb\| \leqq \|A^+\| \max_{b\in\mathcal{U}} \|BR^{-1}Q_1b + CQ_2b\| \\ &\leqq \|A^+\|\kappa\rho_A(1-\kappa\rho_A)^{-2} \max_{b\in\mathcal{U}} (\|Q_1b\| + \|Q_2b\|) \\ &= \sqrt{2}\|A^+\|\kappa\rho_A/(1-\kappa\rho_A)^2\end{aligned}$$

which establishes assertion (b) and completes the proof of Theorem 2.4.2.

Theorem 2.4.2 will now be applied to Problem LS. The resulting bound on $\|dx\|/\|x\|$ is consistent with the estimated bound given by Golub and Wilkinson in [8]. They did not choose to identify the separate roles of ρ_A and ρ_b or the explicit form of $\tilde{\kappa}$.

THEOREM 2.4.3. *Assume* $m \geqq n \geqq 1$, rank $(A) = n$, *and* $\kappa\rho_A < 1$. *If* x *and* $x + dx$ *are the unique vectors minimizing* $\|Ax - b\|$ *and* $\|(A + dA)(x + dx) - (b + db)\|$ *respectively and* $r = b - Ax$, *then*

$$\|dx\| \leqq \frac{\|A^+\|}{1-\kappa\rho_A}\left\{\|db\| + \|dA\|\left[\|x\| + \frac{\|A^+\|}{1-\kappa\rho_A}\|r\|\right]\right\}. \tag{2.4.9}$$

If $x \neq 0$, *then* $b \neq 0$, $b_A \equiv Ax \neq 0$, *and*

$$\frac{\|dx\|}{\|x\|} \leqq \tilde{\kappa}\rho_b\frac{\|b\|}{\|b_A\|} + \tilde{\kappa}\rho_A\left[1 + \tilde{\kappa}\frac{\|r\|}{\|b_A\|}\right] \tag{2.4.10}$$

where $\tilde{\kappa} = \kappa/(1-\kappa\rho_A)$.

Remark. If $b \in \mathcal{R}(A)$ then $\|r\| = 0$ and $b_A = b$ so that Eq. (2.4.10) reduces to

$$\|dx\|/\|x\| \leqq \tilde{\kappa}\rho_A + \tilde{\kappa}\rho_b \tag{2.4.11}$$

which is the standard result [9, p. 177] for square nonsingular matrices.

We now present the proof of Theorem 2.4.3.

By Eq. (2.2.21) the vector dx is uniquely determined by the equation

$$x + dx = (A + dA)^+(b + db)$$

where $x = A^+b$. Thus

$$dx = (A + dA)^+db + [(A + dA)^+ - A^+]b\,. \tag{2.4.12}$$

Using the representation of $[(A + dA)^+ - A^+]$ established in Theorem 2.4.2, Eq. (2.4.12) can be written as

$$dx = (A + dA)^+db + A^+BA^+b + A^+CQ_2b\,.$$

Define the residual vector $r = b - Ax$ and note that $Q_2b = Q_2(r + Ax) = Q_2r$ since $Q_2A = 0$. Thus

$$dx = (A + dA)^+db + A^+Bx + A^+CQ_2r\,.$$

Using the bound for $\|(A + dA)^+\|$ given in Theorem 2.4.1 and the bounds for B, C, and Q_2 given in Theorem 2.4.2 we obtain

$$\|dx\| \leqq \|A^+\|\left[\frac{\|db\|}{1 - \kappa\rho_A} + \frac{\|dA\|\cdot\|x\|}{1 - \kappa\rho_A} + \frac{\|A^+\|\cdot\|dA\|\cdot\|r\|}{(1 - \kappa\rho_A)^2}\right]$$

which establishes Eq. (2.4.9).

If $x \neq 0$ then also $b \neq 0$ and $b_A \equiv Ax \neq 0$. In this case dividing Eq. (2.4.9) by $\|x\|$ and using $\|b_A\| \leqq \|A\|\cdot\|x\|$ gives Eq. (2.4.10) completing the proof of Theorem 2.4.3.

Example. The following example shows that the bound of Eq. (2.4.10) is best possible in the sense that there exists A, dA, b, and db such that it can be approached arbitrarily closely. Let h, s, e, and f be positive numbers with $h \geqq 1$ and define

$$A = \begin{bmatrix} 1 & 0 \\ 0 & h^{-1} \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ s \end{bmatrix},$$

$$dA = \begin{bmatrix} 0 & 0 \\ -e & 0 \\ 0 & e \end{bmatrix}, \qquad db = \begin{bmatrix} 0 \\ f \\ 0 \end{bmatrix}.$$

Then $\|A\| = 1$, $\kappa = h$, $\|dA\| = e$, $\rho_A = e$, $\|b\| = (1 + s^2)^{1/2}$, $\|db\| = f$, $\rho_b = f(1 + s^2)^{-1/2}$.

If x minimizes $\|Ax - b\|$ then

$$x = [1, 0]^T, \qquad b_A = Ax = [1, 0, 0]^T,$$
$$r = b - Ax = [0, 0, s]^T\,.$$

Using Eq. (2.2.19) one can compute

$$(A + dA)^+ = \{1 + h^2e^2(1 + e^2)\}^{-1}\begin{bmatrix} 1 + h^2e^2 & -h^2e^3 & he^2 \\ he & h & h^2e(1 + e^2) \end{bmatrix}.$$

Then

$$\begin{aligned} x + dx &= (A + dA)^+(b + db) \\ &= \{1 + h^2e^2(1 + e^2)\}^{-1}\begin{bmatrix} 1 + (he)^2 - (he)^2(fe) + (he)(se) \\ he + hf + (he)hs + (he)^2(se) \end{bmatrix}. \end{aligned} \tag{2.4.13}$$

Up to this point no approximations have been used. We now assume that e is small in the sense that $he \ll 1$, $fe \ll 1$, and $se \ll 1$. Products containing two or more factors of the form (he), (fe), or (se) will be dropped. Then from Eq. (2.4.13) we obtain

$$x + dx \cong \begin{bmatrix} 1 \\ he + hf + (he)hs \end{bmatrix},$$

$$\|dx\|/\|x\| \cong hf + he(1 + hs)\,.$$

This should be compared with the bound of Eq. (2.4.10) which for this problem can be written as

$$\frac{\|dx\|}{\|x\|} \leqq \frac{h}{(1-he)} \frac{f}{(1+s^2)^{1/2}} \frac{(1+s^2)^{1/2}}{1} + \frac{he}{(1-he)}\left[1 + \frac{h}{(1-he)}\,\frac{s}{1}\right],$$

$$= (hf + he(1+hs))(1+(he)) .$$

For use in the proof of Theorem 2.4.5 we now state the transposed version of Theorem 2.4.2.

Theorem 2.4.4 follows directly from Theorem 2.4.2 and the third identity of Eq. (2.2.12).

THEOREM 2.4.4. *Assume* $n \geqq m \geqq 1$, rank $(A) = m$ *and* $\kappa\rho_A < 1$. *Let* $P = (A + dA)^+ - A^+$. *Then*

$$\text{(2.4.14)} \qquad \text{(a)} \qquad P = (A^+B + Q_2C)A^+$$

where B, Q_2, *and* C *are respectively* $m \times n$, $n \times (n-m)$, *and* $(n-m) \times n$ *matrices satisfying*

$$\|B\| \leqq \|dA\|/(1-\kappa\rho_A), \quad \|C\| \leqq \kappa\rho_A/(1-\kappa\rho_A)^2,$$
$$Q_2{}^TQ_2 = I_{n-m}, \quad AQ_2 = 0$$

and furthermore

$$\text{(2.4.15)} \qquad \text{(b)} \qquad \|P\| \leqq \sqrt{2}\|A^+\|\kappa\rho_A/(1-\kappa\rho_A)^2 .$$

THEOREM 2.4.5. *Assume* $n \geqq m \geqq 1$, rank $(A) = m$, *and* $\kappa\rho_A < 1$. *If* x *and* $x + dx$ *are the minimum length vectors satisfying* $Ax = b$ *and* $(A + dA)(x + dx) = b + db$ *respectively then*

$$\text{(2.4.16)} \qquad \|dx\| \leqq \frac{\|A^+\|}{1-\kappa\rho_A} + \left[\|db\| + \frac{\sqrt{2}}{1-\kappa\rho_A}\|dA\|\cdot\|x\|\right]$$

and if $b \neq 0$ *then* $x \neq 0$ *and*

$$\text{(2.4.17)} \qquad \|dx\|/\|x\| \leqq \kappa\rho_b/(1-\kappa\rho_A) + \sqrt{2}\kappa\rho_A/(1-\kappa\rho_A)^2 .$$

Proof. The vector dx is uniquely determined by the equation

$$x + dx = (A + dA)^+(b + db)$$

where $x = A^+b$. Thus

$$\text{(2.4.18)} \qquad dx = (A + dA)^+db + [(A + dA)^+ - A^+]b .$$

Using the representation of $[(A + dA)^+ - A^+]$ given in Theorem 2.4.4, Eq. (2.4.18) can be written as

$$dx = (A + dA)^+db + (A^+B + Q_2C)A^+b$$
$$= (A + dA)^+db + (A^+B + Q_2C)x .$$

Using the bound for $(A + dA)^+$ given in Theorem 2.4.1, the bounds for B and C given in Theorem 2.4.4 and the orthonormality of Q_2, we obtain

$$\text{(2.4.19)} \qquad \|dx\| \leqq \frac{\|A^+\|}{1-\kappa\rho_A}\left[\|db\| + \frac{\sqrt{2}}{1-\kappa\rho_A}\|dA\|\cdot\|x\|\right].$$

If $b \neq 0$ then $x \neq 0$ and dividing Eq. (2.4.19) by $\|x\|$ gives Eq. (2.4.17) completing the proof of Theorem 2.4.5.

Example. We give an example showing that the constant, $\sqrt{2}$, appearing in Eq. (2.4.16) cannot be reduced. Let

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad dA = \begin{bmatrix} 0 & e & 0 \\ 0 & 0 & e \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then

$$\begin{aligned} A^+ &= A^T[AA^T]^{-1} = A^T \\ (A + dA)^+ &= (A + dA)^T[(A + dA)(A + dA)^T]^{-1} \\ &= \frac{1}{1 + e^2 + e^4} \begin{bmatrix} 1 + e^2 & -e \\ e^3 & 1 \\ -e^2 & e + e^3 \end{bmatrix} \\ x &= A^+b = [0, 1, 0]^T \\ x + dx &= (A + dA)^+b = [-e, 1, e + e^3]^T/(1 + e^2 + e^4) \\ dx &= (x + dx) - x = [-1, -e - e^3, 1 + e^2]^T e/(1 + e^2 + e^4) \\ \|x\| &= 1 \\ \|dx\| &= \sqrt{2}e + O(e^3) \quad \text{for small } e \\ \|A^+\| &= \|A\| = \kappa = 1 \\ \|dA\| &= \rho_A = e\,. \end{aligned}$$

The bound given by Eq. (2.4.16) is

$$\|dx\| \leqq \sqrt{2}e/(1 - e)^2$$

illustrating that the constant, $\sqrt{2}$, cannot be reduced.

In the context of functions having matrix arguments and matrix values, a function $g(x, h)$ is the differential of a function $f(x)$ at x_0 if g is a linear function of h and

$$\lim_{\|dx\| \to 0} \frac{f(x_0 + dx) - f(x_0) - g(x_0, dx)}{\|dx\|} = 0\,.$$

This may be indicated by the notation $df(x) = g(x, dx)|_{x=x_0}$.

THEOREM 2.4.6. *For $m \geqq n \geqq 1$ let $\mathfrak{A}$ denote the set of all $m \times n$ matrices of* rank n. *The pseudoinverse operator is differentiable on $\mathfrak{A}$ with differential*

$$dA^+ = -A^+(dA)A^+ + A^+A^{T+}(dA)^T(I_m - AA^+)\,. \tag{2.4.20}$$

Proof. For $A \in \mathfrak{A}$ we must evaluate the limit as $\|dA\| \to 0$ of

$$\begin{aligned} &Z(A, dA) \\ &\equiv \frac{(A + dA)^+ - A^+ - [-A^+(dA)A^+ + A^+A^{T+}(dA)^T(I - AA^+)]}{\|dA\|}\,. \end{aligned} \tag{2.4.21}$$

Using the expression for $(A + dA)^+ - A^+$ given in Eq. (2.4.6) and writing $A_\delta = A + dA$ Eq. (2.4.21) becomes

$$Z(A, dA) = \frac{-(A_\delta^+ - A^+)(dA)A^+ + (A_\delta^+ A_\delta^T - A^+ A^{T+})(dA)^T(I - AA^+)}{\|dA\|} . \tag{2.4.22}$$

Statement (b) of Theorem 2.4.2 and statement (b) of Theorem 2.4.4 respectively assure the continuity of the pseudoinverse operator for all $A \in \mathfrak{A}$ and for all A^T such that $A \in \mathfrak{A}$. Thus for $A \in \mathfrak{A}$

$$A_\delta^+ - A^+ \to 0 \qquad \text{as } \|dA\| \to 0$$

and

$$A_\delta^+ A_\delta^{T+} - A^+ A^{T+} = (A_\delta^+ - A^+)A_\delta^{T+} + A^+(A_\delta^{T+} - A^{T+}) \to 0 \qquad \text{as } \|dA\| \to 0 \,\text{:}$$

It follows that

$$\lim_{\|dA\| \to 0} Z(A, dA) = 0$$

which establishes Theorem 2.4.6. The transposed form of Theorem 2.4.6 is

Theorem 2.4.7. *For $n \geqq m \geqq 1$ let $\mathfrak{A}$ denote the set of all $m \times n$ matrices of* rank *m. The pseudoinverse operator is differentiable on $\mathfrak{A}$ with differential*

$$dA^+ = -A^+(dA)A^+ + (I_n - A^+A)(dA)^T(A^{T+}A^+) . \tag{2.4.23}$$

3. Two Basic Householder Orthonormal Transformation Algorithms. We remark here that in Sections 1 and 2 the presentation was basically mathematical. In this third section and in Sections 4 and 5 the tempo of the paper will change from mathematical discussions to procedural definition for relevant computation methods and the presentation of related concepts.

An $m \times m$ Householder orthonormal reflection matrix [1] is of the form $Q = I_m + \beta^{-1}uu^T$ where u is an m-vector satisfying $\|u\| \neq 0$ and $\beta = -\|u\|^2/2$. We wish to determine the vector u so that for a particular given m-vector w the transformed m-vector $\tilde{w} = Qw$ has zero components in prescribed positions and leaves certain other prescribed components unchanged.

We have found that for our purposes the required action of the transformation $\tilde{w} = Qw$ can be described by three nonnegative integer parameters l, t, and m, $(l + t + 1 \leqq m)$, as follows:

Components 1 through l are to be left unchanged;
Component $l + 1$ is permitted to change;
Components $l + 2$ through $l + t + 1$ are to be left unchanged;
Components $l + t + 2$ through m are to be zeroed.

In Sections 3.1 and 3.2 algorithms are given for constructing and applying such transformation matrices. Subroutines implementing these two algorithms may be used as basic components with which a systematic set of programs can be constructed for a variety of linear algebraic computations.

In [4] Businger and Golub gave the explicit details for constructing a transformation $H = I_m + \beta^{-1}yy^T$ such that for the vector $\hat{w} = Hw$:

Components 1 through $l + t$ are to be left unchanged;
Components $l + t + 2$ through m are to be zeroed.

It is easy to see that if we let $P_{l,t}$ denote the permutation matrix which exchanges components indexed $(l + 1)$ and $(l + t + 1)$, then we will have the identity $Q = P_{l,t}HP_{l,t}$.

Thus the construction and application could be based on this identity and the algorithm of [4]. But for the purpose of having the present paper contain the operational details which will permit easy coding on a given computer we discuss the construction and application of the matrix $Q = I_m + \beta^{-1}uu^T$ explicitly.

3.1. *Constructing the Matrix* $I_m + \beta^{-1}uu^T$. Suppose that l, t, and m are given integers with $l \geqq 0$, $t \geqq 0$, and $l + t + 1 \leqq m$.

We segment a given m-dimensional column vector

$$w = [\overset{l\text{ terms}}{w_1, \cdots, w_l}, \overset{1\text{ term}}{w_p}, \overset{t\text{ terms}}{w_{l+2}, \cdots, w_{l+t+1}}, \overset{m-l-t-1\text{ terms}}{w_{l+t+2}, \cdots, w_m}]^T \tag{3.1.1}$$

and construct a Householder orthonormal transformation

$$\begin{aligned} Q &= I_m + \beta^{-1}uu^T && \text{if } \beta \neq 0 \\ &= I_m && \text{if } \beta = 0 \end{aligned} \tag{3.1.2}$$

with

$$u = [0, \cdots, 0, u_p, 0, \cdots, 0, u_{l+t+2}, \cdots, u_m]^T, \tag{3.1.3}$$

$$\sigma_u = -\operatorname{sgn}(w_p)(w_p^2 + w_{l+t+2}^2 + \cdots + w_m^2)^{1/2}, \tag{3.1.4}$$

$$u_p = w_p - \sigma_u, \tag{3.1.5}$$

$$\beta = \sigma_u \cdot u_p, \tag{3.1.6}$$

and

$$u_i = w_i, \qquad (i = l + t + 2, \cdots, m). \tag{3.1.7}$$

Then Q is orthonormal and satisfies

$$Qw = \begin{cases} w + [(u^Tw)/\beta]u & \text{if } \beta \neq 0 \\ w & \text{if } \beta = 0 \end{cases} \tag{3.1.8}$$

$$= [w_1, \cdots, w_l, \sigma_u, w_{l+2}, \cdots, w_{l+t+1}, 0, \cdots, 0]^T, \tag{3.1.9}$$

which is the desired transformation.

We now present an algorithm for computing the vector u and σ_u. This algorithm will be denoted by $H1(l, t, m, w, u_p, \sigma_u)$ where the input consists of the integers l, t, and m and the m-vector w. On output the components of w indexed $l + t + 2$ through m will be replaced by the corresponding elements of the m-vector u. The $(l + 1)$st component of u and the number σ_u will be returned as u_p and σ_u.

```
         Type: Integer            $l, t, m, i$
               Real               $w_i\ (i = 1, \cdots, m), \sigma_u, u_p$
               Double Precision   $s$
  Step Number  Description
       1       Set $s: = w_p^2$, $i: = l + t + 2$.
       2       If $i \leqq m$ set $s: = s + w_i^2$ and $i: = i + 1$. Then go
               back to step 2. Else go to step 3.
```

Comment. Some procedure for preventing underflow/overflow must usually be employed in the calculation of s. See [12] for a description of one such procedure.

```
3    Set $\sigma_u = -\mathrm{sgn}\,(w_p)\cdot s^{1/2}$.
4    Set $u_p = w_p - \sigma_u$ .
5    The vector $u$ has been calculated and Eq. (3.1.9)
     is satisfied.
```

Comment. The scalar β is implicitly available as the product $\beta = \sigma_u \cdot u_p$ and hence need not be explicitly saved.

The scalars u_i, $(i = l + t + 2, \cdots, m)$, can reside (remain) in the same storage previously used (known as) w_i, $(i = l + t + 2, \cdots, m)$. Usually it is convenient to store one of the quantities, u_p or σ_u, in the location formerly occupied by w_p, and store the other one in an extra location.

3.2. *Multiplication by* $I_m + \beta^{-1}uu^T$. Assume that Q is given as in Eq. (3.1.2) and that we wish to form the matrix product $Q \cdot c$ for a given m-dimensional column vector c. (Clearly, if we wish to form the matrix product c^TQ for a given m-dimensional row vector c^T, we may compute $c^TQ = [Qc]^T$, since Q is symmetric.)

We first note that

$$Qc = c + [(u^Tc)/\beta]u \,, \tag{3.2.1}$$

where we have already calculated the vector u.

We now describe an algorithm for computing $Q \cdot c$ and placing it in the storage previously occupied by c. This algorithm will be denoted by $H2(l, t, m, u, u_p, \sigma_u, c)$ where the input consists of the integers l, t, and m, the scalar σ_u, the m-vector c, and the (possibly) nonzero components of the m-vector u stored as follows: Components $l + t + 2$ through m of the vector u are stored in the array called u. The $(l + 1)$st component of the vector u is stored in u_p. On output the vector c will have been replaced by Qc.

```
Type: Integer            $l, t, m, i$
      Real               $c_i, (i = 1, \cdots, m), u_p, u_i, (i = l + t + 2, \cdots, m),$
                         $\sigma_u, \beta$
      Double Precision   $s$
  Step Number   Description
      1         Set $s := u_p \cdot c_p$, $i = l + t + 2$.
      2         If $i \leqq m$ set $s := s + u_i \cdot c_i$, $i := i + 1$.
                Then go back to step 2. Else go to step 3.
      3         If $s = 0$ go to step 9. Else go to step 4.
      4         Set $\beta := \sigma_u \cdot u_p$.
      5         If $\beta = 0$ go to step 9. Else go to step 6.
      6         Set $s := s/\beta$.
      7         Set $c_p := c_p + u_p \cdot s$, $i := l + t + 2$.
      8         If $i \leqq m$ set $c_i := c_i + u_i \cdot s$, $i := i + 1$. Then go
                back to step 8. Else go to step 9.
      9         The vector $c$ has been replaced by $Qc$.
```

Comment. Some procedure for preventing underflow/overflow is generally needed when computing the inner product in step 2. See [12] for such a procedure.

4. Extension of the Businger-Golub Algorithm to Permit Rank Deficiency.

4.1. *The Businger-Golub Algorithm.* This algorithm, [1], uses a sequence of

Householder orthonormal transformations H_j and column permutations S_j to reduce an $m \times n$ matrix A of rank $k = n \leqq m$ to upper triangular form. Thus

$$H_n H_{n-1} \cdots H_1 A S_1 \cdots S_n \equiv QAS = \begin{bmatrix} R \\ 0 \end{bmatrix}_{m\times n} \tag{4.1.1}$$

where $Q_{m\times m}$ and $S_{n\times n}$ are orthonormal and $R_{n\times n}$ is upper triangular and nonsingular. We see that

$$A = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix} S^T$$

and by checking conditions (2.2.8) it may be directly verified that

$$A^+ = S[R^{-1}, 0]Q = SR^{-1}[I_n, 0_{n\times(m-n)}]Q\,. \tag{4.1.2}$$

The unique solution, x_0, of Problem LS, is therefore representable as

$$x_0 = SR^{-1}[I_n, 0]Qb\,. \tag{4.1.3}$$

Working from right to left in Eq. (4.1.3) the computation of x_0 proceeds as follows:

(4.1.4)

(a) Compute $c = Qb \equiv H_n(H_{n-1} \cdots (H_1 b) \cdots)$;

(b) Let $\tilde{c}$ denote the first n components of c;

(c) Compute d as the solution of $Rd = \tilde{c}$;

(d) Permute the components of d as indicated by Sd to form x_0.

4.2. *Extension for Rank Deficiency.* We now discuss an extension of this algorithm for the case where rank $(A) = k < n$. Either $m \geqq n$ or $m < n$ is permitted.

We first note that (with exact arithmetic) all elements below the kth row in the product matrix

$$H_k \cdots H_1 A S_1 \cdots S_k \equiv QAS \tag{4.2.1}$$

will be zero. Thus

$$QAS = \begin{array}{cc} \begin{array}{cc} \overset{k}{\text{cols.}} & \overset{n-k}{\text{cols.}} \end{array} & \\ \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} & \begin{array}{l} k \text{ rows} \\ m-k \text{ rows} \end{array} \end{array} \tag{4.2.2}$$

where R_{11} is upper triangular and nonsingular.

By applying specially structured $n \times n$ Householder matrices, K_i, from the right in the order $i = k, k-1, \cdots, 1$ one can obtain

$$[R_{11}, R_{12}]K_k K_{k-1} \cdots K_1 \equiv [R_{11}, R_{12}]T = \begin{array}{cc} \begin{array}{cc} \overset{k}{\text{cols.}} & \overset{n-k}{\text{cols.}} \end{array} & \\ [R, \quad 0 \quad] & k \text{ rows} \end{array} \tag{4.2.3}$$

where $R_{k\times k}$ is upper triangular and nonsingular. The effect of the transformation K_i is to zero elements $k + 1$ through n in row i while not disturbing the zeros present in rows $i + 1$ through k and the subdiagonal zeros in rows 2 through i.

The matrix K_i may be constructed using the procedure

$$H1(i-1, k-i, n, \hat{A}^T_{i.}, *, \hat{a}_{ii}) \tag{4.2.4}$$

(see Section 3). Here $\hat{A} = \{\hat{a}_{ij}\}$ with rows $\hat{A}_{i.}$ denotes the storage array which initially contains A and at the time of this procedure call contains $[R_{11}, R_{12}]K_k \cdots K_{i+1}$. The asterisk in Eq. (4.2.4) denotes an auxiliary computer storage location.

Combining Eqs. (4.2.2) and (4.2.3) gives

$$(4.2.5)\qquad QAST = \begin{array}{c} \begin{array}{cc} k & n-k \\ \text{cols.} & \text{cols.} \end{array} \\ \left[\begin{array}{cc} R & 0 \\ 0 & 0 \end{array}\right] \end{array} \begin{array}{l} \\ k \text{ rows} \\ m-k \text{ rows} \end{array} = \left[\begin{array}{c} I_k \\ 0_{(m-k)\times k} \end{array}\right] R[I_k, 0_{k\times(n-k)}]$$

where $Q_{m\times m}$ and $(ST)_{n\times n}$ are orthonormal and $R_{k\times k}$ is nonsingular and upper triangular. Consequently A has the decomposition

$$(4.2.6)\qquad A = Q^T[I_k, 0]^T R[I_k, 0]T^T S^T$$

and it may be verified, using Eq. (2.2.8) that

$$(4.2.7)\qquad A^+ = ST[I_k, 0]^T R^{-1}[I_k, 0]Q\,.$$

Thus the minimum length solution, x_0, of Problem LS is representable as

$$(4.2.8)\qquad x_0 = ST[I_k, 0]^T R^{-1}[I_k, 0]Qb\,.$$

Working from right to left in Eq. (4.2.8) the vector x_0 can be computed as follows:

(a) Compute $c = Qb \equiv H_k(H_{k-1} \cdots (H_1 b) \cdots)$;

(b) Let $\tilde{c}$ denote the first k components of c ;

(c) Compute $\tilde{d}$ by solving $R\tilde{d} = \tilde{c}$;

(4.2.9) (d) Define $d = \left[\begin{array}{c} \tilde{d} \\ 0 \end{array}\right] \begin{array}{l} k \text{ rows} \\ n-k \text{ rows} \end{array}$;

(e) Compute $e = Td \equiv K_k(K_{k-1} \cdots (K_1 d) \cdots)$;

(f) Permute the components of e as indicated by Se to form x_0 .

It may be verified that the matrices R in Eq. (4.1.1), R_{11} in Eq. (4.2.2), and R in Eq. (4.2.3) have the same set of nonzero singular values as A and thus the same rank and condition number.

4.3. *The Complete Solution.* Using the orthonormality of (ST) and Eq. (4.2.6) it can be verified that the $n \times (n-k)$ matrix H defined by

$$(4.3.1)\qquad H_{n\times(n-k)} = ST[0_{(n-k)\times k}, I_{n-k}]^T$$

satisfies

$$(4.3.2)\qquad H^T H = I_{n-k}$$

and

$$(4.3.3)\qquad AH = 0_{m\times(n-k)}\,.$$

Thus the complete general solution of Problem LS can be written as

$$(4.3.4)\qquad x = x_0 + Hy$$

where y is an arbitrary $(n - k)$-dimensional vector. For any $(n - k) \times (n - k)$ orthonormal matrix B, the matrix

$$\tilde{H} = HB = ST[0_{k\times(n-k)}, B^T]^T$$

also satisfies Eqs. (4.3.2) and (4.3.3) and thus could be used as in Eq. (4.3.4) to provide an alternative parametrization of the general solution of Problem LS.

Generally it is not necessary to compute H explicitly but rather a product of the form HU or UH is needed. Using Eq. (4.3.1) we see that $Y_{n\times q} = H_{n\times(n-k)}U_{(n-k)\times q}$ can be computed as follows:

(a) Define

$$V_{n\times q} = \begin{bmatrix} 0_{k\times q} \\ U \end{bmatrix}.$$

(b) Compute $W_{n\times q} = TV \equiv K_k(K_{k-1} \cdots (K_1 V) \cdots)$.

(c) Permute the rows of W as indicated by SW to form $Y_{n\times q}$.

Similarly $Y_{q\times(n-k)} = U_{q\times n}H_{n\times(n-k)}$ can be computed as follows:

(a) Permute the columns of U as indicated by US to form $V_{q\times n}$.

(b) Compute $W_{q\times n} = VT \equiv (\cdots(VK_k) \cdots K_2)K_1$.

(c) Define $Y_{q\times(n-k)}$ to be the matrix consisting of the last $(n - k)$ columns of W.

4.4 *Matrix Replacement, Pseudorank, and Column Interchanges.* In actual applications in which Problem LS arises it is usually appropriate to regard the matrix A as merely a representative member of some set of matrices, $\mathcal{A}$, which are permissible replacements for A.

We mention three possible reasons for this. It is beyond the scope of this paper to elaborate on these points.

(a) The coefficients of $[A, b]$ may be the result of measurements or previous computations of limited precision.

(b) If the matrix A arises from the linearization of a nonlinear function, f, (as in Newtonian iteration) then the validity of Ax as an approximator for $f(t + x) - f(t)$ generally decreases as $\|x\|$ increases. Thus if a small change in A can produce a relatively large reduction in $\|x\|$ such a change may be desirable.

(c) The columns of A may represent effects (basis functions) which are too highly correlated to be distinguishable using the available data set.

There are various ways in which the set $\mathcal{A}$ may be defined and various criteria according to which one might decide to replace A by a different member of $\mathcal{A}$. In any such method the rank of A is an essentially irrelevant concept. The relevant quantity is the rank of the matrix $\hat{A}$ which replaces A. We define the *pseudorank* of A, relative to a particular computational procedure, to be the rank of the matrix $\hat{A}$ which that procedure selects and uses in place of A. To call attention to the role of a tolerance parameter, say ϵ, in a computational procedure we may use the term *ϵ-pseudorank.*

For an arbitrary $m \times n$ matrix A with column vectors a_j, $(j = 1, \cdots, n)$, let

$$\|A\|_c = \max_j \|a_j\| . \tag{4.4.1}$$

Eq. (4.4.1) defines a matrix norm.

We assume that A satisfies $\|A\|_c = 1$ and that a positive number ϵ is given which determines the set $\mathcal{A}$ according to

$$\mathcal{A} = \{A + dA : \|dA\|_c \leqq \epsilon\} .$$

The arithmetic operations and computer storage will be selected to have a relative precision somewhat smaller than ϵ.

The parameter ϵ will be used in the computation as follows: Suppose the $2q$ transformations $S_1, H_1, \cdots, S_q, H_q$ of Eq. (4.2.1) have been computed and applied yielding

$$\tilde{A} = H_q \cdots H_1 A S_1 \cdots S_q .$$

Let

$$c_j = \left[\sum_{i=q+1}^{m} \tilde{a}_{ij}^2 \right]^{1/2} , \qquad (j = q + 1, \cdots, n) .$$

Let p denote the smallest index such that

$$c_p = \max \{c_j : j = q + 1, \cdots, n\} .$$

If $c_p \leqq \epsilon$ we propose to treat all of the numbers $\tilde{a}_{ij}; i = q + 1, \cdots, m; j = q + 1, \cdots, n$ as zero.

This amounts to replacing $\tilde{A}$ by $\tilde{A} + d\tilde{A}$ with $\|d\tilde{A}\|_c \leqq \epsilon$. This is equivalent to replacing A by $A + dA$ with $\|dA\|_c \leqq \epsilon$ since

$$dA = H_1 \cdots H_q d\tilde{A} S_q \cdots S_1$$

and

$$\|H_1 \cdots H_q d\tilde{A} S_q \cdots S_1\|_c = \|dA\|_c \leqq \epsilon . \tag{4.4.2}$$

The equality in Eq. (4.4.2) follows from the fact that the H_j are orthonormal matrices while the S_j are column permutations.

Furthermore Theorem 2.2.4 may be used to show that Cond $(\tilde{A} + d\tilde{A}) \leqq$ Cond $(\tilde{A})$, and since the condition number is invariant under orthonormal transformations, Cond $(A + dA) \leqq$ Cond (A).

If $c_p \leqq \epsilon$ we define the pseudorank, k, of A to be the integer q and proceed to the backward elimination phase of the computation, indicated by Eq. (4.2.3).

If $c_p > \epsilon$, columns p and $q + 1$ of $\tilde{A}$ are interchanged, the matrix S_{q+1} being defined to represent this operation. The computation then proceeds to the construction of H_{q+1}. Note that the $(q + 1)$st diagonal element of $H_{q+1}\tilde{A}S_{q+1}$ will be either c_p or $-c_p$ and constitutes the $(q + 1)$st diagonal element of the matrix R_{11} of Eq. (4.2.2). It follows that all diagonal elements of R_{11} exceed ϵ in magnitude.

The diagonal elements of R, as defined by Eq. (4.2.3), also exceed ϵ in magnitude since during the operations indicated in Eq. (4.2.3) the magnitude of the ith diagonal element may be increased but not decreased on multiplication by K_i and is unaffected by the matrices $K_j, j \neq i$.

5. Other Linear Algebraic Computations.

5.1. *Multiple Right Sides.* The matrices H_j, S_j, K_i, and R depend on A and ϵ and are independent of b. Once the matrices H_j, S_j, K_i, and R have been computed and stored in compact form the procedure (4.2.9) can be executed for any number of different vectors b.

5.2. *Computation of the Pseudoinverse*, A^+. It is not necessary to compute A^+ explicitly since products of the form A^+B or BA^+ can be computed by appropriate interpretation of Eq. (4.2.7) as for example the computation of A^+b by the procedure of Eq. (4.2.9). However, if one does wish to compute A^+, this can be accomplished by computing A^+I_m using the procedure of Eq. (4.2.9) to compute each column of the product.

5.3. *Triangularization as a Preliminary Step*. The $m \times (n + 1)$ matrix $[A, b]$ can be transformed by orthogonal transformations to a $p \times (n + 1)$ upper triangular matrix $[C, d]$, with $p = \min\,(m, n + 1)$, as a preliminary step before beginning the algorithm of Section 4 or other algorithms related to Problem LS.

The relevance of this preliminary step depends upon the fact that if $[A, b]$ and $[C, d]$ are related by

$$Q_{m\times m}[A, b] = \begin{array}{cl} \begin{array}{cc} \scriptstyle n \atop \scriptstyle \text{cols.} & \scriptstyle 1 \atop \scriptstyle \text{col.} \end{array} & \\ \begin{bmatrix} C & d \\ 0 & 0 \end{bmatrix} & \begin{array}{l} p \text{ rows} \\ m - p \text{ rows} \end{array} \end{array} \tag{5.3.1}$$

where Q is orthonormal, then $[A, b]$ and $[C, d]$ have a number of significant properties in common. Specifically it can be verified that

$$\|Ax - b\| = \|Cx - d\| \qquad \text{for all } x\,,$$

that C and A have the same set of nonzero singular values and thus the same rank and condition number, and that

$$[A, b]^T[A, b] = [C, d]^T[C, d]\,.$$

Therefore $[C, d]$ can be used in place of $[A, b]$ in many algorithms related to Problem LS.

This transformation can be organized so that rows or groups of rows of $[A, b]$ are processed sequentially in forming $[C, d]$. The storage requirement for this processing is just the storage required for $[C, d]$ plus the storage required for the largest block of rows of $[A, b]$ which is to be introduced at any one stage. The minimum storage requirement would be $(n + 1)(n + 2)/2$ locations for the upper triangular matrix $[C, d]$ plus $n + 1$ locations for a single row of $[A, b]$.

This transformation can be used to overcome storage constraints when $m \gg n$ and the main computer storage is large enough to accommodate $(n + 1)(n + 4)/2$ locations but not large enough to accommodate the $m(n + 1)$ elements of $[A, b]$ simultaneously. It is also useful for the reduction of total execution time, with or without the sequential organization, if $m \gg n$ and one wishes to process $[A, b]$ more than once, say in combination with other data sets or by different algorithms in order to obtain different types of information.

We note for comparison that in the more widely known A^TA approach to solving Problem LS it is also possible to process the rows of $[A, b]$ sequentially in computing the $(n + 1) \times (n + 1)$ matrix

$$B \equiv \begin{array}{rc} & \begin{array}{cc} \scriptstyle n \atop \scriptstyle \text{cols.} & \scriptstyle 1 \atop \scriptstyle \text{col.} \end{array} \\ \begin{array}{r} n \text{ rows} \\ 1 \text{ row} \end{array} & \begin{bmatrix} N & v \\ v^T & \gamma \end{bmatrix} \end{array} = [A, b]^T[A, b]\,,$$

and since B is symmetric it suffices to compute and store only the $(n+1)(n+2)/2$ upper triangular submatrix of B, say $\tilde{B}$. However since Cond (C) = Cond (A) and Cond (N) = $[\text{Cond }(A)]^2$, the computer word length which is adequate for storing elements of $[A, b]$ will also be adequate for storing elements of $[C, d]$, *whereas the elements of* $\tilde{B}$ *will require up to twice that word length.*

If the entire matrix $[A, b]$ can be treated as a whole then the transformation can be accomplished by constructing and applying Householder matrices H_i in the order $i = 1, \cdots, p$ with $p = \min(m, n+1)$, as follows:

$$H_p \cdots H_1[A, b] \equiv Q[A, b] = \begin{bmatrix} C & d \\ 0 & 0 \end{bmatrix} \begin{array}{l} p \text{ rows} \\ m - p \text{ rows} \end{array} .$$

For sequential processing let $[A, b]$ be partitioned in the form

(5.3.2)
$$[A, b] = \begin{bmatrix} A_1 & b_1 \\ \vdots & \vdots \\ A_q & b_q \end{bmatrix} \begin{array}{l} m_1 \text{ rows} \\ \vdots \\ m_q \text{ rows} \end{array} .$$

Define

(5.3.3)
$$\mu = \max_t m_t \qquad s_0 = 0$$

and

$$s_t = \sum_{\tau=1}^{t} m_\tau , \qquad t = 1, \cdots, q .$$

Let G denote a computer storage array containing ν rows and $n + 1$ columns where

(5.3.4)
$$\nu = \max_{1 \leq t \leq q} \{m_t + \min[n+1, s_{t-1}]\} \leq n + 1 + \mu .$$

Let $G(i_1, i_2)$ denote the subarray of G consisting of rows i_1 through i_2.

Step 1. Set $t = 1$, and $\beta_0 = 0$.

Step 2. Set $\gamma = \beta_{t-1} + m_t$.
Store $[A_t, b_t]$ into $G(\beta_{t-1} + 1, \gamma)$.
Set $\beta_t = \min\{n + 1, \gamma\}$. The integer β_t can replace β_{t-1} in storage.

(5.3.5) *Step* 3. Construct and apply Householder matrices to reduce the contents of $G(1, \gamma)$ to upper triangular form:

$$H^{(t)}_{\beta t} \cdots H_1^{(t)} G(1, \gamma) \rightarrow G(1, \beta_t) .$$

Step 4. If $t = q$, terminate with the transformed upper triangular matrix $[C, d]$ stored in $G(1, \beta_q)$. Otherwise replace t by $t + 1$ and go to step 2.

This procedure can be interpreted as effecting the orthonormal transformation

(5.3.6)
$$Q_q P_q \cdots Q_2 P_2 Q_1 [A, b] \equiv Q[A, b] \rightarrow \begin{bmatrix} C & d \\ 0 & 0 \end{bmatrix} \begin{array}{l} \beta_q \text{ rows} \\ m - \beta_q \text{ rows} \end{array} ,$$

where

$$\tilde{Q}_t = H_{\beta_t}^{(t)} \cdots H_1^{(t)} \qquad (t = 1, \cdots, q) ,$$

$$Q_t = \begin{array}{c} \\ \left[\begin{matrix} \tilde{Q}_t & 0 \\ 0 & I \end{matrix}\right] \end{array} \begin{array}{l} \beta_{t-1} + m_t \text{ rows} \\ m - \beta_{t-1} - m_t \text{ rows} \end{array} \qquad (t = 1, \cdots, q) ,$$

and P_t is a permutation matrix which moves the submatrix $[A_t, b_t]$ from its position in rows $s_t - m_t + 1$ through s_t to rows $\beta_{t-1} + 1$ through $\beta_{t-1} + m_t$ while moving rows $\beta_{t-1} + 1$ through s_{t-1} (which contain only zeros) to rows $\beta_{t-1} + m_t + 1$ through s_t.

After the first $n + 1$ rows of $[A, b]$ have been processed, the subdiagonal elements in rows 2 through $n + 1$ of G will be zero and will remain zero thereafter. The procedures $H1$ and $H2$ of Section 3 are parametrized so that arithmetic operations involving these zero elements can be avoided.

If $m > n + 1$ and $\mu = 1$ the triangularization procedure described above requires $(n + 2)(n + 1)$ storage locations. This requirement can be reduced to $(n + 4)(n + 1)/2$ by modifying the procedure so that the subdiagonal elements in rows 2 through $n + 1$ of G are always zero (and thus require no storage) rather than only being zero after the first $n + 1$ rows of A have been processed. It suffices to replace steps 1 and 2 of the triangularization procedure by:

Step 1′. Set $t = 1$, and $\gamma = n + 2$. In this modified procedure it is assumed that $m_t = 1$ for all t.

Step 2′. Store $[A_t, b_t]$ into $G(n + 2, n + 2)$. Set $\beta_t = \min(n + 1, t)$. The integer β_t can replace β_{t-1} in storage.

Whenever $m_t = 1$ one could use Jacobi plane rotation (or reflection) matrices instead of Householder matrices to achieve a slight reduction in the number of arithmetic operations and possibly a slight reduction of roundoff error.

5.4. *Triangularization of a Block Diagonal Matrix.* Suppose A has a block diagonal structure in the sense that there exists an integer $w < n$, a partitioning as in Eq. (5.3.2), and a nondecreasing set of integers, $\jmath_1, \cdots, \jmath_q$, such that all nonzero elements in the submatrix A_t occur in columns $\jmath_t$ through $\jmath_t + w - 1$. For this case it can be verified that all nonzero elements in the ith row of the upper triangular matrix C of Eq. (5.3.6) will occur in columns i through $i + w - 1$. Furthermore rows 1 through $\jmath_t - 1$ of C will not be affected by the processing of the submatrices $A_t, \cdots, A_q$ in algorithm (5.3.5).

Taking these observations into account algorithm (5.3.5) can be modified so that the number of columns required in the storage array G is only $w + 1$. At termination of the algorithm the matrix C will be in the array G with

$$g_{ij} = c_{i,i+j-1} , \qquad i = 1, \cdots, n + 1; j = 1, \cdots, w .$$

If the pseudorank of C is n then the solution can be computed by back substitution with no need to occupy additional storage.

This procedure can result in substantial saving of storage. For example the problem of fitting 500 data points using 32 cubic polynomials with continuity of the polynomial values and first derivatives at the 31 prespecified breakpoints leads to a 500 × 67 matrix $[A, b]$ with 33,500 elements. With appropriate parametrization of the polynomials the bandwidth, w, of A will be 4. Therefore this problem can be processed in a $(67 + \mu) \times 5$ storage array ($335 + 5\mu$ elements) where μ is defined by Eq. (5.3.3) and can be taken to be 1 for greatest economy of storage. If each submatrix $[A_t, b_t]$ is taken to have as many rows as possible without forcing an increase in w then the number of arithmetic operations will be the same as in the operation-count-economizing procedure of [15]. (The running time for the above example on an IBM 7094-I using full double precision arithmetic was 1.83 seconds.)

5.5. *The Case of* $m < n$. The extended algorithm of Section 4 can be applied to Problem LS regardless of whether $m \geqq n$ or $m < n$. However if $m < n$, and particularly if it is expected that the pseudorank k will be m, then it is somewhat more natural to apply a transposed form of the extended algorithm.

Applying the extended algorithm to A^T one obtains in place of Eq. (4.2.6)

$$A = ST[I_k, 0]^T R^T [I_k, 0]Q \tag{5.5.1}$$

and thus

$$A^+ = Q^T[I_k, 0]^T(R^T)^{-1}[I_k, 0]T^T S^T . \tag{5.5.2}$$

Then the minimum length solution, x_0, of Problem LS is representable as

$$x_0 = Q^T[I_k, 0]^T(R^T)^{-1}[I_k, 0]T^T S^T b . \tag{5.5.3}$$

Equation (5.5.3) can be translated into an algorithm just as Eq. (4.2.8) was translated into the algorithm (4.2.9). These details are omitted.

Defining

$$H_{n\times(n-k)} = Q^T[0_{(n-k)\times k}, I_{n-k}]^T \tag{5.5.4}$$

the general solution of Problem LS can be written as $x = x_0 + Hy$ where x_0 is defined by Eq. (5.5.3) and y is an arbitrary $(n - k)$-dimensional vector.

This transposed form of the algorithm is particularly appropriate when $k = m$ for then Eq. (5.5.1) reduces to

$$A = SR^T[I_m, 0]Q \tag{5.5.5}$$

with a corresponding simplification of Eqs. (5.5.2) and (5.5.3).

5.6. *Linear Least Squares with Linear Equality Constraints.* Consider Problem LS subject to the additional condition that

$$Cx = d \tag{5.6.1}$$

where C is an $m' \times n$ matrix and d is an m' dimensional vector in the range space of C and the ϵ-pseudorank, k', of C is less than n.

The extended algorithm (preferably in its transposed form if $m' < n$, as will commonly be the case) can be applied to $[C, d]$ to compute an n-vector $\hat{x}$ and an $n \times (n - k')$ matrix H such that the general solution of Eq. (5.6.1) can be expressed as

$$x = \hat{x} + Hy \tag{5.6.2}$$

with y arbitrary. Substituting Eq. (5.6.2) into Problem LS gives

$$\|Ax - b\| = \|Fy - g\|$$

where

$$F = AH \tag{5.6.3}$$

and

$$g = b - A\hat{x}.$$

The extended algorithm can be applied to $[F, g]$ to compute a vector y_0 which minimizes $\|Fy - g\|$. Substituting this y_0 into Eq. (5.6.2) gives the final solution

$$x_0 = \hat{x} + Hy_0\,. \tag{5.6.4}$$

Remarks. (a) If the ϵ-pseudorank, k, of F is less than $n - k'$ then y_0 will be the minimum length solution to the problem of minimizing $\|Fy - g\|$ and x_0 will be the minimum length solution of Problem LS subject to Eq. (5.6.1).

(b) The matrix H need not be computed explicitly. The multiplications involving H in Eqs. (5.6.3) and (5.6.4) can be computed as indicated in Section 4.3 or similarly based on Eq. (5.5.4).

(c) The multiplications involving H are stable with respect to roundoff error since the columns of H are orthonormal. By Theorem 2.2.4, there are conditions, such as A being of rank n, which assure that Cond $(F) \leqq$ Cond (A).

Jet Propulsion Laboratory
MS 125-109a
4800 Oak Grove Drive
Pasadena, California 91103

1. P. Businger & G. Golub, "Linear least squares solutions by Householder transformations," *Numer. Math.*, v. 7, 1965, pp. 269–276. MR **31** #862.

2. E. E. Osborne, "On least squares solutions of linear equations," *J. Assoc. Comput. Mach.*, v. 8, 1961, pp. 628–636. MR **24** #B2543.

3. G. Forsythe & C. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, N. J., 1967. MR **36** #2306.

4. G. H. Golub & P. Businger, *Least Squares, Singular Values and Matrix Approximations; An ALGOL Procedure for Computing the Singular Value Decomposition*, Stanford Computer Sciences Department, Technical Report No. CS73, July 1967, (Mimo., 12 leaves).

5. T. N. E. Greville, "The pseudo-inverse of a rectangular or singular matrix and its application to the solution of systems of linear equations," *SIAM Rev.*, v. 1, 1959, pp. 38–43. MR **21** #424.

6. A. Ben-Israel, "On error bounds for generalized inverses," *SIAM J. Numer. Anal.*, v. 3, 1966, pp. 585–592. MR **35** #6344.

7. J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N. J., 1963. MR **28** #4661.

8. G. H. Golub & J. H. Wilkinson, "Note on the iterative refinement of least squares solution," *Numer. Math.*, v. 9, 1966, pp. 139–148. MR **35** #3849.

9. J. N. Franklin, *Matrix Theory*, Prentice-Hall, Englewood Cliffs, N. J., 1968.

10. G. W. Stewart III, "On the continuity of the generalized inverse," *SIAM J. Appl. Math.*, v. 17, 1969, pp. 33–45.

11. P. J. Davis, *Orthonormalizing Codes in Numerical Analysis. Survey of Numerical Analysis*, McGraw-Hill, New York, 1962, pp. 347–379. MR **25** #734.

12. R. S. Martin, C. Reinsch & J. H. Wilkinson, "Householder's tridiagonalization of a symmetric matrix," *Numer. Math.*, v. 11, 1968, pp. 181–195.

13. E. E. Osborne, "Smallest least squares solutions of linear equations," *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, v. 2, 1965, pp. 300–307. MR **32** #4834.

14. T. L. Jordan, "Experiments on error growth associated with some linear least-squares procedures," *Math. Comp.*, v. 22, 1968, pp. 579–588. MR **37** #4947.

15. J. K. Reid, "A note on the least squares solution of a band system of linear equations by Householder reductions," *Comput. J.*, v. 10, 1967, pp. 188–189. MR **35** #5130.

16. A. Björck, "Solving linear least squares problems by Gram-Schmidt orthogonalization," *Nordisk Tidskr. Informations-Behandling*, v. 7, 1967, pp. 1–21. MR **35** #5126.

17. V. Pereyra, "Stabilizing Linear Least Squares Problems," Proceedings IFIP 68, p. 127, (1968).

18. D. K. Faddeev, V. N. Kublamovskaya & V. N. Faddeeva, *Sur Les Systèmes Linéaires Algébriques de Matrices Rectangulaires et Mal-Conditionnées*, Programmation en Mathématiques Numériques, Editions Centre Nat. Recherche Sci., Paris, 1968, pp. 161–170. MR **37** #6017.