Abstract. Block modified Gram-Schmidt (BMGS) algorithms are developed to factor a full column rank matrix $X \in \mathbb{R}^{m \times n}$, $m \geq n$, into $Q \in \mathbb{R}^{m \times n}$ and upper triangular $R \in \mathbb{R}^{n \times n}$ such that $X = QR$.

and, in exact arithmetic, $Q$ is left orthogonal, i.e., $Q^T Q = I_n$.

The new block algorithms build upon the block Householder representation of Schreiber and Van Loan [R. Schreiber and C.F. Van Loan, A storage-efficient WY representation for products of Householder transformations, SIAM J. Sci. Stat. Computing, 10:53–57, 1989] and an observation by Charles Sheffield analyzed by Paige [C.C. Paige, A useful form of unitary matrix from any sequence of unit 2-norm $n$-vectors, SIAM J. Matrix Anal. Appl., 31(2):565–583, 2009] about the relationship between modified Gram-Schmidt and Householder QR factorization. Using the Sheffield framework, we show that the new BMGS algorithms have a similar relationship to Householder QR factorization and thus similar error analysis properties to modified Gram-Schmidt. Since these algorithms are based upon matrix-matrix operations they are more suitable for cache based computer architectures.

AMS subject classifications. 65F25, 65F20, 65F35

Key words. Q-R decompositions, matrix-matrix operations, block algorithms, orthogonality, residuals, error bounds.

1. Introduction. For given integers $m$ and $n$, $m \geq n$, and a full column rank matrix $X \in \mathbb{R}^{m \times n}$, we develop block modified Gram-Schmidt (BMGS) algorithms that factor $X$ into $Q \in \mathbb{R}^{m \times n}$ and upper triangular, nonsingular $R \in \mathbb{R}^{n \times n}$ such that

$X = QR$ (1.1)

where, in exact arithmetic, $Q$ is left orthogonal, i.e., $Q^T Q = I_n$.

Our BMGS algorithms are built upon rewriting the modified Gram-Schmidt (MGS) [16, §5.2.8] Q-R factorization as a BLAS-3 algorithm [14], that is, one primarily based upon matrix-matrix operations, thus making it more efficient on cache-based computer architectures and more suitable for some distributed computing environments.

The framework for BMGS blends two closely related ideas. The first is a connection between Householder and MGS Q-R factorization first observed by Charles Sheffield and communicated to Gene Golub. That observation states that the matrices $Q$ and $R$ from MGS implicitly produce the Q-R factorization

$X = QR$ (1.2)

where $U \in \mathbb{R}^{(m+n) \times (m+n)}$ is product of Householder matrices. Moreover, this equivalence holds in finite precision arithmetic. We design BMGS algorithms that are able to maintain this structure with small backward error in floating point arithmetic.

The Sheffield structure has been used by Björck and Paige [11] and by Paige, Rožložnik, and Strakoš [24] in the development and analysis of MGS related algorithms. It has also been used by Barlow, Bosner, and Drmač [5], by Bosner and Barlow [12],

The second idea is the Schreiber-Van Loan [27] representation of products of Householder transformations. Since the columnwise version of MGS [16, §5.2.8] is a BLAS-1 [22] (vector operation based) algorithm, this representation is useful in developing BLAS-2 and BLAS-3 versions of MGS. We construct a BLAS-3 version that uses a “tall, skinny Q-R” (TSQR) factorizations [17, 20, 2] other than MGS for a key intermediate computation. This structure allows us to give sufficient conditions for a BMGS algorithm to have similar properties in floating point arithmetic to those of MGS. The BMGS algorithms presented by Jalby and Phillippe [21] and Vanderstraeten [33] can be understood within the structure presented here.

Modified Gram-Schmidt is discussed by Rice [26] in an important experimental paper. The theory behind the algorithm’s error analysis properties was first laid down by Björck [8]. A block classical Gram-Schmidt algorithm with reorthogonalization (BCGS2) is given and analyzed by Barlow and Smoktunowicz [6]. For the case of single columns the algorithm in [6] is that of Abdelmalek [1] and Giraud et al. [15]. BCGS2 is used to develop a block downdating algorithms in [4]. A similar block algorithm based upon classical Gram-Schmidt (CGS), justified only by numerical tests, is proposed by Stewart [32]. Other interesting and closely related block orthogonal decompositions and discussed by Strathopoulos and Wu [31] and Yamazaki, Tomov, and Dongarra [34].

Gram-Schmidt based algorithms are important for the implementation of Krylov space methods such as GMRES [24]. Block GMRES methods for linear systems with multiple right hand sides are discussed by Soodhalter [30] and Simoncini and Gallopoulos [29, 28]. Block Gram-Schmidt algorithms could also be used in implementing the block Lanczos algorithm of Golub and Underwood [18]. Gram-Schmidt algorithms are useful in any use of Q-R decomposition for which the blocks of columns of X are introduced one at a time. For a summary of the role of Gram-Schmidt algorithms is see [10, §2.4, §3.2] or [9].

In the next section, we review the MGS algorithm and relate the Sheffield structure to the Schreiber-Van Loan Householder Q-R representation. In §3, we give three algorithms with MGS-like properties. First, in §3.1, we simply rewrite the MGS algorithm into BLAS-2 and BLAS-3 algorithms using the ideas in §2 whereas in §3.2 we replace a key intermediate Q-R decomposition step that uses MGS with one that uses Householder factorization.

The error analysis of the three algorithms in §3 is presented in §4. That analysis begins with an appropriate perturbation theory in §4.1. The BLAS-2 algorithm given in §3.1 is analyzed in §4.2. The BLAS-3 algorithms in §3.1 and §3.2 are analyzed in §4.3. After bounds three key residuals, all of the algorithms in §3 are shown to fit the Sheffield structure outlined in §2.3 with small backward error. Moreover, we give a set of conditions for which block modified Gram-Schmidt algorithms can be shown to conform to that structure with small backward error. The analysis in §4.3 assume that X is not so ill-conditioned as to be considered “singular to machine precision.”

In §§7, we review the block MGS algorithm of Jalby and Phillippe [21] along with a proposed modification by Vanderstraeten [33] and interpret them in the context of our work in §3 and §4. The proofs of two key theorems from §4.3 are given in §5. A conclusion is given in §6.

2. MGS, the Sheffield Connection, and Block Householder Transformations.

2.1. Some Block Notation. To build the MGS-type algorithms, we partition X and Q into

\[(2.1) \quad X = (X_1, X_2, \ldots, X_s), \quad Q = (Q_1, Q_2, \ldots, Q_s),\]
where $X_k, Q_k \in \mathbb{R}^{m \times p_k}, k = 1, \ldots, s$ and $n = \sum_{k=1}^{s} p_k$. For the sake of simplicity, our block algorithms and their analysis assumes that $p_1 = p_2 = \cdots = p_s = p$ and $n = ps$.

We denote the matrices with the first $k$ blocks of $X, Q$ above as
\[
\hat{X}_k = (X_1, X_2, \ldots, X_k), \quad \hat{Q}_k = (Q_1, Q_2, \ldots, Q_k).
\]

We also use this notation for the special cases when $p = 1$ with $X_k = x_k$ and $Q_k = q_k$ for $x_k, q_k \in \mathbb{R}^m$. Note that through this paper, the integer $k$ indexes the number of blocks. It indexes the number of columns only when $p = 1$.

2.2. The MGS Algorithm. The columnwise version of the MGS algorithm [10, p.62] applied to $X \in \mathbb{R}^{m \times n}$ is given next as Function 2.1. The algorithm as stated below is entirely vector operations, thus it is BLAS-1 [22].

**Function 2.1 ( Modified Gram-Schmidt ).**

\begin{verbatim}
function [Q, R] = MGS(X)
(1) [m, n] = size(X);
(2) r_{11} = \|x_1\|_2; q_1 = x_1 / r_{11};
(3) for k = 2:n
(4) y_k = x_k;
(5) for j = 1:k-1
(6) r_{jk} = q_j^T y_k;
(7) y_k = y_k - r_{jk} q_j;
(8) end;
(9) r_{kk} = \|y_k\|_2; q_k = y_k / r_{kk};
(10) end;
R = (r_{jk}); Q = (q_1, \ldots, q_n);
end;
end;
end;
\end{verbatim}

2.3. Block Householder structure of MGS. To assemble BLAS-2 and BLAS-3 versions of MGS, we rewrite the factorization (1.2) by combining a framework discovered by Charles Sheffield with the block Householder representation given by Schreiber and Van Loan [27].

To begin, for the integers $m$ and $n$ in (1.1), consider an orthogonal matrix $U \in \mathbb{R}^{(m+n) \times (m+n)}$ given by
\[
U = P_1 \cdots P_n
\]
where
\[
P_k = I_{m+n} - w_k w_k^T, \quad \|w_k\|_2 = \sqrt{2}
\]
is a Householder transformation for $k = 1, \ldots, n$.

Using Schreiber and Van Loan’s [27] formulation of block Householder transformations, we let
\[
W = (w_1, \ldots, w_n)
\]
be the matrix of Householder vectors from (2.3)-(2.4), then $U$ may be represented as
\[
U = I_{m+n} - W W^T
\]
for a unit upper triangular $T$. From Pugilisi [25], if we let $S \in \mathbb{R}^{n \times n}$ be the unit upper triangular matrix such that
\[
W^T W = S + S^T
\]
then, in exact arithmetic,

\[ T = S^{-1}. \]  

(2.8)

Schreiber and Van Loan [27] proposed (2.6) as a variant on the BLAS-3 [14] representation of products of Householder transformations from Bischof and Van Loan [7]. They compute the matrix \( T \) according to the recurrence

\[ T_1 = T(1, 1) = (1), \]  

(2.9)

\[ T_k = T(1; k, 1; k) = \begin{pmatrix} k - 1 & 1 \\ 1 & 1 \end{pmatrix}, \]  

(2.10)

\[ g_k = -T_{k-1}W_{k-1}^T w_k, \quad W_{k-1} = (w_1, \ldots, w_{k-1}). \]  

(2.11)

Charles Sheffield first pointed out to Gene Golub that, if Function 2.1 is used to produce the Q-R factorization in (1.1), the resulting matrices \( Q \) and \( R \) satisfy (1.2) where \( U \) is given by (2.3), \( P_k \) is given in (2.4), and

\[ w_k = \begin{pmatrix} -e_k \\ q_k \end{pmatrix}, \quad k = 1, \ldots, n \]

or, equivalently,

\[ W = \begin{pmatrix} -I_n \\ Q \end{pmatrix}. \]  

(2.12)

Some algebra on the expression (2.7) reveals that, for \( W \) in (2.12), \( S \) is given by

\[ S = \text{triu}(Q^T Q) \]  

(2.13)

where \( \text{triu}(\cdot) \) denotes the upper triangular part of the contents.

In the recurrence (2.9)-(2.11), the first statement in (2.11) becomes

\[ g_k = -T_{k-1} \tilde{Q}_{k-1}^T q_k \]  

(2.14)

where \( \tilde{Q}_{k-1} \) is defined by (2.2). Moreover, we have that \( S_k = S(1; k, 1; k) \) and (in exact arithmetic) \( T_k \) in (2.9)-(2.11) satisfy

\[ S_k = \text{triu}(\tilde{Q}_k^T \tilde{Q}_k), \quad T_k = S_k^{-1}. \]  

(2.15)

Using the definition of \( W \) in (2.12), \( U \) in (2.3) has the block form

\[ U = n \begin{pmatrix} I_n - T \\ QT \\ I_m - QTQ^T \end{pmatrix}. \]  

(2.16)

We will also make reference to \( \tilde{Q} \in \mathbb{R}^{m \times n} \) given by

\[ \tilde{Q} = QT = U(n + 1; m + n, 1; n). \]  

(2.17)

In the next section, we given three versions of MGS that implicitly yield a factorization of the form in (1.2) with \( U \) having the structure in (2.16).
### 3. BLAS-2 and BLAS-3 Generalizations of MGS.

#### 3.1. Using the Sheffield Structure and Developing a Block Algorithm.

Using (2.3) to rewrite Function 2.1 in terms of the factorization (1.2), the $k$th step is

$$U_{k-1} = P_1 \cdots P_{k-1}$$

which has the form

$$U_{k-1} = \begin{pmatrix}
I_{k-1} - T_{k-1} & 0 & T_{k-1}Q_{k-1}^T \\
0 & I_{n-k+1} & 0 \\
Q_{k-1}T_{k-1} & 0 & I_m - \tilde{Q}_{k-1}T_{k-1}\tilde{Q}_{k-1}^T
\end{pmatrix}$$

where $\tilde{Q}_{k-1}$ is from (2.2).

The $k$th columns of $Q$ and $R$ are computed according to

$$P_{k-1} \cdots P_1 \begin{pmatrix} 0 \\ x_k \end{pmatrix} = U_{k-1}^T \begin{pmatrix} 0 \\ x_k \end{pmatrix} = \begin{pmatrix} h_k \\ x_k - \tilde{Q}_{k-1}h_k \end{pmatrix}$$

where

$$h_k = T_{k-1}^T\tilde{Q}_{k-1}^Tx_k, \quad y_k = x_k - \tilde{Q}_{k-1}h_k.$$  

We then compute $r_{kk}$ and $q_k$ as in step (9) of Function 2.1.

Combining these operations with the recurrence for $T_k$ given in (2.9)–(2.10), we have a BLAS-2 version of MGS given by Function 3.1 that also outputs $T$.

**Function 3.1 (BLAS-2 MGS Algorithm).**

```matlab
function [Q, R, T] = MGS2(X)

% (1) [m, n] = size(X)
% (2) r_11 = ||x_1||_2; q_1 = x_1/r_11;
% (3) R = (r_11); Q_1 = (q_1), T_1 = (1);
% (4) for k = 2:n
% (5) h_k = \tilde{Q}_{k-1}^Tx_k;
% (6) h_k = T_{k-1}^Tq_k;
% (7) y_k = x_k - \tilde{Q}_{k-1}h_k;
% (8) r_{kk} = ||y_k||_2; q_k = y_k/r_{kk};
% (9) g_k = \tilde{Q}_{k-1}^Tq_k;
% (10) g_k = -T_{k-1}g_k;
% (11) Q_k = \begin{pmatrix} \tilde{Q}_{k-1} & q_k \end{pmatrix}; R_k = \begin{pmatrix} R_{k-1} & h_k \\ 0 & r_{kk} \end{pmatrix};
% (12) T_k = \begin{pmatrix} T_{k-1} & g_k \\ 0 & 1 \end{pmatrix}
% (13) end;
Q = \tilde{Q}_n; R = R_n; T = T_n
```
Function 3.1 gives us $Q$ and $R$ satisfying (1.1) and $T$ satisfying (2.7)–(2.6). The important operations in Function 3.1 are (5),(6),(7),(9), and (10) and these are all matrix-vector products, making this a BLAS-2 algorithm.

To build a BLAS-3 version of MGS, we assume that $X$ is partitioned into blocks of the form as (2.1). Let $Q$ be as in (2.1) and let $R$ be partitioned according to

\[
R = \begin{pmatrix}
R_{11} & R_{12} & \cdots & \cdots & \cdots & R_{1s} \\
0 & R_{22} & R_{23} & \cdots & \cdots & R_{2s} \\
0 & 0 & R_{33} & \cdots & \cdots & R_{3s} \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & R_{s-1,s-1} & R_{s-1,s} \\
0 & \cdots & \cdots & 0 & R_{ss}
\end{pmatrix}, \quad R_{ij} \in \mathbb{R}^{p \times p},
\]

and assume that $T$ and $S$ in (2.13) are partitioned conformally. We let $R_k$ be given by

\[
R_k = \begin{pmatrix}
R_{11} & R_{12} & \cdots & \cdots & \cdots & R_{1k} \\
0 & R_{22} & R_{23} & \cdots & \cdots & R_{2k} \\
0 & 0 & R_{33} & \cdots & \cdots & R_{3k} \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & R_{k-1,k-1} & R_{k-1,k} \\
0 & \cdots & \cdots & 0 & R_{kk}
\end{pmatrix},
\]

and let

\[
H_k = \begin{pmatrix}
R_{1k} \\
R_{2k} \\
\vdots \\
R_{k-1,k}
\end{pmatrix}.
\]

Also, define $T_k$ conformally to $R_k$ in (3.3) and let $G_k$ be given by

\[
G_k = \begin{pmatrix}
T_{1k} \\
T_{2k} \\
\vdots \\
T_{k-1,k}
\end{pmatrix}.
\]

We can then block partition the MGS2 version of MGS as follows

\[
[Q_1, R_{11}, T_{11}] = \text{MGS2}(X_1),
\]

\[
H_k = \bar{T}^T_{k-1} \bar{Q}^T_{k-1} X_k, \quad k = 1, \ldots, s
\]

\[
Y_k = X_k - \bar{Q}_{k-1} H_k,
\]

\[
[Q_k, R_{kk}, T_{kk}] = \text{MGS2}(Y_k).
\]

To obtain a formula for $G_k$ in (3.4) while maintaining the relationship (2.15), we note that the $(k,k)$ block of $S$ in (2.7) is given by

\[
S_{kk} = \text{triu}(Q_k^T Q_k),
\]

we use the standard block inverse formula

\[
S_k^{-1} = \begin{pmatrix}
S_{k-1}^{-1} & -S_{k-1}^{-1} \bar{Q}^T_{k-1} Q_k S_{kk}^{-1} \\
0 & S_{kk}^{-1}
\end{pmatrix},
\]
and we enforce $T_k = S_k^{-1}$ to yield the recurrence

$$T_k = \begin{pmatrix} T_{k-1} & G_k \\ 0 & T_{kk} \end{pmatrix}$$

(3.11)

where

$$G_k = -T_{k-1}Q_{k-1}^TQ_kT_{kk}.$$  

(3.12)

Thus, we have the factorization

$$\tilde{X}_k = \tilde{Q}_kR_k, \quad \tilde{Q}_k = \begin{pmatrix} \tilde{Q}_{k-1} \\ Q_k \end{pmatrix},$$

$$R_k = \begin{pmatrix} R_{k-1} & H_k \\ 0 & R_{kk} \end{pmatrix}.$$  

Using the above development, the following is a formal statement of our algorithm.

**Function 3.2 (MGS3–BLAS-3 Version of the MGS Algorithm).**

```latex
function \[Q, R, T\] =MGS3(X,p)
\begin{enumerate}
\item \[m,n\] = size(X); \(s = n/p\);
\item \[\tilde{Q}_1, R_{11}, T_1\] = MGS2(X_1);
\item \textbf{for} \(k = 2:s\)
\begin{enumerate}
\item \(H_k = \tilde{Q}_{k-1}^T X_k\);
\item \(H_k = T_{k-1}^T H_k\);
\item \(Y_k = X_k - \tilde{Q}_{k-1} H_k\);
\item \([Q_k, R_{kk}, T_{kk}] = \text{MGS2}(Y_k)\);
\item \(F_k = \tilde{Q}_{k-1}^T Q_k\);
\item \(G_k = -T_{k-1} F_k T_{kk}\);
\item \(\tilde{Q}_k = \begin{pmatrix} \tilde{Q} \\ Q_k \end{pmatrix}\), \(R_k = \begin{pmatrix} R_{k-1} & H_k \\ 0 & R_{kk} \end{pmatrix}\)
\item \(T_k = \begin{pmatrix} T_{k-1} & G_k \\ 0 & T_{kk} \end{pmatrix}\)
\end{enumerate}
\item \textbf{end};
\item \(Q = \tilde{Q}_s\), \(R = R_s\), \(T = T_s\)
\end{enumerate}
end. MGS3
```

In the next section, we modify MGS3 by substituting Householder Q-R factorization in steps (2) and (7) for MGS2.

### 3.2. Another block MGS Algorithm.

Since steps (2) and (7) of Function 3.2 compute the factorization of $Y_k$ using Function 3.1, that function is not ideal for modern computing environments. Function 3.1 is BLAS-2 and less easily distributed among processors than the Householder-based TSQR algorithms discussed by Ballard et al. [2]. Discussion of TSQR algorithms in the literature dates back at least as far as a 1988 paper by Golub, Plemmons, and Sameh [17].

We let the statement

$$[Q_k, R_{kk}] = \text{House.QR}(Y_k)$$

(3.13)
denote the result of a Householder-based Q-R factorization applied to $Y_k$ such as those in [2]. We now present a block MGS algorithm called BMGS$_H$ that uses Householder Q-R to factor $Y_k$.

**Function 3.3 (BMGS with Householder-based Q-R).**

```matlab
function [Q, R, T] = BMGS_H(X, p)

[m, n] = size(X); s = n/p;

[Q1, R11] = House_QR(X1); T1 = Ip;

for k = 2:s

Hk = Q_k1 X_k;
Hk = T_k H_k;
Yk = X_k - Hk Q_k1;

[Q_kk, R_kk] = House_QR(Y_k);
F_k = Q_k1 Q_k;
G_k = -T_k F_k; % Note that T_kk = Ip

Q_k = (Q_k1 Q_k);
R_k = (R_k1 H_k 0 R_kk);

T_k = (T_k1 G_k 0 Ip);

end;

Q = Q_s;
R = R_s;
T = T_s;
end. BMGS_H
```

Function 3.3 is similar in structure and data movement to the block classical Gram-Schmidt algorithm BCGS2 by Barlow and Smoktunowicz [6]. However, BCGS2 does four matrix multiplications with $Q_k$ at each step compared with three for BMGS$_H$ and BCGS2 does two calls to HouseQR per step compared to one for BMGS$_H$. BMGS$_H$ also requires two multiplications with $T_k$ per step, but those will cost much less than an extra multiplication with $Q_k$ and an extra call to HouseQR. Under the assumptions of the error analysis in [6], the resulting $Q$ from BCGS2 is near left orthogonal whereas the $Q$ from Function 3.3 is not in general. However, when solving least squares problems, the two algorithms have a similar guarantee of accuracy. For the case $p = 1$, there is a similar tradeoff between Function 3.1 to the cgs2 algorithm in [15].

4. Error Analysis Results. To begin our error analysis results for the factorizations produced by Functions 3.1, 3.2, and 3.3, we develop a perturbation theory in §4.1. For that, we let $Q, R$ and $T$ be computed quantities and let $S$ be the exact matrix defined by (2.13) and define the three residual errors

\begin{align}
TS - I_n &= \Delta_{TS}, \\
QR - X &= \delta X, \\
(I_n - T)R &= \Gamma_{TR}.
\end{align}

We show that if we have bounds on $\|\Delta_{TS}\|_F$, $\|\delta X\|_F$, and $\|\Gamma_{TR}\|_F$, there exist $\tilde{U} \in \mathbb{R}^{(m+n) \times (m+n)}$ exactly orthogonal and $V \in \mathbb{R}^{m \times n}$ exactly left orthogonal such that we have bounds on $\|U - \tilde{U}\|_F$ for $U$ in (2.16), $\|Q - V\|_F$ and $\|Q - V\|_F$ for $\tilde{Q}$ in (2.17), $\|X - \tilde{U}\left( \begin{array}{c} R \\ 0 \end{array} \right)\|_F$ for $X$ in (1.2), and $\|X - VR\|_F$. We also establish bounds
on $\|Q\|_2, \|S\|_2$, and $\|T\|_2$.

In our floating point error analysis in §4.2 and §4.3, we follow a convention in [16, §2.7.7] by producing first order bounds in the machine unit $\varepsilon_M$ and attaching a term of $+O(\varepsilon_M^2)$ where appropriate. Also, to simplify the analysis, many of the perturbation bounds in §4.1 are in the two-norm, but the error analysis bounds in §4.2 and §4.3 are in the Frobenius norm.

Bounds on $\|\Delta_{TS}\|_F$, $\|\delta X\|_F$, and $\|\Gamma_{TR}\|_F$ for Function 3.1 (MGS2) are given in §4.2.

In §4.3, we show that, for Functions 3.2 and 3.3, under restrictions on $R$, for modest sized functions $f_{TR}()$, $f_X()$, and $f_{TS}()$, we have

\begin{align*}
\|\Delta_{TS}\|_F &\leq \varepsilon_M f_{TS}(m,n,p) + O(\varepsilon_M^2), \\
\|\delta X\|_F &\leq \varepsilon_M f_X(m,n,p)\|X\|_F + O(\varepsilon_M^2), \\
\|\Gamma_{TR}\|_F &\leq \varepsilon_M f_{TR}(m,n,p)\|X\|_F + O(\varepsilon_M^2).
\end{align*}

That the recurrences (2.9)-(2.11) and (3.11)-(3.12) produce $T$ satisfying the bound (4.4) is unsurprising. Any reasonable Q-R decomposition algorithm should satisfy a bound such as (4.5). However, equation (4.6) is not necessarily satisfied by all Q-R decomposition algorithms, thus its proof is key to our error analysis.

### 4.1. Perturbation Theory for MGS and BMGS Algorithms

Our first theorem relates the perturbations (4.1)-(4.3) to the distance between $U$ in (2.16) and an orthogonal matrix $\tilde{U}$ and establishes a backward error relationship between $\tilde{U}$, $X$ in (1.2), and $R$.

**Theorem 4.1.** Let $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$, and $T \in \mathbb{R}^{n \times n}$ satisfy (4.1)-(4.3) with $R$ nonsingular and $\|Qe_j\|_2 = 1$ for $j = 1, \ldots, n$. Let $\tilde{X}$ be given by (1.2), $U$ by (2.16), and $S$ by (2.13). Then $S$ is nonsingular, $\|S^{-1}\|_2 \leq 2$,

\begin{align*}
\tilde{U} &= \left( \begin{array}{cc} I_n - S^{-1} & S^{-1}Q^T \\
Q S^{-1} & I_m - QS^{-1}Q^T \end{array} \right), \\
\|U - \tilde{U}\|_F &\leq \sqrt{10}\|Q\|_2\|\Delta_{TS}\|_F \leq \sqrt{10n}\|\Delta_{TS}\|_F,
\end{align*}

is exactly orthogonal,

\begin{align*}
\|U - \tilde{U}\|_F &\leq \sqrt{10}\|Q\|_2\|\Delta_{TS}\|_F \leq \sqrt{10n}\|\Delta_{TS}\|_F,
\end{align*}

and

\begin{align*}
\tilde{X} &= U \left( \begin{array}{c} R \\
0_{m \times n} \end{array} \right), \\
\delta \tilde{X} &= m \left( \begin{array}{c} \delta X_1 \\
\delta X_2 \end{array} \right)
\end{align*}

where

\begin{align*}
\delta X_1 &= \Gamma_{TR} + \Delta_{TS}S^{-1}R \\
\delta X_2 &= \delta X - Q(\delta X_1).
\end{align*}

Thus,

\begin{align*}
\|\delta X_1\|_F &\leq \|\Gamma_{TR}\|_F + 2\|\Delta_{TS}\|_F\|R\|_2, \\
\|\delta X\|_F &\leq \|\delta X\|_F + (1 + \|Q\|_2^2)^{1/2}\|\delta X_1\|_F.
\end{align*}
Proof. From (2.13), $S$ is a unit upper triangular matrix, thus $S^{-1}$ exists. That $\tilde{U}$ in (2.16) is orthogonal can be verified by confirming that $\tilde{U}^T \tilde{U} = I_{m+n}$. Thus

$$\|I_n - S^{-1}\|_2 \leq \|\tilde{U}\|_2 = 1$$

so that

$$\|S^{-1}\|_2 \leq \|I_n\|_2 + \|I_n - S^{-1}\|_2 \leq 2.$$

From (2.16) and (4.7), we note that

$$U - \tilde{U} = \begin{pmatrix}
S^{-1} - T & (T - S^{-1})Q^T \\
Q(T - S^{-1}) & Q(S^{-1} - T)Q^T
\end{pmatrix}$$

(4.14)

Thus

$$\|U - \tilde{U}\|_F \leq \| \begin{pmatrix}
I_n & 0 \\
0 & Q
\end{pmatrix}\|_2 \|\Delta_{TS}\|_F \| \begin{pmatrix}
-S^{-1} & S^{-1}Q^T \\
S^{-1} & -S^{-1}Q^T
\end{pmatrix}\|_2$$

(4.15)

Since $S^{-1}Q^T$ is the $(1,2)$ block of the orthogonal matrix $\tilde{U}$, $\|S^{-1}Q^T\|_2 \leq 1$, thus (4.15) reads

$$\|U - \tilde{U}\|_F \leq \|Q\|_2 \|\Delta_{TS}\|_F \leq \sqrt{10}\|Q\|_2 \|\Delta_{TS}\|_F$$

(4.16)

which is the first inequality in (4.8). Since $Q$ has columns that are unit vectors, $\|Q\|_2 \leq \|Q\|_F = \sqrt{n}$, we have the second inequality in (4.8).

To get (4.9), we use (2.16) and (4.14) and note that

$$\tilde{U} \begin{pmatrix}
R \\
O_{m \times n}
\end{pmatrix} = U \begin{pmatrix}
R \\
0_{m \times n}
\end{pmatrix} + (\tilde{U} - U) \begin{pmatrix}
R \\
0_{m \times n}
\end{pmatrix}$$

$$= \begin{pmatrix}
(I_n - T)R \\
QTR
\end{pmatrix} + \begin{pmatrix}
\Delta_{TS}S^{-1}R \\
-Q\Delta_{TS}S^{-1}R
\end{pmatrix}$$

$$= \begin{pmatrix}
\Gamma_{TR} \\
X + \delta X - Q\Delta_{TS}R
\end{pmatrix} + \begin{pmatrix}
\Delta_{TS}S^{-1}R \\
-Q\Delta_{TS}S^{-1}R
\end{pmatrix}$$

$$= \tilde{X} + \delta \tilde{X}$$

where $\delta \tilde{X}$ and its blocks $\tilde{X}_1$ and $\delta \tilde{X}_2$ satisfy (4.9)–(4.11). Standard norm bounds yield (4.13)-(4.12) $\Box$

To understand the importance of bounding these quantities, we give a version of a theorem from Paige [23] for the case where $R$ is nonsingular.

**Theorem 4.2.** [23] For $X \in \mathbb{R}^{m \times n}$, let $\tilde{X}$ be given by (1.2), let $\delta \tilde{X}$ be given by (4.9) with the partitioning given there, let $\tilde{U} \in \mathbb{R}^{(m+n) \times (m+n)}$ be orthogonal, and let...
\[ Z = \tilde{U}(:, 1:n) \in \mathbb{R}^{(m+n) \times n}, \text{ let } R \in \mathbb{R}^{n \times n} \text{ be nonsingular, and let } \tilde{U} \text{ and } R \text{ satisfy } (4.9)-(4.11). \text{ If } Z \text{ is partitioned into } \]

\[
(4.17) \quad Z = \frac{n}{m} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},
\]

then there exists a left orthogonal matrix \( V \in \mathbb{R}^{m \times n} \) such that

\[
(4.18) \quad X + \delta \tilde{X} = VR
(4.19) \quad \delta \tilde{X} = FZ_1^T (\delta X_1) + \delta X_2
\]

for some \( F \in \mathbb{R}^{m \times n} \) where \( \|F\|_2 \in [0.5, 1] \).

In the context of Theorem 4.1, \( Z_1 \) in (4.17) is

\[
(4.20) \quad Z_1 = I - S^{-1} = (\delta X_1)R^{-1}.
\]

We define the useful quantity

\[
(4.21) \quad \zeta = \|Z_1\|_2 = \|\delta X_1\|_2^{-1}.
\]

The following lemma justifies an assumption about \( \zeta \).

**Lemma 4.3.** Assume the hypothesis and notation of Theorem 4.1, let \( Z \) be as in Theorem 4.2, let \( Z_1 \) and \( Z_2 \) be from (4.17), and let \( \zeta \) be defined by (4.21). Then \( \zeta < 1 \) if and only if \( \text{rank}(Q) = n \).

**Proof.** First assume that \( \text{rank}(Q) = n \). Then \( Z_2 = QS^{-1} \) also has full column rank. For any \( w \in \mathbb{R}^n \) such that \( \|w\|_2 = 1 \), we have that

\[
\|Zw\|_2^2 = \|Z_1w\|_2^2 + \|Z_2w\|_2^2 = 1.
\]

Thus

\[
\zeta^2 = \|Z_1\|_2^2 = \max_{\|w\|_2 = 1} \|Z_1w\|_2^2
= \max_{\|w\|_2 = 1} 1 - \|Z_2w\|_2^2 = 1 - \min_{\|w\|_2 = 1} \|Z_2w\|_2^2.
\]

Since \( Z_2 \) has full column rank, \( \sigma_n(Z_2) \) the smallest singular value of \( Z_2 \), satisfies

\[
\min_{\|w\|_2 = 1} \|Z_2w\|_2 = \sigma_n(Z_2) > 0.
\]

Thus,

\[
\zeta = \left(1 - \sigma_n(Z_2)^2\right)^{1/2} < 1.
\]

If we assume \( \zeta < 1 \), every step of the argument given above can be reversed to conclude that \( \text{rank}(Q) = n \). Thus \( \zeta < 1 \) and \( \text{rank}(Q) = n \) are equivalent conditions.

In Theorem 4.4 and Corollary 4.5, we bound the distance between the exactly left orthogonal matrix \( V \) from Theorem 4.2 and \( \hat{Q} \) from (1.1) and \( \hat{Q} \) in (2.17). We also give bounds on \( \|Q\|_2 \) and \( \|T\|_2 \) necessary for our error analysis proof in §5.

**Theorem 4.4.** Assume the hypothesis and notation of Theorem 4.1, let \( Z = \tilde{U}(:, 1:n) \in \mathbb{R}^{(m+n) \times n} \) be partitioned according to (4.17), let \( \hat{Q} \) be as (2.17), let \( \zeta \) be as in (4.21), and assume that \( Q \) has full column rank.
Then $\zeta < 1$ and for some exactly left orthogonal matrix $V \in \mathbb{R}^{m \times n}$,

\begin{align*}
\|Q - V\|_2 &\leq (\zeta + \zeta^2)/(1 - \zeta), \\
\|\tilde{Q} - V\|_2 &\leq (1 + \zeta)(1 + \zeta^2)\|\Delta_{TS}\|_2/(1 - \zeta) + \zeta^2, \\
\|Q\|_2 &\leq (1 + \zeta^2)/(1 - \zeta), \\
\|S^{-1}\|_2 &\leq 1 + \zeta.
\end{align*}

Proof. First, we note that $Z_1$ satisfies (4.20) and from Lemma 4.3, $\zeta < 1$. Combining (4.2) and (4.18)-(4.19), we have that

\begin{align*}
(Q - V)R &= \delta X - \delta \tilde{X} \\
&= \delta X - (FZ_1^T(\delta \overline{X}_1) + \delta \overline{X}_2) \\
&= \delta X - (FZ_1^T(\delta \overline{X}_1) + \delta X - Q(\delta \overline{X}_1)) \\
&= -FZ_1^T(\delta \overline{X}_1) + Q(\delta \overline{X}_1)
\end{align*}

where $\|F\|_2 \in [0.5, 1]$. Since $R$ is assumed to be nonsingular and $Z_1 = (\delta \overline{X}_1)R^{-1}$, we have

\begin{align*}
Q - V = -FR^{-T}(\delta \overline{X}_1)^T(\delta \overline{X}_1)R^{-1} + Q(\delta \overline{X}_1)R^{-1} \\
&= -FZ_1^T Z_1 + VZ_1 + (Q - V)Z_1.
\end{align*}

Thus,

\begin{align*}
(Q - V)(I_n - Z_1) = -FZ_1^T Z_1 + VZ_1.
\end{align*}

Since $\zeta < 1$, we have that

\begin{align*}
\|Q - V\|_2 &\leq (\|F\|_2\|Z_1\|_2^2 + \|V\|_2\|Z_1\|_2)/(1 - \|Z_1\|_2) \\
&\leq (\zeta + \zeta^2)/(1 - \zeta)
\end{align*}

which is (4.22). The bound on $\|Q\|_2$ follows from

\begin{align*}
\|Q\|_2 = \|V\|_2 + \|Q - V\|_2 &\leq 1 + (\zeta + \zeta^2)/(1 - \zeta) = (1 + \zeta^2)/(1 - \zeta).
\end{align*}

To obtain (4.25), note that

\begin{align*}
\|S^{-1}\|_2 = \|I_n + I_n - S^{-1}\|_2 &\leq \|I_n\|_2 + \|I_n - S^{-1}\|_2 = 1 + \zeta.
\end{align*}

To get (4.23), we use the relationship

\begin{align*}
\tilde{Q}R = Z_2R + (\tilde{Q} - Z_2)R = Z_2R + Q(\Delta_{TS})S^{-1}R \\
&= X + \delta \overline{X}_2 + Q\Delta_{TS}S^{-1}R.
\end{align*}

Thus we have

\begin{align*}
(\tilde{Q} - V)R &= \delta \overline{X}_2 + Q(\Delta_{TS})S^{-1}R - \delta \tilde{X} \\
&= \delta \overline{X}_2 + Q(\Delta_{TS})S^{-1}R - (FZ_1^T(\delta \overline{X}_1) + \delta \overline{X}_2) \\
&= Q(\Delta_{TS})S^{-1}R - FZ_1^T(\delta \overline{X}_1)
\end{align*}

which yields

\begin{align*}
\tilde{Q} - V = Q(\Delta_{TS})S^{-1} - FZ_1^T(\delta \overline{X}_1)R^{-1} \\
&= Q(\Delta_{TS})S^{-1} - FZ_1^T Z_1.
\end{align*}
Using standard norm bounds gives us
\[
\|\widetilde{Q} - V\|_2 \leq \|Q\|_2 \|\Delta_{TS}\|_2 \|S^{-1}\|_2 + \|F\|_2 \|Z_1\|_2^2 \leq (1 + \zeta)(1 + \zeta^2)\|\Delta_{TS}\|_2/(1 - \zeta) + \zeta^2
\]
which is (4.23).

**Remark 4.1.** Since the distance between \(Q\) and the left orthogonal matrix \(V\) is \(O(\zeta)\) and the distance between \(\widetilde{Q}\) in (2.17) and the same \(V\) is \(O(\max\{\|\Delta_{TS}\|_2, \zeta^2\})\), we expect \(\widetilde{Q}\) to be more nearly left orthogonal than \(Q\).

**Corollary 4.5.** Assume the hypotheses and terminology of Theorems 4.1 and 4.4. Then
\[
\|T\|_2 \leq (1 + \|\Delta_{TS}\|_2)(1 + \eta),
\]
\[
\|I_n - S\|_2 \leq \zeta/(1 - \zeta), \quad \|S\|_2 \leq 1/(1 - \zeta).
\]

**Proof.** We have that
\[
T = S^{-1} + T - S^{-1} = (I_n + \Delta_{TS})S^{-1}
\]
Thus
\[
\|T\|_2 = \|(I_n + \Delta_{TS})S^{-1}\|_2 \leq (1 + \|\Delta_{TS}\|_2)|S^{-1}|_2 \leq (1 + \|\Delta_{TS}\|_2)(1 + \zeta)
\]
which is (4.26).

From (4.20), \(\|I_n - S^{-1}\|_2 = \|Z_1\|_2 = \zeta < 1\),
\[
\|S\|_2 = \|[I + (I - S^{-1})]^{-1}\|_2 \leq 1/(1 - \|I - S^{-1}\|_2) \leq 1/(1 - \zeta)
\]
which is the second part of (4.27). Thus,
\[
I_n - S = (S^{-1} - I_n)S
\]
so that
\[
\|I_n - S\|_F \leq \|S\|_2\|I_n - S^{-1}\|_2 \leq \|S\|_2\|Z_1\|_2 \leq \zeta/(1 - \zeta)
\]
which is the first part of (4.27).

Motivated by (4.22), we make the assumption
\[
\|Q - V\|_2 \leq (\zeta + \zeta^2)/(1 - \zeta) \leq 1.
\]
Equation (4.28) holds if \(Q\) is remotely close to an orthogonal matrix and is equivalent to the assumption that
\[
\zeta \leq \sqrt{2} - 1 \approx 0.41421.
\]
Using the definition (4.21), the bounds (4.6) and (4.4) on \(\|\Gamma_{TR}\|_F\) and \(\|\Delta_{TS}\|_F\), and the bound (4.12), we can say that
\[
\zeta \leq \hat{\zeta} = \varepsilon_M f_{\zeta}(m, n, p)\|X\|_F\|R^{-1}\|_2
\]
\[
f_{\zeta}(m, n, p) = f_{TR}(m, n, p) + 2f_{TS}(m, n, p),
\]
and thus, we make the assumption that
\[
\zeta \leq \hat{\zeta} \leq \sqrt{2} - 1.
\]
The assumption (4.32) is false only if \( X \) is sufficiently ill-conditioned to be \( \mathcal{O}(\varepsilon_M) \) distance from a rank deficient matrix. We also make the assumption that \( \|\Delta_{TS}\|_2 \leq \sqrt{2} - 1 \).

Equation (4.32) and Corollary 4.5 allows us to say that
\[
\|Q\|_2, \|S\|_2 \leq 2, \quad \|T\|_2 \leq \sqrt{2}(1 + \|\Delta_{TS}\|_2) \leq 2.
\]

Also, for each \( k \) in Functions 3.1, 3.2, and 3.3 we can say that
\[
\|\hat{Q}_k\|_2, \|Q_k\|_2, \|S_k\|_2, \|T_k\|_2, \|T_{kk}\|_2 \leq 2.
\]

The bounds in (4.34) simplify our error analysis.

Equation (4.34) allows us to replace (4.8) and (4.10) with
\[
\|U - \tilde{U}\|_F \leq 2\sqrt{10}\|\Delta_{TS}\|_F,
\]
\[
\|\delta X\|_F \leq \|\delta X\|_F + \sqrt{5} (\|\Gamma_{TR}\|_F + 2\|\Delta_{TS}\|_F\|R\|_2).
\]

The last corollary establishes a bound on orthogonality of \( Q \) similar to that from [8].

**Corollary 4.6.** Assume the hypothesis and terminology of Theorem 4.4. Also assume (4.32). Then
\[
\|I_n - Q^TQ\|_F \leq 2\hat{\zeta}/(1 - \hat{\zeta}).
\]

**Proof.** Following the argument in Corollary 4.5 to prove (4.27) and bounding the Frobenius norm instead of the two-norm, we have
\[
\|I_n - S\|_F \leq \|S\|_2\|Z_1\|_F \leq \hat{\zeta}/(1 - \zeta) \leq \hat{\zeta}/(1 - \hat{\zeta}).
\]

From the definition of \( S \) in (2.13), we have
\[
I_n - Q^TQ = I_n - S + I_n - S^T,
\]
thus
\[
\|I_n - Q^TQ\|_F \leq 2\|I_n - S\|_F \leq 2\hat{\zeta}/(1 - \hat{\zeta})
\]
which is (4.37). \( \Box \)

For the remainder of §4, we bound \( \|\Delta_{TS}\|_F, \|\delta X\|_F \) and \( \|\Gamma_{TR}\|_F \) for Functions 3.2 and 3.3 and thus use the bounds (4.35)-(4.36) to establish the conditional backward error relationship (4.9).

**4.2. Error Analysis of MGS2.** Function 3.1 is Schreiber and Van Loan’s [27] algorithm for representing products of Householder transformations applied to the Q-R factorization of \( X \) in (1.2). From [27] and the analysis by Bischof and Van Loan [7], we may conclude that
\[
\bar{X} + \Delta \bar{X} = U \begin{pmatrix} R \\ 0_{m \times n} \end{pmatrix}
\]
where
\[
\|\Delta \bar{X}\|_F \leq \varepsilon_M g_1(m, p)\|X\|_F + \mathcal{O}(\varepsilon_M^2),
\]
for a modestly growing function $g_1(\cdot)$ (not specified in [7, 27]). The backward error $\Delta X$ in (4.38) is distinct from $\delta X$ in (4.9). The authors of [7, 27] also show that $U$ satisfies

\begin{equation}
\|I_m - U^T U\|_F \leq \varepsilon_M g_2(m, n) + O(\varepsilon_M^2)
\end{equation}

for a modest sized function $g_2(\cdot)$ (also unspecified in [7, 27]).

Equation (4.39) is sufficient to show (4.5) and (4.6) as shown below. To show (4.6), we note that $\Gamma_TR = (I_n - T)R = \Delta X(1:n,:)$

thus

\begin{equation}
\|\Gamma_TR\|_F = \|\Delta X(1:n,:)|\|_F \leq \varepsilon_M g_1(m, n)\|X\|_F + O(\varepsilon_M^2).
\end{equation}

Likewise, to show (4.5), we note that $\delta_X = QR - X = Q(I_n - T)R + QTR - X = Q\Delta X(1:n,:) + \Delta X(n+1:m+n,:) = (Q I_{m}) \Delta X$

Thus

\begin{equation}
\|\delta X\|_F \leq \|(Q I_{m})\|_2\|\Delta X\|_F \leq (1 + \|Q\|_2^2)^{1/2} \|\Delta X\|_F \leq \varepsilon_M(n+1)^{1/2} g_1(m,n)\|X\|_F.
\end{equation}

The bound (4.40) is sufficient to establish that Function 3.1 produces a stable factorization. However, because of the role Function 3.1 plays in Function 3.2, we need the following theorem.

**Theorem 4.7.** Let $T_k, k = 1, ..., n$, let $R$ be computed as in Function 3.1, let $S_k$ be given by (2.15), and let $R$ satisfy (4.32). Then

\begin{equation}
T_k S_k = I_k + \Delta_k, \quad \|\Delta_k\|_F \leq L_{TS}(m,k)\varepsilon_M + O(\varepsilon_M^2)
\end{equation}

where $L_{TS}(m,k) = \sqrt{1.5}mk$. Thus, for Function 3.1, $\Delta_{TS}$ in (4.1) satisfies

\begin{equation}
\|\Delta_{TS}\|_F \leq \varepsilon_M L_{TS}(m,n) + O(\varepsilon_M^2).
\end{equation}

**Proof.** This is a simple induction argument. For $k = 1$, we note that

$T_1 = S_1 = (1),$

thus (4.43) holds with $\Delta_1 = (0)$. For $k = 2, ..., s$, using error analysis bounds on basic operations in [19, pp.67-73], we have that

$g_k = -T_{k-1}\hat{Q}_{k-1}^T q_k - T_{k-1}(\delta g^{(1)}_k) - \delta g^{(2)}_k$

where

$\|\delta g^{(1)}_k\|_2 \leq m\varepsilon_M\|\hat{Q}_{k-1}\|_F\|q_k\|_2 + O(\varepsilon_M^2) = m\sqrt{k-1}\varepsilon_M + O(\varepsilon_M^2),$

$\|\delta g^{(2)}_k\|_2 \leq k\varepsilon_M\|T_{k-1}\|_F\|\hat{Q}_{k-1}^T q_k\|_2 + O(\varepsilon_M^2) \leq 2(k-1)^{3/2}\varepsilon_M + O(\varepsilon_M^2).$
Thus

$$\mathbf{g}_k + \delta \mathbf{g}_k = -T_{k-1} \tilde{Q}_{k-1}^T \mathbf{q}_k$$

where

$$\|\delta \mathbf{g}_k\|_2 \leq \|T_{k-1}\|_2 \|\delta \mathbf{g}_k^{(1)}\|_2 + \|\delta \mathbf{g}_k^{(2)}\|_2 + \mathcal{O}(\varepsilon_M^2)$$

$$\leq (m + 2(k-1))\sqrt{k-1} \varepsilon_M + \mathcal{O}(\varepsilon_M^2).$$

Using the induction hypothesis,

$$T_k S_k = \begin{pmatrix} T_{k-1} & \mathbf{g}_k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_{k-1} & \tilde{Q}_{k-1}^T \mathbf{q}_k \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} I_{k-1} + \Delta_k & T_{k-1} \tilde{Q}_{k-1}^T \mathbf{q}_k + \mathbf{g}_k \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} I_{k-1} + \Delta_k - \delta \mathbf{g}_k & 0 \\ 0 & 1 \end{pmatrix}.$$}

Therefore,

$$T_k S_k - I_k = \Delta_k = \begin{pmatrix} \Delta_{k-1} & \delta \mathbf{g}_k \\ 0 & 0 \end{pmatrix}$$

implying that

$$\|\Delta_k\|_F^2 = \|\Delta_{k-1}\|_F^2 + \|\delta \mathbf{g}_k\|_2^2.$$}

That recursion yields

$$\|\Delta_k\|_F^2 \leq \left( \sum_{j=1}^{k-1} m^2(j-1) + 4(j-1)^3 \varepsilon_M^2 \right) = \left[ m^2(k-2)(k-1)/2 + k^2(k-1)^2 \varepsilon_M^2 \right] + \mathcal{O}(\varepsilon_M^3)$$

$$< m^2 k^2 [1/2 + k^2/m^2] \varepsilon_M^2 + \mathcal{O}(\varepsilon_M^3) \leq 1.5 m^2 k^2 \varepsilon_M^2 + \mathcal{O}(\varepsilon_M^3).$$

Taking square roots yields the bound (4.43). Equation (4.44) is merely the case $k = n$. □

We note that Function 3.1 is just Function 3.2 for $p = 1$, and equations (4.41), (4.42), and (4.44) establish the bounds (4.4)-(4.6) hold with

$$f_{TS}(m, n, 1) = L_{TS}(m, p) = \sqrt{1.5 mn}, \ f_X(m, n, 1) = (n + 1)^{1/2} g_1(m, n) \text{ and}$$

$$f_{TR}(m, n, 1) = g_1(m, n).$$

### 4.3. Backward Error Bounds on MGS3 and BMGS_H

The first two theorems in this section prove the bounds (4.4)-(4.6) for Functions 3.2 and 3.3. Our last theorem uses the perturbation theory in §4.1 to establish that these algorithms produce $Q$ and $R$ that—for an exactly left orthogonal matrix $V$—satisfy bounds on $\|I_n - Q^T Q\|_F$ and $\|X - VR\|_F$ similar to those for MGS given in [8, 11].

The argument to produce these results works because steps (2) and (7) of Functions 3.2 and 3.3 produce $Q_k$, $R_{kk}$, and $T_{kk}$ that satisfy bounds of the form producing $Q_k$, $R_{kk}$, and $T_{kk}$ such that

(4.45) \[ T_{kk} S_{kk} - I_p = \Delta_{kk}, \quad \|\Delta_{kk}\|_F \leq \varepsilon_M L_{TS}(m, p) + \mathcal{O}(\varepsilon_M^2), \]

(4.46) \[ Y_k + \Delta Y_k = Q_k R_{kk}, \quad \|\Delta Y_k\|_F \leq \varepsilon_M L_Y(m, p) \|Y_k\|_F + \mathcal{O}(\varepsilon_M^2), \]

(4.47) \[ (I_p - T_{kk}) R_{kk} = \Gamma_{kk}, \quad \|\Gamma_{kk}\|_F \leq \varepsilon_M L_{TR}(m, p) \|Y_k\|_F + \mathcal{O}(\varepsilon_M^2). \]
where \( L_{TS}(m, p) \) in \( L_Y(m, p) \), and \( L_{TR}(m, p) \) are modestly growing functions and \( S_{kk} \) is the matrix defined by (3.9).

For Function 3.2, these steps are performed by Function 3.1 (MGS2). In section 3.2, we state the appropriate bounds based on analysis in [27, 7] and Theorem 4.7. From Theorem 4.7, equation (4.45) holds with \( L_{TS}(m, p) = \sqrt{1.5}mp \), from (4.42), equation (4.46) holds with \( L_Y(m, p) = (p + 1)^{1/2} g_1(m, p) \), from (4.41), equation (4.47) holds with \( L_{TR}(m, p) = g_1(m, p) \).

For Function 3.3, these steps are performed with Householder Q-R factorization and \( T_{kk} = I_p \). From [19, Theorem 19.4], (4.45) holds with \( L_{TS}(m, p) = d_Y mp \) where \( d_Y \) is a constant. Since \( T_{kk} = I_p \), we have
\[
\Gamma_{kk} = (I_p - T_{kk}) R_{kk} = 0,
\]
so that (4.47) holds with \( L_{TR}(m, p) \equiv 0 \).

This leads to Theorems 4.8 and 4.9. The hypotheses of these theorems specifically refer to Function 3.2 or 3.3, but their results hold if any factorization method that satisfies (4.45)-(4.47) is substituted for the Q-R factorization in steps (2) and (7) of Functions 3.2. Both theorems are proved in §5.

The following integer functions will be used to establish the backward error bounds for Functions 3.2 and 3.3. In the definitions below \( g_1(m, p) \) is the function discussed after (4.39).

The first three functions are for equations (4.45)-(4.47).
\[
\begin{align*}
L_{TS}(m, p) & = \begin{cases} \sqrt{1.5}mp & \text{for Function 3.2} \\ d_Y mp^{3/2} & \text{for Function 3.3} \end{cases}, \\
L_Y(m, p) & = \begin{cases} (p + 1)^{1/2} g_1(m, p) & \text{for Function 3.2} \\ d_Y mp & \text{for Function 3.3} \end{cases}, \\
L_{TR}(m, p) & = \begin{cases} g_1(m, p) & \text{for Function 3.2} \\ 0 & \text{for Function 3.3} \end{cases}.
\end{align*}
\]

The next three functions are for error bounds on matrix operations in the main loop of Functions 3.2 and 3.3.
\[
\begin{align*}
L_1(m, t) & = (m + 8t)t^{1/2}, \\
L_2(m, t) & = \sqrt{2}mt^{1/2}, \\
L_3(m, t, p) & = 8t^2 + 4mt^{1/2}p + 8p^2.
\end{align*}
\]

The function \( f_{TS}() \) is for bounding the backward error as in (4.4).
\[
f_{TS}(m, n, p) = \sqrt{69s} \max \{ L_3(m, n - p, p), L_{TS}(m, p) \}, \quad s = n/p.
\]

The next three functions are for intermediate quantities in the error analysis proofs.
\[
\begin{align*}
L_4(m, t, p) & = L_3(m, t, p) + 8L_{TR}(m, p) + 4L_Y(m, p) \\
& \quad + 4L_2(m, t) + 5f_{TS}(m, t, p) + 4L_1(m, t), \\
L_{XR}(m, n, p) & = 2\sqrt{10}f_{TS}(m, n, p) + \sqrt{5}L_1(m, t) + L_2(m, t), \\
L_{YR}(m, n, p) & = L_Y(m, p) + (p + 1)^{1/2} (L_{TR}(m, p) + 2L_{TS}(m, p)).
\end{align*}
\]
The last two functions are for bounding the backward errors in (4.5)-(4.6) and \( s = n/p \).

\[
\begin{align*}
(4.58) \quad f_X(m, n, p) &= \sqrt{5}(L_Y(m, p) + L_2(m, p)). \\
(4.59) \quad f_{TR}(m, n, p) &= \sqrt{2}S\max\{L_4(m, n - p), L_{TR}(m, p)\}; \\
(4.60) \quad f_{VR}(m, n, p) &= \sqrt{2}f_X(m, n, p) + \sqrt{10}f_{TR}(m, n, p) + 2f_{TS}(m, n, p) \\
(4.61) \quad f_{UR}(m, n, p) &= f_X(m, n, p) + \sqrt{5}(f_{TR}(m, n, p) + 2f_{TS}(m, n, p))
\end{align*}
\]

The first theorem establishes (4.4).

**Theorem 4.8.** Suppose that \( X \in \mathbb{R}^{m \times n} \) is partitioned according to (2.1) and that its Q-R factorization is computed by Function 3.2 or 3.3. For \( k = 1, \ldots, s \), let \( T_k \) be as produced by Function 3.2 or 3.3 in floating point arithmetic with machine unit \( \varepsilon_M \). Let \( S_k \) be given by (2.15), and assume that, at each step, the Q-R factorization of \( Y_k \) produces \( T_{kk} \) satisfying (4.47). Then

\[
(4.62) \quad T_kS_k = I_{pk} + \Delta_k, \quad \|\Delta_k\|_2 \leq \varepsilon_Mf_{TS}(m, k, p) + \mathcal{O}(\varepsilon_M^2)
\]

where \( f_{TS}(m, k, p) \) is given by (4.54). Thus, for \( k = s \) we have the inequality (4.4). For the matrix \( U \) in (2.16) there exists an exactly orthogonal matrix \( \tilde{U} \) such that

\[
(4.63) \quad \|U - \tilde{U}\|_F \leq \varepsilon_M\sqrt{10nf_{TS}(m, n, p)} + \mathcal{O}(\varepsilon_M^2).
\]

The second theorem establishes (4.5)-(4.6).

**Theorem 4.9.** Assume the hypothesis and notation of Theorem 4.8. Assume that \( R \) is nonsingular and that (4.32) holds. For \( k = 1, \ldots, s \), let \( \tilde{X}_k \) and \( \tilde{Q}_k \) be given by (2.2), and let \( R_k \) be given by (3.3). Then \( \tilde{Q}_k, R_k \) and \( T_k \) satisfy

\[
\begin{align*}
(4.64) \quad \tilde{X}_k + \delta\tilde{X}_k &= \tilde{Q}_kR_k, \\
(4.65) \quad \|\delta\tilde{X}_k\|_F &\leq \varepsilon_Mf_X(m, k, p)\|\tilde{X}_k\|_F + \mathcal{O}(\varepsilon_M^2). \\
(4.66) \quad (I_{nk} - T_k)R_k &= \Gamma_k, \quad \|\Gamma_k\|_F \leq \varepsilon_Mf_{TR}(m, k, p)\|\tilde{X}_k\|_F + \mathcal{O}(\varepsilon_M^2),
\end{align*}
\]

Thus, from (4.65)-(4.66) for \( k = s \), we have (4.5) and (4.6).

The final theorem establishes bounds for the loss of orthogonality in \( Q \) and loss of orthogonal similarity between \( X \) and \( R \).

**Theorem 4.10.** Assume the hypothesis and terminology of Theorem 4.9. Then \( Q \) satisfies (4.37) with \( \tilde{\zeta} \) given in (4.30)-(4.31). There exists a left orthogonal matrix \( V \) such that

\[
(4.67) \quad \|X - VR\|_F \leq \varepsilon_Mf_{VR}(m, n, p)\|X\|_F + \mathcal{O}(\varepsilon_M^2)
\]

where \( f_{VR}(\cdot) \) is given by (4.60). Also, there exists an orthogonal \( \tilde{U} \) satisfying (4.63) and (4.9) with

\[
(4.68) \quad \|\delta\tilde{X}\|_F \leq \varepsilon_Mf_{UR}(m, n, p)\|X\|_F + \mathcal{O}(\varepsilon_M^2)
\]

where \( f_{UR}(\cdot) \) is given by (4.61).

**Proof.** From (4.19),

\[
X - VR = -FZ_1^T(\delta\tilde{X}_1) - \delta\tilde{X}_2
\]
so that
\[
\|X - VR\|_F \leq \|F\|_2\|Z_1\|_2\|\delta X_1\|_F + \|\delta X_2\|_F
\]
\[
\leq \|\delta X_1\|_F + \|\delta X_2\|_F
\]
(4.69)
Using (4.36) and Theorems 4.8 and 4.9, this becomes
\[
\|X - VR\|_F \leq \sqrt{2}(\|\delta X\|_F + \sqrt{5}(\|\Gamma TR\|_F + 2\|\Delta TS\|_F))\|R\|_2
\]
\[
\leq \varepsilon_M\sqrt{2}f_X(m,n,p)\|X\|_F + \varepsilon_M(\sqrt{10}f_TR(m,n,p)\|X\|_F + 2\sqrt{2}f_TS(m,n,p)\|R\|_2) + O(\varepsilon_M^2).
\]
Using (4.69) and orthogonal equivalence, we have that
\[
\|R\|_2 \leq \|X\|_F + \|X - VR\|_F
\]
so that
\[
\|X - VR\|_F \leq \varepsilon_M f_{VR}(m,n,p)\|X\|_F + \varepsilon_M^2 2\sqrt{2}f_{TS}(m,n,p)\|X - VR\|_F + O(\varepsilon_M^3)
\]
Solving for \(\|X - VR\|_F\) yields (4.67).

Equation (4.68) is just the result of (4.9) and (4.36) coupled with the bounds on \(\|\Delta TS\|_F\), \(\|\Gamma TR\|_F\), and \(\|\delta X\|_F\) from Theorems 4.8 and 4.9. \[\square\]

5. Proof of Theorems 4.8 and 4.9.

5.1. Error Bounds for Matrix Operations. The first lemma needed to prove Theorems 4.8 and 4.9 uses the two simple floating bound error bounds for matrices \(A, C \in \mathbb{R}^{m \times n}\) and \(B \in \mathbb{R}^{n \times p}\). These are

\[
\text{fl}(AB) = AB + E, \quad \|E\|_F \leq \alpha \|A\|_F\|B\|_F \varepsilon_M + O(\varepsilon_M^3),
\]
(5.1)
\[
\text{fl}(A + C) = A + C + E, \quad \|E\|_F \leq \|A + C\|_F \varepsilon_M.
\]
(5.2)
Equation (5.1) is a version the bound in [19, p.71], equation (5.2) is just the error in one floating point addition applied to all of the entries of \(A\) and \(C\). The matrix operations whose error bounds are given in Lemma 5.1 are from the main loop of Functions 3.2 and 3.3.

Lemma 5.1. Assume the hypothesis of Theorem 4.9. Also let \(t = (k - 1)p\) be the number of columns in \(\hat{Q}_{k-1}\). The intermediate quantities computed in the main loop of Function 3.3 satisfy the following error bounds:

\[
H_k + \delta H_k = T_{k-1}^T \hat{Q}_{k-1}^TX_k, \quad \|\delta H_k\|_F \leq \varepsilon_M L_1(m,t)\|X_k\|_F + O(\varepsilon_M^2),
\]
(5.3)
\[
Y_k + \delta Y_k = X_k - \hat{Q}_{k-1}H_k, \quad \|\delta Y_k\|_F \leq \varepsilon_M L_2(m,t)\left\| \begin{pmatrix} H_k \\ Y_k \end{pmatrix} \right\|_F + O(\varepsilon_M^2),
\]
(5.4)
\[
G_k + \delta G_k = -T_{k-1}\hat{Q}_{k-1}^TQ_kT_{kk}, \quad \|\delta G_k\|_F \leq \varepsilon_M L_3(m,t,p) + O(\varepsilon_M^2)
\]
(5.5)
where \(L_1(m,t), L_2(m,t),\) and \(L_3(m,t,p)\) are given in (4.51)-(4.53).

Proof. The operations from (5.3)-(5.5) are combinations of (5.1) and (5.2). To prove (5.3), we have that
\[
H_k = T_{k-1}^T \hat{Q}_{k-1}^TX_k + T_{k-1}^T(\delta H_k^{(1)} + \delta H_k^{(2)})
\]
where \(\delta H_k^{(1)}\) is the error from the operation \(\hat{Q}_{k-1}^TX_k\) and \(\delta H_k^{(2)}\) is the error from multiplying \(T_{k-1}^T\) times the computed product of \(\hat{Q}_{k-1}^TX_k\). Thus
\[
\|\delta H_k^{(1)}\|_F \leq m\varepsilon_M\|\hat{Q}_{k-1}\|_F\|X_k\|_F + O(\varepsilon_M^2) \leq m_{H_{k-1}}\|X_k\|_F + O(\varepsilon_M^2),
\]
(5.1)
$\|\delta H_k^{(2)}\|_2 \leq t \varepsilon_M \|T_{k-1}\|_F \|\hat{Q}_{k-1}^T X_k\|_F + O(\varepsilon_M^2)$
\begin{equation}
\leq t^{3/2} \varepsilon_M \|T_{k-1}\|_2 \|\hat{Q}_{k-1}\|_2 \|X_k\|_F + O(\varepsilon_M^2)
\leq 4t^{3/2} \varepsilon_M \|X_k\|_F + O(\varepsilon_M^2).
\end{equation}

Thus,
\begin{equation}
\|\delta H_k\|_F = \|T_{k-1}^T (\delta H_k^{(1)}) + \delta H_k^{(2)}\|_F
\leq \|T_{k-1}\|_2 \|\delta H_k^{(1)}\|_F + \|\delta H_k^{(2)}\|_F
\leq (8t + m)t^{1/2} \|X_k\|_F \varepsilon_M + O(\varepsilon_M^2).
\end{equation}

To prove (5.4), we have that the matrix multiply and add satisfies

$Y_k = X_k - \hat{Q}_{k-1} H_k + \delta Y_k^{(1)} + \delta Y_k^{(2)}$

where
\begin{equation}
\|\delta Y_k^{(1)}\|_F \leq m \varepsilon_M \|\hat{Q}_{k-1}\|_F \|H_k\|_F + O(\varepsilon_M^2) \leq mt^{1/2} \|H_k\|_F + O(\varepsilon_M^2)
\end{equation}

and
\begin{equation}
\|\delta Y_k^{(2)}\|_F \leq \varepsilon_M \|Y_k\|_F.
\end{equation}

Thus
\begin{equation}
\|\delta Y_k\|_2 \leq (mt^{1/2} \|H_k\|_F + \|Y_k\|_F)\varepsilon_M + O(\varepsilon_M^2)
\leq (m^2 t + 1)^{1/2} \left( \frac{\|H_k\|_F}{\|Y_k\|_F} \right) \|2 \varepsilon_M + O(\varepsilon_M^2) \leq \sqrt{2} m t^{1/2} \left( \frac{H_k}{Y_k} \right) \|F + O(\varepsilon_M^2)
\end{equation}

which is (5.4).

To show (5.5), we note that the three matrix multiplications are
\begin{equation}
G_k = -T_{k-1} \hat{Q}_{k-1}^T Q_k T_{kk} + \delta G_k^{(1)} - T_{k-1} (\delta G_k^{(2)}) - T_{k-1} \hat{Q}_{k-1}^T (\delta G_k^{(3)})
\end{equation}

where $\delta G_k^{(j)}$ are the errors from the three matrix multiplications involved in computing $G_k$. Thus, since $\|T_{kk}\|_F \leq \sqrt{p} \|T_{kk}\|_2 \leq \sqrt{p} \|T\|_2 \leq 2 \sqrt{p}$, we have
\begin{equation}
\|\delta G_k^{(1)}\|_F \leq t \varepsilon_M \|T_{k-1}\|_F \|\hat{Q}_{k-1}^T Q_k T_{kk}\|_F + O(\varepsilon_M^2)
\leq t \varepsilon_M \|T_{k-1}\|_2 \|T_{kk}\|_2 \|\hat{Q}_{k-1}\|_F \|Q_k\|_2 + O(\varepsilon_M^2)
\leq 8t^2 \varepsilon_M + O(\varepsilon_M^2),
\end{equation}
\begin{equation}
\|\delta G_k^{(2)}\|_F \leq m \varepsilon_M \|\hat{Q}_{k-1}\|_F \|Q_k T_{kk}\|_F + O(\varepsilon_M^2)
\leq m \varepsilon_M \|\hat{Q}_{k-1}\|_F \|Q_k\|_F \|T_{kk}\|_2 + O(\varepsilon_M^2)
\leq 2mt^{1/2} \varepsilon_M + O(\varepsilon_M^2),
\end{equation}
\begin{equation}
\|\delta G_k^{(3)}\|_2 \leq p \varepsilon_M \|Q_k\|_F \|T_{kk}\|_F \leq 2p^2 \varepsilon_M + O(\varepsilon_M^2).
\end{equation}

Thus, we have (5.5) with $\|\delta G_k\|_2$ bounded by
\begin{equation}
\|\delta G_k\|_2 \leq \|\delta G_k^{(1)}\|_2 + \|T_{k-1}\|_2 \|\delta G_k^{(2)}\|_2 + \|T_{k-1}\|_2 \|\hat{Q}_{k-1}\|_2 \|\delta G_k^{(3)}\|_F
\leq (8t^2 + 4mt^{1/2} p^{1/2} + 8p^2) \varepsilon_M + O(\varepsilon_M^2) = L_3(m, t, p) \varepsilon_M + O(\varepsilon_M^2).
\end{equation}
5.2. Proof of Theorem 4.8. The bound on $\|\delta G_k\|_F$ in Lemma 5.1 allows us to prove Theorem 4.8.

Proof. [of Theorem 4.8] This induction proof makes use of the definition of $\Delta_{kk}$ in (4.45). For $k = 1$, we have that

$$\Delta_1 = T_1S_1 - I_p$$

$$= T_{11}S_{11} - I_p = \Delta_{11}.$$ 

Thus,

$$\|\Delta_1\|_F = \|\Delta_{11}\|_F \leq \varepsilon_M L_{TS}(m, p) + O(\varepsilon_M^2).$$

For $k > 1$, let $t = (k - 1)p$. We have

$$T_k S_k = \begin{pmatrix} T_k & G_k \\ 0 & T_{kk} \end{pmatrix} \begin{pmatrix} S_{k-1} & \tilde{Q}_{k-1}^T Q_k \\ 0 & S_{kk} \end{pmatrix}$$

$$= \begin{pmatrix} T_{k-1}S_{k-1} & T_{k-1}\tilde{Q}_{k-1}^T Q_k + G_kS_{kk} \\ 0 & T_{kk}S_{kk} \end{pmatrix}$$

$$= \begin{pmatrix} I_t + \Delta_{k-1} & T_{k-1}\tilde{Q}_{k-1}^T Q_k - T_{k-1}\tilde{Q}_{k-1}^T Q_kT_{kk}S_{kk} - (\delta G_k)S_{kk} \\ 0 & I_p + \Delta_{kk} \end{pmatrix}$$

$$= \begin{pmatrix} I_t + \Delta_{k-1} & -T_{k-1}\tilde{Q}_{k-1}^T Q_k\Delta_{kk} - (\delta G_k)S_{kk} \\ 0 & I_p + \Delta_{kk} \end{pmatrix}.$$ 

Thus,

$$\Delta_k = T_k S_k - I_{kp}$$

$$= \begin{pmatrix} \Delta_{k-1} & -T_{k-1}\tilde{Q}_{k-1}^T Q_k\Delta_{kk} - (\delta G_k)S_{kk} \\ 0 & \Delta_{kk} \end{pmatrix}.$$ 

so that for $t = (k - 1)p$,

$$\|\Delta_k\|_F^2 \leq \|\Delta_{k-1}\|_F^2 + \|\delta G_k\|_F^2 + \|\Delta_{kk}\|_F^2 + \|\tilde{Q}_{k-1}\|_F^2 + \|Q_k\|_F^2$$

$$\leq \varepsilon_M^2 \|f_{TS}(m, t, p)\|^2 + 4L_3^2(m, t, p) + 65L_{TS}^2(m, p) + O(\varepsilon_M^3).$$

This recurrence is bounded by

$$\|\Delta_k\|_F^2 \leq 69k\varepsilon_M^2 \max\{L_3^2(m, t, p), L_{TS}^2(m, p)\} + O(\varepsilon_M^3).$$

Taking square roots yields (4.62). If we let $k = s = n/p$, then we have (4.4). If we just apply (4.8) to (4.4), we have (4.63). \(\Box\)

5.3. Proof of Theorem 4.9. We now give necessary bounds on the norm of \(\begin{pmatrix} H_k \\ Y_k \end{pmatrix}\) in (2.1) and \(\|R_{kk}\|_F\) from (3.3). These bounds simplify the proof of Theorem 4.9 by allowing us to state all of our error bounds in terms of \(\|X_k\|_F\).

Lemma 5.2. Assume the hypothesis and terminology of Lemma 5.1. Let \(X_k\) and \(R_k\) be given by the partition (2.1). Then

$$\|H_k\|_F \leq (1 + \varepsilon_M L_{XR}(m, t, p))\|X_k\|_F + O(\varepsilon_M^2), \quad t = (k - 1)p$$

where \(L_{XR}(m, t, p)\) is given by (4.56). Thus,

$$\|\delta Y_k\|_F \leq L_2(m, t)\|X_k\|_F + O(\varepsilon_M^2).$$
If we assume that the Q-R factorization of $Y_k$ satisfies (4.45)-(4.47), then

$$
\|R_{kk}\|_F \leq (1 + L_{YH}(m,p)\varepsilon_M)\|Y_k\|_F + O(\varepsilon_M^2)
$$

where $L_{YH}(m,p)$ is defined by (4.57).

Proof. Using the definition of $U_{k-1}$, the results of Lemma 5.1 can be written

$$
U_{k-1}^T \begin{pmatrix} 0 \\ X_k \end{pmatrix} = \begin{pmatrix} H_k \\ Y_k \end{pmatrix} + \begin{pmatrix} \delta H_k \\ \delta Y_k + \bar{Q}_{k-1}\delta H_k \end{pmatrix}.
$$

By the induction hypothesis and the use of Theorem 4.1, there is an exactly orthogonal matrix $\bar{U}_{k-1}$ such that

$$
\|U_{k-1} - \bar{U}_{k-1}\|_F \leq \sqrt{10}\|Q_{k-1}\|_F \|\Delta_{k-1}\|_F \leq 2\sqrt{10}f_{TS}(m,t,p)\varepsilon_M + O(\varepsilon_M^2).
$$

Thus, we may write (5.9) as

$$
\begin{pmatrix} H_k \\ Y_k \end{pmatrix} = \bar{U}_{k-1}^T \begin{pmatrix} 0 \\ X_k \end{pmatrix} - \begin{pmatrix} \delta H_k \\ \delta Y_k - \bar{Q}_{k-1}\delta H_k \end{pmatrix} + (U_{k-1} - \bar{U}_{k-1})^T \begin{pmatrix} 0 \\ X_k \end{pmatrix}.
$$

Bounding the norm of the left side of this equation with the right, we have

\[
\| \begin{pmatrix} H_k \\ Y_k \end{pmatrix} \|_F \leq \|X_k\|_F + \|\delta Y_k\|_F + \|\begin{pmatrix} I_t \\ \bar{Q}_{k-1} \end{pmatrix}\|_2\|\delta H_k\|_F + \|U_{k-1} - \bar{U}_{k-1}\|_F\|X_k\|_F
\]

\[
\leq \|X_k\|_F + \varepsilon_M L_2(m,t)\|H_k\|_F + \varepsilon_M \left(1 + \|\bar{Q}_{k-1}\|_2^2\right)^{1/2} L_1(m,t)\|X_k\|_F
\]

\[
+ \sqrt{10}\|f_{TS}(m,t,p)\|\|X_k\|_F + O(\varepsilon_M^2)
\]

\[
\leq \|X_k\|_F + \varepsilon_M L_2(m,t)\|H_k\|_F + \varepsilon_M \sqrt{5} L_1(m,t)\|X_k\|_F
\]

\[
+ 2\sqrt{10}\varepsilon_M f_{TS}(m,t,p)\|X_k\|_F + O(\varepsilon_M^2).
\]

Solving for $\| \begin{pmatrix} H_k \\ Y_k \end{pmatrix} \|_F$ yields

\[
\| \begin{pmatrix} H_k \\ Y_k \end{pmatrix} \|_F \leq \|X_k\|_F (1 + \varepsilon_M \sqrt{5} L_1(m,t) + 2\varepsilon_M \sqrt{10} f_{TS}(m,t,p)) / (1 - \varepsilon_M L_2(m,t)) + O(\varepsilon_M^2)
\]

\[
= (1 + \varepsilon_M \sqrt{5} L_1(m,t) + L_2(m,t) + 2\sqrt{10} f_{TS}(m,t,p))\|X_k\|_F + O(\varepsilon_M^2)
\]

Equation (5.7) is just (5.4) combined with (5.6).

If we invoke Theorem 4.1 and use (4.45)-(4.47), there is an exactly orthogonal matrix $\bar{U}_{kk}$ such that

$$
\begin{pmatrix} 0_{p\times p} \\ Y_k \end{pmatrix} + \delta Y_k = \bar{U}_{kk} \begin{pmatrix} R_{kk} \\ 0_{m\times p} \end{pmatrix}
$$

where

$$
\|\delta Y_k\|_F \leq \|\Delta_k\|_F + (1 + \|Q_{k}\|_2)\left(\|\Gamma_{kk}\|_F + 2\|\Delta_{kk}\|_F\|R_{kk}\|_2\right)
\]

$$

$$
\leq \|\Delta_k\|_F + (p + 1)\left(\|\Gamma_{kk}\|_F + 2\|\Delta_{kk}\|_F\|R_{kk}\|_F\right)
\]

$$

$$
\leq \varepsilon_M (L_{YH}(m,p) + (p + 1)\left(\|\Gamma_{kk}\|_F + 2\|\Delta_{kk}\|_F\|R_{kk}\|_F\right)
\]

$$

$$
+ 2\varepsilon_M (p + 1)\|L_{TS}(m,p)\|\|R_{kk}\|_F + O(\varepsilon_M^2).
$$

(5.11)
Thus, using (5.11) to bound \(\|R_{kk}\|_F\) we have

\[
\|R_{kk}\|_F \leq \|Y_k\|_F + \|\delta Y_k\|_F \\
\leq (1 + \varepsilon_M ((L_Y(m,p) + (p + 1)^{1/2} L_{TR}(m,p))\|Y_k\|_F) \\
+ 2\varepsilon_M (p + 1)^{1/2} L_{TS}(m,p)\|R_{kk}\|_F + O(\varepsilon_M^2).
\]

(5.12)

Solving for \(\|R_{kk}\|_F\) yields

\[
\|R_{kk}\|_F \leq \|Y_k\|_F(1 + \varepsilon_M ((L_Y(m,p) + (p + 1)^{1/2} L_{TR}(m,p))\|Y_k\|_F)/(1 - 2\varepsilon_M (p + 1)^{1/2} L_{TS}(m,p)) + O(\varepsilon_M^2) \\
= (1 + \varepsilon_M L_Y(m,p) + (p + 1)^{1/2} L_{TR}(m,p) + 2L_{TS}(m,p))\|Y_k\|_F + O(\varepsilon_M^2) \\
= (1 + \varepsilon_M L_Y(m,p))\|Y_k\|_F + O(\varepsilon_M^2)
\]

(5.13)

\[\|I_T - T_{k-1}\|H_k - G_k R_{kk}\|_F \leq \varepsilon_M L_4(m, t, p)\|X_k\|_F + O(\varepsilon_M^2),\]

where \(L_4(\cdot)\) is given by (4.55).

**Proof.** This proof is a combination of our assumptions about the Q-R factorization in step (2) and (7) of Functions 3.2 and 3.3, the error bounds in Lemma 5.1, and Theorem 4.8.

Before proving Theorem 4.9, we prove a key technical lemma that links the computation of \(T\) with the computation of \(R\) in Functions 3.2 and 3.3.

**Lemma 5.3.** Assume the hypothesis and notation of Lemma 5.1. Assume that \(T_{kk}, R_{kk} \in \mathbb{R}^{p \times p}\) satisfy (4.47) and that \(Q_k\) and \(R_{kk}\) satisfy (4.46). Then, letting \(t = (k - 1)p\), we have

\[
\|(I - T_{k-1})H_k - G_k R_{kk}\|_F \leq \varepsilon_M L_4(m, t, p)\|X_k\|_F + O(\varepsilon_M^2),
\]

where \(L_4(\cdot)\) is given by (4.55).

We then use our assumptions about \(T_{kk}\) and \(R_{kk}\) to show that

\[
G_k R_{kk} = -T_{k-1} \hat{Q}_{k-1}^T Q_k T_{kk} + E_1,
\]

\[
E_1 = -\delta G_k R_{kk}.
\]

We then use our assumptions about \(T_{kk}\) and \(R_{kk}\) to show that

\[
G_k R_{kk} = -T_{k-1} \hat{Q}_{k-1}^T Q_k R_{kk} + T_{k-1} \hat{Q}_{k-1}^T Q_k (I_p - T_{kk}) R_{kk} + E_1
\]

\[
= -T_{k-1} \hat{Q}_{k-1}^T Q_k R_{kk} + E_2 + E_1,
\]

\[
E_2 = T_{k-1} \hat{Q}_{k-1}^T Q_k \Gamma_{kk}.
\]

Using our assumption about the backward error in the Q-R factorization of \(Y_k\) and the error (5.4), we have

\[
G_k R_{kk} = -T_{k-1} \hat{Q}_{k-1}^T (X_k - \hat{Q}_{k-1} H_k) + E_3 + E_2 + E_1
\]

\[
E_3 = T_{k-1} \hat{Q}_{k-1} (\delta Y_k - \Delta Y_k).
\]

The definition of \(S_{k-1}\) in (2.15) expands this into

\[
G_k R_{kk} = -T_{k-1} \hat{Q}_{k-1}^T X_k + T_{k-1} (S_{k-1} + S_{k-1}^T - I) H_k + E_3 + E_2 + E_1
\]

\[
= -T_{k-1} S_{k-1}^T T_{k-1} \hat{Q}_{k-1}^T X_k + T_{k-1} (S_{k-1} + S_{k-1}^T - I) H_k + E_4 + E_3 + E_2 + E_1,
\]

\[
E_4 = T_{k-1} \Delta X_k + \hat{Q}_{k-1}^T X_k.
\]
Finally, we make use of the bound (5.3) to reveal the desired relationship between the computation of $G_k$ to form $T_k$ and the computation of $H_k$ to form $R_k$. We have

$$G_k R_{kk} = -T_{k-1} S^T \tilde{S}_{k-1} (H_k + \delta H_k) + T_{k-1} (S_{k-1} + S^T \tilde{S}_{k-1} - I) H_k + E_4 + E_3 + E_2 + E_1$$

$$(5.14)$$

$$E_0 = -T_{k-1} S^T \tilde{S}_{k-1} (\delta H_k), \quad E_6 = (T_{k-1} S_{k-1} - I) H_k = \Delta_{k-1} H_k.$$

Thus

$$G_k R_{kk} - (I_t - T_{k-1}) H_k = E$$

where

$$E = E_0 + E_5 + E_4 + E_3 + E_2 + E_1.$$

Our bounds on the six error matrices are

$$\|E_1\|_F \leq \|\delta G_k\|_F \|R_{kk}\|_F \leq \varepsilon_M L_3(m, t, p) \|R_{kk}\|_F + O(\varepsilon_M^2)$$

$$\leq \varepsilon_M L_2(m, t, p) \|Y_k\|_F + O(\varepsilon_M^2) \leq \varepsilon_M L_3(m, t, p) \|X_k\|_F + O(\varepsilon_M^2),$$

$$\|E_2\|_F \leq \|T_{k-1}\|_2 \|\tilde{Q}_{k-1}\|_2 \|\tilde{Q}_k\|_2 \|\Gamma_{kk}\|_F \leq 8 \|\Gamma_{kk}\|_F$$

$$\leq 8 \varepsilon_M L_{TR}(m, p) \|Y_k\|_F + O(\varepsilon_M^2) \leq 8 \varepsilon_M L_{TR}(m, p) \|X_k\|_F + O(\varepsilon_M^2),$$

$$\|E_3\|_F \leq \|T_{k-1}\|_2 \|\tilde{Q}_{k-1}\|_2 \|\tilde{Q}_k\|_2 \|\Delta_{k-1}\|_2 \|X_k\|_F \leq 4 \varepsilon_M L_Y(m, p) \|Y_k\|_F + L_2(m, t) \|X_k\|_F + O(\varepsilon_M^2)$$

$$\leq 4 \varepsilon_M L_Y(m, p) + L_2(m, t) \|X_k\|_F + O(\varepsilon_M^2),$$

$$\|E_4\|_F \leq \|T_{k-1}\|_2 \|\tilde{Q}_{k-1}\|_2 \|\tilde{Q}_k\|_2 \|X_k\|_F \leq 4 \varepsilon_M f_{TS}(m, t, p) \|X_k\|_F + O(\varepsilon_M^2),$$

$$\|E_5\|_F \leq \|T_{k-1}\|_2 \|S_{k-1}\|_2 \|\delta H_k\|_F \leq 4 \varepsilon_M L_1(m, t) \|X_k\|_F + O(\varepsilon_M^2),$$

$$\|E_6\|_F \leq \|H_k\|_2 \|\Delta_{k-1}\|_2 \|X_k\|_F \leq \varepsilon_M f_{TS}(m, t, p) \|X_k\|_F + O(\varepsilon_M^2).$$

Thus,

$$\|E\|_F \leq \|E_1\|_F + \|E_2\|_F + \|E_3\|_F + \|E_4\|_F + \|E_5\|_F + \|E_6\|_F$$

$$= \varepsilon_M L_4(m, t, p) \|X_k\|_F + O(\varepsilon_M^2)$$

where $L_4(m, t, p)$ is given by (4.55). $\square$

The cancellation of the term $T_{k-1} S^T \tilde{S}_{k-1} H_k$ in (5.14), necessary to complete the proof of Lemma 5.3, occurs because $Y_k$ is computed according to (3.7) and is the only argument in the proof of either Theorem 4.8 or Theorem 4.9 that cannot be made for classical Gram-Schmidt.

**Proof.** [of Theorem 4.9] First, we prove the simpler bound, (4.64)-(4.65) by a simple induction argument. For $k = 1$, we have

$$\tilde{Q}_1 R_1 = Q_1 R_1 = X_1 + \delta X_1, \quad \|\delta X_1\|_F \leq \varepsilon_M L_Y(m, p) \|X_1\|_F + O(\varepsilon_M^2).$$

Thus, from the definition of $f_{TR}(-)$ in (4.59) and $L_4(-)$ in (4.55), $f_{TR}(m, p) \geq L_Y(m, p)$, leading to (4.64)-(4.65) for $k = 1$.

For $k = 2, \ldots, s$, let $t = (k - 1)p$. We have that

$$\tilde{Q}_k R_{kk} = \left( \begin{array}{c} \tilde{Q}_{k-1} \quad Q_k \\ \tilde{Q}_{k-1} \quad 0 \end{array} \right) \left( \begin{array}{cc} R_{k-1} & H_k \\ \tilde{Q}_{k-1} & R_{kk} \end{array} \right)$$

$$= \left( \begin{array}{cc} \tilde{X}_{k-1} + \delta \tilde{X}_{k-1} & \tilde{Q}_{k-1} H_k + Q_k R_{kk} \end{array} \right).$$
From (5.4), (4.46), and the induction hypothesis, we have
\[ \hat{Q}_kR_k = \hat{X}_k + \delta \hat{X}_k \]
where
\[ \delta \hat{X}_k = \begin{pmatrix} \delta \hat{X}_{k-1} & \Delta Y_k - \delta Y_k \end{pmatrix}. \]

We note that
\[ \| \delta \hat{X}_k \|_F \leq \left\| \left( \| \delta \hat{X}_{k-1} \|_F + \| \Delta Y_k \|_F + \| \delta Y_k \|_F \right) \right\|_F. \]

Using the induction hypothesis, (5.4), and (4.46), we have
\[ \| \delta \hat{X}_k \|_F \leq \varepsilon_M \| \left( f_X(m, t, p) \ L_Y(m, p) + L_1(m, p) \right) \hat{X}_k \|_F + \mathcal{O}(\varepsilon_M^2) \]
which is (4.64)-(4.65). If we take \( k = s \), we have (4.5).

To prove (4.66), we do another induction argument. For \( k = 1 \), the assumption (4.47) gives us
\[ \| (I - T_1)R_1 \|_F \leq \varepsilon_M L_{TR}(m, p) \| \hat{X}_1 \|_F + \mathcal{O}(\varepsilon_M^2). \]

From (4.55), \( L_4(m, 0, p) \geq L_{TR}(m, p) \), we have the bound (4.66).

For the induction step, we have that
\[ (I - T_k)R_k = \begin{pmatrix} I - T_{k-1} & -G_k \\ 0 & I - T_{kk} \end{pmatrix} \begin{pmatrix} R_{k-1} & H_k \\ 0 & R_{kk} \end{pmatrix} \]
\[ = \begin{pmatrix} (I - T_{k-1})R_{k-1} & (I - T_{k-1})H_k - G_k R_{kk} \\ 0 & (I - T_{kk})R_{kk} \end{pmatrix}. \]

Thus,
\[ \| (I - T_k)R_k \|_F = \| \Gamma_k \|_F \]
\[ = \| \left( \| \Gamma_{k-1} \|_F \| (I - T_{k-1})H_k - G_k R_{kk} \|_F \| \Gamma_{kk} \|_F \right) \|_F. \]

Thus, we can bound the (1, 2) block of the in the above matrix from Lemma 5.3 as (5.13). We use the induction hypothesis to bound the (1, 1) block and the assumption (4.47) to bound the (2, 2) block. That yields
\[ \| \Gamma_k \|_F \leq \varepsilon_M \left( f_{TR}(m, t, p) \ L_4(m, t, p) \right) \| \hat{X}_k \|_F + \mathcal{O}(\varepsilon_M^2) \]
\[ \leq \varepsilon_M f_{TR}(m, kp, p) \| \hat{X}_k \|_F + \mathcal{O}(\varepsilon_M^2) \]
where \( f_{TR}(\cdot) \) is defined by (4.59). If we let \( k = s = n/p \), we have the bound (4.6).

6. Conclusion. The MGS algorithm produces good least squares solutions because it has small bounds on the residual (4.2) and of the residuals (4.1) and (4.3) for the implied Householder Q-R factorization in §2.3.

BMGS algorithms inherit the favorable error analysis properties of the MGS algorithm provided that they produce \( Q, R \), and \( T \) such that all three residuals (4.1)-(4.3)
satisfy bounds of the form (4.4)-(4.6). Such bounds assure that the Sheffield structure from [23] is satisfied with backward error of $O(\varepsilon M \|X\|_F)$ as shown in Theorem 4.1.

The BMGS algorithms in Functions 3.2 and 3.3 are shown to satisfy the bounds (4.4)-(4.6) because of how $Y_k$ is computed from $X_k$ in lines (4)-(6) of these procedures and because the Q-R factorization of $X_1$ in line (2) and $Y_k$ in line (7) is done by a procedure which satisfies the bounds (4.45)-(4.47). These procedure are entirely based upon matrix-matrix (i.e., BLAS-3) operations. The structure described was shown in §?? to be applicable to variants of the algorithm of Jalby and Phillipe [21].

Acknowledgements. The author acknowledges helpful conversations about topics related to this work with Ichitaro Yamazaki and Kamesh Madduri.

REFERENCES