Numer. Math. (2013) 123:395–423
DOI 10.1007/s00211-012-0496-2

Numerische
Mathematik

Reorthogonalized block classical Gram–Schmidt

Jesse L. Barlow · Alicja Smoktunowicz

Received: 17 August 2011 / Revised: 14 May 2012 / Published online: 12 September 2012

Abstract A reorthogonalized block classical Gram–Schmidt algorithm is proposed that factors a full column rank matrix A into $A = QR$ where Q is left orthogonal (has orthonormal columns) and R is upper triangular and nonsingular. This block Gram–Schmidt algorithm can be implemented using matrix–matrix operations making it more efficient on modern architectures than orthogonal factorization algorithms based upon matrix-vector operations and purely vector operations. Gram–Schmidt orthogonal factorizations are important in the stable implementation of Krylov space methods such as GMRES and in approaches to modifying orthogonal factorizations when columns and rows are added or deleted from a matrix. With appropriate assumptions about the diagonal blocks of R, the algorithm, when implemented in floating point arithmetic with machine unit ε_M, produces Q and R such that $\|I - Q^T Q\| = O(\varepsilon_M)$ and $\|A - QR\| = O(\varepsilon_M \|A\|)$. The first of these bounds has not been shown for a block Gram–Schmidt procedure before. As consequence of these results, we provide a different analysis, with a slightly different assumption, that re-establishes a bound of Giraud et al. (Num Math, 101(1):87–100, 2005) for the CGS2 algorithm.

Mathematics Subject Classification Primary 65F25; Secondary 65F35

The research of J. L. Barlow was sponsored by the National Science Foundation under contract no. CCF-1115704.

J. L. Barlow (✉)
Department of Computer Science and Engineering, University Park, PA 16802-6822, USA
e-mail: barlow@cse.psu.edu

A. Smoktunowicz
Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
e-mail: smok@mini.pw.edu.pl

1 Introduction

For a matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$, assumed to have full column rank, we consider the computation of the Q-R decomposition

$$A = QR \tag{1.1}$$

where $Q \in \mathbb{R}^{m\times n}$ is left orthogonal (i.e., $Q^T Q = I_n$) and $R \in \mathbb{R}^{n\times n}$ is upper triangular.

The approach considered is block classical Gram–Schmidt with reorthogonalization (BCGS2) which operates on groups of columns of A instead of columns in order to create a BLAS-3 [6] compatible algorithm, that is, an algorithm built upon matrix–matrix operations rather than matrix-vector operations or vector operations. Such algorithms tend to be more efficient on modern computer architectures because they make more effective use of caching.

We present a block Gram–Schmidt algorithm that is a generalization of the classical Gram–Schmidt method with reorthogonalization (CGS2) which was first analyzed by Abdelmalek [1]. However, our analysis and development follow the flavor of a more recent analysis given by Giraud et al. [7]. Reorthogonalized Gram–Schmidt is discussed by Rice [15] in an important 1966 experimental paper, although it is not clear whether it is the same algorithm as in [1,7]. A similar block algorithm based upon classical Gram–Schmidt (CGS), justified only by numerical tests, is proposed by Stewart [17]. Other block Gram–Schmidt algorithms are presented by Jalby and Phillippe [13], and Vanderstaeten [18]. An interesting and closely related block orthogonal decomposition is discussed by Strathopoulos and Wu [16].

An excellent summary of the role of Gram–Schmidt algorithms is given in [4, Sects. 2.4, 3.2] and in the 1994 survey paper [3]. Gram–Schmidt algorithms have a prominent role in two contexts in scientific computing: the implementation of Krylov space methods such as GMRES [9,14] and the modification of Q-R factorizations when rows or columns are added or deleted [2,5].

To construct the BCGS2 algorithm, A is partitioned into blocks $A = (A_1, \ldots, A_s)$, where each A_k has p_k columns, i.e., $A_k \in \mathbb{R}^{m\times p_k}$ for $k = 1, \ldots, s$, with $n = p_1 + p_2 + \cdots + p_s$. BCGS2 is based upon matrix–matrix operations on these blocks as opposed to the matrix–vector or vector operations on columns.

Throughout the paper, $\|\cdot\|$ denotes the matrix or vector two-norm depending upon context. We also briefly use $\|\cdot\|_F$ in the proof of Theorem 3.4 to denote the Frobenius norm.

We prove that BCGS2 is numerically stable under conditions outlined in Sect. 3, thus producing computed Q and R by BCGS2 in floating point arithmetic with machine unit ε_M that satisfy

$$\|I_n - Q^T Q\| \leq \varepsilon_M f_1(m, n, p), \tag{1.2}$$

$$\|A - QR\| \leq \varepsilon_M f_2(m, n, p)\|A\| \tag{1.3}$$

for modestly growing functions $f_1(\cdot)$ and $f_2(\cdot)$, where $p = \max_{1\leq i\leq s} p_i$ is the maximum block size of A. We say that a matrix satisfying a bound similar to (1.2) is *near left orthogonal.* The property (1.3) is true of the Householder, Givens, modified Gram–Schmidt, and classical Gram–Schmidt orthogonal factorization algorithms [10, Chapter 19], but the property (1.2) is true of the Householder and Givens algorithms but not of these two Gram–Schmidt algorithms. Moreover, it has not been shown for any of the block algorithms cited above [13,16–18].

In Sect. 2, to construct a step of the BGCS2 algorithm, we first consider an important subproblem where we are given a near left orthogonal matrix $U \in \mathbb{R}^{m\times t}$, and a matrix $B \in \mathbb{R}^{m\times p}$ $t + p \leq n$. Similar to the bound (1.2), U satisfies

$$\|I_t - U^T U\| \leq \varepsilon_M f_1(m, t, p) \ll 1. \tag{1.4}$$

The subproblem is to find $Q_B \in \mathbb{R}^{m\times p}$ left orthogonal, $R_B \in \mathbb{R}^{p\times p}$ upper triangular, and $S_B \in \mathbb{R}^{t\times p}$ such that

$$B = US_B + Q_B R_B, \tag{1.5}$$

$$U^T Q_B = 0. \tag{1.6}$$

For $p = 1$, the problem (1.5)–(1.6) is discussed in [2,5] for its role in modifying Q-R decompositions.

A block Gram–Schmidt step with reorthogonalization, called a BCGS2 step and given as Function 2.2, produces Q_B, R_B, and S_B that approximately solve the problem (1.5)–(1.6). In the context of developing the BCGS2 algorithm (Function 2.3) to obtain (1.1), $U = Q(:\, , 1: t)$ is the near left orthogonal factor from the Q-R decomposition of $A(:\, , 1: t)$, while $B = A(:\, , t+1: t+p)$ is the next p columns to be added to the decomposition. The resulting matrices in (1.5)–(1.6) are $Q_B = Q(:\, , t+1: t+p)$, $S_B = R(1: t, t+1: t+p)$ and $R_B = R(t+1: t+p, t+1: t+p)$. Thus Q_B makes up the new columns of Q whereas S_B and R_B make up the new columns of R.

The proof of (1.2)–(1.3) is then constructed in Sect. 3 from analysis of how well a BCGS2 step (Function 2.2) satisfies (1.5)–(1.6). In Sects. 3.1 and 3.2, we show that Function 2.2 computes Q_B, R_B and S_B such that $\|B - US_B - Q_B R_B\| = \mathcal{O}(\varepsilon_M \|B\|)$ (with only standard error analysis preconditions) and $\|U^T Q_B\| = \mathcal{O}(\varepsilon_M)$ (provided $\varepsilon_M \|B\| \|R_B^{-1}\| \ll 1$). In Sect. 3.3, we build upon those results to establish the bounds (1.2)–(1.3) for BCGS2 algorithm.

In Sect. 4, we show that our analysis has implications for the CGS2 algorithm in [1,7], in particular, that the CGS2 algorithm produces a near orthogonal matrix under weaker assumptions than given in [7]. Our conclusion is in Sect. 5.

2 Reorthogonalized block Gram–Schmidt with reorthogonalization (BCGS2)

Our development of a block classical Gram–Schmidt algorithm with reorthogonalization (BGCS2) builds the algorithm from simple matrix functions, so that the error analysis in Sect. 3 may be constructed in the same manner.

Throughout this paper, we use the notation $\bar{Q}$, $\bar{R}$ *and, later,* $\bar{S}$, *with subscripts when necessary, to denote computed submatrices from intermediate stages of the computation of* Q *and* R *in* (1.1).

The first matrix function in our development of BCGS2, called **block_CGS_step** *(for block classical Gram–Schmidt step)*, inputs a near left orthogonal $U \in \mathbb{R}^{m\times t}$ satisfying (1.4), and $B \in \mathbb{R}^{m\times p}$ $t + p \leq n \leq m$, to produce a near left orthogonal $\bar{Q} \in \mathbb{R}^{m\times p}$, an upper triangular $\bar{R} \in \mathbb{R}^{p\times p}$, and $\bar{S} \in \mathbb{R}^{t\times p}$ such that (in exact arithmetic)

$$Y = (I_m - UU^T)B \tag{2.1}$$

$$Y = \bar{Q}\bar{R}, \quad \bar{Q}^T\bar{Q} = I_p, \quad \text{(Q-R factorization)} \tag{2.2}$$

$$B = U\bar{S} + \bar{Q}\bar{R}, \quad \bar{S} = U^T B. \tag{2.3}$$

The computation in (2.2) is assumed to be done by an orthogonal factorization function **local_qr**, which inputs the intermediate variable $Y \in \mathbb{R}^{m\times p}$, $p \leq n \leq m$, and produces

$$[\bar{Q}, \bar{R}] = \textbf{local\_qr}(Y) \tag{2.4}$$

where $\bar{R} \in \mathbb{R}^{p\times p}$ is upper triangular, and $\bar{Q} \in \mathbb{R}^{m\times p}$ is near left orthogonal. In the context of our development of BCGS2, Y results from computation with a block within a larger matrix, hence the name **local_qr**. In floating point arithmetic, with machine unit ε_M, we expect the computed $\bar{Q}$ and $\bar{R}$ to satisfy

$$\|I_p - \bar{Q}^T\bar{Q}\| \leq \varepsilon_M L_1(m, p) < 1, \tag{2.5}$$

$$Y + \Delta Y = \bar{Q}\bar{R}, \quad \|\Delta Y\| \leq \varepsilon_M L_1(m, p) \quad \|\bar{Y}\|, \tag{2.6}$$

for some modest function $L_1(\cdot)$, thus (2.5) says that $\bar{Q}$ is near left orthogonal. A function **local_qr** satisfying (2.5)–(2.6) may be produced using Householder or Givens factorization. In our MATLAB codes for this algorithm, (2.4) is just the statement $[\bar{Q}, \bar{R}] = \textbf{qr}(Y, 0)$. An interpretation of [10, Sect. 19.3] on the error analysis of Householder factorization yields a function $L_1(m, p) = d_1 mp^{3/2}$ where d_1 is a constant. For appropriate BLAS-3 speed [6], that is, to take advantage of caching, the implementation of **local_qr** may be done using the "tall, skinny" Q-R (TSQR) discussed in the recent Ph.D. thesis by Hoemmen [11, Sect. 2.3].

In all of our function descriptions, lines beginning with % denote comments.

Function 2.1 (One block CGS step)

```
function $[\bar{Q}, \bar{R}, \bar{S}]$ = block_CGS_step($U$, $B$)
%
% Uses local_qr function defined by (2.4).
%
$\bar{S} = U^T B$;
$Y = B - U\bar{S}$;
$[\bar{Q}, \bar{R}]$=local_qr($Y$);
end block_CGS_step
```

The error analysis of Function 2.1 is given in Sect. 3.1.

Although (2.3) shows that $\bar{Q}$, $\bar{R}$ and $\bar{S}$ satisfy (1.5), from (2.1) if $\bar{R}$ is nonsingular, in exact arithmetic, the residual of (1.6) is given by

$$U^T \bar{Q} = (I - U^T U)U^T B \bar{R}^{-1} \tag{2.7}$$

so

$$\|U^T \bar{Q}\| \le \|I - U^T U\| \|U\| \|B\| \|\bar{R}^{-1}\|. \tag{2.8}$$

Even in the unlikely event that $\|B\| \|\bar{R}^{-1}\|$ is $\mathcal{O}(1)$, the residual (2.7) will almost certainly be too large for our purposes. Therefore, we introduce Function 2.2 below, called **block_CGS2_step** *(for two block classical Gram–Schmidt steps)*, which consists of two applications of **block_CGS_step** with the same near left orthogonal matrix $U \in \mathbb{R}^{m \times t}$. For a given $B \in \mathbb{R}^{m \times p}$, the first application of **block_CGS_step** produces near left orthogonal $\bar{Q}_1$, upper triangular $\bar{R}_1$, and $\bar{S}_1$ such that

$$B = U\bar{S}_1 + \bar{Q}_1 \bar{R}_1, \quad \bar{S}_1 = U^T B, \tag{2.9}$$

while the second application of **block_CGS_step** takes $\bar{Q}_1$ and yields near left orthogonal Q_B, upper triangular $\bar{R}_2$, and $\bar{S}_2$ satisfying

$$\bar{Q}_1 = U\bar{S}_2 + Q_B \bar{R}_2, \quad \bar{S}_2 = U^T \bar{Q}_1. \tag{2.10}$$

From (2.9)–(2.10), it follows that

$$B = U\bar{S}_1 + (Q_B \bar{R}_2 + U\bar{S}_2)\bar{R}_1 = U(\bar{S}_1 + \bar{S}_2 \bar{R}_1) + Q_B(\bar{R}_2 \bar{R}_1), \tag{2.11}$$

so that (1.5) holds with

$$S_B = \bar{S}_1 + \bar{S}_2 \bar{R}_1, \quad R_B = \bar{R}_2 \bar{R}_1. \tag{2.12}$$

We may also assert (in exact arithmetic) that

$$Q_B R_B = (I - UU^T)^2 B. \tag{2.13}$$

However, if R_B is nonsingular, the residual from (1.6) for Q_B in (2.13) is

$$U^T Q_B = (I - U^T U)^2 U^T B R_B^{-1}$$

thus

$$\|U^T Q_B\| \le \|I - U^T U\|^2 \|U\| \|B\| \|R_B^{-1}\|. \tag{2.14}$$

From (1.4), (2.11), (2.12), and (2.14), if $\varepsilon_M \|B\| \|R_B^{-1}\| \ll 1$, then $\|U^T Q_B\| = \mathcal{O}(\varepsilon_M)$, giving us a small residual for (1.5)–(1.6).

Function 2.2 (Two steps of block CGS)

```
**function** $[Q_B, R_B, S_B]$ = block_CGS2_step $(U, B)$
%
% Uses Function 2.1 as two calls to
% **block_CGS_step**
%
$[\bar{Q}_1, \bar{R}_1, \bar{S}_1]$ = **block_CGS_step**$(U, B)$;
$[Q_B, \bar{R}_2, \bar{S}_2]$ = **block_CGS_step**$(U, \bar{Q}_1)$;
$S_B = \bar{S}_1 + \bar{S}_2 \bar{R}_1$;
$R_B = \bar{R}_2 \bar{R}_1$;
**end** block_CGS2_step
```

The behavior of Function 2.2 in floating point arithmetic is discussed in Sect. 3.2. Theorem 3.2 in Sect. 3.2 gives a bound for $\|U^T Q_B\|$ with a precise condition on $\|B\|\|R_B^{-1}\|$. Also, in that section, Theorem 3.3 bounds the residual for the Eq. (1.5).

To produce the Q-R factorization of $A \in \mathbb{R}^{m\times n}$ in (1.1), building upon **block_CGS2_step**, we partition A into

$$A = (A_1, A_2, \ldots, A_s) \tag{2.15}$$

where $A_k \in \mathbb{R}^{m\times p_k}$ for $k = 1, \ldots, s$. In the input to Function 2.3, we define the parameter **blocks** as the integer vector

$$\mathbf{blocks} = (p_1, \ldots, p_s)^T. \tag{2.16}$$

We partition Q in (1.1) into

$$Q = (Q_1, Q_2, \ldots, Q_s), \tag{2.17}$$

and let

$$\widehat{Q}_k = (Q_1, Q_2, \ldots, Q_k), \quad \widehat{A}_k = (A_1, A_2, \ldots, A_k). \tag{2.18}$$

We write the upper triangular matrix R in (1.1) in the block form

$$R = \begin{pmatrix} R_{11} & R_{12} & \cdots\cdots\cdots & R_{1s} \\ & R_{22} & \cdots\cdots\cdots & R_{2s} \\ & & \cdots\cdots\cdots\cdots & \cdots \\ & & & R_{ss} \end{pmatrix}, \tag{2.19}$$

and let

$$R_k = \begin{pmatrix} R_{11} & R_{12} & \cdots\cdots\cdots & R_{1k} \\ & R_{22} & \cdots\cdots\cdots & R_{2k} \\ & & \cdots\cdots\cdots\cdots & \cdots \\ & & & R_{kk} \end{pmatrix}. \tag{2.20}$$

The initial step of factoring the first block is $[Q_1, R_1]$=**local_qr**(A_1) with $R_{11} = R_1$ and $\widehat{Q}_1 = Q_1$. Then for $k = 1, \ldots, s-1$, using **block_CGS2_step** (Function 2.2), we compute $S_{k+1} = S_B$, $Q_{k+1} = Q_B$, and $R_{k+1,k+1} = R_B$ with input arguments $B = A_{k+1}$ and $U = \widehat{Q}_k$. Thus

$$R_{k+1} = \begin{pmatrix} R_k & S_{k+1} \\ 0 & R_{k+1,k+1} \end{pmatrix}, \quad \widehat{Q}_{k+1} = \begin{pmatrix} \widehat{Q}_k & Q_{k+1} \end{pmatrix}, \quad \widehat{A}_{k+1} = \begin{pmatrix} \widehat{A}_k & A_{k+1} \end{pmatrix}. \tag{2.21}$$

In exact arithmetic, the Eq. (2.11)–(2.12) establishes that

$$A_{k+1} = \widehat{Q}_k S_{k+1} + Q_{k+1} R_{k+1,k+1}. \tag{2.22}$$

Induction on (2.21)–(2.22) yields

$$\widehat{A}_k = \widehat{Q}_k R_k, \quad k = 1, \ldots, s, \tag{2.23}$$

thus we have (1.1) with $Q = \widehat{Q}_s$ and $R = R_s$. The bound (2.5) reads as a bound on $\|I_{p_{k+1}} - Q_{k+1}^T Q_{k+1}\|$ and Eq. (2.14) reads as a bound on $\|\hat{Q}_k^T Q_{k+1}\|$. These two bounds, by induction, achieve a conditional bound on $\|I_n - Q^T Q\|$ as discussed later in Sect. 3. The above discussion motivates the function **BCGS2** defined next.

Function 2.3 (Block Classical Gram–Schmidt with Reorthogonalization (BCGS2))

```
function [$Q$, $R$]=BCGS2 ($A$, blocks)
%
% Uses block_CGS2_step from Function 2.2
% and local_qr defined by (2.4)
%
[$m$, $n$]=size($A$); s=length(blocks); $high$ = blocks(1);
[$Q$, $R$]=local_qr($A(:, 1: high)$);
%
% In the notation of (2.18) and (2.20), this is
% [$Q_1$, $R_1$]=local_qr($A_1$)
%
for $k = 1 : s - 1$
    $low = high + 1$; $high = high +$ blocks$(k + 1)$;
    [$Q_{new}$, $R_{new}$, $S_{new}$]=block_CGS2_step($Q$, $A(:, low: high)$);
%
% In the notation of (2.21), this is
% [$Q_{k+1}$, $R_{k+1,k+1}$, $S_{k+1}$]=block_CGS2_step($\widehat{Q}_k$, $A_{k+1}$);
%
    $R = \begin{pmatrix} R & S_{new} \\ 0 & R_{new} \end{pmatrix}$; $Q = \begin{pmatrix} Q & Q_{new} \end{pmatrix}$;
end;
end; BCGS2
```

The error analysis of **BCGS2** (Function 2.3) is given in Sect. 3.3 and is an induction argument on the bounds on Function 2.2 in Sect. 3.2. The bound (1.2) in Theorem 3.4 is established with assumption (3.14) about quantities computed during Function 2.3, whereas the bound (1.3) is established, with no preconditions other than those for Eqs. (3.10)–(3.13), in Theorem 3.5.

3 Error analysis of BCGS2

To produce the bounds (1.2)–(1.3) on **BCGS2** (Function 2.3), we make the simplifying assumption that all of the blocks $A_1, A_2, \ldots, A_s$ have the same dimension, that is, in (2.16), $p = p_1 = p_2 = \cdots = p_s$. To have blocks of differing size, we could just let $p = \max_{1 \leq i \leq s} p_i$ and make some minor, but tedious, adjustments to the induction arguments in Sect. 3.3.

In addition to the function $L_1(m, p)$ from (2.5)–(2.6) characterizing the error analysis properties assumed for **local_qr** defined in (2.4), the following functions of the integers m, t, and p with the modest constants d_i, $i = 2, 3, 4, 5$ are used to characterize error bounds in this paper:

$$L_2(m, t, p) = d_2 m t^{1/2} p^{1/2}, \quad L_3(t, p) = d_3 p^{1/2}(1 + 2t^{3/2}), \tag{3.1}$$

$$L_4(p) = d_4 p^{1/2}(1 + 2p^{3/2}), \quad L_5(p) = d_5 p^2, \tag{3.2}$$

$$L_F(m, t, p) = L_1(m, p) + \sqrt{2} L_2(m, t, p) + L_3(t, p), \tag{3.3}$$

$$L_6(m, t, p) = [4(1 + \sqrt{2}) L_F(m, t, p) + 5 L_1(m, p)], \tag{3.4}$$

$$f_{resid}(m, t, p) = (1 + \sqrt{2}) L_1(m, p) + 2 L_3(t, p) + L_4(p) + \sqrt{2} L_5(p). \tag{3.5}$$

Assuming that $m \geq t \geq p$, $L_F(\cdot)$,$L_6(\cdot)$, and $f_{resid}(\cdot)$ are $\mathcal{O}(m t^{1/2} p^{1/2})$.
If we let

$$k = \lceil t/p \rceil,$$

then, we also use the following functions of m, t, and p:

$$\gamma(k) = (51k + 1)^{-1/2}, \tag{3.6}$$

$$f_1(m, t, p) = \begin{cases} \gamma(k)^{-1} L_F(m, t, p), & t \leq n - p_s \\ f_1(m, t - p_s, p). & t > n - p_s \end{cases} \tag{3.7}$$

$$f_2(m, t, p) = \begin{cases} (k + 1)^{1/2} f_{resid}(m, t, p), & t \leq n - p_s \\ f_2(m, t - p_s, p). & t > n - p_s \end{cases} \tag{3.8}$$

$$f_{sing}(m, t, p) = \sqrt{2}(f_1(m, t, p) + L_F(m, t, p)) + \gamma(k) L_5(p). \tag{3.9}$$

Again, assuming $m \geq t \geq p$, the functions $f_1(\cdot)$,$f_2(\cdot)$ and $f_{sing}(\cdot)$ are $\mathcal{O}(mt)$.

Along with these functions, we also define

$$c_U(\varepsilon_M) = 1 + 0.5\varepsilon_M f_1(m,n,p), \quad c_Q(\varepsilon_M) = 1 + 0.5\varepsilon_M L_1(m,p), \tag{3.10}$$

$$c_6(\varepsilon_M) = (1 - \varepsilon_M L_6(m,n,p))^{-1/2}, \tag{3.11}$$

$$c_R(\varepsilon_M) = 1 + \varepsilon_M (L_F(m,n,p) + c_F(\varepsilon_M) L_1(m,p)), \tag{3.12}$$

$$c_S(\varepsilon_M) = 1 + \varepsilon_M (0.5 f_1(m,n,p) + L_2(m,n,p)). \tag{3.13}$$

To avoid second order terms in ε_M in our analysis in Sect. 3 *, we assume that all of the functions* (3.10)–(3.13) *are bounded and that all are less than $\sqrt{2}$, a conservative assumption. Realistically, we could choose any bound in the interval* $(1, 2)$, *but $\sqrt{2}$ works nicely in the arithmetic in our error analysis.*

In Theorem 3.4, using the notation from (2.15)–(2.19), the definition (3.9), and $t_k = \sum_{j=1}^{k} p_j = kp$, the bound (1.2) holds under the assumption

$$f_{sing}(m, t_{k-1}, p) \|A_k\| \|R_{kk}^{-1}\| \le \gamma(k), \quad k = 2, \ldots, s. \tag{3.14}$$

The complications of constructing the detailed error analysis proof in this section require us to break the proof into sections and several small technical lemmas. The proofs of some of these technical lemmas are in the appendix.

The flow of the proof follows the development of Function 2.3 from Functions 2.1 and 2.2. Function 2.1 is analyzed in Sect. 3.1, the primary goal being the residual bound in Theorem 3.1. The analysis of Function 2.2 in Sect. 3.2 applies Theorem 3.1 to the two calls to Function 2.1 to produce Theorem 3.2 bounding $\|U^T Q_B\|$ and Theorem 3.3 bounding $\|B - US_B - Q_B R_B\|$. In Sect. 3.3, induction arguments on Theorems 3.2 and 3.3 establish Theorems 3.4 and 3.5 thereby giving us (1.2)–(1.3).

3.1 Error bounds for Function 2.1

We consider Function 2.1 with input matrices $U \in \mathbb{R}^{m\times t}$ and $B \in \mathbb{R}^{m\times p}$ where U satisfies (1.4). A simple eigenvalue/singular value analysis and our assumption about $c_U(\cdot)$ yields the bound

$$\|U\| \le 1 + 0.5\varepsilon_M f_1(m,t,p) \le c_U(\varepsilon_M) \le \sqrt{2}. \tag{3.15}$$

Our assumptions about the function **local_qr** allow us to use the bounds (2.5)–(2.6) and, again, from eigenvalue/singular value bounds to conclude that

$$\|\bar{Q}\| \le 1 + 0.5\varepsilon_M L_1(m,p) = c_Q(\varepsilon_M) \le \sqrt{2}. \tag{3.16}$$

Also, since by our assumption $\varepsilon_M L_1(m,p) \le \sqrt{2} - 1 < 0.7$ then

$$\begin{aligned} \|\bar{Q}^{\dagger}\| &\le (1 - \varepsilon_M L_1(m,p))^{1/2} \\ &\le 1 + \varepsilon_M L_1(m,p) \le \sqrt{2}, \end{aligned} \tag{3.17}$$

where $\bar{Q}^{\dagger}$ is the Moore–Penrose pseudoinverse of $\bar{Q}$.

Our bounds on the effects of floating point arithmetic on Function 2.1 depend upon Lemmas 3.1 and 3.2, two technical lemmas that we prove in the appendix.

Lemma 3.1 *In floating point arithmetic with machine unit ε_M, Function 2.1 with input matrices $U \in \mathbb{R}^{m\times t}$ satisfying (1.4) and $B \in \mathbb{R}^{m\times p}$ produces $\bar{Q}, \bar{R}, \bar{S}$ and Y such that for $L_1(\cdot)$ defined in (2.5)–(2.6), and $L_2(\cdot)$ and $L_3(\cdot)$ defined in (3.1), we have*

$$\bar{S} + \delta\bar{S} = U^T B, \quad \|\delta\bar{S}\| \le \varepsilon_M L_2(m, t, p)\|B\|, \tag{3.18}$$

$$Y + \delta Y = B - U\bar{S}, \quad \|\delta Y\| \le \varepsilon_M L_3(t, p)\|B\|. \tag{3.19}$$

We also have that $\bar{Q}, \bar{R}$ and Y satisfy (2.5)–(2.6).

Lemma 3.2 establishes a norm bound on the projection $I_m - UU^T$ in floating point arithmetic.

Lemma 3.2 *If $U \in \mathbb{R}^{m\times t}$ satisfies (1.4), then*

$$\|I_m - UU^T\| \le 1. \tag{3.20}$$

The main theorem of this section is the following backward error bound on Function 2.1.

Theorem 3.1 *Assume the hypothesis and notation of Lemma 3.1, let ΔY be defined by (2.6), then the computed $\bar{Q}$ and $\bar{R}$ from Function 2.1 satisfy*

$$\bar{Q}\bar{R} = (I_m - UU^T)B + F \tag{3.21}$$

where

$$F = \Delta Y - \delta Y + U(\delta\bar{S}). \tag{3.22}$$

Thus

$$\|F\| \le \varepsilon_M L_F(m, t, p)\|B\|, \tag{3.23}$$

where $L_F(\cdot)$ is defined in (3.3).

Proof We simply unwind the relationships from Lemma 3.1 and Eq. (2.6) to obtain

$$\begin{aligned}
\bar{Q}\bar{R} &= Y + \Delta Y \\
&= B - U\bar{S} - \delta Y + \Delta Y \\
&= (I_m - UU^T)B + U(\delta\bar{S}) - \delta Y + \Delta Y \\
&= (I_m - UU^T)B + F
\end{aligned}$$

which is (3.21)–(3.22). To get (3.23), standard norm inequalities lead to

$$\begin{aligned}\|F\| &\le \|U\|\|\delta\bar{S}\| + \|\delta Y\| + \|\Delta Y\|\\ &\le \varepsilon_M[c_U(\varepsilon_M)L_2(m,t,p) + L_1(m,p) + L_3(t,p)]\|B\|\\ &\le \varepsilon_M[\sqrt{2}L_2(m,t,p) + L_1(m,p) + L_3(t,p)]\|B\|\\ &= \varepsilon_M L_F(m,t,p)\|B\|\end{aligned}$$

which is (3.23). □

A norm bound that results from Theorem 3.1 is necessary for our error analysis.

Corollary 3.1 *Assume the hypothesis and notation of Lemma 3.1, then the computed $\bar{R}$ from Function 2.1 satisfies*

$$\|\bar{R}\| \le c_R(\varepsilon_M)\|B\| \tag{3.24}$$

where $c_R(\cdot)$ is defined by (3.12).

Proof Taking (3.21), multiplying on the left by $\bar{Q}^\dagger$ in (3.17), and reorganizing terms, we have

$$\bar{R} = \bar{Q}^\dagger[(I_m - UU^T)B + F].$$

Thus

$$\|\bar{R}\| \le \|\bar{Q}^\dagger\|[\|I_m - UU^T\|\|B\| + \|F\|]. \tag{3.25}$$

To bound $\|\bar{R}\|$ in (3.25), we note from Lemma 3.20 that $\|I_m - UU^T\| \le 1$. Thus, using the definition of $c_R(\varepsilon_M)$ in (3.12) and the bound (3.17), (3.25) becomes

$$\begin{aligned}\|\bar{R}\| &\le (1 + \varepsilon_M L_F(m,t,p))(1 - \varepsilon_M L_1(m,p))^{-1/2}\|B\|\\ &\le (1 + \varepsilon_M L_F(m,t,p))(1 + \varepsilon_M L_1(m,p))\|B\|\\ &= (1 + \varepsilon_M[L_F(m,t,p) + c_F(\varepsilon_M)L_1(m,p)])\|B\|\\ &= c_R(\varepsilon_M)\|B\|\end{aligned}$$

which is (3.24). □

3.2 Error bounds for Function 2.2

Function 2.2 inputs $U \in \mathbb{R}^{m\times t}$ satisfying (1.4) and $B \in \mathbb{R}^{m\times p}$ and produces output $Q_B \in \mathbb{R}^{m\times p}$, upper triangular $R_B \in \mathbb{R}^{p\times p}$ and $S_B \in \mathbb{R}^{t\times p}$ with the objective that they satisfy (1.5)–(1.6). At the end of this section, in Theorem 3.3, we show that (1.5) is satisfied to the extent that

$$\|B - US_B - Q_B R_B\| \le \varepsilon_M f_{resid}(m,t,p)\|B\| \tag{3.26}$$

where $f_{resid}(\cdot)$ is given by (3.5).

The following theorem bounds $\|U^T Q_B\|$, establishing that (1.6) is satisfied with small residual.

Theorem 3.2 *Let Function 2.2 input $U \in \mathbb{R}^{m\times t}$ satisfying (1.4) and $B \in \mathbb{R}^{m\times p}$ and output $Q_B \in \mathbb{R}^{m\times p}$, $S_B \in \mathbb{R}^{t\times p}$ and upper triangular $R_B \in \mathbb{R}^{p\times p}$. If, for $f_{sing}(\cdot)$ defined in (3.9), B and R_B satisfy*

$$\varepsilon_M f_{sing}(m,t,p)\|B\|\|R_B^{-1}\| \leq \gamma(k), \quad k = \lceil t/p \rceil, \tag{3.27}$$

then U and Q_B satisfy

$$\|U^T Q_B\| \leq 5\varepsilon_M L_F(m,t,p), \tag{3.28}$$

$$\|I_p - Q_B^T Q_B\| \leq \varepsilon_M L_1(m,p). \tag{3.29}$$

Equation (3.29) holds because Q_B is the near left orthogonal factor from **local_qr**. The proof of (3.28), outlined below, requires several small steps resulting from interpreting Theorem 3.1 and Corollary 3.1 and from using a few technical lemmas.

First, interpreting Corollary 3.1 from the two calls to Function 2.1 and combining it with (3.16) yields the norm bounds for $\bar{R}_1, \bar{R}_2 \in \mathbb{R}^{p\times p}$ given by

$$\|\bar{R}_1\| \leq c_R(\varepsilon_M)\|B\| \tag{3.30}$$

$$\begin{aligned}\|\bar{R}_2\| &\leq c_R(\varepsilon_M)\|\bar{Q}_1\| \\ &\leq c_R(\varepsilon_M)c_Q(\varepsilon_M).\end{aligned} \tag{3.31}$$

The computations of S_B and R_B in the last part of Function 2.2 require the bounds establish in the following lemma which is proved in the appendix.

Lemma 3.3 *Let Function 2.2 input $U \in \mathbb{R}^{m\times t}$ satisfying (1.4) and $B \in \mathbb{R}^{m\times p}$ and output $Q_B \in \mathbb{R}^{m\times p}$, $S_B \in \mathbb{R}^{t\times p}$ and upper triangular $R_B \in \mathbb{R}^{p\times p}$. Then in terms of the intermediate values $\bar{S}_1, \bar{S}_2 \in \mathbb{R}^{t\times p}$ and $\bar{R}_1, \bar{R}_2 \in \mathbb{R}^{p\times p}$, we have*

$$S_B + \delta S_B = \bar{S}_1 + \bar{S}_2\bar{R}_1, \quad \|\delta S_B\| \leq \varepsilon_M L_4(p)\|B\|, \tag{3.32}$$

$$R_B + \delta R_B = \bar{R}_2\bar{R}_1, \quad \|\delta R_B\| \leq \varepsilon_M L_5(p)\|B\|, \tag{3.33}$$

where $L_4(\cdot)$ and $L_5(\cdot)$ are given by (3.2).

To prove the next lemma, we use Weyl's inequality for singular values [8, Corollary 8.6.2] on Eq. (3.33) to obtain the bound

$$|\sigma_\ell(R_B) - \sigma_\ell(\bar{R}_2\bar{R}_1)| \leq \|\delta R_B\| \leq \varepsilon_M L_5(p)\|B\|, \quad \ell = 1,\ldots,p, \tag{3.34}$$

where $\sigma_\ell(\cdot)$ is the ℓth singular value of the contents. Lemma 3.4 shows that (3.34) combined with assumption (3.27) establishes a needed bound on $\sigma_p(\bar{R}_2\bar{R}_1)$.

Lemma 3.4 *Assume that $\bar{R}_1$, $\bar{R}_2$, $R_B \in \mathbb{R}^{p\times p}$ are the upper triangular matrices produced by Function 2.2 in floating point arithmetic with machine unit ε_M and (3.27). Then $\bar{R}_1$ and $\bar{R}_2$ are nonsingular and, $\sigma_p(\bar{R}_2\bar{R}_1)$, the smallest singular value of $\bar{R}_2\bar{R}_1$ satisfies*

$$\varepsilon_M\sqrt{2}[f_1(m,t,p)+L_F(m,t,p)]\|B\| \le \gamma(k)\sigma_p(\bar{R}_2\bar{R}_1). \tag{3.35}$$

Proof To get the singular value bound (3.35), note that the smallest singular value of R_B satisfies

$$\sigma_p(R_B) = \|R_B^{-1}\|^{-1}$$

thus assumption (3.27) may be written

$$\gamma(k)\sigma_p(R_B) \ge \varepsilon_M f_{sing}(m,t,p)\|B\| > 0.$$

From (3.34),

$$\sigma_p(R_B) - \varepsilon_M L_5(p)\|B\| \le \sigma_p(\bar{R}_2\bar{R}_1)$$

thus, from the definition of $f_{sing}(m,t,p)$, we have (3.35).

From (3.35), the square matrix $\bar{R}_2\bar{R}_1$ must be nonsingular, thus both of the square matrices $\bar{R}_2$ and $\bar{R}_1$ are nonsingular. □

We are now ready for the core argument in proving (3.28) which requires invoking Theorem 3.1 for the two calls to Function 2.1 in Function 2.2. For those calls, Eqs. (3.21) and (3.23) read

$$\bar{Q}_1\bar{R}_1 = (I-UU^T)B + F_1, \quad \|F_1\| \le \varepsilon_M L_F(m,t,p)\|B\|, \tag{3.36}$$

$$\bar{Q}_B\bar{R}_2 = (I-UU^T)\bar{Q}_1 + F_2, \quad \|F_2\| \le \varepsilon_M L_F(m,t,p)\|\bar{Q}_1\|. \tag{3.37}$$

Since we have the bound (3.16) on $\|\bar{Q}_1\|$, using the assumption about (3.10), we write

$$\|F_2\| \le \varepsilon_M c_Q(\varepsilon_M) L_F(m,t,p) \le \sqrt{2}\varepsilon_M L_F(m,t,p). \tag{3.38}$$

To solve for $U^T Q_B$, multiply Eq. (3.37) on the left by U^T and on the right by $\bar{R}_2^{-1}$ and Eq. (3.36) on the left by U^T and on the right by $\bar{R}_1^{-1}\bar{R}_2^{-1}$ to obtain

$$U^T\bar{Q}_1\bar{R}_2^{-1} = [(I_n - U^TU)U^TB + U^TF_1]\bar{R}_1^{-1}\bar{R}_2^{-1}, \tag{3.39}$$

$$U^TQ_B = (I_n - U^TU)U^T\bar{Q}_1\bar{R}_2^{-1} + U^TF_2\bar{R}_2^{-1}. \tag{3.40}$$

The use of standard norm inequalities, assumptions (1.4) and (3.27) our assumption about $c_U(\varepsilon_M)$ in (3.10), and the definition (3.7) on (3.40) yields

$$\begin{aligned}\|U^T Q_B\| &\leq \|I_n - U^T U\|\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| + \|U\|\|F_2\|\|\bar{R}_2^{-1}\| \\ &\leq \varepsilon_M[f_1(m,t,p)\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| + \sqrt{2}c_U(\varepsilon_M)L_F(m,t,p)\|R_2^{-1}\|] \\ &\leq \varepsilon_M L_F(m,t,p)[\gamma(k)^{-1}\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| + 2\|R_2^{-1}\|]. \qquad (3.41)\end{aligned}$$

Thus the bound on $\|U^T Q_B\|$ depends upon bounding $\|U^T \bar{Q}_1 \bar{R}_2^{-1}\|$ and $\|R_2^{-1}\|$.

A bound on $\|U^T \bar{Q}_1 \bar{R}_2^{-1}\|$ comes from bounding the two-norm from (3.39) yielding

$$\begin{aligned}\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| &\leq [\|I_n - U^T U\|\|B\| + \|F_1\|]\|U\|\|(\bar{R}_2\bar{R}_1)^{-1}\| \\ &\leq \varepsilon_M[f_1(m,t,p) + L_F(m,t,p)]c_U(\varepsilon_M)\|B\|\|(\bar{R}_2\bar{R}_1)^{-1}\| \\ &\leq \sqrt{2}\varepsilon_M[f_1(m,t,p) + L_F(m,t,p)]\|B\|\|(\bar{R}_2\bar{R}_1)^{-1}\| \qquad (3.42)\end{aligned}$$

If we combine (3.35) from Lemma 3.4 with (3.42), we have

$$\begin{aligned}\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| &\leq \gamma(k)\sigma_p(\bar{R}_2\bar{R}_1)\|(\bar{R}_2\bar{R}_1)^{-1}\| \\ &\leq \gamma(k). \qquad (3.43)\end{aligned}$$

Thus, to bound $\|U^T Q_B\|$ in (3.41) we need only bound $\|\bar{R}_2^{-1}\|$. Fortunately, it is related to $\|U^T \bar{Q}_1 \bar{R}_2^{-1}\|$ as shown in the following lemma which is proved in the appendix.

Lemma 3.5 *Let $\bar{R}_2 \in \mathbb{R}^{p\times p}$ and $\bar{Q}_1 \in \mathbb{R}^{m\times p}$ be result of implementing Function 2.2 in floating point arithmetic with machine unit ε_M. Then,*

$$\|\bar{R}_2^{-1}\|^2(1-\xi) = 1 + \|U^T \bar{Q}_1 \bar{R}_2^{-1}\|^2. \qquad (3.44)$$

where

$$|\xi| \leq \varepsilon_M L_6(m,t,p)$$

with $L_6(\cdot)$ given by (3.4).

We use the assumption (3.27) to obtain a bound on $\|U^T \bar{Q}_1 R_2^{-1}\|$ and show that $\|R_2^{-1}\|$ has an appropriate bounds.

Lemma 3.6 *Assume (3.27) and assume the hypothesis and notation of Lemma 3.1. Let $\gamma(\cdot)$ be defined by (3.6). Then Function 2.2 produces $\bar{Q}_1$ and $\bar{R}_2$ such that*

$$\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| \leq \gamma(k), \qquad (3.45)$$

$$\|\bar{R}_2^{-1}\| \leq c_6(\varepsilon_M)(1+\gamma^2(k))^{1/2} \leq 2 \qquad (3.46)$$

where $c_6(\cdot)$ is defined in (3.12).

Proof Equation (3.45) is just (3.43). Substituting (3.43) into (3.44) in Lemma 3.5, using the definition of $c_6(\varepsilon_M)$ in (3.12), and taking square roots yields

$$\begin{aligned}\|\bar{R}_2^{-1}\| &\leq \left(1 + \|U^T \bar{Q}_1 \bar{R}_2^{-1}\|^2\right)^{1/2} (1-\xi)^{-1/2}, \quad |\xi| \leq \varepsilon_M L_6(m,t,p)\\ &\leq \left(1 + \|U^T \bar{Q}_1 \bar{R}_2^{-1}\|^2\right)^{1/2} (1-\varepsilon_M L_6(m,t,p))^{-1/2}\\ &\leq c_6(\varepsilon_M)\left(1 + \|U^T \bar{Q}_1 \bar{R}_2^{-1}\|^2\right)^{1/2}\\ &\leq c_6(\varepsilon_M)\left(1 + \gamma(k)^2\right)^{1/2} \qquad (3.47)\end{aligned}$$

which is the first inequality in (3.46). The second inequality in Eq. (3.46) comes from $\gamma(k) \leq 1$ for all k and $c_6(\varepsilon_M) \leq \sqrt{2}$ by the assumption below (3.13). □

We can now complete the proof of Eq. (3.28) in Theorem 3.2 and thereby obtain our conditional bound on $\|U^T Q_B\|$ from Function 2.2.

Proof of Theorem 3.2 Applying (3.41) with the bounds from Lemma 3.6, using the fact that $\gamma(k) \leq 1$ for all $k \geq 1$, and the assumptions $c_U(\varepsilon_M), c_6(\varepsilon_M) \leq \sqrt{2}$ below (3.10)–(3.12), we obtain

$$\begin{aligned}\|U^T Q_B\| &\leq \varepsilon_M L_F(m,t,p)[\gamma(k)^{-1}\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| + c_U(\varepsilon_M)\|R_2^{-1}\|]\\ &\leq \varepsilon_M L_F(m,t,p)[\gamma(k)^{-1}\gamma(k) + \sqrt{2} c_U(\varepsilon_M) c_6(\varepsilon_M)(1+\gamma^2(k))^{1/2}]\\ &= \varepsilon_M L_F(m,t,p)[1 + 2\sqrt{2}(1+\gamma^2(k))^{1/2}] \leq 5\varepsilon_M L_F(m,t,p)\end{aligned}$$

Thus, we have (3.28). □

Remark 3.1 (on Theorem 3.2) Instead of the assumption (3.27), we could just as easily have made the assumption

$$\|\bar{R}_2^{-1}\| \leq \left(1 + \gamma(k)^2\right)^{1/2}.$$

This assumption is closely related to a bound given in [1]. Using Lemma 3.6, would have given us the bound

$$\|U^T \bar{Q}_1 \bar{R}_2^{-1}\| \leq \gamma(k) - \xi(1+\gamma(k))^{1/2}$$

with $\xi = \mathcal{O}(\varepsilon_M)$. Although the bound on ξ would have been different from the bound in Lemma 3.6, we would still arrive at a bound much like (3.28).

On the other hand, $\bar{R}_2$ is an intermediate quantity in Function 2.2 whereas the factor $\|B\|\|R_B^{-1}\|$ is based upon an input argument and an output argument from Function 2.2. Either $\|B\|\|R_B^{-1}\|$ or $\|\bar{R}_2^{-1}\|$ could be approximated by norm and condition estimators (see, for instance, [10, Chapter 15]) and those approximations could be computed along with the quantities in Function 2.2.

We now prove the residual bound (3.26).

Theorem 3.3 *Let $U \in \mathbb{R}^{m\times t}$ and $B \in \mathbb{R}^{m\times p}$ be input to Function 2.2 and let $Q_B \in \mathbb{R}^{m\times p}$, $S_B \in \mathbb{R}^{t\times p}$, and $R_B \in \mathbb{R}^{p\times p}$. Then, in floating point arithmetic with machine unit ε_M, the computed Q_B, R_B and S_B from Function 2.2 satisfy (3.26).*

Proof Multiplying on the right on (3.37) by $\bar{R}_1$ yields

$$Q_B\bar{R}_2\bar{R}_1 = (I_m - UU^T)\bar{Q}_1\bar{R}_1 + F_2\bar{R}_1,$$

thus from (3.33),

$$Q_BR_B = (I_m - UU^T)\bar{Q}_1\bar{R}_1 + F_2\bar{R}_1 - Q_B(\delta R_B).$$

If we use the fact that $F_2 = \Delta Y_2 - \delta Y_2 + U(\delta\bar{S}_2)$ from Theorem 3.1, then we have

$$Q_BR_B = \bar{Q}_1\bar{R}_1 - U\bar{S}_2\bar{R}_1 + (\Delta Y_2 - \delta Y_2)\bar{R}_1 - Q_B(\delta R_B).$$

Expanding $\bar{Q}_1\bar{R}_1$ using Theorem 3.1 yields

$$\begin{aligned} Q_BR_B &= (I_m - UU^T)B - U\bar{S}_2\bar{R}_1 + F_1 + (\Delta Y_2 - \delta Y_2)\bar{R}_1 - Q_B(\delta R_B)\\ &= B - U\bar{S}_1 - U\bar{S}_2\bar{R}_1 - U(\delta\bar{S}_1) + F_1 + (\Delta Y_2 - \delta Y_2)\bar{R}_1 - Q_B(\delta R_B). \end{aligned}$$

Using the definition of F_1 in Theorem 3.1 and the backward error for S_B in (3.32) we have

$$Q_BR_B = B - US_B - U(\delta S_B) + \Delta Y_1 - \delta Y_1 + (\Delta Y_2 - \delta Y_2)\bar{R}_1 - Q_B(\delta R_B).$$

Using norm bounds and assumption below (3.10)–(3.12) yields

$$\begin{aligned} \|B - US_B - Q_BR_B\| &\le \|U\|\|\delta S_B\| + \|\Delta Y_1\| + \|\delta Y_1\| + \|\Delta Y_2\|\|\bar{R}_1\|\\ &\quad +\|\delta Y_2\|\|\bar{R}_1\| + \|Q_B\|\|\delta R_B\|\\ &\le \varepsilon_M[c_U(\varepsilon_M)L_4(p) + L_1(m,p) + L_3(t,p)\\ &\quad +c_Q(\varepsilon_M)L_5(p)]\|B\| + [L_1(m,p) + L_3(t,p)]\|\bar{R}_1\|)\\ &\le \varepsilon_M((1 + c_R(\varepsilon_M))(L_1(m,p) + 2L_3(t,p))\\ &\quad +L_4(p) + \sqrt{2}L_5(p))\|B\|\\ &\le \varepsilon_M[(1+\sqrt{2})L_1(m,p) + 3L_3(t,p) + \sqrt{2}L_5(p)]\|B\|)\\ &= \varepsilon_M f_{resid}(m,t,p)\|B\| \end{aligned}$$

establishing (3.26). □

3.3 Error bounds for BCGS2

Obtaining the bounds (1.2) and (1.3) is simply the result of induction arguments on Theorems 3.2 and 3.3. We begin with (1.2).

Theorem 3.4 *Let $\widehat{Q}_k \in \mathbb{R}^{m \times t_k}$, $t_k = kp$, $k = 1, \ldots, s$, as defined by (2.18), be the result of k steps of Function 2.3 in floating point arithmetic with machine unit ε_M and assume (3.14). Then for $k = 1, \ldots, n$,*

$$\|I_{t_k} - \widehat{Q}_k^T \widehat{Q}_k\| \le \varepsilon_M f_1(m, t_{k-1}, p) \tag{3.48}$$

where $f_1(\cdot)$ is given by (3.7), and $\gamma(\cdot)$ by (3.6). Interpreting (3.48) and the definition of $f_1(\cdot)$ in (3.7) for $k = s$ yields (1.2).

Proof This is a proof by induction on Theorem 3.2. For $k = 1$, we note that $\widehat{Q}_1 = Q_1$ which just results from $[Q_1, R_{11}] = \textbf{local\_qr}(A_1)$. Since $\gamma(0) = 1$ and $L_F(m, 0, p) \ge L_1(m, p)$, we have

$$\begin{aligned} \|I_p - \widehat{Q}_1^T \widehat{Q}_1\| &= \|I_p - Q_1^T Q_1\| \\ &\le \varepsilon_M L_1(m, p) \\ &\le \varepsilon_M \gamma(0)^{-1} L_F(m, 0, p) \\ &\le \varepsilon_M f_1(m, 0, p). \end{aligned}$$

For the induction step, assume $t_k \le n - p_s = n - p$, and that (3.48) holds for k. Then

$$I_{t_{k+1}} - \widehat{Q}_{k+1}^T \widehat{Q}_{k+1} = \begin{pmatrix} I_{t_k} - \widehat{Q}_k^T \widehat{Q}_k & \widehat{Q}_k^T Q_{k+1} \\ Q_{k+1}^T \widehat{Q}_k & I_p - Q_{k+1}^T Q_{k+1} \end{pmatrix},$$

so that, by a classical bound on the two-norm,

$$\begin{aligned} \|I_{t_{k+1}} - \widehat{Q}_{k+1}^T \widehat{Q}_{k+1}\| &= \left\| \begin{pmatrix} I_{t_k} - \widehat{Q}_k^T \widehat{Q}_k & \widehat{Q}_k^T Q_{k+1} \\ Q_{k+1}^T \widehat{Q}_k & I_p - Q_{k+1}^T Q_{k+1} \end{pmatrix} \right\| \\ &\le \left\| \begin{pmatrix} \|I_{t_k} - \widehat{Q}_k^T \widehat{Q}_k\| & \|\widehat{Q}_k^T Q_{k+1}\| \\ \|Q_{k+1}^T \widehat{Q}_k\| & \|I_p - Q_{k+1}^T Q_{k+1}\| \end{pmatrix} \right\|. \end{aligned} \tag{3.49}$$

Invoking the induction hypothesis, applying Theorem 3.2 to $\widehat{Q}_k$ and $Q_{k+1}$with the assumption (3.14) substituted for (3.27), yields that $\|I_{t_k} - \widehat{Q}_k^T \widehat{Q}_k\|$ satisfies (3.48) and that

$$\|\widehat{Q}_k^T Q_{k+1}\| \le 5\varepsilon_M L_F(m, t_k, p).$$

Thus (3.49) becomes

$$\begin{aligned} \|I_{t_{k+1}} - \widehat{Q}_{k+1}^T \widehat{Q}_{k+1}\| &\le \varepsilon_M \left\| \begin{pmatrix} f_1(m, t_{k-1}, p) & 5L_F(m, t_k, p) \\ 5L_F(m, t_k, p) & L_1(m, p) \end{pmatrix} \right\| \\ &\le \varepsilon_M \left\| \begin{pmatrix} f_1(m, t_{k-1}, p) & 5L_F(m, t_k, p) \\ 5L_F(m, t_k, p) & L_1(m, p) \end{pmatrix} \right\|_F. \end{aligned} \tag{3.50}$$

From our definition of $L_F(\cdot)$ in (3.3), $L_F(m, t_k, p) \geq L_1(m, p)$, thus (3.50) may be interpreted

$$\|I_{t_{k+1}} - \widehat{Q}_{k+1}^T \widehat{Q}_{k+1}\| \leq \varepsilon_M \left\| \begin{pmatrix} f_1(m, t_{k-1}, p) & 5L_F(m, t_k, p) \\ 5L_F(m, t_k, p) & L_F(m, t_k, p) \end{pmatrix} \right\|_F . \tag{3.51}$$

Using the definitions of $f_1(\cdot)$ in (3.7) and of $\gamma(\cdot)$ in (3.6), and the fact that $L_F(\cdot)$ is non-decreasing in all of its arguments, (3.51) becomes

$$\begin{aligned} \|I_{t_{k+1}} - \widehat{Q}_{k+1}^T \widehat{Q}_{k+1}\| &\leq \varepsilon_M \left\| \begin{pmatrix} \gamma(k-1)^{-1} L_F(m, t_{k-1}, p) & 5L_F(m, t_k, p) \\ 5L_F(m, t_k, p) & L_F(m, t_k, p) \end{pmatrix} \right\|_F . \\ &\leq \varepsilon_M L_F(m, t_k, p) \left\| \begin{pmatrix} \gamma(k-1)^{-1} & 5 \\ 5 & 1 \end{pmatrix} \right\|_F \\ &= \varepsilon_M L_F(m, t_k, p) \left(\gamma(k-1)^{-2} + 51 \right)^{1/2} \\ &= \varepsilon_M \gamma(k)^{-1} L_F(m, t_k, p) & (3.52) \\ &= \varepsilon_M f_1(m, t_k, p) & (3.53) \end{aligned}$$

which is the induction step for (3.48). □

Remark 3.2 (on Theorem 3.4) We note that Theorem 3.4 depends upon the condition (3.14) which, in turn, depends upon

$$\max_{2 \leq k \leq s} \|A_k\| \|R_{kk}^{-1}\|. \tag{3.54}$$

The value in (3.54) can be computed by computing the quantity $\|A_k\| \|R_{kk}^{-1}\|$ during each call to Function 2.2 as discussed in Remark 3.1. Then (3.54) could be computed in the main loop of Function 2.3. We note that (in exact arithmetic),

$$\max_{2 \leq k \leq s} \|A_k\| \|R_{kk}^{-1}\| \leq \|R\| \|R^{-1}\| = \|A\| \|A^{\dagger}\|$$

which is the standard condition number of A (or R). Moreover, the quantity in (3.54) is at least as easy to estimate or approximate as $\|R\| \|R^{-1}\|$.

Our last theorem proves (1.3).

Theorem 3.5 *Let $\widehat{Q}_k \in \mathbb{R}^{m \times t_k}$ in (2.18) and $R_k \in \mathbb{R}^{t_k \times t_k}$ in (2.20), for $t_k = kp$, be the result of k steps of Function 2.3 in floating point arithmetic with machine unit ε_M. Then for $\widehat{A}_k \in \mathbb{R}^{m \times t_k}$, $k = 1, \ldots, s$ in (2.18) satisfies*

$$\|\widehat{A}_k - \widehat{Q}_k R_k\| \leq \varepsilon_M f_2(m, t_{k-1}, p) \|\widehat{A}_k\| \tag{3.55}$$

where $f_2(\cdot)$ is given in (3.8). Thus (1.3) follows from (3.55) by taking $k = s$ and using the definition of $f_2(\cdot)$ in (3.8).

Proof By induction on k based upon Theorem 3.3. For $k = 1$, $\widehat{A}_1 = A_1$ $\widehat{Q}_1 = \bar{Q}_1$ and $\bar{R}_1 = R_{11}$, by our assumption (2.6), and the definition of $f_{resid}(\cdot)$ in (3.5) and $f_2(\cdot)$ in (3.8),

$$\begin{aligned} \|\widehat{A}_1 - \widehat{Q}_1\bar{R}_1\| &\le \varepsilon_M L_1(m, p)\|\widehat{A}_1\| \\ &\le \varepsilon_M f_{resid}(m, 0, p)\|\widehat{A}_1\| \\ &\le \varepsilon_M f_2(m, 0, p)\|\widehat{A}_1\|. \end{aligned}$$

For the induction step, assume that (3.55) is true for step k and prove (3.55) for step $k + 1$. For $k < s$, we have that

$$\widehat{A}_{k+1} - \widehat{Q}_{k+1}R_{k+1,k+1} = \begin{pmatrix} \widehat{A}_k & A_{k+1} \end{pmatrix} - \begin{pmatrix} \widehat{Q}_k & Q_{k+1} \end{pmatrix} \begin{pmatrix} R_k & S_{k+1} \\ 0 & R_{k+1,k+1} \end{pmatrix}$$

where $S_{k+1} \in \mathbb{R}^{t_k \times p}$ and $R_{k+1,k+1} \in \mathbb{R}^{p \times p}$. Thus,

$$\|\widehat{A}_{k+1} - \widehat{Q}_{k+1}R_{k+1}\|^2 \le \|\widehat{A}_k - \widehat{Q}_k R_k\|^2 + \|A_{k+1} - \widehat{Q}_k S_{k+1} - Q_{k+1}R_{k+1,k+1}\|^2. \tag{3.56}$$

The first term on the right side of (3.56) is bounded by the induction hypothesis, whereas the second is bounded by applying Theorem 3.3 to A_{k+1}, $\widehat{Q}_k$, S_{k+1}, Q_{k+1} and $R_{k+1,k+1}$ leading to the bounds

$$\begin{aligned} \|\widehat{A}_k - \widehat{Q}_k R_k\| &\le \varepsilon_M k^{1/2} f_{resid}(m, t_{k-1}, p)\|\widehat{A}_k\|, \\ \|A_{k+1} - \widehat{Q}_k S_{k+1} - Q_{k+1}R_{k+1,k+1}\| &\le \varepsilon_M f_{resid}(m, t_k, p)\|A_k\|. \end{aligned}$$

Combining (3.56), (3.57), and (3.57) yields

$$\|\widehat{A}_{k+1} - \widehat{Q}_{k+1}R_{k+1}\|^2 \le \varepsilon_M^2[k f_{resid}^2(m, t_{k-1}, p)\|\widehat{A}_k\|^2 + f_{resid}^2(m, t_k, p)\|A_{k+1}\|^2], \tag{3.57}$$

Since $f_{resid}(\cdot)$ is monotone nondecreasing in all of its arguments and $\|\widehat{A}_k\|$, $\|A_{k+1}\| \le \|\widehat{A}_{k+1}\|$, (3.57) becomes

$$\begin{aligned} \|\widehat{A}_{k+1} - \widehat{Q}_{k+1}R_{k+1}\|^2 &\le (k+1)\varepsilon_M^2 f_{resid}^2(m, t_k, p)\|\widehat{A}_{k+1}\|^2 \\ &\le \varepsilon_M^2 f_2^2(m, t_k, p)\|\widehat{A}_{k+1}\|^2. \end{aligned}$$

Taking square roots establishes the induction step of the argument. □

4 Interpreting the bounds for CGS2

Our results in Theorems 3.4 and 3.5 for Function 2.3 in Sect. 3.3 have implications for the CGS2 algorithm in [1,7]. First, we summarize that algorithm. Let $A \in \mathbb{R}^{m \times n}$

be given column-wise as

$$A = (\mathbf{a}_1, \ldots, \mathbf{a}_n).$$

In this context, **local_qr**, rather than being in Householder or Givens orthogonal factorization, is the two-line function

```
**function** $[\bar{\mathbf{q}}, \bar{r}]$= local_qr($\mathbf{y}$)
$\bar{r} = \|\mathbf{y}\|$;
$\bar{\mathbf{q}} = \mathbf{y}/\bar{r}$;
**end** *local_qr*
```

To generate the decomposition (1.1) with $R = (r_{ij})$ and $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$, a step in the CGS2 algorithm takes a near left orthogonal matrix $U \in \mathbb{R}^{m \times p}$, a vector $\mathbf{b} \in \mathbb{R}^m$, and produces $r_b \in \mathbb{R}$, $\mathbf{s}_b \in \mathbb{R}^p$ and $\mathbf{q}_b \in \mathbb{R}^m$ such that

$$\mathbf{q}_b r_b = (I_m - UU^T)^2 \mathbf{b}, \quad \|\mathbf{q}_b\| = 1, \tag{4.1}$$

$$\mathbf{b} = U\mathbf{s}_b + \mathbf{q}_b r_b. \tag{4.2}$$

Thus our columnwise version of Function 2.2 is next.

Function 4.1 (One step of CGS2)

```
**function** $[\mathbf{q}_b, r_b, \mathbf{s}_b] = cgs2\_step(U, \mathbf{b})$
$\bar{\mathbf{s}}_1 = U^T\mathbf{b}$; $\mathbf{y}_1 = \mathbf{b} - U\bar{\mathbf{s}}_1$;
$[\bar{\mathbf{q}}_1, \bar{r}_1] = \mathbf{local\_qr}(\mathbf{y}_1)$;
$\bar{\mathbf{s}}_2 = U^T\bar{\mathbf{q}}_1$; $\mathbf{y}_2 = \bar{\mathbf{q}}_1 - U\bar{\mathbf{s}}_2$;
$[\mathbf{q}_b, \bar{r}_2] = \mathbf{local\_qr}(\mathbf{y}_2)$;
$\mathbf{s}_b = \bar{\mathbf{s}}_1 + \bar{\mathbf{s}}_2\bar{r}_1$; $r_b = \bar{r}_2\bar{r}_1$;
**end** *cgs2_step*
```

We now give the CGS2 algorithm from [1,7] for computing the decomposition (1.1).

Function 4.2 (Classical Gram–Schmidt with Reorthogonalization (CGS2))

```
**function** $[Q, R]$=$cgs2(A)$
$[m, n]$=**size**$(A)$;
$R = \|A(:, 1)\|$; $Q = A(:, 1)/R$;
**for** $k = 2 : n$
    $[\mathbf{q}_{new}, r_{new}, \mathbf{s}_{new}]$=**cgs2_step**$(Q, A(:, k))$;
    $R = \begin{matrix} R & \mathbf{s}_{new} \\ 0 & r_{new} \end{matrix}$;   $Q = \; Q \; \mathbf{q}_{new}$ ;
**end**;
**end**; *cgs2*
```

Unlike the version of CGS2 in [7], we use **local_qr** in Function 4.1 for scaling $\mathbf{y}_1$ to get $\bar{\mathbf{q}}_1$. It is done here to reinforce the connection between the development Functions 4.1 and 4.2 and that of Function 2.2 and 2.3. If we did not compute $\bar{\mathbf{q}}_1$, for the first orthogonalization step in Function 4.1, the relationships (4.3)–(4.6) would *(implicitly)* be replaced by the exact relationships

$$\begin{aligned}\mathbf{y}_1 &= \bar{\mathbf{q}}_1\bar{r}_1,\\ \bar{r}_1 &= \|\bar{\mathbf{y}}_1\|, \quad \bar{\mathbf{q}}_1^T\bar{\mathbf{q}}_1 = 1\end{aligned}$$

thereby making our interpretation of Theorems 3.4 and 3.5, given below, pessimistic.

When $p = 1$, then $t_k = k$. In floating point arithmetic, the computed values from the function **local_qr** in this section satisfy

$$\bar{r} = (1+\delta)\|\mathbf{y}\|, \quad |\delta| \le d_0 m\varepsilon_M \tag{4.3}$$

for a modest constant d_0, and

$$\bar{\mathbf{q}} = (I_m + E)\mathbf{y}/\bar{r}, \quad E = \operatorname{diag}(\epsilon_i), \quad \|E\| \le \varepsilon_M. \tag{4.4}$$

From (4.3)–(4.4), it follows that

$$\|\mathbf{y} - \bar{\mathbf{q}}\bar{r}\| \le \varepsilon_M\|\mathbf{y}\| \tag{4.5}$$

and

$$|1 - \bar{\mathbf{q}}^T\bar{\mathbf{q}}| \le d_1 m\varepsilon_M \tag{4.6}$$

for a modest constant d_1 which are (2.5)–(2.6) with $L_1(m,1) = \max\{d_1 m, 1\} = d_1 m$.

The other operations of a CGS step are

$$\bar{\mathbf{s}} + \delta\bar{\mathbf{s}} = U^T\mathbf{b}$$

where for a modest constant d_2,

$$\|\delta\bar{\mathbf{s}}\| \le d_2\varepsilon_M m k^{1/2}\|\mathbf{b}\|.$$

Thus, $L_2(m,k,1) = d_2 m k^{1/2}$.

We also have

$$\mathbf{y}_1 + \delta\mathbf{y}_1 = \mathbf{b} - U\bar{\mathbf{s}}_1$$

where

$$\|\delta\mathbf{y}_1\| \le d_3\varepsilon_M(1 + k^{3/2})\|\mathbf{b}\|.$$

Therefore $L_3(k, 1) = d_3(1 + k^{3/2})$. Thus one step is

$$\bar{\mathbf{q}}_1 \bar{r}_1 = (I_m - UU^T)\mathbf{b} + \mathbf{f}_1$$

where

$$\|\mathbf{f}_1\| \le \varepsilon_M L_F(m, k, 1)\|\mathbf{b}\|$$

and

$$\begin{aligned} L_F(m, k, 1) &= L_1(m, 1) + \sqrt{2}L_2(m, k, 1) + L_3(k, 1) \\ &= d_1 m + \sqrt{2}d_2 m k^{1/2} + d_3(1 + k^{3/2}) \\ &\le d_F m k^{1/2} \end{aligned}$$

for some modest constant d_F.

The two operations

$$\begin{aligned} \mathbf{s}_B + \delta\mathbf{s}_B &= \bar{\mathbf{s}}_1 + \bar{\mathbf{s}}_2 \bar{r}_1, \\ r_B + \delta r_B &= \bar{r}_2 \bar{r}_1, \end{aligned}$$

satisfy

$$\begin{aligned} \|\delta\mathbf{s}_B\| &\le d_4 \varepsilon_M \|\mathbf{b}\|, \\ |\delta r_B| &\le \varepsilon_M \|\mathbf{b}\| \end{aligned}$$

for a modest constant d_4. Thus $L_4(1) = d_4$ and $L_5(1) = d_5 = 1$ so that from the definition in (3.5),

$$\begin{aligned} f_{resid}(m, k, 1) &= (1 + \sqrt{2})d_1 m + 2d_3(1 + k^{3/2}) + d_4 + \sqrt{2} \\ &\le d_{resid} \max\{m, k^{3/2}\} \end{aligned}$$

for some modest constant d_{resid}.

Taking $t = s = n$, we have that Function 4.2 obtains the orthogonal factorization satisfying

$$\begin{aligned} \|A - QR\| &\le \varepsilon_M n^{1/2} f_{resid}(m, n, 1)\|A\| \\ &\le d_{resid}\varepsilon_M \max\{mn^{1/2}, n^2\}\|A\|. \end{aligned}$$

The condition for near orthogonality of Q has a nice interpretation. Assumption (3.14) may be written

$$\varepsilon_M f_{sing}(m, k, 1)\|\mathbf{a}_k\| \le \gamma(k)|r_{kk}| \tag{4.7}$$

where again, $\gamma(\cdot)$ is given in (3.6).

Finally,

$$\begin{aligned}
f_{sing}(m,k,1) &= \sqrt{2}(f_1(m,k,1) + L_F(m,k,1)) + \gamma(k)L_5(1) \\
&\le 5[(51k+1)^{1/2}]L_F(m,k,1) + d_5 \\
&= 5[(51k+1)^{1/2}](d_1 m + d_2 m k^{1/2} + d_3(1+k^{3/2})) + d_5 \\
&\le d_{sing}\max\{mk, k^2\} = d_{sing} mk
\end{aligned}$$

for some constant d_{sing}.

The assumption (4.7) leads to the bound

$$\|I - Q^T Q\| \le \varepsilon_M f_1(m,n,1)$$

where

$$\begin{aligned}
f_1(m,n,1) &= \gamma(n)^{-1} L_F(m,n,1) \\
&= \beta \max\{mn, n^2\} = \beta mn
\end{aligned}$$

for some modest constant β.

Notice that (4.7) is merely an assumption that each of the diagonals of R is sufficiently bounded away from zero. There is no assumption on the condition number of R (or A) and (4.7) is weaker than the assumption given by Giraud et al. [7, Theorem 2 and Lemma 2] for Function 4.2.

5 Conclusions

The block Gram–Schmidt procedure proposed as Function 2.3 is a BLAS-3 [6] compatible algorithm that provides a near orthogonal Q provided that the diagonal blocks on R are not too ill-conditioned. As with several Q-R factorization algorithms, Function 2.3 produces a small residual according to the criterion (1.3). Such an algorithm could be used in the implementation of Krylov space methods such as GMRES.

The intermediate step given as Function 2.2 solves the problem (1.5)–(1.6) effectively provided that the condition number $\|B\|\|R_B^{-1}\|$ is bounded as in (3.27). This is important in the development of block generalizations of the update/downdate algorithms in [5,2].

Finally, if we consider the special case of block size 1, using different assumptions, we produce a bound of Giraud et al. [7] for Function 4.2 showing that a near left orthogonal Q is produced if diagonals of R are bounded sufficiently away from zero as in (4.7).

Acknowledgments Ilse Ipsen, Miro Rozložnik, Hasan Erbay and two tenacious referees made helpful suggestions on an earlier version of this manuscript. Rich Lehoucq told the first author about reference [16].

6 Appendix. Proofs of technical lemmas from Sect. 3

Two of the technical lemmas in Sect. 3, Lemmas 3.1 and 3.3 are based upon matrix multiply and add operations, thus those proofs are presented first. The proof of Lemma 3.2 and the proof of Lemma 3.5 are verifications that two norm bounds in exact arithmetic are not significantly altered in floating point arithmetic.

Using a result from Higham [10, p.71] on matrix multiplication in floating point arithmetic and putting in an additional term for a matrix add, the operation

$$M = C + GH, \quad \begin{array}{l} M, C \in \mathbb{R}^{m\times p} \\ G \in \mathbb{R}^{m\times t}, H \in \mathbb{R}^{t\times p}, \\ m \geq t \geq p \end{array} \tag{6.1}$$

satisfies

$$M + \Delta M = C + GH \tag{6.2}$$

where

$$|\Delta M| \leq d_{mat}\varepsilon_M t(|C| + |G||H|)$$

for some modest constant d_{mat}. If we use the bound on the two-norm

$$\|M\| \leq \||M|\| \leq \min\{m^{1/2}, p^{1/2}\}\|M\|$$

for an $M \in \mathbb{R}^{m\times p}$, $m \geq p$, then we have

$$\|\Delta M\| \leq d_{mat}\varepsilon_M t p^{1/2}(\|C\| + t^{1/2}\|G\|\|H\|). \tag{6.3}$$

Thus the bound on these operations for Functions 2.1 and 2.2 in Lemmas 3.1 and 3.3 are just a matter of bounding $\|C\|$, $\|G\|$ and $\|H\|$ in the appropriate context.

First, we prove Lemma 3.1 which concerns the operations of the form (6.1) in Function 2.1.

Proof of Lemma 3.1 Interpreting (6.3) for the computation in (3.18), we have

$$\|\delta\bar{S}\| \leq d_{mat}\varepsilon_M m p^{1/2} t^{1/2}\|U\|\|B\|.$$

By the assumption (3.15), $\|U\| \leq c_U(\varepsilon_M) \leq \sqrt{2}$, thus

$$\begin{aligned}\|\delta\bar{S}\| &\leq \sqrt{2}d_{mat}\varepsilon_M m p^{1/2} t^{1/2}\|B\| \\ &= d_2\varepsilon_M m p^{1/2} t^{1/2}\|B\|, \quad d_2 = \sqrt{2}d_{mat}, \\ &= \varepsilon_M L_2(m, t, p)\|B\|\end{aligned}$$

which is the bound in (3.18).

The second such operation is given in (3.19). The interpretation of (6.3) for that computation is

$$\|\delta Y\| \leq d_{mat}\varepsilon_M t p^{1/2}[\|B\| + t^{1/2}\|U\|\|\bar{S}\|].$$

By our assumptions below (3.13),

$$\begin{aligned}\|\bar{S}\| &\leq \|U^T B\| + \|\delta\bar{S}\|\\ &\leq \|U\|\|B\| + \varepsilon_M L_2(m,t,p)\|B\|\\ &\leq (1+\varepsilon_M(0.5 f_1(m,t,p) + L_2(m,t,p))\|B\|\\ &\leq c_S(\varepsilon_M)\|B\|\\ &\leq \sqrt{2}\|B\|.\end{aligned}$$

By (3.15), we have that $\|U\| \leq \sqrt{2}$, thus

$$\begin{aligned}\|\delta Y\| &\leq d_{mat}\varepsilon_M t p^{1/2}(1+2t^{1/2})\|B\|\\ &= d_3\varepsilon_M t p^{1/2}(1+2t^{1/2})\|B\|\\ &= \varepsilon_M L_3(t,p)\|B\|\end{aligned}$$

which is the bound in (3.19). □

Lemma 3.3 concerns two computations in Function 2.2 that are of the form (6.2).

Proof of Lemma 3.3 Interpreting (6.3) for the computation in (3.32) yields

$$\|\delta S_B\| \leq d_{mat}\varepsilon_M p^{3/2}[\|\bar{S}_1\| + p^{1/2}\|\bar{S}_2\|\|\bar{R}_1\|].$$

To bound the norms, we have that

$$\begin{aligned}\|\bar{S}_1\| &\leq c_B(\varepsilon_M)\|B\| \leq \sqrt{2}\|B\|,\\ \|\bar{S}_2\| &\leq c_B(\varepsilon_M)\|\bar{Q}_1\| \leq c_B(\varepsilon_M)c_Q(\varepsilon_M) \leq 2,\\ \|\bar{R}_1\| &\leq c_R(\varepsilon_M)\|B\| \leq \sqrt{2}\|B\|.\end{aligned}$$

Thus

$$\begin{aligned}\|\delta S_B\| &\leq d_{mat}\varepsilon_M p^{3/2}[\sqrt{2} + 2\sqrt{2}p^{1/2}]\|B\|\\ &\leq d_4\varepsilon_M p^{3/2}[1+2p^{1/2}]\|B\|, \quad d_4 = \sqrt{2}d_{mat}\\ &= \varepsilon_M L_4(p)\|B\|.\end{aligned}$$

The final such operation is that in (3.33). The interpretation of (6.3) is

$$\|\delta R_B\| \leq d_{mat}\varepsilon_M p^2\|\bar{R}_1\|\|\bar{R}_2\|.$$

Using the bounds in (3.30)–(3.31), this is

$$\begin{aligned}\|\delta R_B\| &\le d_{mat}\varepsilon_M p^2 c_R^2(\varepsilon_M)c_Q(\varepsilon_M)\|B\| \\ &\le 2\sqrt{2}d_{mat}\varepsilon_M p^2\|B\| \\ &= d_5\varepsilon_M p^2\|B\|, \quad d_5 = 2\sqrt{2}d_{mat} \\ &= \varepsilon_M L_5(p)\|B\|. \end{aligned}$$

□

Lemma 3.2, our next technical lemma, is a bound on $\|I_m - UU^T\|$ where $U \in \mathbb{R}^{m\times t}$ is near left orthogonal.

Proof of Lemma 3.2 Let U have the Q-R decomposition

$$U = Z\begin{pmatrix} R_U \\ 0_{(m-t)\times t}\end{pmatrix}$$

where Z is orthogonal and R_U is upper triangular. Then

$$\|I_t - U^T U\| = \|I_t - R_U^T R_U\| \le \varepsilon_M f_1(m,t,p)$$

and

$$I_m - UU^T = Z\begin{pmatrix} I_t - R_U R_U^T & 0 \\ 0 & I_{m-t}\end{pmatrix} Z^T.$$

Since $R_U^T R_U$ and $R_U R_U^T$ have the same eigenvalues,

$$\|I_t - R_U R_U^T\| = \|I_t - R_U^T R_U\| \le \varepsilon_M f_1(m,t,p).$$

Using the assumption (3.15) and the results of Theorem 3.1, we have

$$\begin{aligned}\|I_m - UU^T\| &\le \max\{\|I_t - R_U R_U^T\|, 1\} \\ &= \max\{\|I_t - R_U^T R_U\|, 1\} \\ &\le \max\{\varepsilon_M f_1(m,t,p), 1\} = 1 \end{aligned} \tag{6.4}$$

by our assumption about $f_1(m,t,p)$ in (1.4). □

Lemma 3.5, our final technical lemma, is a result relating $\|\bar{R}_2^{-1}\|$ to $\|U^T\bar{Q}_1\bar{R}_2^{-1}\|$. It has the elementary but long proof given next.

Proof of Lemma 3.5 We start with interpreting Lemma 3.1 for the second CGS step in Function 2.2 which leads to

$$Q_B\bar{R}_2 = (I_m - UU^T)\bar{Q}_1 + F_2.$$

Taking the normal equations matrices of both sides yields

$$\begin{aligned}\bar{R}_2^T Q_B^T Q_B\bar{R}_2 = \bar{Q}_1^T(I_m - UU^T)^2\bar{Q}_1 + F_2^T(I_m - UU^T)\bar{Q}_1 \\ +\bar{Q}_1^T(I_m - UU^T)F_2 + F_2^T F_2. \end{aligned} \tag{6.5}$$

An expansion of $\bar{Q}_1^T(I_m - UU^T)^2\bar{Q}_1$ produces

$$\begin{aligned}\bar{Q}_1^T(I_m - UU^T)^2\bar{Q}_1 &= \bar{Q}_1^T\bar{Q}_1 - \bar{Q}_1^T UU^T\bar{Q}_1 - \bar{Q}_1^T U(I_t - U^T U)U^T\bar{Q}_1 \\ &= I_p - \bar{Q}_1^T UU^T\bar{Q}_1 + \bar{Q}_1^T\bar{Q}_1 - I_p \\ &\quad -\bar{Q}_1^T U(I_t - U^T U)U^T\bar{Q}_1 \end{aligned} \tag{6.6}$$

so that the combination of (6.5) and (6.6) is

$$\bar{R}_2^T\bar{R}_2 = I_p - \bar{Q}_1^T UU^T\bar{Q}_1 + E \tag{6.7}$$

where

$$\begin{aligned} E &= E_1 + E_2 + E_3, \\ E_1 &= F_2^T(I_m - UU^T)\bar{Q}_1 + \bar{Q}_1^T(I_m - UU^T)F_2 + F_2^T F_2, \\ E_2 &= \bar{Q}_1^T\bar{Q}_1 - I_p - \bar{Q}_1^T U(I_t - U^T U)U^T\bar{Q}_1, \\ E_3 &= \bar{R}_2^T(Q_B^T Q_B - I_p)\bar{R}_2. \end{aligned}$$

Standard norm bounds for E_1 and E_3 yields

$$\begin{aligned} \|E_1\| &\leq 2\|F_2\|\|\bar{Q}_1\|\|I_m - UU^T\| + \|F_2\|^2 \\ &\leq 2\|F_2\|(\|\bar{Q}_1\| + 0.5\|F_2\|) \\ &\leq 2\varepsilon_M c_Q(\varepsilon_M)L_F(m,t,p)(c_Q(\varepsilon_M) + 0.5\varepsilon_M c_Q(\varepsilon_M)L_F(m,t,p)) \\ &= 2c_Q^2(\varepsilon_M)L_F(m,t,p)(1 + 0.5L_F(m,t,p)) \\ &\leq 2c_Q^2(\varepsilon_M)c_F(\varepsilon_M)L_F(m,t,p) \\ &\leq 4\sqrt{2}L_F(m,t,p) \\ \|E_3\| &= \|I_p - Q_B^T Q_B\|\|\bar{R}_2\|^2 \leq \varepsilon_M c_R^2(\varepsilon_M)c_Q^2(\varepsilon_M)L_1(m,p) \leq 4\varepsilon_M L_1(m,p) \end{aligned}$$

whereas for E_2 if we use the fact that

$$\begin{aligned} \|U^T\bar{Q}_1\| &\leq \|U^T\bar{Q}_1\bar{R}_2^{-1}\|\|\bar{R}_2\| \\ &\leq c_R(\varepsilon_M)c_Q(\varepsilon_M)\gamma(k), \end{aligned} \tag{6.8}$$

then

$$\begin{aligned} \|E_2\| &\leq \|I_p - \bar{Q}_1^T\bar{Q}_1\| + \|I_t - U^T U\|\|U^T\bar{Q}_1\|^2 \\ U &\leq \varepsilon_M[L_1(m,p) + f_1(m,t,p)\gamma^2(k)c_R^2(\varepsilon_M)c_Q^2(\varepsilon_M)] \\ &\leq \varepsilon_M[L_1(m,p) + \gamma(k)c_R^2(\varepsilon_M)c_Q^2(\varepsilon_M)L_F(m,t,p)], \\ &\leq \varepsilon_M[L_1(m,p) + 4L_F(m,t,p)] \end{aligned}$$

since $\gamma(k) \leq 1$, $c_R(\varepsilon_M) \leq \sqrt{2}$, and $c_Q(\varepsilon_M) \leq \sqrt{2}$. Thus, using (3.4), we have

$$\begin{aligned}\|E\| &\leq \|E_1\| + \|E_2\| + \|E_3\| \\ &\leq \varepsilon_M[4(1+\sqrt{2})L_F(m,t,p) + 5L_1(m,p)] \\ &= \varepsilon_M L_6(m,t,p).\end{aligned}$$

Now to show the equivalence between (3.45) and (3.46). Since we assume that $\bar{R}_2$ is nonsingular, we can rewrite (6.7) as

$$I_p = \bar{R}_2^{-T}\bar{R}_2^{-1} - \bar{R}_2^{-T}\bar{Q}_1^T U U^T \bar{Q}_1 \bar{R}_2^{-1} + \bar{R}_2^{-T} E \bar{R}_2^{-1}$$

so that

$$\bar{R}_2^{-T}\bar{R}_2^{-1} = I_p + \bar{R}_2^{-T}\bar{Q}_1^T U U^T \bar{Q}_1 \bar{R}_2^{-1} - \bar{R}_2^{-T} E \bar{R}_2^{-1}. \tag{6.9}$$

If $\lambda_{max}(\cdot)$ is the maximum absolute eigenvalue of the contents, then (6.9) plus Weyl's inequality for eigenvalues [8, Corollary 8.1.6] and a norm bound yield

$$\lambda_{max}(\bar{R}_2^{-T}\bar{R}_2^{-1}) = 1 + \lambda_{max}(\bar{R}_2^{-T}\bar{Q}_1 U U^T \bar{Q}_1 \bar{R}_2^{-1}) - \xi\|\bar{R}_2^{-1}\|^2 \tag{6.10}$$

where

$$|\xi| \leq \|E\|. \tag{6.11}$$

Using the relationship, $\lambda_{max}(C^T C) = \|C\|^2$ on (6.10) yields (3.44). □

References

1. Abdelmalek, N.I.: Roundoff error analysis for Gram–Schmidt method and solution of linear least squares problems. BIT **11**(4), 354–367 (1971)
2. Barlow, J.L., Smoktunowicz, A., Erbay, H.: Improved Gram–Schmidt downdating methods. BIT **45**, 259–285 (2005)
3. Björck, Å.: Numerics of Gram–Schmidt orthogonalization. Linear Algebra Appl. **197–198**, 297–316 (1994)
4. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM Publications, Philadelphia (1996)
5. Daniel, J.W., Gragg, W.B., Kaufman, L., Stewart, G.W.: Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization. Math. Comp. **30**(136), 772–795 (1976)
6. Dongarra, J.J., DuCroz, J.J., Duff, I.S., Hammarling, S.J.: A set of level 3 basic linear algebra subprograms. ACM Trans. Math. Softw. **16**, 1–17 (1990)
7. Giraud, L., Langou, J., Rozložnik, M.: Rounding error analysis of the classical Gram–Schmidt orthogonalization process. Numer. Math. **101**(1), 87–100 (2005)
8. Golub, G.H., Van Loan, C.F.: Matrix Computations. The Johns Hopkins Press, Baltimore (1996)
9. Greenbaum, A., Rozložnik, M., Strakoš, Z.: Numerical behavior of the modified Gram–Schmidt GMRES implementation. BIT **37**, 706–719 (1997)
10. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM Publications, Philadelphia (2002)
11. Hoemmen, M.F.: Communication-avoiding Krylov subspace methods PhD thesis. University of California, Berkeley, CA, USA (2010)
12. Horn, R.A., Johnson, C.A.: Matrix Analysis. Cambridge University Press, Cambridge (1985)

13. Jalby, W., Phillippe, B.: Stability analysis and improvement of the block Gram–Schmidt algorithm. SIAM J. Sci. Stat. Comput. **12**, 1058–1073 (1991)
14. Paige, C.C., Rozložnik, M., Strakoš, Z.: Modified Gram–Schmidt (MGS), least squares and the backward stability of MGS-GMRES. SIAM J. Matrix Anal. Appl. **28**(1), 264–284 (2006)
15. Rice, J.R.: Experiments on Gram–Schmidt orthogonalization. Math. Comput. **20**(94), 325–328 (1966)
16. Stathopoulos, A., Wu, K.: A block orthogonalization procedure with constant synchronization requirements. SIAM J. Sci. Comput. **23**(6), 2165–2182 (2002)
17. Stewart, G.W.: Block Gram–Schmidt orthogonalization. SIAM J. Sci. Comput. **31**(1), 761–775 (2008)
18. Vanderstaeten, D.: An accurate parallel block Gram–Schmidt algorithm without reorthogonalization. Numer. Linear Algebra Appl. **7**, 219–236 (2000)