Experiments on Error Growth Associated with Some Linear Least-Squares Procedures

By T. L. Jordan

Abstract. Some numerical experiments were performed to compare the performance of procedures for solving the linear least-squares problem based on Gram-Schmidt, Modified Gram-Schmidt, and Householder transformations, as well as the classical method of forming and solving the normal equations. In addition, similar comparisons were made of the first three procedures and a procedure based on Gaussian elimination for solving an $n \times n$ system of equations. The results of these experiments suggest that: (1) the Modified Gram-Schmidt procedure is best for the least-squares problem and that the procedure based on Householder transformations performed competitively; (2) all the methods for solving least-squares problems suffer the effects of the condition number of A^TA, although in a different manner for the first three procedures than for the fourth; and (3) the procedure based on Gaussian elimination is the most economical and essentially, the most accurate for solving $n \times n$ systems of linear equations. Some effects of pivoting in each of the procedures are included.

1. Introduction. As part of a periodic review of basic subroutines issued for general use at the Los Alamos Scientific Laboratory, three common methods and a variant of one of these methods were compared for solving linear least-squares problems. Because of the well-known numerical difficulties encountered with least-squares problems, the primary test problem used a matrix made up of columns from the inverse of a Hilbert segment. This difficult test problem was selected anticipating that differences in methods and implementation would be magnified. The results verify that this is the case.

The calculations were performed on the Stretch (IBM-7030) computer, a 64-bit binary machine with a 48-bit floating point mantissa. All the calculations were done in single precision with the exception of certain inner products (computed with double-precision accumulation and rounded to single-precision). The test data had to be representable exactly in the machine and the results had to be known, because the effects due to error in the input data can completely overshadow the effect of rounding errors [1].

The problem is that of finding the least-squares solution to $Ax = b$. The example discussed at greatest length is that in which A is taken as the first five columns of the inverse of the 6×6 segment of the Hilbert matrix. The right-hand column, b, is taken so that the solution is 1, 1/2, 1/3, 1/4, 1/5. In this example the matrices A and b have exact representations in the Stretch computer, thereby ensuring that all error is generated in performing the least-squares procedures. Reinforcement of the solution by iterating on the error vector is not pertinent to this presentation [2].

Received August 30, 1967. Revised November 17, 1967.

2. Methods.

I. *Solution of Normal Equations* (*NE*). For completeness, the more classical approach to solving least-squares problems, that of forming the normal equations and solving the resultant linear system by Gaussian elimination, has been included in the experiments. Thus, one forms $A^TAx = A^Tb$, where $A^TA = [(A_i, A_j)]$, $A^Tb = [(A_i, b)]$, A_i is the ith column vector of A, and (x, y) is the inner product of vectors x and y. The matrices A^TA and A^Tb are computed with double-precision accumulation and rounded to single-precision. The normal equations are then solved by LSS, a Los Alamos subroutine for solving systems of linear equations. The pertinent characteristics of the LSS subroutine are:

(A) At each stage of the reduction to upper triangular form, the subroutine performs maximal pivoting only within a column (partial pivoting). This is a limitation imposed by doing the extra precision described in the following paragraph.

(B) Extra precision is achieved in the reduction to upper triangular form by saving the necessary coefficients for each reduction and then calculating and storing the reduced elements only once. The new elements are linear combinations of preceding reduced elements and are accumulated in double precision. More precisely, the procedure for the triangular reduction, ignoring pivot determination, is:

For $k = 2, 3, \cdots, n$

For $i = 2, 3, \cdots, m$

$$\begin{aligned} s_{i,k-1} &= 0 && \text{if } i < k \\ &= a'_{i,k-1}/a'_{k-1,k-1} && \text{if } i \geqq k \end{aligned}$$

$$a'_{i,k} = a_{i,k} - \sum_{j=1}^{k-1} s_{i,j}a'_{j,k},$$

where $a'_{1,j} = a_{1,j}$ for $j = 1, 2, \cdots, n$, $a'_{i,1} = a_{i,1}$ for $i = 1, 2, \cdots, m$, and m and n are the numbers of rows and columns of A, respectively.

(C) Extra precision is achieved in the back substitution by doing similar linear combinations in double precision.

II. *Householder Orthogonal Transformations* (*HH*). The method described in [3] was programmed without automatic pivoting; however, various column arrangements of A were tried including least- and most-optimal pivoting using the strategy defined in [3]. In fact, although little difference appears due to pivoting, the least-optimal pivoting produced the better results on the primary test problem.

The method may be summarized by the following procedure:

For $k = 1, 2, \cdots, n$

$$\sigma^{(k)} = \left[\sum_{i=k}^{m} (a_{i,k}^{(k)})^2\right]^{1/2}$$

$$\beta^{(k)} = 1/[\sigma^{(k)}(\sigma^{(k)} + |a_{k,k}^{(k)}|)]$$

For $i = 1, 2, \cdots, m$

$$\begin{aligned} u_i^{(k)} &= 0 && i < k \\ &= \operatorname{sgn}(a_{k,k}^{(k)})[\sigma^k + |a_{k,k}^{(k)}|] && i = k \\ &= a_{i,k}^{(k)} && i > k \end{aligned}$$

$$A^{(k+1)} = A^{(k)} - u^{(k)}(\beta^{(k)}u^{(k)T}A^{(k)})$$

$$b^{(k+1)} = b^{(k)} - u^{(k)}(\beta^{(k)}u^{(k)T}b^{(k)}),$$

where m is the number of rows and n is the number of columns.

When applied to A and b, this orthogonal transformation produces an $n \times n$ upper triangular system $Rx = c$, whose solution is the least-squares solution to $Ax = b$. The upper triangular system $Rx = c$ is solved by back substitution.

Concerning implementation, it should be pointed out where double-precision inner products are computed. These places are

(1) in the calculation of $\sigma^{(k)}$,

(2) in the calculation of $u^{(k)T}A^{(k)}$ and $u^{(k)T}b^{(k)}$, and

(3) in the back substitution in the same fashion as done in the linear system subroutine LSS.

The pivoting strategy described in [3] chooses at the kth stage the column of $A^{(k)}$ which will maximize $|A_{k,k}^{(k+1)}|$. This is equivalent to an interchange of columns k to m, such that $\sum_{i=k}^{m} (a_{i,j}^{(k)})^2$ is maximized. The invariance of column lengths under orthogonal transformations makes this a simple calculation once the original column lengths are computed.

III. *Column Orthogonalization.* This method transforms the column vectors of A into an orthogonal set and then orthogonalizes b with respect to this new set of orthogonal vectors [4]. A geometrical interpretation of the classical description of the Gram-Schmidt orthogonalization procedure is the following: at the rth stage make the rth column vector orthogonal to each of the $r - 1$ previously orthogonalized columns vectors, and do this for column vectors indexed $r = 2, 3, \cdots, n$. A variant procedure for obtaining the same set of vectors has the following geometrical interpretation. At the rth stage, make the $(n - r + 1)$ column vectors indexed r, $r + 1, \cdots, n$ orthogonal to the $(r - 1)$th column vector, and do this for column indices $r = 2, 3, \cdots, n$. Since there are large numerical differences in the experimental results, depending upon which interpretation is implemented, the procedures shall be referred to as Gram-Schmidt and Modified Gram-Schmidt, respectively.

A recent paper by Björck [5] includes an error analysis of the Modified Gram-Schmidt orthogonalization procedure. The pertinent numerical results of those that follow support nicely the conclusions of Björck.

Gram-Schmidt Orthogonalization (*GS*). The code written to demonstrate the classical Gram-Schmidt approach assumes as input an augmented matrix

$$Q = \left\| \begin{array}{c|c} A & -b \\ \hline I & 0 \end{array} \right\|, \tag{1}$$

where Q is $(m + n) \times (n + 1)$. Let $Q_i^{(1)}$ and $A_i^{(1)}$ designate the columns of Q and A, respectively. The transformation performed on Q is given by the procedure:

$Q_1' = Q_1^{(1)}$
For $j = 2, 3, \cdots, n + 1$
  $l_{j-1}^2 = (A_{j-1}', A_{j-1}')$
  For $i = 1, 2, \cdots, j - 1$

$$Q_j^{(i+1)} = Q_j^{(i)} - \frac{(A_i', A_j^{(1)})}{l_i^2} Q_i' \tag{2}$$

$Q_j' = Q_j^{(j)}$.

Double-precision accumulation was implemented for each indicated inner product. By reserving storage for $c_i = (A_i', A_j^{(1)})/l_i^2$, $i = 1, 2, \cdots, j - 1$, each component

of Q_j' can be computed with double-precision accumulation. Thus, we have an additional inner-product calculation that was done in extra precision. This extra-precision work does not appear convenient in the Modified Gram-Schmidt method. It was originally assumed that this extra precision would make this method more accurate than Modified Gram-Schmidt. The results of the experiments indicate this is not the case. The resultant Q' matrix, obtained by applying the above transformation, is

$$Q' = \left\| \begin{array}{c|c} A' & \epsilon \\ \hline U & \hat{x} \end{array} \right\|,$$

where A' is a matrix of orthogonal columns, ϵ is the residual error $b - Ax$, U is an upper triangular matrix such that $A' = AU$, and $\hat{x}$ is the least-square solution.

Modified Gram-Schmidt Orthogonalization (*MGS*). Let Q be the matrix defined by Eq. (1). Again, let $Q_i^{(1)}$ and $A_i^{(1)}$ designate the columns of Q and A, respectively. Then we transform Q by the following procedure:

$Q_1' = Q_1^{(1)}$
For $j = 2, 3, \cdots, n + 1$
  $l_{j-1}^2 = (A_{j-1}', A_{j-1}')$
  For $i = 1, 2, \cdots, j - 1$

$$Q_j^{(i+1)} = Q_j^{(i)} - \frac{(A_i', A_j^{(i)})}{l_i^2} Q_i' \tag{3}$$

$Q_j' = Q_j^{(j)}$.

Pivoting in the Gram-Schmidt or Modified Gram-Schmidt procedure consists of a column interchange at the kth stage such that the length of $A_k^{(k)}$ relative to the original length of $A_k^{(1)}$ is largest among the candidates $A_k^{(k)}, A_{k+1}^{(k)}, \cdots, A_n^{(k)}$.

It should be emphasized that the two methods produce approximations to the same orthogonal set of vectors, that the amount of arithmetic required is the same, and that the procedures differ in only one detail. The results of these experiments leaves no doubt as to the preferable method for the least-squares problem. The superiority of the Modified Gram-Schmidt over Gram-Schmidt has been established by Rice [6] for the orthogonalization problem.

The above presentation of Gram-Schmidt and Modified Gram-Schmidt procedures gives some insight as to why the Modified Gram-Schmidt is superior. The only difference in the two procedures appears in the factors $(A_i', A_j^{(1)}) = c$ and $(A_i', A_j^{(i)}) = c$ in Eqs. (2) and (3), respectively. Since the $A_j^{(i)}$ are successive images of $A_j^{(1)}$ obtained by subtracting the components of $A_j^{(1)}$ parallel to each A_i', $i = 1, 2, \cdots, j - 1$, from $A_j^{(1)}$, then $\|A_j^{(i)}\| \leqq \|A_j^{(1)}\|$. The error in c due to the error in A_i' is magnified by $\|A_j^{(i)}\|$ and $\|A_j^{(1)}\|$ in the two approximations to c.

3. Pivoting. Optimum pivoting is defined as that permissible arrangement of rows and columns which produces the most accurate results when the algorithms are applied without further row or column interchange. Such an arrangement is rarely known in advance, and consequently some strategy to approximate this is required as the calculation proceeds. For singular problems (rank $r < n$) some strategy is a necessity in order to refrain using an approximate zero column as a pivot column. For nonsingular problems there are also known strategies to avoid as a general procedure. However, when some of the effects of the above pivoting

strategies are analyzed, it will be observed that this is indeed a very complicated problem.

Let ${}^6H^{-1}_{6\times5}$ denote the first five columns of the inverse of the 6 × 6 Hilbert segment. Define the column arrangement (1, 2, 3, 4, 5), abbreviated (1, 5) in the graphs,* as the natural order. For each of the methods, (5, 4, 3, 2, 1), abbreviated (5, 1) in the graphs,** represents a "best"arrangement of columns for each strategy, whereas (1, 2, 3, 4, 5) represents a "poorest" arrangement of columns for each strategy.

A summary of results reveals the following conclusions for this example and each method.

(A) Householder—results were mixed and close, therefore no conclusions should be drawn. The method does not appear sensitive to pivoting.

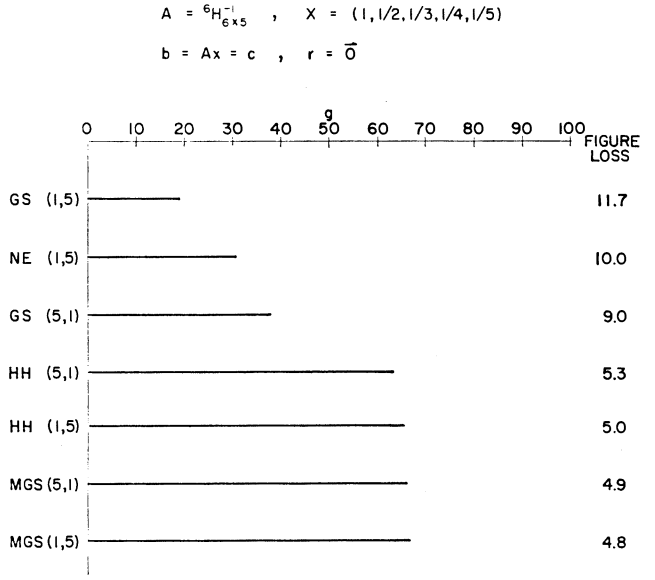

FIGURE 1a. *Least squares: Inverse of the Hilbert segment—0 residual.*

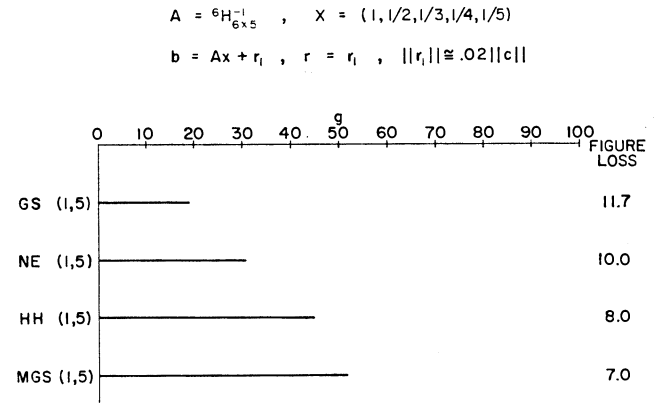

FIGURE 1b. *Least squares: Inverse of the Hilbert segment—2% residual.*

* See Fig. 1a through 1e and 3.

** See Fig. 1a and 3.

$A = {}^{6}H^{-1}_{6\times5}$, $X = (1, 1/2, 1/3, 1/4, 1/5)$

$b = Ax + r_2$, $r_2 = 3r_1$, $\|r_2\| \cong .06\|c\|$

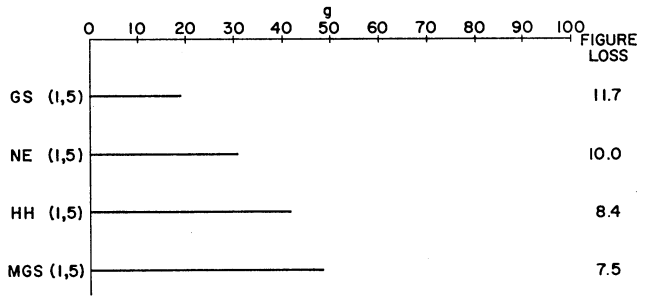

FIGURE 1c. *Least squares: Inverse of the Hilbert segment—6% residual.*

$A = {}^{6}H^{-1}_{6\times5}$, $X = (1, 1/2, 1/3, 1/4, 1/5)$

$b = Ax + r_3$, $r_3 = 12r_1$, $\|r_3\| \cong .24\|c\|$

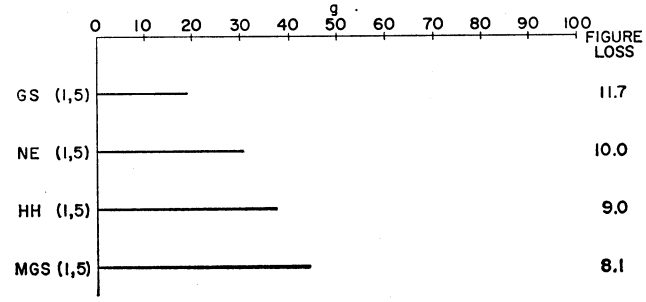

FIGURE 1d. *Least squares: Inverse of the Hilbert segment—24% residual.*

$A = {}^{6}H^{-1}_{6\times5}$, $X = (1, 1/2, 1/3, 1/4, 1/5)$

$b = Ax + r_4$, $r_4 = 120r_1$, $\|r_4\| \cong 2.4\|c\|$

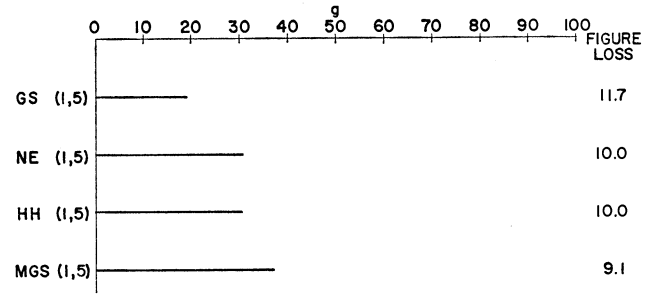

FIGURE 1e. *Least squares: Inverse of the Hilbert segment—240% residual.*

(B) Modified Gram-Schmidt—the results slightly favor poorest pivoting, but results were so close no conclusions should be drawn. This method does not appear sensitive to pivoting.

(C) Gram-Schmidt—the number of good digits was nearly doubled. This strategy makes Gram-Schmidt compete with the normal equations.

(D) Gaussian elimination—the cases tested here were (1) Partial Pivoting (PP) with column arrangements (a) (1, 2, 3, 4, 5) and (b) (5, 4, 3, 2, 1), the latter providing maximal pivots at each stage; (2) complete pivoting, i.e., pivoting with a maximum element (M.P.); and (3) pivoting with a minimum element (m.p.). Since it is known that no zero elements occur in this example, pivoting with a minimal element is possible. If we order the results according to maximal accuracy, we obtain 1.(b), 3, 2, 1.(a), corresponding to maximum errors 1.5, 2.2, 4.5, and 6.2—each $\times 10^{-11}$. (This study is included in Fig. 3.)

$$A = (a_{ij}) = \left[(i-1)2^{-7}\right]^{j-1}$$

$$1 \le i \le 129 \quad , \quad 1 \le j \le 7$$

$$X = (1,1,1,1,1,1,1)$$

$$b = Ax \quad , \quad r = \vec{0}$$

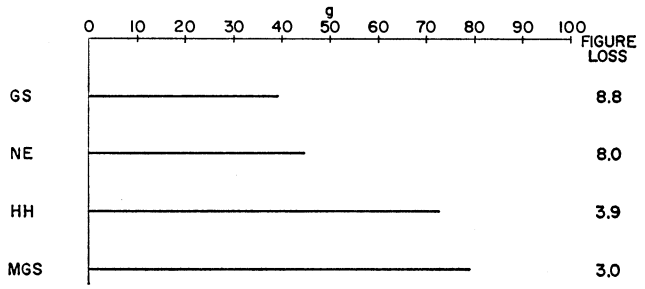

FIGURE 2a. *Least squares: Polynomial—degree* 6—129 *points on* [0, 1].

$$A = (a_{ij}) = \left[(i-1)2^{-10}\right]^{j-1}$$

$$1 \le i \le 1025 \quad , \quad 1 \le j \le 5$$

$$X = (1,1,1,1,1)$$

$$b = Ax \quad , \quad r = \vec{0}$$

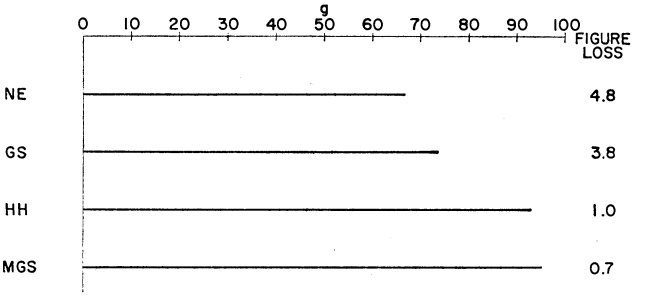

FIGURE 2b. *Least squares: Polynomial—degree* 4—1025 *points on* [0, 1].

If one draws any conclusions about pivoting in this study, it is that the Hilbert segment is not very sensitive to pivoting. This investigation of pivoting revealed

that a generally unworkable strategy produces almost the best results. This serves to emphasize the complexity of optimum pivoting.

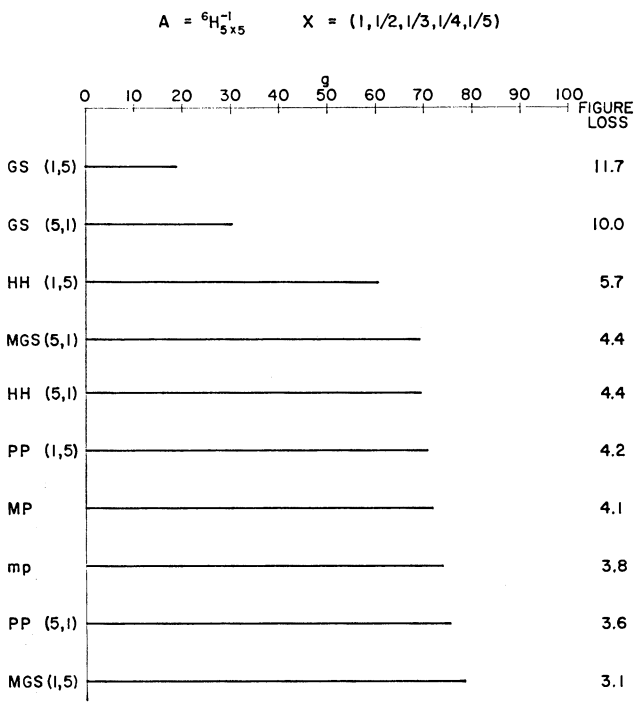

FIGURE 3. *Linear system: Inverse of the Hilbert segment.*

4. Scope of Numerical Study. Results shown here come from three examples. In each case A and b have exact representations in the computer and the solution is known. The examples are

(A) The least-squares problem for six rows and five columns of the inverse of the 6×6 Hilbert segment denoted by $^6H^{-1}_{6\times5}$. The solution vector was taken as (1, 1/2, 1/3, 1/4, 1/5) with various error components. The comparative performances of the four procedures described above are shown for increasing residuals in Fig. 1a through 1e.

(B) The least-squares problem for a polynomial of degree $n - 1$ with $2^m + 1$ equidistant data points (i.e. $\Delta x = 2^{-m}$) on [0, 1]. The values m and n are constrained such that x_i^r, $0 \leqq r \leqq n - 1$ and $0 \leqq i \leqq m$, is exactly representable in the computer. The solution vector has components all 1's. Two polynomial cases were studied. The comparative performances are shown in Fig. 2a and 2b for ($m = 7$, $n = 7$) and ($m = 10$, $n = 5$), respectively.

(C) The linear equation problem where A is the first five rows and five columns of the inverse of the Hilbert segment. The solution vector is taken as (1, 1/2, 1/3, 1/4, 1/5). Fig. 3 shows comparative performances of the three methods based on orthogonal transformations and Gaussian elimination with various pivot arrangements.

The error measure is taken as the maximum relative error. The results remain comparatively the same when other conventional error norms are used. The results will show the maximum figure loss associated with each experiment and a percent of full accuracy g given by $g = \epsilon/14.4$, where, if

$$\begin{aligned} \epsilon_i &= \log_{10} \frac{X_i^T}{X_i^T - X_i^C} \quad \text{if} \quad X_i^T \neq X_i^C \\ &= 14.4 \qquad\qquad\quad \text{if} \quad X_i^T = X_i^C , \end{aligned}$$

X_i^T is the true value of the ith component of the solution vector and X_i^C is the corresponding computed value, $\epsilon = \max_i \epsilon_i$. The 14.4 comes from the decimal equivalent of the Stretch 48-bit mantissa. The figure loss is the more nearly invariant measure for computers with differing word size.

Finally, it is important to observe the performance of the various methods as a function of the length of the error vector. If we write $Ax = b = c + r$, where c lies in the space S_A spanned by A and r is the residual vector which is a vector orthogonal to S_A, then we will note differing behaviors as the lengths of r varies and c remains fixed. The cases studied include

$$\begin{aligned} \|r_0\| &= 0 , \\ \|r_1\| &\cong .02 \, \|c\| , \\ \|r_2\| &= \quad 3 \, \|r_1\| \cong 0.06 \, \|c\| , \\ \|r_3\| &= \quad 12 \, \|r_1\| \cong 0.24 \, \|c\| , \\ \|r_4\| &= 120 \, \|r_1\| \cong 2.4 \, \|c\| , \end{aligned}$$

where $c = {}^6H_{6\times 5}^{-1}$ $(1, 1/2, 1/3, 1/4, 1/5)^T$ and $r_1 = (4620, 3960, 3465, 3080, 2772, 2520)^T$. The lengths of r_1 and c are, respectively, 8.517×10^3 and 4.18×10^5.

5. Conclusions. Modified Gram-Schmidt performed better than the other methods in all the examples studied. The Householder method performed competitively.

The dramatic difference in performance between Gram-Schmidt and Modified Gram-Schmidt illustrates the need to worry about the details of implementation and associated error analysis. It can be the difference between good and nonsensical results.

The insensitivity of the normal equations to the size of the error vector makes it more competitive if $m \gg n$ and the normal equations can be formed economically with double-precision accumulation and solved completely with double-precision arithmetic.

For the linear-equation problem, Gaussian elimination with partial pivoting and double-precision accumulation generally is the best. Modified Gram-Schmidt and Householder compete, but are also more expensive in numerical operations.

This example also illustrates that single-precision iterative techniques would not do any better for the case where the solution is compatible. The error vector had a relative error of 10^{-14} when compared to the b vector. It is, of course, well known that the residual vector should be computed with double-precision accumulation.

It is well known that the figure loss associated with problems involving A^TA is essentially double that of A since the condition number of A^TA is the square of

that of A. It is not so well known that the square of the condition number enters into those methods based on orthogonal transformations—presumably as a factor of the length of the residual vector r.

Acknowledgments. Programming was done by Joe Duran, Bertha Fagan, and Josephine Powers. Helpful suggestions defining the experiments were made by Blair Swartz.

Los Alamos Scientific Laboratory
Los Alamos, New Mexico 87544

1. J. H. WILKINSON, "Error analysis of direct methods of matrix inversion," *J. Assoc. Comput. Mach.*, v. 8, 1961, pp. 281–330. MR **31** #874.
2. G. GOLUB & J. H. WILKINSON, "Note on the iterative refinement of least squares solution," *Numer. Math.*, v. 9, 1966, pp. 139–206.
3. G. GOLUB, "Numerical methods for solving linear least squares problems," *Numer. Math.*, v. 7, 1965, pp. 206–216. MR **31** #5323.
4. E. E. OSBORNE, "Smallest least squares solutions of linear equations," *SIAM J. Numer. Anal.*, v. 2, 1965, pp. 300–307. MR **32** #4834.
5. A. BJÖRCK, "Solving linear least squares problems by Gram-Schmidt orthogonalization," *Nordisk Tidskr. Informations-Behandling*, v. 7, 1967, pp. 1–21.
6. J. R. RICE, "Experiments on Gram-Schmidt orthogonalization," *Math. Comp.*, v. 20, 1966, pp. 325–328. MR **33** #898.