Complete Orthogonal Decomposition for Weighted Least Squares*

Patricia D. Hough[†]
Stephen A. Vavasis[‡]

January 9, 1995

Abstract

Consider a full-rank weighted least-squares problem in which the weight matrix is highly ill-conditioned. Because of the ill-conditioning, standard methods for solving least-squares problems, QR factorization and the nullspace method for example, break down. G. W. Stewart established a norm bound for such a system of equations, indicating that it may be possible to find an algorithm that gives an accurate solution. S. A. Vavasis proposed a new definition of stability that is based on this result. He also defined the NSH algorithm for solving this least-squares problem and showed that it satisfies his definition of stability. In this paper, we propose a complete orthogonal decomposition algorithm to solve this problem and show that it is also stable. This new algorithm is simpler and more efficient than the NSH method.

1 Introduction

We consider solving the problem

$$\min_{\mathbf{y}\in R^n} \|D^{-1/2}(A\mathbf{y}-\mathbf{b})\| \tag{1}$$

for $\mathbf{y}$, where D is a symmetric positive definite $m \times m$ matrix, A is an $m \times n$ matrix, $\mathbf{y}$ is an n-vector, and $\mathbf{b}$ is an m-vector. An equivalent way to write this problem is

$$\begin{bmatrix} D & -A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix},$$

which is a special case of an equilibrium system [9]. Applications include optimization involving a barrier function, finite elements, structural analysis, and electrical networks. The following assumptions are made throughout the paper.

*This work supported by an NSF Presidential Young Investigator grant, with matching funds received from AT&T and Xerox Corp.

†Center for Applied Mathematics, Cornell University, Ithaca, New York 14853

‡Department of Computer Science, Cornell University, Ithaca, New York 14853.

A1. *A has rank n, i.e. full column rank.*

A2. *D is diagonal.*

A1 and A2 imply that (1) is a full-rank weighted least-squares problem with a unique solution, and they allow us to use the norm bound obtained by Stewart. That bound is given in the following theorem.

Theorem 1 *[8] Let $\mathcal{D}$ denote the set of all positive definite $m \times m$ real diagonal matrices. Let A be an $m \times n$ real matrix of rank n. Then there exist finite constants χ_A and $\bar{\chi}_A$ such that*

$$\begin{aligned} a)\ \chi_A &= \sup\{\| \left(A^T D^{-1} A\right)^{-1} A^T D^{-1} \| : D \in \mathcal{D}\}, \text{ and} \\ b)\ \bar{\chi}_A &= \sup\{\| A \left(A^T D^{-1} A\right)^{-1} A^T D^{-1} \| : D \in \mathcal{D}\}. \end{aligned}$$

A similar result was obtained independently by Todd [10]. In this theorem, the norm can be any matrix norm induced by a vector norm. In this paper, $\|\cdot\| = \|\cdot\|_2$. Similarly, the condition number of a matrix M is the condition number of M in the 2-norm, i.e. $\kappa(M) = \kappa_2(M)$. We make one more assumption.

A3. *D is very ill-conditioned.*

The ill-conditioning of D arises in certain classes of finite element problems [11], electrical networks, and it *always* occurs in the barrier method for optimization [14]. It also indicates that the coefficient matrix of the least-squares problem is ill-conditioned. For this reason, the methods typically used to solve least-squares problems can give highly inaccurate solutions $\mathbf{y}$, as argued by Vavasis [12]. Since D is the source of ill-conditioning, we use the definition of stability given by Vavasis.

Definition 1 *[12] An algorithm for* (1) *is* **stable** *if, in the presence of finite-precision arithmetic, an error bound of the form*

$$\|\mathbf{y} - \hat{\mathbf{y}}\| \leq \epsilon \cdot f(A) \cdot \|\mathbf{b}\| \qquad (2)$$

is satisfied, where $\mathbf{y}$ is the true solution, $\hat{\mathbf{y}}$ is the computed solution, $f(A)$ is some function of A not depending on D, and $\epsilon > 0$ is machine roundoff.

Other standard terminology is modified analogously. For example, a well-conditioned matrix is one for which there is an upper bound on the condition number that does not depend on D. In order to show that our algorithm is stable, then, we strive to obtain bounds on norms, condition numbers, and errors that do not depend on D.

We now present the algorithm.

Algorithm: Orthogonal Decomposition

Step 1: QR factor, with column pivoting, $\left(D^{-1/2}A\right)^T$ to get

$$\left(D^{-1/2}A\right)^T = A^T D^{-1/2} = QRP, \tag{3}$$

where Q is an $n \times n$ orthogonal matrix, R is an $n \times m$ upper triangular matrix, and P is an $m \times m$ permutation matrix.

Step 2: Apply reduced QR factorization (without pivoting) to R^T to get

$$R^T = Z_1 U_1, \tag{4}$$

where Z_1 is an $m \times n$ matrix with orthonormal columns, and U_1 is an $n \times n$ upper triangular matrix.

Step 3: Solve the following system, via back substitution, for $\bar{\mathbf{y}}$:

$$U_1 \bar{\mathbf{y}} = Z_1^T P D^{-1/2} \mathbf{b}. \tag{5}$$

Step 4: To get $\mathbf{y}$, multiply the result of Step 3 by Q:

$$\mathbf{y} = Q\bar{\mathbf{y}}. \tag{6}$$

Solution of least-squares problems via QR factorization with column pivoting was introduced by Golub [3]. Note that the QR factorization for the least-squares problem occurs in Step 2. The QR factorization in Step 1 is to make the algorithm stable.

First, we compare this algorithm to Vavasis's NSH method [12]. The NSH method is the only algorithm in the literature known to stably (in the sense of Definition 1) solve (1). The NSH algorithm employs nonstandard techniques, particularly when choosing the nullspace basis for $A^T D^{-1}$. In contrast, the complete orthogonal decomposition algorithm uses standard techniques that are well understood, namely QR decomposition and back substitution. Also, our algorithm requires less work than the NSH algorithm. The NSH method requires solving an $m \times m$ system of equation and thus requires $O(m^3)$ flops. The work for the QR factorizations dominates the work required for the complete orthogonal algorithm, so this algorithm requires $O(mn^2)$ flops. Since $n < m$ (and n could be much smaller than m), the complete orthogonal decomposition algorithm requires less work.

The rest of this paper is devoted to analysis of the stability of the orthogonal decomposition algorithm. Before giving a rigorous stability analysis of the algorithm, we offer an intuitive explanation of why this algorithm finds an accurate solution. The first step is a QR factorization of a matrix that is well-conditioned up to a scaling of the columns. So, the result is a computed upper triangular matrix that is close to the exact upper triangular matrix. It would be useful to know something about the condition number of this matrix as well. To minimize confusion assume, without loss of generality, that $A^TD^{-1/2}$ has been "pre-pivoted." This means that the columns of $A^TD^{-1/2}$ are ordered in such a way that the norms of the first n columns are, loosely speaking, in decreasing order. In addition, the norms of the first n columns are larger than those of the last $m-n$ columns. One might suspect that this implies that the entries of $D^{-1/2}$ are ordered in the same way. In other words, some inequality similar to

$$d_i^{-1/2} \geq d_j^{-1/2} \text{ for } i \leq j, 1 \leq i \leq n, 1 \leq j \leq m$$

might hold. This ordering becomes significant in the second step of the algorithm.

Recall that

$$R = Q^TA^TD^{-1/2}.$$

Notice the Q^TA^T is upper triangular. So let

$$\bar{R} = Q^TA^T.$$

Notice also that R is ill-conditioned and that the ill-conditioning arises from $D^{-1/2}$. We try to "offset" the effects of $D^{-1/2}$ in the following naive way. Let $\bar{D} = D(1:n, 1:n)$, and consider the following:

$$\bar{D}^{1/2}R = \bar{D}^{1/2}\bar{R}D^{-1/2} = \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^{1/2}\bar{r}_{11} & \cdots & \left(\frac{d_1}{d_n}\right)^{1/2}\bar{r}_{1n} & \cdots & \left(\frac{d_1}{d_m}\right)^{1/2}\bar{r}_{1m} \\ & \ddots & \vdots & & \vdots \\ & & \left(\frac{d_n}{d_n}\right)^{1/2}\bar{r}_{nn} & \cdots & \left(\frac{d_n}{d_m}\right)^{1/2}\bar{r}_{nm} \end{bmatrix}.$$

It is clear that $\bar{R}$ is well-conditioned, so if the weights are indeed in the order described above, then it is not difficult to show that there are upper bounds on all entries. It can also be shown that there are upper bounds on the entries of $\left(\bar{D}^{1/2}R(:,1:n)\right)^{-1}$. Using this information, it is not difficult to show that $\bar{D}^{1/2}R$ (and hence $R^T\bar{D}^{1/2}$) is well-conditioned, i.e. R^T is well-conditioned up to a scaling of the columns. In the second step, then, we have a least-squares problem with a coefficient matrix that is well-conditioned up to a scaling of the columns, namely

$$\min_{\bar{\mathbf{y}} \in R^n} \|R^T\bar{\mathbf{y}} - D^{-1/2}\mathbf{b}\|.$$

In traditional analysis, R^T being well-conditioned up to a scaling of the columns indicates that the standard algorithms give an accurate solution. In this analysis, therefore, one might expect a parallel result.

In the remainder of the paper, we present a rigorous stability analysis of our proposed algorithm. The next section contains a discussion of numerical tolerance. Section 3 examines the condition number and the accuracy of the upper triangular matrix R computed in the first step of the algorithm. Section 4 does the same for U_1, the upper triangular matrix computed in Step 2. An analysis of Steps 3 and 4 of the algorithm is presented in Section 5.

2 A note on numerical tolerance

In the upcoming analysis we assume throughout that any occurrence of exact linear dependence among the columns of A^T is always determined correctly in Step 1 of the algorithm (QR factorization with pivoting). This requires the use of a numerical tolerance. To illustrate this point, consider applying the algorithm when $D^{-1/2} = \operatorname{diag}(1, 1, 1, 10^{-20})$ and

$$A^T = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 1 & 7 \end{pmatrix}.$$

Observe that the third column of A^T is dependent on the first two. If the QR factorization of $A^T D^{-1/2}$ were done in exact arithmetic, this dependence would be manifested as a '0' in the $(3,3)$ position of the factored $A^T D^{-1/2}$ after the first two QR factorization steps, and column 4 would be chosen for the third pivot.

In finite-precision arithmetic, however, we would expect the $(3,3)$ entry to be on the order of machine-epsilon rather than 0. Because column 4 is weighted by 10^{-20}, the unwanted residual in the $(3,3)$ position could cause column 3 to be chosen for the next column pivot instead of column 4. Thus, without modification, ordinary QR-factorization with column pivoting procedure has missed a linear dependence.

We address this problem as follows: after the kth QR factorization step, we check whether the residual portion (that is, positions $k+1, \ldots, n$) of any uneliminated column has become very small (according to some tolerance level) with respect to the original norm of that column. If so, we change those entries to zeros. Notice that this test requires very little additional work because the usual QR factorization algorithm with column pivoting already monitors the norms of the residual portions of the columns [3]. In this way, if there are exact dependences among the rows of A, our algorithm does not miss them.

If this numerical test fails to detect exact dependence, then our stability analysis no longer holds. It can be shown that the test we have proposed will fail only in the case that there is near-dependence among the columns of A^T. However, in this case, the parameter χ_A given by Theorem 1 is large, and so our stability bound (which depends on χ_A and $\bar{\chi}_A$) is not practically applicable.

3 The First QR Factorization

The intuitive discussion in §1 asserted that the ordering of the weights produced in Step 1 of the algorithm is important in stabilizing the algorithm. It is necessary, then, to determine how the weights of the basis rows compare to those of other basis rows and to those of the nonbasis rows. We begin by comparing the norms of the rows A.

Lemma 1 *Let B be an $n \times n$ matrix whose columns are an arbitrary set of n rows $\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_n}$ of A that form a basis for the rowspace of A. Then*

$$\max_{1 \le j \le n} \|\mathbf{a}_{i_j}\| \le (\chi_A \|A\|) \cdot \min_{1 \le j \le n} \|\mathbf{a}_{i_j}\|. \tag{7}$$

Proof: Let B be the basis defined above. Without loss of generality, suppose that $\|\mathbf{a}_{i_n}\| = \min_{1 \le j \le n} \|\mathbf{a}_{i_j}\|$. Then, we can write

$$B = \left[\hat{B}, \mathbf{a}_{i_n}\right]$$

and partition B^{-1} as:

$$B^{-1} = \left[\begin{array}{c} X \\ \mathbf{v}^T \end{array} \right].$$

Then

$$B^{-1}B = I = \left[\begin{array}{cc} X\hat{B} & X\mathbf{a}_{i_n} \\ \mathbf{v}^T\hat{B} & \mathbf{v}^T\mathbf{a}_{i_n} \end{array} \right].$$

This means $\mathbf{v}^T\mathbf{a}_{i_n} = 1$. By the Cauchy-Schwarz inequality, $\|\mathbf{v}\| \ge 1/\|\mathbf{a}_{i_n}\|$. Also,

$$\|\mathbf{v}\| \le \|B^{-1}\| \le \chi_A.$$

The first inequality is because $\mathbf{v}$ is a row of B^{-1}, and the second is from [12]. Thus,

$$1/\|\mathbf{a}_{i_n}\| \le \chi_A,$$

i.e.

$$\min_{1 \le j \le n} \|\mathbf{a}_{i_j}\| \ge 1/\chi_A.$$

Multiply both sides of this inequality by the inequality $\|A\| \ge \max_{1 \le j \le n} \|\mathbf{a}_{i_j}\|$ to obtain

$$\max_{1 \le j \le n} \|\mathbf{a}_{i_j}\| \le (\chi_A \|A\|) \cdot \min_{1 \le j \le n} \|\mathbf{a}_{i_j}\|,$$

as required. □

Notice that we have proved that there is a lower bound on any column of *any* basis for the rowspace of A.

Suppose now that $k > 0$ steps of the factorization have been completed. Partition the resulting matrix by rows as follows:

$$AQ_1 \cdots Q_k = \bar{A} = \left[\begin{array}{cccc} \alpha_{11} & 0 & \cdots & \mathbf{0}^T \\ \vdots & & & \vdots \\ \alpha_{k1} & \cdots & \alpha_{kk} & \mathbf{0}^T \\ \alpha_{k+1,1} & \cdots & \alpha_{k+1,k} & \bar{\mathbf{a}}_{k+1}^T \\ \vdots & & \vdots & \vdots \\ \alpha_{m,1} & \cdots & \alpha_{m,k} & \bar{\mathbf{a}}_m^T \end{array} \right].$$

Lemma 1 can be extended so that is applies to the residual portions of the rows of A, i.e. $\bar{\mathbf{a}}_{k+1}^T, \ldots, \bar{\mathbf{a}}_n^T$.

Lemma 2 *Let B be an set of n columns of A^T such that the first k columns of B are $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_k}\}$ and B is a basis. Suppose that $\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_k}$ are the first k columns chosen by the column pivoting in the QR factorization. As with $\bar{A}$ above, write*

$$Q_k^T \cdots Q_1^T B = \begin{bmatrix} \alpha_{i_1,1} & \cdots & \alpha_{i_k,1} & \alpha_{i_{k+1},1} & \cdots & \alpha_{i_n,1} \\ 0 & & \vdots & \vdots & & \vdots \\ \vdots & & \alpha_{i_k k} & \alpha_{i_{k+1} k} & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{a}}_{i_{k+1}} & \cdots & \bar{\mathbf{a}}_{i_n} \end{bmatrix}.$$

Then

$$\max_{k+1 \le j \le n} \|\bar{\mathbf{a}}_{i_j}\| \le (\chi_A \|A\|) \cdot \min_{k+1 \le j \le n} \|\bar{\mathbf{a}}_{i_j}\|. \tag{8}$$

Proof: Let $Q = Q_1 \cdots Q_k$, and let B be defined as above. Then

$$Q^T B = \begin{bmatrix} R & X \\ 0 & \bar{B} \end{bmatrix},$$

where R is upper triangular. Comparing this matrix to the partitioned one above, we see that the columns of $\bar{B}$ are $\bar{\mathbf{a}}_{i_{k+1}}, \cdots, \bar{\mathbf{a}}_{i_n}$. To prove Lemma 2, then, we need a lower bound on a typical column of $\bar{B}$. Notice that

$$(Q^T B)^{-1} = \begin{bmatrix} R^{-1} & -R^{-1} X \bar{B}^{-1} \\ 0 & \bar{B}^{-1} \end{bmatrix}.$$

Therefore,

$$\|\bar{B}^{-1}\| \le \|B^{-1}\| \le \chi_A.$$

Now, as above, turn an upper bound on $\|\bar{B}^{-1}\|$ into a lower bound on any column of $\bar{B}$. □

Using the previous two lemmas, we can determine the relationships between the weights of the basis rows and those of the nonbasis rows. For the remainder of the paper assume, as in the intuitive discussion in §1, that the columns of $A^T D^{-1/2}$ have been reordered so that no pivoting is necessary. This implies that not only do the first n rows of A form a basis for the rowspace of A, but there is also an order that has been imposed on the rows of A. Let $B = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and let $d_1, \ldots, d_m$ denote the (reordered) entries of D.

Theorem 2 *Suppose the first $k \ge 0$ steps of the QR factorization have been completed. If $d_{k+1}^{-1/2}$ is the weight assigned to $\mathbf{a}_{k+1} \in B$, and $d_j^{-1/2}$ is the weight assigned to $\mathbf{a}_j \notin B$, then*

$$\frac{d_{k+1}}{d_j} \le (\chi_A \|A\|)^4, \tag{9}$$

provided $\mathbf{a}_j$ is linearly independent of the first k basis vectors.

Proof: Let $\bar{A}$ be as defined before Lemma 2 for $k \geq 0$. (Notice that when $k = 0$, this is just the matrix A, partioned into rows.) Since B forms a basis for the rowspace of A,

$$\mathbf{a}_j = \sum_{i=1}^{n} c_i \mathbf{a}_i.$$

Assuming $\mathbf{a}_j$ is linearly independent of the first k basis rows implies

$$\mathbf{a}_j - \sum_{i=1}^{k} c_i \mathbf{a}_i = \sum_{i=k+1}^{n} c_i \mathbf{a}_i \neq \mathbf{0},$$

which means $c_i \neq 0$ for at least one i such that $k+1 \leq i \leq n$. Let us take $Q = Q_1 \cdots Q_k$. Then

$$Q^T \mathbf{a}_j - \sum_{i=1}^{k} c_i Q^T \mathbf{a}_i = \sum_{i=k+1}^{n} c_i Q^T \mathbf{a}_i,$$

and

$$\bar{\mathbf{a}}_j - \sum_{i=1}^{k} c_i \bar{\mathbf{a}}_i = \sum_{i=k+1}^{n} c_i \bar{\mathbf{a}}_i.$$

Notice that $\bar{\mathbf{a}}_i = \mathbf{0}$ for $1 \leq i \leq k$. So

$$\bar{\mathbf{a}}_j = \sum_{i=k+1}^{n} c_i \bar{\mathbf{a}}_i,$$

where $c_i \neq 0$ for at least one i. Let l be such that $k+1 \leq l \leq n$ and $c_l \neq 0$. Then

$$\bar{\mathbf{a}}_l = \frac{1}{c_l}\left(\bar{\mathbf{a}}_j - \sum_{i=k+1, i\neq l}^{n} c_i \bar{\mathbf{a}}_i\right).$$

So, $\bar{B} = \{\bar{\mathbf{a}}_j, \bar{\mathbf{a}}_{k+1}, \ldots, \bar{\mathbf{a}}_{l-1}, \bar{\mathbf{a}}_{l+1}, \ldots, \bar{\mathbf{a}}_n\}$ is a basis for $\{\bar{\mathbf{a}}_{k+1}, \ldots, \bar{\mathbf{a}}_m\}$. Since (7) and (8) hold for any basis for the rowspace of A,

$$\|\bar{\mathbf{a}}_j\| \geq \frac{\max\left\{\|\bar{\mathbf{a}}_i\| : \bar{\mathbf{a}}_i \in \bar{B}\right\}}{\chi_A \|A\|}.$$

Recall that the columns of $A^T D^{-1/2}$ have been reordered so that no pivoting is necessary. This means that there is an order imposed on the columns of $A^T D^{-1/2}$. More specifically, since we are at step $k+1$,

$$\left(\frac{1}{d_j}\right)^{1/2} \|\bar{\mathbf{a}}_j\| \leq \left(\frac{1}{d_{k+1}}\right)^{1/2} \|\bar{\mathbf{a}}_{k+1}\|.$$

Thus,

$$\frac{d_{k+1}}{d_j} \quad \leq \quad \left(\frac{\|\bar{\mathbf{a}}_{k+1}\|}{\|\bar{\mathbf{a}}_j\|}\right)^2$$

$$
\begin{aligned}
&\leq \left(\frac{\chi_A \cdot \|A\| \cdot \max_{k+1\leq i\leq n} \|\bar{\mathbf{a}}_i\|}{\max\left\{\|\bar{\mathbf{a}}_i\| : \bar{\mathbf{a}}_i \in \bar{B}\right\}}\right)^2 \\
&\leq \left(\frac{\chi_A \cdot \|A\| \cdot \max_{k+1\leq i\leq n} \|\bar{\mathbf{a}}_i\|}{\min_{k+1\leq i\leq n} \|\bar{\mathbf{a}}_i\|}\right)^2 \\
&\leq (\chi_A\|A\|)^4,
\end{aligned}
$$

which is (9).

□

We also need to know the relationships between the weights of the basis rows of A. Suppose $\mathbf{a}_i, \mathbf{a}_j \in B$, where $i < j$. It follows from (7), (8), and the implicit order indicated by the absence of column pivoting that

$$\frac{d_i}{d_j} \leq (\chi_A\|A\|)^2 \leq (\chi_A\|A\|)^4. \tag{10}$$

Recall that the intuitive argument given in §1 relied on the weights being in the following order:

$$d_i^{-1/2} \geq d_j^{-1/2} \text{ for } i \leq j, 1 \leq i \leq n, 1 \leq j \leq m.$$

What we have found, however, is that they are not ordered in exactly this way. Instead, this ordering holds up to scaling by a constant, i.e.

$$d_i^{-1/2} \geq \frac{d_j^{-1/2}}{(\chi_A\|A\|)^2} \text{ for } i \leq j, 1 \leq i \leq n, 1 \leq j \leq m.$$

This bound is sufficient for our arguments.

The second step of the algorithm performs a QR factorization on R^T. In order to analyze that step, then, it is necessary to know something about the condition of R. The relationships between the weights of the rows of A are used in the following theorem to show that R is well-conditioned up to a scaling of the rows or the columns. Recall that for a rectangular matrix M, $\kappa(M)$ is the condition number (in the 2-norm) of M, i.e.

$$\kappa(M) = \|M\| \cdot \| \left(M^T M\right)^{-1} M^T\|.$$

Theorem 3 *Let* $C = \bar{D}^a R D^{1/2-a}$, *where* $a \geq 0$ *and* $\bar{D} = D(1:n, 1:n)$. *If* $\tilde{C} = C(1:k,:)$, *then*

$$\kappa(\tilde{C}) \leq n^4 \cdot (\chi_A\|A\|)^{16a+2} \tag{11}$$

for any $1 \leq k \leq n$.

Proof: First, we must find an upper bound on $\|\tilde{C}\|$. Since $\tilde{C}$ is a submatrix of C, then $\|\tilde{C}\| \leq \|C\|$. Therefore, it is sufficient show that there is an upper bound on $\|C\|$. Write C as follows:

$$C = \bar{D}^a R D^{1/2-a} = \bar{D}^a Q^T A^T D^{-1/2} D^{1/2-a} = \bar{D}^a \bar{R} D^{-a},$$

where $\bar{R} = Q^T A^T$. If the entries of C are written explicitly,

$$C = \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^a \bar{r}_{11} & \cdots & \left(\frac{d_1}{d_n}\right)^a \bar{r}_{1n} & \cdots & \left(\frac{d_1}{d_m}\right)^a \bar{r}_{1m} \\ & \ddots & \vdots & & \vdots \\ & & \left(\frac{d_n}{d_n}\right)^a \bar{r}_{nn} & \cdots & \left(\frac{d_n}{d_m}\right)^a \bar{r}_{nm} \end{bmatrix}.$$

Consider $\bar{R}_1 = \bar{R}(:, 1:n)$. Again, let B be the basis consisting of the first n rows of A. Then $\bar{R}_1 = Q^T B$. So, $\left|\frac{1}{\bar{r}_{ii}}\right| \leq \|B^{-1}\| \leq \chi_A$ [12] for $1 \leq i \leq n$. If $\bar{\mathbf{r}}_i^T$ is the ith row of $\bar{R}$, then $\|\bar{\mathbf{r}}_i\| \leq \|A\|$ for all $1 \leq i \leq m$. These facts, (9), and (10) imply the following:

$$\begin{aligned} \frac{1}{\chi_A} &\leq |\bar{r}_{ii}| \leq \|A\|, 1 \leq i \leq n \text{ and} && (12) \\ \left\| d_i^a \bar{\mathbf{r}}_j^T D^{-a} \right\| &\leq \|A\| \cdot (\chi_A \|A\|)^{4a}, i \leq j, 1 \leq i \leq n, 1 \leq j \leq m. && (13) \end{aligned}$$

Recall that Theorem 2, and thus (13), holds only when $\mathbf{a}_j$ is linearly independent of the first $i-1$ basis vectors. We must now consider the case not covered by Theorem 2. Suppose that B and D are defined as before. For each nonbasis row $\mathbf{a}_j$ there is a $1 \leq k \leq n$ such that $\mathbf{a}_j$ is linearly independent of the first $k-1$ basis vectors, but is linearly dependent on the first k basis vectors. So,

$$\mathbf{a}_j = \sum_{i=1}^{k} c_i \mathbf{a}_i,$$

where $c_k \neq 0$. Now, suppose that k steps of the QR factorization have been completed. Then

$$Q_k^T \cdots Q_1^T \mathbf{a}_j = \sum_{i=1}^{k} c_i Q_k^T \cdots Q_1^T \mathbf{a}_i = \sum_{i=1}^{k} c_i \begin{bmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ki} \\ \mathbf{0} \end{bmatrix}$$

So, $\bar{\mathbf{a}}_j = \mathbf{0}$ (where $\bar{\mathbf{a}}_j$ is as in Lemma 2). After this point, transformations act only on $\bar{\mathbf{a}}_j$. This gives

$$Q^T \mathbf{a}_j = \sum_{i=1}^{k} c_i \begin{bmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ii} \\ \mathbf{0} \end{bmatrix},$$

telling us that $\bar{r}_{ij} = 0$ for $i > k$, so these entries do not contribute to the norm in (13). Thus, (13) holds even in the case where $\mathbf{a}_j$ is not linearly independent of the first $i-1$ basis vectors.

If $\mathbf{c}_i^T$ is i^{th} row of C, then

$$\|C\| \leq \sum_{i=1}^{n} \|\mathbf{c}_i\|$$

$$
\begin{aligned}
&\le n \cdot \max_{1\le i\le n} \|\mathbf{c}_i\| \\
&= n \cdot (\chi_A\|A\|)^{4a} \cdot \max_{1\le i\le n} \|\bar{\mathbf{r}}_i\| \\
&\le n \cdot \|A\| \cdot (\chi_A\|A\|)^{4a}.
\end{aligned}
$$

The next step is to find an upper bound on $\left\| \left(\tilde{C}\tilde{C}^T\right)^{-1} \right\|$. Let $\tilde{C}_1 = \tilde{C}(:, 1:k)$. Recall that

$$\| \left(\tilde{C}\tilde{C}^T\right)^{-1} \| \le \|\tilde{C}_1^{-T}\|^2.$$

If $C_1 = C(:, 1:n)$, it is easy to show that $\tilde{C}_1^{-T}$ is a submatrix of C_1^{-T}. To obtain an upper bound on $\|C_1^{-T}\|$, we use the following fact, which will be proved after the current proof.

Fact: If $C_1 = C(1:n, 1:n)$, then

$$\|C_1^{-T}\| \le n \cdot \chi_A \cdot (\chi_A\|A\|)^{4a}.$$

So,

$$
\begin{aligned}
\| \left(\tilde{C}\tilde{C}^T\right)^{-1} \| &= \|\tilde{C}_1^{-T}\|^2 \\
&\le \|C_1^{-T}\|^2 \\
&\le \left[n \cdot \chi_A \cdot (\chi_A\|A\|)^{4a}\right]^2,
\end{aligned}
$$

and the bound on the condition number is

$$
\begin{aligned}
\kappa(\tilde{C}) &= \|\tilde{C}\| \cdot \|\tilde{C}^T \left(\tilde{C}\tilde{C}^T\right)^{-1} \| \\
&\le \|\tilde{C}\|^2 \cdot \| \left(\tilde{C}\tilde{C}^T\right)^{-1} \| \\
&\le n^4 \cdot (\chi_A\|A\|)^{16a+2}.
\end{aligned}
$$

Thus, we have proved our claim. □

Recall that we must still prove the fact used in the above proof. We state it in the form of the following lemma and give the proof below.

Lemma 3 *Let $C_1 = C(:, 1:n)$, where C is defined as in the previous theorem. Then*

$$\|C_1^{-T}\| \le n \cdot \chi_A \cdot (\chi_A\|A\|)^{4a}. \tag{14}$$

Proof: Recall that

$$C = \bar{D}^a \bar{R} D^{-a},$$

where $\bar{D} = D(1:n, 1:n)$ and $\bar{R} = Q^T A^T$. So,

$$C_1 = \bar{D}^a \bar{R}(:, 1:n) \bar{D}^{-a}.$$

Notice that

$$\bar{R}(:,1:n) = Q^T B,$$

where B is the rowspace basis consisting of the first n rows of A. If $L = Q^T B^{-T}$, then

$$\begin{aligned} C_1^{-T} &= \bar{D}^{-a} L \bar{D}^{a} \\ &= \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^a l_{11} & & & \\ \left(\frac{d_1}{d_2}\right)^a l_{21} & \left(\frac{d_2}{d_2}\right)^a l_{22} & & \\ \vdots & \vdots & \ddots & \\ \left(\frac{d_1}{d_n}\right)^a l_{n1} & \left(\frac{d_2}{d_n}\right)^a l_{n2} & \cdots & \left(\frac{d_n}{d_n}\right)^a l_{nn} \end{bmatrix}. \end{aligned}$$

If $\mathbf{c}_i^T$ is the i^{th} row of C_1^{-T}, then

$$\begin{aligned} \|C_1^{-T}\| &\le \sum_{i=1}^{n} \|\mathbf{c}_i\| \\ &\le n \cdot \max_{1\le i\le n} \|\mathbf{c}_i\| \\ &\le n \cdot (\chi_A \|A\|)^{4a} \cdot \max_{1\le i\le n} \|\mathbf{l}_i\| \\ &\le n \cdot (\chi_A \|A\|)^{4a} \cdot \|Q^T B^{-T}\| \\ &= n \cdot (\chi_A \|A\|)^{4a} \cdot \|B^{-T}\| \\ &\le n \cdot \chi_A \cdot (\chi_A \|A\|)^{4a}, \end{aligned}$$

as claimed. □

We now know that R is not only well-conditioned up to a scaling of the rows, but it is well-conditioned up to a scaling of both the rows and the columns. This result will be useful later in the analysis.

Standard results tell us that the differences between the columns of the computed matrix $\hat{R}$ and those of the exact matrix R are small. Since the second step of the algorithm is a QR factorization of R^T, we need to know that the same is true of the rows. This is given by the following theorem.

Theorem 4 *Let $\mathbf{r}_j^T$ and $\hat{\mathbf{r}}_j^T$ be the j^{th} rows of R and $\hat{R}$, respectively. Then*

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|}{\|\mathbf{r}_j^T\|} \le \epsilon \cdot n \cdot \left\{ 1 + (j-1)\left(\chi_A\|A\|\right)^2 + \sum_{i=j+1}^{m} \left[1 + (j-1)\left(\chi_A\|A\|\right)^6 \right] \right\}^{1/2}. \tag{15}$$

Proof: Recall that

$$R = Q^T A^T D^{-1/2} = Q^T X,$$

where $X = A^T D^{-1/2}$. So, $\mathbf{r}_i = Q^T \mathbf{x}_i$. According to Wilkinson [13],

$$\|\mathbf{r}_i - \hat{\mathbf{r}}_i\| \le \epsilon \cdot n \cdot \|\mathbf{r}_i\|, 1 \le i \le m,$$

where $\mathbf{r}_i$ is the i^{th} column of R, and $\hat{\mathbf{r}}_i$ is the i^{th} column of $\hat{R}$. This gives us a bound on the elementwise error, namely

$$|r_{ji} - \hat{r}_{ji}| \leq \|\mathbf{r}_i - \hat{\mathbf{r}}_i\| \leq \epsilon \cdot n \cdot \|\mathbf{r}_i\|, 1 \leq i \leq m, 1 \leq j \leq n.$$

We can now find a bound on the normwise error of the rows of R, namely

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|^2}{\|\mathbf{r}_j^T\|^2} \leq \frac{\sum_{i=j}^m (r_{ji} - \hat{r}_{ji})^2}{\sum_{i=j}^m r_{ji}^2} \leq \frac{\epsilon^2 \cdot n^2 \cdot \sum_{i=j}^m \|\mathbf{r}_i\|^2}{r_{jj}^2}.$$

Consider $\frac{\|\mathbf{r}_i\|^2}{r_{jj}^2}$. Suppose that $k \geq 0$ steps of the QR factorization have been completed. Partition X in a manner similar to that of A for Lemma 2. Then $r_{jj}^2 = \|\bar{\mathbf{x}}_j\|^2$. Since no pivoting is necessary, the columns of X are ordered such that $\|\bar{\mathbf{x}}_i\|^2 \leq \|\bar{\mathbf{x}}_j\|^2$ for $i \geq j$. So,

$$\begin{aligned}
\frac{\|\mathbf{r}_i\|^2}{r_{jj}^2} &\leq \frac{\|\mathbf{r}_i(j:n)\|^2 + r_{1i}^2 + \cdots + r_{j-1,i}^2}{r_{jj}^2} \\
&\leq 1 + \frac{d_i^{-1}\left(\bar{r}_{1i}^2 + \cdots + \bar{r}_{j-1,i}^2\right)}{d_j^{-1}\bar{r}_{jj}^2} \\
&\leq \begin{cases} 1 + (j-1)\left(\chi_A\|A\|\right)^2 & \text{for } i = j \\ 1 + (j-1)\left(\chi_A\|A\|\right)^6 & \text{for } j+1 \leq i \leq m \end{cases}
\end{aligned}$$

Thus,

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|^2}{\|\mathbf{r}_j^T\|^2} \leq \epsilon^2 \cdot n^2 \cdot \left[1 + (j-1)\left(\chi_A\|A\|\right)^2 + \sum_{i=j+1}^m \left[1 + (j-1)\left(\chi_A\|A\|\right)^6\right]\right].$$

Taking the square root of both sides gives the required result. □

We have now established that the QR factorization in the first step of the algorithm gives an upper triangular matrix R that is well-conditioned up to a scaling of the rows or the columns. It also yields a computed matrix, $\hat{R}$, whose rows are close to those of R. With these results in hand, we move on to the analysis of the second step of the algorithm.

4 The Second QR Factorization

Recall that in this step we use the "skinny" QR factorization,

$$R^T = Z_1 U_1,$$

where Z_1 is an $m \times n$ matrix with orthonormal columns and U_1 is an $n \times n$ upper triangular matrix. The results of §3 tell us that R^T is well-conditioned up to a scaling of the columns. The QR factorization in this step, then, gives an upper triangular matrix $\hat{U}_1$ that is close to the exact upper triangular matrix U_1. In addition, the results of §3 can be used to prove similar results about U_1. Again, let $D = \text{diag}(d_1, \ldots, d_m)$ and $\bar{D} = D(1:n, 1:n)$. Since U_1 is the coefficient matrix of the system of equations in the back substitution step, the following theorem concerning the condition of U_1 will be quite useful.

Theorem 5 *Let U_1 and $\bar{D}$ be defined as above. Then*

$$\kappa\left(\bar{D}^a U_1 \bar{D}^{1/2-a}\right) \leq n^{10} \cdot (\chi_A \|A\|)^{8a+24} \tag{16}$$

for $-\frac{1}{2} \leq a \leq \frac{1}{2}$.

Proof: First, notice that

$$R^T U_1^{-1} = Z_1.$$

Let $\mathbf{v}_k = U_1^{-1}(1:k, k)$, $\tilde{R} = R(1:k,:)$, and $\tilde{D} = D(1:k, 1:k)$. It follows from the fact that $U_1^{-T} R R^T U_1^{-1} = I$ that

$$\frac{1}{u_{kk}} \tilde{R}\tilde{R}^T \mathbf{v}_k = \mathbf{e}_k,$$

where $\mathbf{e}_k$ is the k^{th} column of the $k \times k$ identity matrix. So,

$$\begin{aligned}
\mathbf{v}_k &= u_{kk}\left(\tilde{R}\tilde{R}^T\right)^{-1} \mathbf{e}_k \\
&= \left(\mathbf{z}_k^T \mathbf{r}_k\right) \tilde{D}^{1/2} \left(\tilde{D}^{1/2}\tilde{R}\tilde{R}^T\tilde{D}^{1/2}\right)^{-1} \tilde{D}^{1/2}\mathbf{e}_k \\
&= \left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2} \tilde{D}^{1/2}\mathbf{x},
\end{aligned}$$

where $\mathbf{z}_k$ is the kth column of Z_1, $\mathbf{r}_k$ is the kth column of R^T and $\mathbf{x}$ is the last column of $\left(\tilde{D}^{1/2}\tilde{R}\tilde{R}^T\tilde{D}^{1/2}\right)^{-1}$. Multiplying both sides by $d_k^{-a}\tilde{D}^{a-1/2}$ yields

$$d_k^{-a}\tilde{D}^{a-1/2}\mathbf{v}_k = \left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2-a}\tilde{D}^a\mathbf{x}.$$

We show that there is an upper bound on the right-hand side as follows:

$$\begin{aligned}
\|\left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2-a}\tilde{D}^a\mathbf{x}\| &= d_k^{1/2} \cdot \left|\mathbf{z}_k^T\mathbf{r}_k\right| \cdot \|d_k^{-a}\tilde{D}^a\mathbf{x}\| \\
&\leq d_k^{1/2} \cdot \|\mathbf{r}_k\| \cdot \|\tilde{D}^a \left(\tilde{D}^{1/2}\tilde{R}\tilde{R}^T\tilde{D}^{1/2}\right)^{-1} \tilde{D}^{-a}\| \\
&\leq \|A\| \cdot (\chi_A\|A\|)^{4a} \cdot \|\left(\tilde{D}^{1/2+a}\tilde{R}\tilde{R}^T\tilde{D}^{1/2-a}\right)^{-1}\| \\
&= \|A\| \cdot (\chi_A\|A\|)^{4a} \cdot \frac{\kappa\left(\tilde{D}^{1/2+a}\tilde{R}\tilde{R}^T\tilde{D}^{1/2-a}\right)}{\|\tilde{D}^{1/2+a}\tilde{R}\tilde{R}^T\tilde{D}^{1/2-a}\|} \\
&\leq \|A\| \cdot (\chi_A\|A\|)^{4a} \cdot \frac{\kappa\left(\tilde{D}^{1/2+a}\tilde{R}D^{-a}D^a\tilde{R}^T\tilde{D}^{1/2-a}\right)}{\bar{r}_{11}^2} \\
&\leq \chi_A \cdot (\chi_A\|A\|)^{4a+1} \cdot \kappa\left(\tilde{D}^{1/2+a}\tilde{R}D^{-a}\right) \cdot \kappa\left(D^a\tilde{R}^T\tilde{D}^{1/2-a}\right).
\end{aligned}$$

In the fifth line of the above inequality, $\bar{r}_{11}$ is the (1,1) entry of $\bar{R} = Q^T A^T$. Notice that if $-\frac{1}{2} \leq a \leq \frac{1}{2}$, then Theorem 3 applies. So,

$$\begin{aligned}
\|d_k^{-a} \tilde{D}^{1/2-a} \mathbf{v}_k\| &\leq \chi_A \cdot (\chi_A \|A\|)^{4a+1} \cdot \kappa\left(\tilde{D}^{1/2+a} \tilde{R} D^{-a}\right) \cdot \kappa\left(D^a \tilde{R}^T \tilde{D}^{1/2-a}\right) \\
&\leq n^8 \cdot \chi_A \cdot (\chi_A \|A\|)^{4a+21}.
\end{aligned}$$

Now,

$$\begin{aligned}
\|\bar{D}^{a-1/2} U_1^{-1} \bar{D}^{-a}\| &\leq \sum_{i=1}^{n} \|d_i^{-a} \bar{D}^{a-1/2} \mathbf{v}_i\| \\
&= n \cdot \max_{1 \leq i \leq n} \|d_i^{-a} \tilde{D}^{a-1/2} \mathbf{v}_i\| \\
&\leq n^9 \cdot \chi_A \cdot (\chi_A \|A\|)^{4a+21},
\end{aligned}$$

for $-\frac{1}{2} \leq a \leq \frac{1}{2}$. In order to find and upper bound on $\|\bar{D}^a U_1 \bar{D}^{1/2-a}\|$, we do the following:

$$\begin{aligned}
\|\bar{D}^a U_1 \bar{D}^{1/2-a}\| &\leq \sum_{i=1}^{n} \|d_i^{1/2-a} \bar{D}^a \mathbf{u}_i\| \\
&\leq n \cdot \max_{1 \leq i \leq n} \|d_i^{1/2-a} \bar{D}^a \mathbf{u}_i\| \\
&\leq n \cdot (\chi_A \|A\|)^{4a} \cdot \max_{1 \leq i \leq n} \|d_i^{1/2} Z_1^T \mathbf{r}_i\| \\
&= n \cdot (\chi_A \|A\|)^{4a} \cdot \|d_i^{1/2} D^{-1/2} \bar{\mathbf{r}}_i\| \\
&\leq n \cdot \|A\| \cdot (\chi_A \|A\|)^{4a+2},
\end{aligned}$$

where $\mathbf{u}_i$ is the ith column of U_1, $\mathbf{r}_i$ is the ith column of R^T, and $\bar{\mathbf{r}}_i$ is the ith column of $\bar{R}$. The third line of the inequality is derived from the second using (10), the fact that U_1 is upper triangular, and $U_1 = Z^T R^T$. The last line is obtained by applying Theorem 3. Thus for $-\frac{1}{2} \leq a \leq \frac{1}{2}$,

$$\begin{aligned}
\kappa\left(\bar{D}^a U_1 \bar{D}^{1/2-a}\right) &= \|\bar{D}^a U_1 \bar{D}^{1/2-a}\| \cdot \|\bar{D}^{a-1/2} U_1^{-1} \bar{D}^{-a}\| \\
&\leq n^{10} \cdot (\chi_A \|A\|)^{8a+24}.
\end{aligned}$$

$\square$

Now that we know that $\bar{D}^a U_1 \bar{D}^{1/2-a}$ is well-conditioned for $-\frac{1}{2} \leq a \leq \frac{1}{2}$, we move on to the analysis of the remainder of the algorithm.

5 Finding the Solution y

In analyzing the remainder of the algorithm, we first show that a small error introduced in the back substitution step. In Step 3 of the algorithm, the upper triangular system

$$U_1 \bar{\mathbf{y}} = Z_1^T D^{-1/2} \mathbf{b}$$

is solved for $\bar{\mathbf{y}}$. (Note that this is slightly different from the system given in Step 3 of the algorithm as presented in §1 because we are assuming that the columns of $A^T D^{-1/2}$ have been "pre-pivoted.") Instead of working with the system given above, consider the following system:

$$\bar{D}^{1/2} U_1 \bar{\mathbf{y}} = \bar{D}^{1/2} Z_1^T D^{-1/2} \mathbf{b},$$

where $D = \mathrm{diag}(d_1, d_2, \ldots, d_m)$ and $\bar{D} = D(1:n, 1:n)$ as before. In working through the steps of back substitution, we find that solving this system is equivalent to solving the original one, even in floating point arithmetic. (In other words, a rescaling of the rows does not change the numerical bounds.) Recall from the last section that $\bar{D}^a U_1 \bar{D}^{1/2-a}$ is well-conditioned for $-\frac{1}{2} \leq a \leq \frac{1}{2}$. Therefore, standard back substitution results apply when showing that the error at this step is small. The following theorem states that error bound.

Theorem 6 *Let $\bar{\mathbf{y}}$ be the exact solution to $\bar{D}^{1/2} U_1 \bar{\mathbf{y}} = \bar{D}^{1/2} Z_1^T D^{-1/2} \mathbf{b}$, and let $\check{\mathbf{y}}$ be the computed solution. Then*

$$\|\bar{\mathbf{y}} - \check{\mathbf{y}}\| \leq \epsilon \cdot n^{29} \cdot \chi_A \cdot (\chi_A \|A\|)^{75} \cdot \|\mathbf{b}\| + O(\epsilon^2). \tag{17}$$

Proof: Let $\check{\mathbf{y}}$ be the computed solution to the above system. Then $\check{\mathbf{y}}$ is the exact solution to the nearby system of equations,

$$(\bar{D}^{1/2} U_1 + E) \check{\mathbf{y}} = \bar{D}^{1/2} Z_1^T D^{-1/2} \mathbf{b}.$$

The matrix E accounts for errors during the back substitution and $|E| \leq \epsilon \cdot \left|\bar{D}^{1/2} U_1\right|$, where ϵ is machine roundoff [4]. So,

$$\bar{D}^{1/2} U_1 \bar{\mathbf{y}} - (\bar{D}^{1/2} U_1 + E) \check{\mathbf{y}} = \mathbf{0},$$

or

$$\bar{\mathbf{y}} - \check{\mathbf{y}} = (\bar{D}^{1/2} U_1)^{-1} E \check{\mathbf{y}}.$$

Substituting for $\check{\mathbf{y}}$ on the right-hand side yields

$$\bar{\mathbf{y}} - \check{\mathbf{y}} = (D^{1/2} U)^{-1} E (D^{1/2} U + E)^{-1} \bar{D}^{1/2} Z_1^T D^{-1/2} \mathbf{b}.$$

Thus,

$$\begin{aligned}
\|\bar{\mathbf{y}} - \check{\mathbf{y}}\| &\leq \|(\bar{D}^{1/2} U_1)^{-1}\| \cdot \|E\| \cdot \|(\bar{D}^{1/2} U_1 + E)^{-1}\| \cdot \|\bar{D}^{1/2} Z_1^T D^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\leq \epsilon \cdot \|(\bar{D}^{1/2} U_1)^{-1}\| \cdot \|\bar{D}^{1/2} U_1\| \cdot \|(\bar{D}^{1/2} U_1 + E)^{-1}\| \cdot \|\bar{D}^{1/2} U_1^{-T} R D^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\leq \epsilon \cdot \|(\bar{D}^{1/2} U_1)^{-1}\| \cdot \|\bar{D}^{1/2} U_1\| \cdot (\|I\| + \|(\bar{D}^{1/2} U_1)^{-1} E\| + \|(\bar{D}^{1/2} U_1)^{-1} E\|^2 \\
&\quad + \|(\bar{D}^{1/2} U_1)^{-1} E\|^3 + \cdots) \cdot \|(\bar{D}^{1/2} U_1)^{-1}\| \cdot \|\bar{D}^{1/2} U_1^{-T} \bar{D}^{-1} \bar{D} R D^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\leq \epsilon \cdot \kappa(\bar{D}^{1/2} U_1) \cdot \|(\bar{D}^{1/2} U_1)^{-1}\| \cdot \|\bar{D}^{1/2} U_1^{-T} \bar{D}^{-1}\| \cdot \|\bar{D} R D^{-1/2}\| \cdot \|\mathbf{b}\| + \mathrm{O}(\epsilon^2) \\
&\leq \epsilon \cdot n^{29} \cdot \chi_A \cdot (\chi_A \|A\|)^{75} \cdot \|\mathbf{b}\| + \mathrm{O}(\epsilon^2),
\end{aligned}$$

as claimed. □

In the theorem above, the errors in the computation of U_1 itself (which also contribute to the error in $\bar{\mathbf{y}}$) are not included, but be could accounted for as a somewhat larger perturbation matrix E. As we we have already argued at the end of §3, the errors in computing U_1 are small. In fact, the errors made in forming each row of U_1 are small with respect to the norm of that row. Therefore, the perturbation matrix E is small with respect to $\bar{D}^{1/2}U_1$. Explicitly including this analysis in the previous theorem would make the proof more complicated, but the bound would be qualitatively the same.

The final step is to obtain $\mathbf{y}$ by multiplying $\bar{\mathbf{y}}$ by Q. Let $\hat{\mathbf{y}}$ be the computed result. We assume that $\hat{\mathbf{y}}$ accounts for the errors during both this step and the previous step. The error bound is obtained as follows:

$$\begin{aligned}
\|\mathbf{y}-\hat{\mathbf{y}}\| &\leq \epsilon\cdot n\cdot\|\mathbf{y}\|+\|\bar{\mathbf{y}}-\check{\mathbf{y}}\| \\
&\leq \epsilon\cdot n\cdot\left\|\left(\bar{D}^{1/2}U_1\right)^{-1}\right\|\cdot\|\bar{D}^{-1}U_1^{-1}\bar{D}^{1/2}\|\cdot\|\bar{D}RD^{-1/2}\|\cdot\|\mathbf{b}\|+\|\bar{\mathbf{y}}-\check{\mathbf{y}}\| \\
&\leq \epsilon\cdot\left[n^{19}\cdot\chi_A\cdot(\chi_A\|A\|)^{47}+n^{29}\cdot\chi_A\cdot(\chi_A\|A\|)^{75}\right]\cdot\|\mathbf{b}\|+\mathrm{O}(\epsilon^2).
\end{aligned}$$

Notice that the error bound is of the form

$$\|\mathbf{y}-\hat{\mathbf{y}}\|\leq\epsilon\cdot f(A)\cdot\|\mathbf{b}\|.$$

Thus, our algorithm satisfies the definition of stability.

6 Summary and Open Questions

The least-squares problem

$$\min_{\mathbf{y}\in R^n}\|D^{-1/2}\left(A\mathbf{y}-\mathbf{b}\right)\|,$$

where D is an $m\times m$ diagonal, positive definite, ill-conditioned matrix; A is an $m\times n$ full-rank matrix; $\mathbf{y}$ and $\mathbf{b}$ are n-vectors, has a unique solution. Because of the ill-conditioning of D, the standard methods for solving least-squares problems do not find an accurate solution. In an attempt to find a standard algorithm that will solve this problem accurately, we have employed complete orthogonal decomposition. This involves four steps, given in §1. We then proceeded to show that this algorithm is stable, as defined in §1.

The first step is a "stabilizing" QR factorization that gives an upper triangular matrix that is well-conditioned up to a scaling of the rows or the columns. Using this, we are able to show that the result of the second step is an upper triangular matrix that is also well-conditioned up to a scaling of the rows or the columns. We were then able to show that there is a bound on the error in the back substitution step. It was then easy to show that there is a small error introduced in the last step. Thus, this algorithm gives an accurate solution to this least-squares problem.

Now that we know that the algorithm is stable, there are a couple of further questions.

1. This algorithm has been implemented using dense methods. In many applications, the matrix A is sparse. Can we implement this algorithm in such a way that it takes advantage of that sparsity?

2. The results thus far are theoretical. This algorithm has not yet been extensively tested in applications. The question, then, is whether or not this algorithm is effective in applications. We are currently beginning tests of our algorithm in interior point methods [6].

The problem of stably solving the ill-conditioned equilibrium system in barrier methods for optimization has received a fair amount of attention [2]. In the case of barrier methods for linear programming (that is, interior point methods), the equibrium system reduces to weighted least squares, which is the problem addressed by this paper. Other authors have recently looked at ill-conditioning in barrier methods including Coleman and Liu [1], M. Wright [15], S. Wright [16], Nash and Sofer [7], and Gould [5].

The differences between these other works and ours may be summarized as follows. These other works typically look at the more general problem $\min \|H^{-1/2}(A\mathbf{y}-\mathbf{b})\|$ where H is symmetric, positive definite, but not necessarily diagonal, which is a problem that we currently cannot address with our techniques. On the other hand, these other authors all make an assumption that the large and small entries on the diagonal of H have some correlation with the columns of A^T. This corresponds to a nondegeneracy assumption about the underlying optimization problem. In contrast, our method does not involve any restrictions about where "large" versus "small" entries of D can appear, and thus our method has no difficulty when there is degeneracy or near-degeneracy in the underlying optimization problem.

References

[1] T. F. Coleman and J. Liu. An interior Newton method for quadratic programming. Technical Report CTC93TR153, Advanced Computing Research Institute, Cornell University, Ithaca, NY, 1993.

[2] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London, 1981.

[3] G. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7:206–216, 1965.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations, 2nd Edition.* Johns Hopkins Univ. Press, Baltimore, 1989.

[5] N. Gould. On the accurate determination of search directions for simple differentiable penalty functions. *IMA Journal of Numerical Analysis*, 6:357–372, 1986.

[6] P. D. Hough. Complete orthogonal decomposition in interior point methods. Work in progress, 1995.

[7] S. G. Nash and A. Sofer. A barrier method for large-scale constrained optimization. *ORSA Journal on Computing*, 5:40–53, 1993.

[8] G. W. Stewart. On scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 112:189–193, 1989.

[9] G. Strang. A framework for equilibrium equations. *SIAM Review*, 30:283–297, 1988.

[10] M. J. Todd. A Dantzig-Wolfe-like variant of Karmarkar's interior-point linear programming algorithm. *Operations Research*, 38:1006–1018, 1990.

[11] S. A. Vavasis. Stable finite elements for problems with wild coefficients. Technical Report TR-93-1364, Department of Computer Science, Cornell University, Ithaca, N. Y., 1993.

[12] S. A. Vavasis. Stable numerical algorithms for equilibrium systems. *SIAM J. Matrix Anal. Appl.*, 15:1108–1131, 1994.

[13] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, New York, 1965.

[14] M. H. Wright. Interior methods for constrained optimization. In *Acta Numerica 1992*. Cambridge University Press, Cambridge, 1992.

[15] M. H. Wright. Determining subspace information from the Hessian of a barrier function. Technical Report 92–02 AT&T Bell Laboratories Numerical Analysis Manuscript, 1992.

[16] S. J. Wright. Stability of linear algebra computations in interior-point methods for linear programming. Technical Report MCS-P446-0694, Mathematics and Computer Science Division, Argonne National Laboratory, Chicago, IL, 1994.