A ROBUST CRITERION FOR THE MODIFIED GRAM–SCHMIDT ALGORITHM WITH SELECTIVE REORTHOGONALIZATION[*]

LUC GIRAUD[†] AND JULIEN LANGOU[†]

Abstract. A new criterion for selective reorthogonalization in the modified Gram–Schmidt algorithm is proposed. We study its behavior in the presence of rounding errors. We give some counterexample matrices which prove that the standard criteria might fail. Through numerical experiments, we illustrate that our new criterion seems to be suitable also for the classical Gram–Schmidt algorithm with selective reorthogonalization.

Key words. Gram–Schmidt algorithm with selective reorthogonalization

AMS subject classifications. 65F25, 65G50, 15A23

DOI. 10.1137/S106482750340783X

Introduction. Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be a real $m \times n$ matrix $(m > n)$ whose columns are linearly independent. In many applications, it is required to have an orthonormal basis for the space spanned by the columns of $\mathbf{A}$. This amounts to a matrix $\mathbf{Q} \in \mathbb{R}^{m \times n}$ with orthonormal columns such that $\mathbf{A} = \mathbf{QR}$, $\mathbf{R} \in \mathbb{R}^{n \times n}$. Moreover, it is possible to require $\mathbf{R}$ to be triangular; we then end up with the so-called QR-factorization. For all j, the first j columns of $\mathbf{Q}$ are an orthonormal basis for the space spanned by the first j columns of $\mathbf{A}$.

Starting from $\mathbf{A}$, there are many algorithms that build such a factorization. In this paper, we focus on the Gram–Schmidt algorithm [1] that consists of projecting successively the columns of $\mathbf{A}$ on the space orthogonal to the space spanned by the columns of $\mathbf{Q}$ already constructed. Depending on how the projections are performed, there are two main versions of this algorithm [3]: the classical Gram–Schmidt algorithm (CGS) and the modified Gram–Schmidt algorithm (MGS). In exact arithmetic, both algorithms produce exactly the same results and the resulting matrix $\mathbf{Q}$ has orthonormal columns. In the presence of round-off errors, $\mathbf{Q}$ computed by CGS differs from that computed by MGS. In both cases, the columns of $\mathbf{Q}$ may be far from orthogonal. To remedy this problem, a solution is to iterate the procedure and to project each column of $\mathbf{A}$ several times instead of only once on the space orthogonal to the space spanned by the constructed columns of $\mathbf{Q}$. Giraud, Langou, and Rozložník [17] have shown that, when using floating-point arithmetic, either for CGS or MGS, two iterations are enough when the initial matrix $\mathbf{A}$ is numerically nonsingular. This confirms what was already experimentally well known for $n = 2$ vectors (see Parlett [10]). In this paper, we focus mainly on the Gram–Schmidt algorithms, where the number of projections for each column of $\mathbf{A}$ is either 1 or 2. When the number of reorthogonalizations performed is exactly 2, we call the resulting algorithm the classical (resp., modified) Gram–Schmidt algorithm with reorthogonalization and denote it by CGS2 (resp., MGS2); the MGS2 algorithm is given in Algorithm 1.

The use of either CGS2 or MGS2 guarantees a reliable result in terms of orthogonality [17]; however, the computational cost is twice as much as for CGS or

†CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France (giraud@cerfacs.fr, langou@cerfacs.fr).

MGS. In many applications, we observe that either CGS or MGS is good enough; the additional reorthogonalizations performed in CGS2 or MGS2 are then useless. A good compromise in terms of orthogonality quality and time is to use a selective reorthogonalization criterion to check whether for each column of $\mathbf{A}$ an extra reorthogonalization is needed. Historically, it is known that Rutishauser [5] introduced the first criterion in a Gram–Schmidt algorithm with reorthogonalization. We refer to it as the K-criterion. It is dependent on a single parameter $K \geq 1$. The resulting algorithms are called the classical and modified Gram–Schmidt algorithms with selective reorthogonalization and K-criterion; they are denoted by CGS2(K) and MGS2(K), respectively. We give below the MGS2(K).

Algorithm 1 (MGS2)	**Algorithm 2** (MGS2(K))
for $j = 1$ to n **do**	**for** $j = 1$ to n **do**
$\quad \mathbf{a}_j^{(1)(1)} = \mathbf{a}_j$	$\quad \mathbf{a}_j^{(1)(1)} = \mathbf{a}_j$
$\quad$ **for** $k = 1$ to $j - 1$ **do**	$\quad$ **for** $k = 1$ to $j - 1$ **do**
$\quad\quad r_{kj}^{(1)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(1)}$	$\quad\quad r_{kj}^{(1)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(1)}$
$\quad\quad \mathbf{a}_j^{(k+1)(1)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k r_{kj}^{(1)}$	$\quad\quad \mathbf{a}_j^{(k+1)(1)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k r_{kj}^{(1)}$
$\quad$ **end for**	$\quad$ **end for**
	$\quad$ **if** $\left(\frac{\|\mathbf{a}_j\|_2}{\|\mathbf{a}_j^{(j)(1)}\|_2} \leq K \right)$ **then**
	$\quad\quad r_{jj} = \|\mathbf{a}_j^{(j)(1)}\|_2$
	$\quad\quad \mathbf{q}_j = \mathbf{a}_j^{(j)(1)} / r_{jj}$
	$\quad\quad r_{kj} = r_{kj}^{(1)}, 1 \leq k \leq j - 1$
	$\quad$ **else**
$\quad \mathbf{a}_j^{(1)(2)} = \mathbf{a}_j^{(j)(1)}$	$\quad\quad \mathbf{a}_j^{(1)(2)} = \mathbf{a}_j^{(j)(1)}$
$\quad$ **for** $k = 1$ to $j - 1$ **do**	$\quad\quad$ **for** $k = 1$ to $j - 1$ **do**
$\quad\quad r_{kj}^{(2)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(2)}$	$\quad\quad\quad r_{kj}^{(2)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(2)}$
$\quad\quad \mathbf{a}_j^{(k+1)(2)} = \mathbf{a}_j^{(k)(2)} - \mathbf{q}_k r_{kj}^{(2)}$	$\quad\quad\quad \mathbf{a}_j^{(k+1)(2)} = \mathbf{a}_j^{(k)(2)} - \mathbf{q}_k r_{kj}^{(2)}$
$\quad$ **end for**	$\quad\quad$ **end for**
$\quad r_{jj} = \|\mathbf{a}_j^{(j)(2)}\|_2$	$\quad\quad r_{jj} = \|\mathbf{a}_j^{(j)(2)}\|_2$
$\quad \mathbf{q}_j = \mathbf{a}_j^{(j)(2)} / r_{jj}$	$\quad\quad \mathbf{q}_j = \mathbf{a}_j^{(j)(2)} / r_{jj}$
$\quad r_{kj} = r_{kj}^{(1)} + r_{kj}^{(2)}, 1 \leq k \leq j - 1$	$\quad\quad r_{kj} = r_{kj}^{(1)} + r_{kj}^{(2)}, 1 \leq k \leq j - 1$
	$\quad$ **end if**
end for	**end for**

Using floating-point arithmetic, Parlett [10] showed that for two vectors the orthogonality obtained (measured by $|q_1^T q_2|$) is bounded by a constant times $K\varepsilon$, where ε denotes the machine precision. This gives a way of computing K to ensure a satisfactory level of orthogonality. For n vectors, the choice of the parameter K is not so clear. Giraud, Langou, and Rozložník [17] showed that if K is greater than the condition number of $\mathbf{A}$, $\kappa(\mathbf{A})$, then neither CGS2($K = \kappa(\mathbf{A})$) nor MGS2($K = \kappa(\mathbf{A})$) performs any reorthogonalization. Interesting values for K therefore range from 1 (this corresponds to CGS2 or MGS2) to $\kappa(\mathbf{A})$ (this corresponds to CGS or MGS). If K is high, then we have few reorthogonalizations, so we could expect a lower level of orthogonality than if K were smaller, where more reorthogonalizations are performed. To reach orthogonality at the machine precision level, Rutishauser [5] chose the value $K = 10$. We find the following explanation of this value in Gander [9, p. 12]: *"In particular*

one may state the rule of thumb that at least one decimal digit is lost by cancellation if $10\|\mathbf{a}_j^{(1)}\|_2 \le \|\mathbf{a}_j\|_2$. *This equation is the criterion used by Rutishauser to decide whether reorthogonalization is necessary.*" The value $K = \sqrt{2}$ is also very often used since the publication of the paper of Daniel et al. [7] (e.g., by Ruhe [11] and Reichel and Gragg [13]). More exotic values like $K = 100.05$ [8] or $K = \sqrt{5}$ [16] have also been implemented. Hoffmann [12] tested a wide range of values $K = 2, 10, \ldots, 10^{10}$. The conclusion of his experiments is that the K-criterion is always satisfied at either the first or second loop and the final level of orthogonality is proportional to the parameter K and to machine precision, exactly as is the case for two vectors.

The goal of this paper is to present new ideas on the subject of selective reorthogonalization. In section 1, we show that MGS2 applied to numerically nonsingular matrices gives a set of vectors orthogonal to machine precision. This is summarized in Theorem 1.1. The proof given in section 1 is strongly related to the work of Björck [4]. In fact we extend his result for MGS to MGS2. Sections 1.1–1.5 use his results directly with modifications adapted to a second loop of reorthogonalization. In sections 1.5–1.11, we develop special results that aim to show that the R-factor corresponding to the second loop is well conditioned. To work at step p of the algorithm, an assumption on the level of orthogonality at the previous step is necessary; this is done in section 1.8 using an induction assumption. In section 1.12, we adapt the work of Björck [4] to conclude that the level of orthogonality at step p is such that the induction assumption holds. During this proof, several assumptions are made; in section 1.13, for sake of clarity, we encompass all these assumptions into one. Finally, in section 1.14, we conclude the proof by induction. In section 2.1, we give a new criterion for the MGS algorithm. This criterion is dependent on a single parameter L. We call this the L-criterion and call the resulting algorithm MGS2(L). This criterion appears naturally from the proof in section 1, and the result of Theorem 1.1 for MGS2 holds also for MGS2(L) when $L < 1$. Therefore, we state that MGS2(L) with $L < 1$ applied to numerically nonsingular matrices gives a set of vectors orthogonal to machine precision. In section 2.2, we give a counterexample matrix for which, if $L = 1.03$, MGS2(L) provides a set of vectors that are far from orthogonal. Concerning the K-criterion, first of all we notice that the K-criterion makes sense for $K > 1$; otherwise MGS2(K) reduces to MGS2. In section 3, we give counterexample matrices for which MGS2(K), with K ranging from 1.43 down to 1.05, provides a set of vectors that are far from orthogonal. These examples illustrate that the K-criterion may not be robust.

The result established in section 1 for MGS2 is similar to that given in [17]. Both results establish with two different proofs that MGS2 gives a set of vectors orthogonal to machine precision. However, the proof given in this paper is different and applies only to the MGS algorithm, whereas the CGS algorithm is covered by the proof in [17]. The advantage of our new proof is that it enables us to derive the L-criterion for the MGS algorithm. Moreover, this paper extends the work of Björck [4] directly from MGS to MGS2(L).

In the error analysis, we shall assume that floating-point arithmetic is used and will follow the technique and notation of Wilkinson [2] and Björck [4]. Let 'op' denote any of the four operators $+ - * /$. Then an equation of the form

$$z = \mathrm{fl}(x\text{'op'}y)$$

will imply that x, y, and z are floating-point numbers and z is obtained from x and y using the appropriate floating-point operation. We assume that the rounding errors

in these operations are such that

$$\mathrm{fl}(x\text{`op'}y) = (x\text{`op'}y)(1 + \varepsilon), |\varepsilon| \leq 2^{-t},$$

where 2^{-t} is the unit round-off.

In sections 1 and 2.1, to distinguish computed quantities from exact quantities, we use an overbar on the computed quantities. For the sake of readability in sections 2.2 and 3, which are dedicated to numerical experiments, the overbars are no longer used. Throughout this paper, the matrices are denoted by bold capitals, e.g., $\mathbf{A}$; vectors are denoted by bold characters, e.g., $\mathbf{x}$; scalars are denoted in a normal font, e.g., η. The entry (i, j) of $\mathbf{A}$ is denoted by a_{ij}. However, when there may be ambiguity, we use a comma, e.g., the entry $(j - 1, j)$ of $\mathbf{A}$ is denoted by $a_{j-1,j}$. The jth column of $\mathbf{A}$ is the vector $\mathbf{a}_j$. The paper is written for real matrices, the Euclidean scalar product is denoted by $\mathbf{x}^T\mathbf{y}$, $\| \ \|_2$ stands for the 2-norm for vectors and for the induced norm for the matrix, and $\| \ \|_F$ stands for the Frobenius norm. $\sigma_{\min}(\mathbf{A})$ is the minimum singular value of $\mathbf{A}$ in the 2-norm. $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$ in the 2-norm. $\mathbf{I}_p$ is the identity matrix of dimension p. Finally, we shall mention that our results also extend to complex arithmetic calculations.

1. Adaptation of the work by Björck [4] for the MGS algorithm to the MGS2 algorithm.

1.1. Description of the MGS2 algorithm without square roots.
In this section, we use the same approach as Björck in [4]. In his paper, he considers the MGS algorithm without square roots to study its numerical behavior in floating-point arithmetic. In order to keep most of our work in agreement with his, we also study the MGS2 algorithm without square roots instead of the MGS2 algorithm (Algorithm 1). The MGS2 algorithm without square roots is described by Algorithm 3.

Algorithm 3 (MGS2 without square roots)

for $j = 1$ to n **do**
$\quad \mathbf{a}_j^{(1)(1)} = \mathbf{a}_j$
$\quad$ **for** $k = 1$ to $j - 1$ **do**
$\quad\quad r_{kj}^{'(1)} = \mathbf{q}_k^{'T} \mathbf{a}_j^{(k)(1)}/d_k$
$\quad\quad \mathbf{a}_j^{(k+1)(1)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k' r_{kj}^{'(1)}$
$\quad$ **end for**
$\quad \mathbf{a}_j^{(1)(2)} = \mathbf{a}_j^{(j)(1)}$
$\quad$ **for** $k = 1$ to $j - 1$ **do**
$\quad\quad r_{kj}^{'(2)} = \mathbf{q}_k^{'T} \mathbf{a}_j^{(k)(2)}/d_k$
$\quad\quad \mathbf{a}_j^{(k+1)(2)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k' r_{kj}^{'(2)}$
$\quad$ **end for**
$\quad \mathbf{q}_j' = \mathbf{a}_j^{(j)(2)}$
$\quad d_j = \|\mathbf{q}_j'\|_2^2$
$\quad r_{kj}' = r_{kj}^{'(1)} + r_{kj}^{'(2)}, 1 \leq k \leq j - 1$
$\quad r_{jj} = 1$
end for

The factorization resulting from MGS2 without square roots is denoted by

$$\mathbf{A} = \mathbf{Q}'\mathbf{R}',$$

where $\mathbf{R}'$ is a unit upper triangular matrix and $(\mathbf{Q}')^T \mathbf{Q}'$ is diagonal. The main interest in that approach is to avoid the square root operation ($\sqrt{\ }$) in floating-point arithmetic. The associated algorithm only requires the four basic operations $+$, $-$, $*$, and $/$. In exact arithmetic, the link between the QR-factors $\mathbf{Q}'$ and $\mathbf{R}'$ of Algorithm 3 and the QR-factors $\mathbf{Q}$ and $\mathbf{R}$ of Algorithm 1 is

$$\mathbf{q}_j = \mathbf{q}'_j / \|\mathbf{q}'_j\|_2 \quad \text{and} \quad r_{kj} = r'_{kj} \|\mathbf{q}'_j\|_2, \qquad k = 1, \ldots, j-1, \quad j = 1, \ldots, n.$$

1.2. Basic definitions for the error analysis. Following Björck [4], we define for $j = 1, \ldots, n$ the computed quantities for Algorithm 3,

$$
\begin{aligned}
\bar{r}'^{(r)}_{kj} &= \mathrm{fl}(\bar{\mathbf{q}}'^{T}_k \bar{\mathbf{a}}^{(k)(r)}_j / \bar{d}_k) && \text{for } k = 1, \ldots, j-1 \text{ and } r = 1, 2, \\
\bar{\mathbf{a}}^{(k+1)(r)}_j &= \mathrm{fl}(\bar{\mathbf{a}}^{(k)(r)}_j - \bar{\mathbf{q}}'_k \bar{r}'^{(r)}_{kj}) && \text{for } k = 1, \ldots, j-1 \text{ and } r = 1, 2, \\
\bar{\mathbf{q}}'_j &= \bar{\mathbf{a}}^{(j)(2)}_j, \\
\bar{d}_j &= \mathrm{fl}(\|\bar{\mathbf{q}}'_j\|^2_2), \\
\bar{r}'_{kj} &= \mathrm{fl}(\bar{r}'^{(1)}_{kj} + \bar{r}'^{(2)}_{kj}) && \text{for } k = 1, \ldots, j-1, \\
\bar{r}'_{jj} &= \mathrm{fl}(1).
\end{aligned}
$$

The initialization is

$$\bar{\mathbf{a}}^{(1)(1)}_j = \mathbf{a}_j.$$

At the end of the first loop (i.e., $r = 1$) the following copy is performed before starting the next loop (i.e., $r = 2$):

$$\bar{\mathbf{a}}^{(j)(2)}_j = \bar{\mathbf{a}}^{(1)(1)}_j.$$

We also introduce the normalized quantities for $j = 1, \ldots, n$,

$$
(1.1) \qquad \forall j = 1, \ldots, k-1, \quad
\begin{aligned}
\bar{\mathbf{q}}_j &= d^{-1/2}_j \bar{\mathbf{q}}'_j, & \bar{r}_{jj} &= d^{1/2}_j, \\
\bar{r}^{(r)}_{kj} &= d^{1/2}_j \bar{r}'^{(r)}_{kj}, & \bar{r}_{kj} &= \bar{r}^{(1)}_{kj} + \bar{r}^{(2)}_{kj},
\end{aligned}
$$

where

$$
d^{1/2}_j = \begin{cases} \|\bar{\mathbf{q}}'_j\|_2, & \bar{\mathbf{q}}'_j \neq 0, \\ 1, & \bar{\mathbf{q}}'_j = 0. \end{cases}
$$

Note that these latter quantities are never computed by the MGS2 algorithm without square roots—they are defined a posteriori. Thus expressions in (1.1) are exact relations.

From (1.1), the following relations also hold:

$$\|\bar{\mathbf{q}}_j\|_2 = 1, \quad \bar{r}_{jj} = \|\bar{\mathbf{a}}^{(j)(2)}_j\|_2, \quad \text{and} \quad \bar{\mathbf{a}}^{(j)(2)}_j = \bar{\mathbf{q}}_j \bar{r}_{jj}.$$

The first relation implies that $\mathbf{I} - \bar{\mathbf{q}}_j \bar{\mathbf{q}}^T_j$ is an orthogonal projection.

This section aims to prove the following theorem.

THEOREM 1.1. *Let $\mathbf{A}$ be an $m \times n$ matrix on which MGS2 without square roots is run using a well-designed floating-point arithmetic to obtain the computed Q-factor $\bar{\mathbf{Q}}$.*

Let 2^{-t} be the unit round-off.

Let L be a real such that $0 < L < 1$. If

$$(1.2) \qquad \frac{1}{L(1-L)} \times 10n^{5/2}(4.5m + 2)2^{-t} \cdot \kappa_2(\mathbf{A}) \leq 1,$$

then $\bar{\mathbf{Q}}$ is such that

$$(1.3) \qquad \|\mathbf{I} - \bar{\mathbf{Q}}^T\bar{\mathbf{Q}}\|_2 \leq \frac{2.61}{1-L} \cdot n^{3/2}(n + 1 + 2.5m)2^{-t}.$$

Notice that (1.3) indicates that the level of orthogonality reached with MGS2 is of the order of machine precision and that assumption (1.2) implies that $\mathbf{A}$ is numerically nonsingular. In the remainder of this section, we make a series of assumptions on $\mathbf{A}$ that hold until the end of the section. In section 1.13, we combine all these assumption into one to finally obtain (1.2).

1.3. Errors in an elementary projection. The complete MGS2 algorithm is based on a sequence of elementary projections. In that respect, it is important to fully understand what is happening for each of them. In exact arithmetic, we have the following relations:

$$\mathbf{a}_j^{(k+1)(r)} = \mathbf{a}_j^{(k)(r)} - \mathbf{q}_k r_{kj}^{(r)},$$
$$\mathbf{a}_j^{(k+1)(r)} = (\mathbf{I} - \mathbf{q}_k\mathbf{q}_k^T)\mathbf{a}_j^{(k)(r)},$$
$$\mathbf{q}_k^T\mathbf{a}_j^{(k)(r)} = r_{kj}^{(r)},$$
$$\|\mathbf{a}_j^{(k+1)(r)}\|_2 \leq \|\mathbf{a}_j^{(k)(r)}\|_2.$$

Björck [4], in his error analysis of an elementary projection, gives the equivalent of these four relations in floating-point arithmetic. We recall his results. In this section, the set of indices j for the column, r for the loop, and k for the projection are frozen. Following Björck [4], we assume

$$(1.4) \qquad m \geq 2 \quad \text{and} \quad 2n(m+2)2^{-t_1} < 0.01,$$

where $t_1 = t - \log_2(1.06)$.

If $\bar{\mathbf{q}}_k^{'} \neq \mathbf{0}$, we define the related errors $\delta_j^{(k)(r)}$ and $\eta_j^{(k)(r)}$ by

$$(1.5) \qquad \bar{\mathbf{a}}_j^{(k+1)(r)} = \bar{\mathbf{a}}_j^{(k)(r)} - \bar{\mathbf{q}}_k\bar{r}_{kj}^{(r)} + \delta_j^{(k)(r)},$$
$$(1.6) \qquad \bar{\mathbf{a}}_j^{(k+1)(r)} = (\mathbf{I} - \bar{\mathbf{q}}_k\bar{\mathbf{q}}_k^T)\bar{\mathbf{a}}_j^{(k)(r)} + \eta_j^{(k)(r)}.$$

In the singular situation, that is, when $\bar{\mathbf{q}}_k^{'} = \mathbf{0}$, these relations are satisfied with

$$(1.7) \qquad \bar{\mathbf{a}}_j^{(k+1)(r)} = \bar{\mathbf{a}}_j^{(k)(r)} \quad \text{and} \quad \delta_j^{(k)(r)} = \eta_j^{(k)(r)} = \mathbf{0}.$$

In the nonsingular case, Björck [4] shows that

$$(1.8) \quad \|\delta_j^{(k)(r)}\|_2 \leq 1.45 \cdot 2^{-t}\|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 \quad \text{and} \quad \|\eta_j^{(k)(r)}\|_2 \leq (2m+3) \cdot 2^{-t_1}\|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.$$

The error between $\bar{\mathbf{q}}_k^T\bar{\mathbf{a}}_j^{(k)(r)}$ and the computed value $\bar{r}_{kj}^{(r)}$ is given by

(1.9)
$$|\bar{\mathbf{q}}_k^T\bar{\mathbf{a}}_j^{(k)(r)} - \bar{r}_{kj}^{(r)}| < ((m+1) \cdot |\bar{\mathbf{q}}_k^T\bar{\mathbf{a}}_j^{(k)(r)}| + m\|\bar{\mathbf{a}}_j^{(k)(r)}\|_2)2^{-t_1} \leq (2m+1)2^{-t_1} \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.$$

In exact arithmetic, we have $\mathbf{a}_j^{(k+1)(r)} = (I_m - \mathbf{q}_k\mathbf{q}_k^T)\mathbf{a}_j^{(k)(r)}$, and thus $\|\mathbf{a}_j^{(k+1)(r)}\|_2 \leq \|\mathbf{a}_j^{(k)(r)}\|_2$. In floating-point arithmetic, it can happen that the norm of the vector $\mathbf{a}_j^{(k+1)(r)}$ becomes larger than $\mathbf{a}_j^{(k)(r)}$ due to the rounding errors. It is therefore important to have an upper bound to control $\mathbf{a}_j^{(k+1)(r)}$. After k projections, $k < n$, Björck [4] shows that

$$(1.10) \qquad \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 < 1.006\|\mathbf{a}_j^{(1)(r)}\|_2.$$

The constant 1.006 comes from assumption (1.4). For more details, we refer directly to Björck [4]. Since $1.006^2 < 1.013$,

$$(1.11) \qquad \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 < 1.013\|\mathbf{a}_j\|_2.$$

1.4. Errors in the factorization. We define

$$(1.12) \qquad \mathbf{E} = \bar{\mathbf{Q}}\bar{\mathbf{R}} - \mathbf{A}.$$

We shall prove the inequality

$$(1.13) \qquad \|\mathbf{E}\|_F < 2.94(n-1) \cdot 2^{-t}\|\mathbf{A}\|_F.$$

Summing (1.5) for $k = 1, 2, \ldots, j-1$ and $r = 1, 2$ and using the relations

$$\bar{\mathbf{a}}_j^{(1)(1)} = \mathbf{a}_j, \quad \bar{\mathbf{a}}_j^{(j)(1)} = \bar{\mathbf{a}}_j^{(1)(2)}, \quad \bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{q}}_j\bar{r}_{jj}, \quad \bar{r}_{kj} = \bar{r}_{kj}^{(1)} + \bar{r}_{kj}^{(2)},$$

we get

$$(1.14) \qquad \sum_{k=1}^{j} \bar{\mathbf{q}}_k \cdot \bar{r}_{kj} - \mathbf{a}_j = \sum_{k=1}^{j-1}(\delta_j^{(k)(1)} + \delta_j^{(k)(2)}).$$

Let us define $\delta_j = \sum_{k=1}^{j-1}(\delta_j^{(k)(1)} + \delta_j^{(k)(2)})$. Then, from inequalities (1.8), we have

$$\|\delta_j\|_2 < 1.45 \cdot 2^{-t} \sum_{r=1}^{2} \sum_{k=1}^{j-1} \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.$$

Using both inequality (1.11) and the fact that $1.013 \times 1.45 \times 2 < 2.94$, we have

$$\|\delta_j\|_2 < 2.94 \cdot 2^{-t}(j-1)\|\mathbf{a}_j\|_2.$$

Finally, we obtain

$$\|\mathbf{E}\|_F = \left(\sum_{j=1}^{n} \|\delta_j\|_2^2\right)^{1/2} < 2.94 \cdot 2^{-t}(n-1)\left(\sum_{j=1}^{n} \|\mathbf{a}_j\|_2^2\right)^{1/2} = 2.94(n-1) \cdot 2^{-t}\|\mathbf{A}\|_F.$$

1.5. Nonsingularity of $\bar{\mathbf{A}}$. From (1.12), a sufficient condition for $\bar{\mathbf{A}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}$ to be full rank is given by Björck [4]. If the exact factorization of $\mathbf{A}$ is $\mathbf{A} = \mathbf{Q}\mathbf{R}$, then $\bar{\mathbf{A}}$ has rank n if

$$(1.15) \qquad 2.94(n-1) \cdot 2^{-t}\|\mathbf{A}\|_F\|\mathbf{R}^{-1}\|_2 \leq \sqrt{2} - 1.$$

We assume in the following that inequality (1.15) is satisfied. This ensures that, for all $r = 1, 2$ and for all $j = 1, \ldots, n$,

$$\|\bar{\mathbf{a}}_j^{(j)(r)}\|_2 \neq 0.$$

1.6. Theorem of Pythagoras. The purpose of this section is to exhibit an upper bound for

$$(1.16) \qquad \sqrt{\sum_{i=1}^{j-1} (\bar{r}_{ij}^{(r)})^2}$$

that will be used later in sections 1.9, 1.10, and 1.11. In what follows, we are interested in each step r individually. Therefore, for the sake of readability, we no longer use the superscript $^{(r)}$ to label the index loop.

In exact arithmetic, after the jth step of the MGS algorithm, we have

$$\mathbf{a}_j = \sum_{k=1}^{j-1} (\mathbf{q}_k r_{kj}) + \mathbf{a}_j^{(j)},$$

and as the vectors $\mathbf{q}_k, k = 1, \ldots, j-1$, are orthonormal, we have

$$(1.17) \qquad \sum_{k=1}^{j-1} (r_{kj})^2 + \|\mathbf{a}_j^{(j)}\|_2^2 = \|\mathbf{a}_j\|_2^2.$$

Equation (1.17) is nothing but the theorem of Pythagoras. Still in exact arithmetic, let $\mathbf{Q}_{j-1}$ be such that $\|\mathbf{q}_k\|_2 = 1, k = 1, \ldots, j-1$, without any additional assumption. Then, from the column $\mathbf{a}_j$ running step j of the MGS algorithm, we get

$$\mathbf{a}_j^{(1)} = (\mathbf{I} - \mathbf{q}_1 \mathbf{q}_1^T)\mathbf{a}_j, \qquad \text{with } r_{1j} = \mathbf{q}_1^T \mathbf{a}_j \qquad \Rightarrow \|\mathbf{a}_j\|_2^2 = (r_{1j})^2 + \|\mathbf{a}_j^{(1)}\|_2^2,$$

$$\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots$$

$$\mathbf{a}_j^{(j)} = (\mathbf{I} - \mathbf{q}_{j-1} \mathbf{q}_{j-1}^T)\mathbf{a}_j^{(j-1)}, \quad \text{with } r_{j-1,j} = \mathbf{q}_{j-1}^T \mathbf{a}_j^{(j-1)} \Rightarrow \|\mathbf{a}_j^{(j-1)}\|_2^2 = (r_{j-1,j})^2 + \|\mathbf{a}_j^{(j)}\|_2^2,$$

$$\Rightarrow \overline{\|\mathbf{a}_j\|_2^2 = \sum_{k=1}^{j-1} (r_{kj})^2 + \|\mathbf{a}_j^{(j)}\|_2^2}.$$

We recover property (1.17). Therefore we have the following statement: *When step j of MGS is performed in exact arithmetic with $\|\mathbf{q}_k\|_2 = 1, k = 1, \ldots, j-1$, property (1.17) is true.* We apply the same idea in floating-point calculations. From (1.5),

$$\bar{\mathbf{a}}_j^{(k+1)} = \bar{\mathbf{a}}_j^{(k)} - \bar{\mathbf{q}}_k \bar{r}_{kj} + \delta_j^{(k)},$$

$$\Rightarrow \quad \bar{\mathbf{a}}_j^{(k)} + \delta_j^{(k)} = \bar{\mathbf{a}}_j^{(k+1)} + \bar{\mathbf{q}}_k \bar{r}_{kj},$$

$$(1.18) \qquad \Rightarrow \|\bar{\mathbf{a}}_j^{(k)}\|_2^2 + \alpha_j^{(k)} = \|\bar{\mathbf{a}}_j^{(k+1)}\|_2^2 + (\bar{r}_{kj})^2,$$

where

$$\alpha_j^{(k)} = (\delta_j^{(k)})^T \delta_j^{(k)} + 2(\delta_j^{(k)})^T \bar{\mathbf{a}}_j^{(k)} - 2\bar{r}_{kj} (\bar{\mathbf{q}}_k)^T \bar{\mathbf{a}}_j^{(k+1)}.$$

Therefore we can get the following upper bound for $|\alpha_j^{(k)}|$:

$$(1.19) \qquad |\alpha_j^{(k)}| \le \|\delta_j^{(k)}\|_2^2 + 2\|\delta_j^{(k)}\|_2 \|\bar{\mathbf{a}}_j^{(k)}\|_2 + 2|\bar{r}_{kj}||\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k+1)}|.$$

From (1.6) it follows that

$$(1.20) \qquad (\bar{\mathbf{q}}_k)^T \bar{\mathbf{a}}_j^{(k+1)} = (\bar{\mathbf{q}}_k)^T \eta_j^{(k)},$$

and therefore

$$(1.21) \qquad |\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k+1)}| \leq \|\eta_j^{(k)}\|_2.$$

For $|\bar{r}_{kj}|$, (1.9) gives

$$(1.22) \qquad |\bar{r}_{kj}| \leq (1 + (2m+1)2^{-t_1}) \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 \leq 1.01 \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.$$

Using (1.4), (1.8), (1.10), (1.21), and (1.22) in inequality (1.19), we get

$$(1.23)$$
$$\begin{aligned}
|\alpha_j^{(k)}| &\leq (1.006)^2 \times [1.45^2 \times 2^{-t} + 2 \times 1.45 + 2 \times 1.06 \times 1.01 \times (2m+3)] \cdot 2^{-t} \|\bar{\mathbf{a}}_j\|_2^2 \\
&\leq (4.34m + 9.33)2^{-t} \cdot \|\mathbf{a}_j\|_2^2,
\end{aligned}$$

where we use inequality (1.4) to bound 2^{-t} with 0.0016.

Summing equality (1.18) for $k = 1, \ldots, j-1$ gives

$$\|\mathbf{a}_j\|_2^2 + \sum_{k=1}^{j-1} \alpha_j^{(k)} = \|\bar{\mathbf{a}}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{r}_{kj})^2,$$

and then using inequality (1.23), we obtain

$$\left| \left(\|\bar{\mathbf{a}}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{r}_{kj})^2 \right) - \|\mathbf{a}_j\|_2^2 \right| \leq (4.34m + 9.33)(j-1)2^{-t_1} \cdot \|\mathbf{a}_j\|_2^2.$$

Using the fact that $\sqrt{1+x} \leq 1 + x/2$ for all $x \geq -1$, we have

$$(1.24) \qquad \sqrt{\|\bar{\mathbf{a}}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{r}_{kj})^2} \leq [1 + (2.17m + 4.67)(j-1)2^{-t_1}] \cdot \|\mathbf{a}_j\|_2.$$

Let us assume that

$$(1.25) \qquad (2.04m + 4.43)(j-1)2^{-t_1} \leq 0.01;$$

then we get

$$(1.26) \qquad \sqrt{\|\bar{\mathbf{a}}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{r}_{kj})^2} \leq 1.01 \cdot \|\mathbf{a}_j\|_2.$$

We remark that (1.26) and assumption (1.25) are satisfied without any assumption on the orthogonality of the columns of $\bar{\mathbf{Q}}_{j-1}$.

1.7. Condition number of A and maximum value of $K_j^{(1)} = \|\mathbf{a}_j\|_2 / \|\bar{a}_j^{(j)(1)}\|_2$ for $j = 1, \ldots, n$. We define

$$(1.27) \qquad K_j^{(1)} = \frac{\|\mathbf{a}_j\|_2}{\|\bar{a}_j^{(j)(1)}\|_2} \quad \text{and} \quad K_j^{(2)} = \frac{\bar{\mathbf{a}}_j^{(1)(2)}}{\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2}.$$

Notice that $\|\bar{a}_j^{(j)(1)}\|_2 \neq 0$ and $\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \neq 0$ because we make the assumption (1.15) on the numerical nonsingularity of $\mathbf{A}$. We have seen in the introduction that the quantity $K_j^{(1)}$ plays an important role for checking the quality of the orthogonality for the computed vector $\bar{\mathbf{q}}_j$ with respect to the previous $\bar{\mathbf{q}}_i$, $i = 1, \ldots, n$. In this section, we derive an upper bound for $K_j^{(1)}$.

In exact arithmetic, if MGS is run on $\mathbf{A}$ to obtain the QR-factors $\mathbf{Q}$ and $\mathbf{R}$, then

$$\sigma_{\min}(\mathbf{A}) = \sigma_{\min}(\mathbf{R}) \leq |r_{jj}| = \|\mathbf{a}_j^{(j)(1)}\|_2 \quad \text{and} \quad \|\mathbf{A}\|_2 \geq \|\mathbf{a}_j\|_2;$$

thus

$$(1.28) \qquad K_j^{(1)} = \frac{\|\mathbf{a}_j\|_2}{\|\mathbf{a}_j^{(j)(1)}\|_2} \leq \kappa(\mathbf{A}).$$

Inequality (1.28) indicates that, in exact arithmetic, $K_j^{(1)}$ is always less than the condition number of $\mathbf{A}$, $\kappa_2(\mathbf{A})$. With rounding errors, we can establish a bound similar to inequality (1.28).

We recall (1.14), that is,

$$\mathbf{a}_k = \sum_{i=1}^{k} \bar{\mathbf{q}}_k \cdot \bar{r}_{ik} - \delta_k, \quad k = 1, \ldots, j-1.$$

For $k = j$, we consider only the first loop (i.e., $r = 1$). This gives

$$\mathbf{a}_j = \sum_{i=1}^{j} \bar{\mathbf{q}}_i \cdot \bar{r}_{i,j}^{(1)} + \bar{\mathbf{a}}_j^{(j)(1)} - \delta_j^{(1)}$$

with $\delta_j^{(1)} = \sum_{k=1}^{j-1} \delta_j^{(k)(1)}$. In matrix form, this can be written as

$$\mathbf{A}_j = \bar{\mathbf{Q}}_{j-1} \bar{\mathbf{R}}_{(j-1,j)} - \boldsymbol{\Delta}_j$$

with

$$\bar{\mathbf{Q}}_{j-1} \in \mathbb{R}^{m \times j-1}, \quad \bar{\mathbf{Q}}_{j-1} = [\bar{\mathbf{q}}_1, \ldots, \bar{\mathbf{q}}_{j-1}], \quad \text{and} \quad \bar{\mathbf{R}}_{(j-1,j)} \in \mathbb{R}^{j-1 \times j}$$

such that

$$\bar{\mathbf{R}}_{(j-1,j)} = \begin{pmatrix} \bar{r}_{1,1} & \cdots & \bar{r}_{1,j-1} & \bar{r}_{1,j}^{(1)} \\ & \ddots & \vdots & \vdots \\ & & \bar{r}_{j-1,j-1} & \bar{r}_{j-1,j}^{(1)} \end{pmatrix}.$$

Finally, $\boldsymbol{\Delta}_j \in \mathbb{R}^{m \times j}$ is defined by

$$\boldsymbol{\Delta}_j = [\delta_1, \ldots, \delta_{j-1}, \delta_j^{(1)} - \bar{\mathbf{a}}_j^{(j)(1)}]$$

with

$$0 < \|\boldsymbol{\Delta}_j\|_F \leq 2.94(j-1) \cdot 2^{-t} \|\mathbf{A}_j\|_F + \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2.$$

Notice that, by construction, the matrix $\bar{\mathbf{Q}}_{j-1} \bar{\mathbf{R}}_{(j-1,j)}$ is of rank $j-1$. Therefore the matrix $\mathbf{A}_j + \boldsymbol{\Delta}_j$ is singular, whereas we assume that the matrix $\mathbf{A}_j$ is nonsingular. The

distance to singularity for a matrix $\mathbf{A}_j$ can be related to its minimum singular value. Some theorems on relative distance to singularity can be found in many textbooks (e.g., [14, p. 73] and [15, p. 111]). Although the textbooks usually assume that the matrices are square, this statement is also true for rectangular matrices. In our case, we have

$$\sigma_{\min}(\mathbf{A}_j) = \min\{\|\mathbf{\Delta}\|_2, \mathbf{\Delta} \in \mathbb{R}^{m \times j} \text{ so that } \mathbf{A}_j + \mathbf{\Delta} \text{ is singular } \} \le \|\mathbf{\Delta}_j\|_2.$$

Dividing by $\|\mathbf{A}_j\|_2$, we get

$$\frac{1}{\kappa_2(\mathbf{A}_j)} \le \frac{\|\mathbf{\Delta}_j\|_2}{\|\mathbf{A}_j\|_2} \le \frac{\|\mathbf{\Delta}_j\|_F}{\|\mathbf{A}_j\|_2},$$

and since we know that $\|\mathbf{\Delta}_j\|_F \ne 0$, this gives

$$\kappa_2(\mathbf{A}) \ge \kappa_2(\mathbf{A}_j) \ge \frac{\|\mathbf{A}_j\|_2}{\|\mathbf{\Delta}_j\|_F} \ge \frac{1}{2.94(n-1) \cdot 2^{-t} \frac{\|\mathbf{A}_j\|_F}{\|\mathbf{A}_j\|_2} + \frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{\|\mathbf{A}_j\|_2}};$$

however,

$$\frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{\|\mathbf{a}_j\|_2} = K_j^{(1)}, \qquad \frac{\|\mathbf{a}_j\|_2}{\|\mathbf{A}_j\|_2} \le 1, \quad \text{and} \quad \frac{\|\mathbf{A}_j\|_F}{\|\mathbf{A}_j\|_2} < \sqrt{j},$$

and therefore

$$\kappa_2(\mathbf{A}) \ge \frac{1}{2.94(j-1)\sqrt{j} \cdot 2^{-t} + \frac{1}{K_j^{(1)}}}.$$

For instance, if we assume that

(1.29) $$2.94(n-1)n^{1/2} \cdot 2^{-t}.\kappa_2(\mathbf{A}) < 0.09,$$

where the value 0.09 is taken arbitrarily but another value leads to a final similar result, we have the inequality

$$K_j^{(1)} \le \frac{1}{1 - 2.94(j-1)j^{1/2}2^{-t} \cdot \kappa_2(\mathbf{A})}\kappa_2(\mathbf{A}).$$

Using assumption (1.29) we get

(1.30) $$K_j^{(1)} \le 1.1 \cdot \kappa_2(\mathbf{A}).$$

We remark that (1.30) and assumption (1.29) are independent of the orthogonality of the previously computed $\bar{\mathbf{Q}}_{j-1}$; it is just a consequence of (1.12).

Note that the value 0.09 of the right-hand side in (1.29) is arbitrary. We point out that since the numerical properties of the Gram–Schmidt algorithm are invariant under column scaling (without consideration of underflow), instead of the condition number $\kappa(\mathbf{A})$ one can use

$$\kappa_D(\mathbf{A}) = \min_{\mathbf{D} \text{ diagonal matrix}} \kappa(\mathbf{A}\mathbf{D}).$$

1.8. Induction assumption. We want to show that the orthogonality of the computed vectors $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \ldots, \bar{\mathbf{q}}_n$ is of the order of machine precision.

In exact arithmetic, at step j, to show that the vector $\mathbf{q}_j$ generated by the MGS algorithm is orthogonal to the previous ones, we use the fact that the previous $\mathbf{q}_i$, $i = 1, \ldots, j-1$, are already orthogonal to each other. Therefore to show the orthogonality at step j in floating-point arithmetic, we make an assumption on the orthogonality at step $j-1$.

The orthogonality of the computed vectors $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \ldots, \bar{\mathbf{q}}_n$, can be measured by the norm of the matrix $(\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}})$. Let $\mathbf{U}_p$, $p = 1, \ldots, n$, be the strictly upper triangular matrix of size (p, p) with entries

$$u_{ij} = \bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_j, \quad 1 \leq i < j \leq p \quad \text{and} \quad u_{ij} = 0, \quad 1 \leq j \leq i \leq p.$$

We note $\mathbf{U} = \mathbf{U}_n$ and have

$$(1.31) \qquad \mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}} = -(\mathbf{U} + \mathbf{U}^T).$$

We construct a proof by induction to show that $\|\mathbf{U}\|_2$ is small at step n. Therefore, we assume that at step $p - 1$,

$$(1.32) \qquad \|\mathbf{U}_{p-1}\|_2 \leq \lambda.$$

Our aim is to show that at step p, we still have $\|\mathbf{U}_p\|_2 \leq \lambda$. The value of λ is exhibited during the proof.

In the following, the index variables i, j, k, and p are such that

$$1 \leq j \leq p \leq n, \quad 1 \leq i \leq j, \quad \text{and} \quad 1 \leq k \leq j.$$

1.9. Bound for $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}|$ for $k = 1, \ldots, j-1$ and $j = 1, \ldots, p$. The expression $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}|$ represents the orthogonality between $\bar{\mathbf{q}}_k$, $k = 1, \ldots, j-1$, and the vector $\bar{\mathbf{a}}_j^{(j)(1)}$ given by the first step of MGS ($r = 1$). In exact arithmetic, this quantity is zero. Following Björck [4], we sum (1.5) for $i = k+1, k+2, \ldots, j-1$ and $r = 1$ to get

$$\bar{\mathbf{a}}_j^{(j)(1)} = \bar{\mathbf{a}}_j^{(k+1)(1)} - \sum_{i=k+1}^{j-1} \bar{\mathbf{q}}_i \bar{r}_{ij}^{(1)} + \sum_{i=k+1}^{j-1} \delta_j^{(i)(1)}.$$

Hence, multiplying this relation by $\bar{\mathbf{q}}_k^T$ and using (1.20), we get

$$\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)} = - \sum_{i=k+1}^{j-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i) \bar{r}_{ij}^{(1)} + \bar{\mathbf{q}}_k^T \left(\bar{\eta}_j^{(k)(1)} + \sum_{i=k+1}^{j-1} \delta_j^{(i)(1)} \right).$$

Therefore,

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| \leq \sqrt{\sum_{i=k+1}^{j-1} (\bar{r}_{ij}^{(1)})^2} \sqrt{\sum_{i=k+1}^{j-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i)^2} + \|\bar{\eta}_j^{(k)(1)}\|_2 + \sum_{i=k+1}^{j-1} \|\delta_j^{(i)(1)}\|_2.$$

We can interpret the terms of the right-hand side as follows.

1. The orthogonalization of $\bar{\mathbf{a}}_j^{(k)(1)}$ against $\bar{\mathbf{q}}_k$ is not performed exactly; this corresponds to the second term.

2. The resulting vector $\bar{\mathbf{a}}_j^{(k+1)(1)}$ is orthogonalized on $\bar{\mathbf{q}}_i$, $i = k+1, \ldots, j-1$, and, since $\bar{\mathbf{Q}}$ is not orthogonal, we also lose orthogonality here; this corresponds to the first term.

3. Moreover, all projections $i = k + 1, \ldots, j - 1$ are also done inaccurately; this corresponds to the third term.

Using inequalities (1.8) and (1.10), we have

$$\|\bar{\eta}_j^{(k)(1)}\|_2 + \sum_{i=k+1}^{j-1} \|\delta_j^{(i)(1)}\|_2 \leq (2.14m + 3.20 + 1.46(j - k - 1))2^{-t} \cdot \|\mathbf{a}_j\|_2.$$

Finally, using inequalities (1.26) and (1.32), we get

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| \leq [1.01\lambda + (2.14m + 1.46(j - k - 1) + 3.20)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

1.10. Bounds for $|\bar{r}_{kj}^{(2)}|$, $k = 1, \ldots, j - 1$, and $j = 1, \ldots, p$. Having a bound for the orthogonality of the first step, we now study its influence in the second step by computing $|\bar{r}_{kj}^{(2)}|$. Again summing (1.5) for $i = 1, 2, \ldots, k - 1$ and $r = 2$ we get

$$\bar{\mathbf{a}}_j^{(k)(2)} = \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{i=1}^{k-1} \bar{\mathbf{q}}_i \bar{r}_{ij}^{(2)} + \sum_{i=1}^{k-1} \delta_j^{(i)(2)}.$$

Hence multiplying by $\bar{\mathbf{q}}_k^T$, we get

$$\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)} = \bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{i=1}^{k-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i) \bar{r}_{ij}^{(2)} + \bar{\mathbf{q}}_k^T \sum_{i=1}^{k-1} \delta_j^{(i)(2)}.$$

Taking moduli, we have

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)}| \leq |\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| + \sqrt{\sum_{i=1}^{k-1} (\bar{r}_{ij}^{(2)})^2} \sqrt{\sum_{i=1}^{k-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i)^2} + \sum_{i=1}^{k-1} \|\delta_j^{(i)(2)}\|_2.$$

Similarly as in section 1.9, we bound each term in the right-hand side and get

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)}| \leq [2.02\lambda + (2.14m + 1.46(j - 2) + 3.20)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

Using inequalities (1.9) and (1.10), we know that $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)} - \bar{r}_{kj}^{(2)}| \leq (2.15m + 1.08) \cdot 2^{-t}\|\mathbf{a}_j\|_2$; therefore

$$|\bar{r}_{kj}^{(2)}| \leq [2.02\lambda + (4.29m + 1.46(j - 2) + 4.28)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

This expression can be simplified to obtain

(1.33) $$|\bar{r}_{kj}^{(2)}| \leq [2.02\lambda + 5.75(m + 1)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

1.11. Bound for $K_j^{(2)} = \|\bar{\mathbf{a}}_j^{(1)(2)}\|_2/\|\bar{a}_j^{(j)(2)}\|_2,\ j = 1,\ldots,p$. While the quantity $K_j^{(1)}$ is important for the level of orthogonality after the first orthogonalization loop, the quantity $K_j^{(2)}$ is important for the level of orthogonality after the second orthogonalization loop. In exact arithmetic, we have $\mathbf{a}_j^{(1)(2)} = \mathbf{a}_j^{(j)(2)}$ and therefore $K_j^{(2)} = 1$. In this section, we show that $K_j^{(2)}$ in floating-point arithmetic is close to one.

Let us again sum (1.5) for $r = 2$, $k = 1, 2, \ldots, j - 1$, to get

$$\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)} + \sum_{k=1}^{j-1} \delta_j^{(k)(2)};$$

then

$$(1.34) \qquad \|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \geq \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2 - \left\|\sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)}\right\|_2 - \sum_{k=1}^{j-1} \|\delta_j^{(k)(2)}\|_2.$$

The induction assumption (1.31) implies that $\|\bar{\mathbf{Q}}\|_2 \leq \sqrt{1 + \lambda^2}$. From this, we can get an upper bound for $\|\sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)}\|_2$; that is,

$$\left\|\sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)}\right\|_2 \leq \|\bar{\mathbf{Q}}\|_2 \left\|\begin{pmatrix} \bar{r}_{1j}^{(2)} \\ \vdots \\ \bar{r}_{j-1,j}^{(2)} \end{pmatrix}\right\|_2 \leq \sqrt{1 + \lambda^2} \left\|\begin{pmatrix} \bar{r}_{1j}^{(2)} \\ \vdots \\ \bar{r}_{j-1,j}^{(2)} \end{pmatrix}\right\|_2.$$

Using inequality (1.33) we get

$$\left\|\sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)}\right\|_2 \leq \sqrt{1 + \lambda^2} \cdot \sqrt{j - 1}[2.02\lambda + 5.75(m + 1)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

With inequalities (1.8) and (1.34) we have

$$\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \geq \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2 - \left[\sqrt{1 + \lambda^2} \cdot \sqrt{j - 1}(2.02\lambda + 5.75(m + 1)2^{-t}) \right.$$
$$\left. + 1.47(j - 1)2^{-t}\right] \|\mathbf{a}_j\|_2.$$

Dividing by $\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2$ we have

$$1/K_j^{(2)} \geq 1 - \left[\sqrt{1 + \lambda^2} \cdot \sqrt{j - 1}[2.02\lambda + 5.75(m + 1)2^{-t}] - 1.47(j - 1)2^{-t}\right] K_j^{(1)}.$$

Let us assume that

(1.35)

$$1.1\kappa_2(\mathbf{A}) \left[\sqrt{1 + \lambda^2} \cdot \sqrt{j - 1}[2.02\lambda + 5.75(m + 1)2^{-t}] + 1.47(j - 1)2^{-t}\right] \leq 0.67 < 1,$$

where the value 0.67 is taken arbitrarily, but another value leads to a final, similar result. We obtain

$$K_j^{(2)} \leq \frac{1}{1 - K_j^{(1)}\sqrt{1 + \lambda^2} \cdot \sqrt{j - 1}[2.02\lambda + 5.75(m + 1)2^{-t}]} \leq \frac{1}{0.67}.$$

This gives

$$(1.36) \qquad\qquad K_j^{(2)} \le 1.5.$$

We remark that assumption (1.35) is dependent on the parameter λ that is still not yet known.

1.12. Bound for the orthogonality of the vectors. Summing (1.5) from $k = i+1, i+2, \dots, j-1$ and $r = 2$ we get

$$(1.37) \qquad \bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{a}}_j^{(i+1)(2)} - \sum_{k=i+1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)}.$$

From (1.20), we have $\bar{\mathbf{q}}_i^T \bar{\mathbf{a}}_j^{(i+1)(2)} = \bar{\mathbf{q}}_i^T \eta_j^{(i)(2)}$ and $\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{q}}_j \bar{r}_{jj}^{(2)}$. Therefore multiplying (1.37) by $\bar{\mathbf{q}}_i^T$ we get

$$\sum_{k=i+1}^{j} \bar{r}_{kj}^{(2)} (\bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_k) = \bar{\mathbf{q}}_i^T \left(\eta_j^{(i)(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)} \right).$$

We divide this by $|\bar{r}_{jj}^{(2)}|$ (which is different from 0) to get

$$(1.38) \qquad \sum_{k=i+1}^{j} \frac{\bar{r}_{kj}^{(2)}}{|\bar{r}_{jj}^{(2)}|} (\bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_k) = \frac{\bar{\mathbf{q}}_i^T (\eta_j^{(i)(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)})}{|\bar{r}_{jj}^{(2)}|}.$$

We recall that this equality is true for all $j = 1, \dots, p$ and $i = 1, \dots, j-1$.

Define $\mathbf{M}_p$ as the unit upper triangular matrix with the (k,j) entry, m_{kj}, given by

$$(1.39) \qquad\qquad m_{kj} = \frac{\bar{r}_{kj}^{(2)}}{|\bar{r}_{jj}^{(2)}|} \quad \text{for } k < j,$$

and let $\mathbf{S}_p$ be the strictly upper triangular matrix, where the (i,j) entry, s_{ij}, is

$$s_{ij} = \frac{\bar{\mathbf{q}}_i^T (\eta_j^{(i)(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)})}{|\bar{r}_{jj}^{(2)}|} \quad \text{for } i < j.$$

Since the entry (i,k) of $\mathbf{U}_p$ is $u_{ik} = \bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_k$, (1.38) can be rewritten as

$$\forall j = 1, \dots, p, \quad \forall i = 1, \dots, j-1, \qquad s_{ij} = \sum_{k=i+1}^{j} u_{ik} m_{kj}.$$

Taking into account the facts that $\mathbf{U}_p$ and $\mathbf{S}_p$ are strictly upper triangular and $\mathbf{M}_p$ is upper triangular, we obtain

$$(1.40) \qquad\qquad \mathbf{S}_p = \mathbf{U}_p \mathbf{M}_p.$$

In [4], Björck gives an upper bound for the 2-norm of each column of $\mathbf{S}_p$ as

$$\|\mathbf{s}_j\|_2 \le 0.87 \cdot n^{1/2} (n + 1 + 2.5m) 2^{-t} \frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{|\bar{r}_{jj}^{(2)}|}.$$

Since $|\bar{r}_{jj}^{(2)}| = \|\bar{\mathbf{a}}_j^{(j)(2)}\|_2$, we obtain

$$\|\mathbf{s}_j\|_2 \leq 0.87 K_j^{(2)} \cdot n^{1/2}(n + 1 + 2.5m)2^{-t}.$$

Using inequality (1.36) and the fact that $0.87 \times 1.5 = 1.305$, we get

(1.41) $$\|\mathbf{S}_p\|_2 \leq 1.305n(n + 1 + 2.5m)2^{-t}.$$

$\mathbf{M}_p$ is nonsingular. Therefore from (1.40) we have

(1.42) $$\|\mathbf{U}_p\|_2 \leq \|\mathbf{M}_p^{-1}\|_2\|\mathbf{S}_p\|_2.$$

At this stage, the quantity of interest is $\|\mathbf{M}_p^{-1}\|_2$.

It is interesting to relate this proof to that of Björck [4], who shows an inequality similar to inequality (1.42) for MGS, with $\|\mathbf{S}_p\|_2$ of the order of machine precision, $\mathbf{U}_p$ as defined in section 1.8, but with $\bar{\mathbf{q}}$ coming from MGS and $\mathbf{M}_p$ as defined in (1.39) but with $\bar{r}_{kj}$ coming from MGS. Since he proves that, for MGS, $\|\mathbf{M}_p^{-1}\|_2$ is of the order of $\kappa(\mathbf{A})$, he obtains the result that the final orthogonality obtained with MGS is of the order of $\kappa(\mathbf{A})2^{-t}$. Our goal is to show that $\|\mathbf{M}_p^{-1}\|_2$ is independent of $\kappa(\mathbf{A})$ and is of the order of 1.

An idea for controlling the 2-norm of $\mathbf{M}_p$ is to show that $\mathbf{M}_p$ is diagonally dominant by columns. Following Varah [6], we say that $\mathbf{M}_p$ is diagonally dominant by columns if

(1.43) $$\forall j = 1, \ldots, n, \quad |m_{jj}| > \sum_{k \neq j} |m_{kj}|.$$

In our case, since $\mathbf{M}_p$ is unit triangular it would be diagonally dominant by columns if

$$\forall j = 1, \ldots, n, \quad 1 > \sum_{k=1}^{j-1} |m_{kj}|.$$

It then becomes natural to look for an upper bound for $\sum_{k=1}^{j-1} |m_{kj}|$ that is lower than 1.

From (1.33) we have

$$\sum_{k=1}^{j-1} |m_{kj}| \leq (j-1)[2.02\lambda + 5.75(m+1)2^{-t}]\frac{\|\mathbf{a}_j\|_2}{|\bar{r}_{jj}^{(2)}|}.$$

Therefore

$$\sum_{k=1}^{j-1} |m_{kj}| \leq (j-1)[2.02\lambda + 5.75(m+1)2^{-t}]K_j^{(1)}K_j^{(2)}.$$

Using (1.30) and (1.36), we get as $1.1 \times 1.5 = 1.65$,

$$\sum_{k=1}^{j-1} |m_{kj}| \leq 1.65(j-1)[2.02\lambda + 5.75(m+1)2^{-t}]\kappa_2(A).$$

We assume that

(1.44) $$1.65(n-1)[2.02\lambda + 5.75(m+1)2^{-t}]\kappa_2(A) \leq L,$$

where L is a real number such that $0 < L < 1$. With inequality (1.44), we obtain

$$(1.45) \qquad \sum_{k=1}^{j-1} |m_{kj}| \le L.$$

This means that $\mathbf{M}_p$ is diagonally dominant by columns.

Let us decompose $\mathbf{M}_p$ as

$$\mathbf{M}_p = \mathbf{I}_p + \mathbf{C}_p,$$

where $\mathbf{C}_p$ is strictly upper triangular. Inequality (1.45) means that

$$(1.46) \qquad \|\mathbf{C}_p\|_1 = \max_{j=1,\dots,p} \sum_{k=1}^{j-1} |m_{kj}| \le L.$$

In addition, we also have

$$(\mathbf{I}_p + \mathbf{C}_p)(\mathbf{I}_p - \mathbf{C}_p + \cdots + (-1)^n \mathbf{C}_p^{n-1}) = \mathbf{I}_p + (-1)^n \mathbf{C}_p^n.$$

Since $\mathbf{C}_p$ is strictly upper triangular, it is nilpotent (i.e., we have $\mathbf{C}_p^n = 0$) so that

$$\mathbf{M}_p(\mathbf{I}_p - \mathbf{C}_p + \cdots + (-1)^n \mathbf{C}_p^{n-1}) = \mathbf{I}_p.$$

Therefore

$$\mathbf{M}_p^{-1} = \mathbf{I}_p - \mathbf{C}_p + \cdots + (-1)^n \mathbf{C}_p^{n-1}.$$

In norm this implies that

$$\|\mathbf{M}_p^{-1}\|_2 \le 1 + \|\mathbf{C}_p\|_1 + \|\mathbf{C}_p\|_1^2 + \cdots + \|\mathbf{C}_p\|_1^{n-1}$$
$$\le 1 + L + L^2 + \cdots + L^{n-1} = \frac{1 - L^n}{1 - L}.$$

Finally we get

$$(1.47) \qquad \|\mathbf{M}_p^{-1}\|_1 \le \frac{1}{1 - L},$$

which implies that

$$(1.48) \qquad \|\mathbf{M}_p^{-1}\|_2 \le \frac{\sqrt{n}}{1 - L}.$$

Notice that inequality (1.47) is nothing other than the result of Corollary 1 of Varah [6] applied to matrices with unit diagonal. The parameter L has to be chosen between 0 and 1. It should be neither too close to 0, so that assumption (1.44) does not become too strong, nor too close to 1, so that the bound (1.48) on $\|M_p^{-1}\|_1$ does not become too large. With inequalities (1.41), (1.42), and (1.48), we get

$$(1.49) \qquad \|\mathbf{U}_p\|_2 \le \frac{1.305}{1 - L} \cdot n^{3/2}(n + 1 + 2.5m)2^{-t}.$$

A natural choice for λ is then

$$(1.50) \qquad \lambda = \frac{1.305}{1 - L} \cdot n^{3/2}(n + 1 + 2.5m)2^{-t},$$

so that the induction assumption (1.32) is verified at step p.

1.13. Assumptions on A. Since λ is defined, it is possible to explicitly state the assumptions made on **A**. The assumptions made are (1.4), (1.15), (1.25), (1.29), (1.35), and (1.44). We focus here on the main assumption, that is, (1.44). We replace λ by its value and get

$$\frac{1}{L} \times 1.65(n-1) \left[2.02 \frac{1.305}{1-L} \times n^{3/2}(n+1+2.5m) + 5.75(m+1) \right] 2^{-t} \cdot \kappa_2(\mathbf{A}) \leq 1.$$

For the sake of simplicity we replace it with

$$\frac{1}{L(1-L)} \times 10 n^{5/2}(4.5m+2) 2^{-t} \cdot \kappa_2(\mathbf{A}) \leq 1.$$

1.14. Conclusion of the proof by induction. We have shown that, if we assume (1.2) and define λ with (1.50), then, if at step $(p-1)$ we have $\|\mathbf{U}_{p-1}\|_2 \leq \lambda$, then at step p we also have $\|\mathbf{U}_p\|_2 \leq \lambda$. At step $n = 1$, $\mathbf{U}_1$ is defined as $\|\mathbf{U}_1\|_2 = 0$ and thus $\|\mathbf{U}_1\|_2 \leq \lambda$. From this, we conclude that at step n, we have

$$\|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \frac{2.61}{1-L} \cdot n(n+1+2.5m) 2^{-t}.$$

This completes the proof of Theorem 1.1.

Theorem 1.1 involves a parameter L while MGS2 is parameter free. We can nevertheless use the result of that theorem to assess the quality of the orthogonality of the set of vectors generated by MGS2 by setting $L = 0.5$. The value 0.5 is chosen to relax to the maximum the assumption (1.2) on the nonsingularity of **A**.

2. Link with selective reorthogonalization.

2.1. Sufficiency of the condition $L < 1$ for robust reorthogonalization. The key property of the matrix **M** is given by inequality (1.45). The main effort in the proof given in section 1 consists of showing that for all $j = 1, \ldots, n$ we have, after the reorthogonalization loop,

$$L_j^{(2)} = \sum_{k=1}^{j-1} \frac{|r_{kj}^{(2)}|}{r_{jj}^{(2)}} \leq L < 1.$$

However, this property may already occur after the first orthogonalization, that is,

$$(2.1) \qquad L_j^{(1)} = \sum_{k=1}^{j-1} \frac{|r_{kj}^{(1)}|}{\|\mathbf{a}_j^{(1)}\|_2} \leq L < 1.$$

In this case, we do not need to reorthogonalize $\mathbf{a}_j^{(1)}$ to comply with inequality (1.45) at the second loop since it is already satisfied at the first loop. From this, we propose a new algorithm that checks whether or not inequality (2.1) is satisfied at step j, $r = 1$. We call the resulting criterion the L-criterion and the corresponding algorithm MGS2(L). MGS2(L) is the same as the MGS2(K) algorithm except that line 7 is replaced by

$$\text{if} \quad \frac{\sum_{k=1}^{j-1} |r_{kj}^{(1)}|}{\|\mathbf{a}_j^{(1)}\|_2} \leq L \quad \text{then.}$$

Since we have derived MGS2 without square roots from MGS2, we derive MGS2(L) without square roots from MGS2(L). The proof established in section 1 for MGS2 without square roots basically needs inequality (1.13) to be satisfied and $\|\mathbf{U}_p\|_2 \leq \lambda$ assuming $\|\mathbf{U}_{p-1}\|_2 \leq \lambda$, $p \geq 1$. Whether one loop or two are performed, inequality (1.13) holds. If the L-criterion is satisfied at step p for the first loop, then we can state that $\|\mathbf{U}_p\|_2 \leq \lambda$. If not, at the second loop, we have $\|\mathbf{U}_p\|_2 \leq \lambda$. Therefore Theorem 1.1 holds also for MGS2(L) without square roots. We recall that Theorem 1.1 is true for $0 < L < 1$.

From the point of view of Theorem 1.1, the optimal value of L having the weaker assumption on $\mathbf{A}$ is 0.5. With respect to the orthogonality, the lower L is, the better the orthogonality. To minimize the computational cost of the algorithm, a large value of L would imply performing only a few reorthogonalizations. Therefore, in Theorem 1.1, the value for L between 0 and 1 is a trade-off between the computational cost and the expected orthogonality quality. In our experiments, we choose the value $L = 0.99$.

2.2. Necessity of the condition $L < 1$ to ensure the robustness of the selective reorthogonalization. In this section we exhibit some counterexample matrices $\mathbf{A}$ such that for any given value $L > 1$ the orthogonality obtained by MGS2(L) may be very poor. Our strategy is to find a matrix such that the following properties hold.

> *Property* 1. The matrix is numerically nonsingular but ill-conditioned.
> *Property* 2. MGS2(L) applied to this matrix performs no reorthogonalization and so reduces to MGS.

Let us define the matrix $\mathbf{A}(n, \alpha) \in \mathbb{R}^{n \times n}$ as

$$(2.2) \qquad \mathbf{A}(n, \alpha) = \mathbf{V}\mathbf{T}_A(n, \alpha) = \mathbf{V} \begin{pmatrix} \alpha & 1 & & \\ & \ddots & \ddots & \\ & & \alpha & 1 \\ & & & \alpha \end{pmatrix},$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ is such that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

Matrices $\mathbf{A}(n, \alpha)$ have the property that if we apply MGS2(L) (in exact arithmetic), we get

$$(2.3) \qquad L_j^{(1)} = \sum_{k=1}^{j-1} \frac{|r_{kj}^{(1)}|}{\|\mathbf{a}_j^{(1)}\|_2} = \frac{1}{\alpha}.$$

If we set α such that $L_j^{(1)} > L$, that is, $1/\alpha > L$, then the L-criterion is always satisfied. In this case, no reorthogonalization is performed, and then Property 2 is satisfied.

Moreover for all α, $0 < \alpha < 1$, the condition number of the matrix $\kappa(\mathbf{A}(n, \alpha))$ can be made arbitrarily large by choosing an appropriate n. We justify this claim by studying the matrix $\mathbf{T}_A(n, \alpha)$. First of all, we have

$$\mathbf{T}_A(n, \alpha)\mathbf{x}_1 = \begin{pmatrix} \alpha & 1 & & & & \\ & \alpha & 1 & & & \\ & & \alpha & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \alpha & 1 \\ & & & & & \alpha \end{pmatrix} \begin{pmatrix} 1 \\ -\alpha \\ \alpha^2 \\ \vdots \\ (-1)^{n-2}\alpha^{n-2} \\ (-1)^{n-1}\alpha^{n-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ (-1)^{n-1}\alpha^n \end{pmatrix};$$

therefore

$$\sigma_{\min}(\mathbf{T}_A(n,\alpha)) \leq \frac{\|\mathbf{T}_A(n,\alpha)\mathbf{x}_1\|_2}{\|\mathbf{x}_1\|_2} \leq \alpha^n \sqrt{\frac{1-\alpha^{2n}}{1-\alpha^2}}.$$

On the other hand, we also have

$$\mathbf{T}_A(n,\alpha)\mathbf{x}_2 = \begin{pmatrix} \alpha & 1 & & & & \\ & \alpha & 1 & & & \\ & & \alpha & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \alpha & 1 \\ & & & & & \alpha \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ \alpha \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

and therefore

$$\sigma_{\max}(\mathbf{T}_A(n,\alpha)) \geq \frac{\|\mathbf{T}_A(n,\alpha)\mathbf{x}_2\|_2}{\|\mathbf{x}_2\|_2} = \sqrt{1+\alpha^2}.$$

From (2.2), the condition number of $\mathbf{A}(n,\alpha)$ is the same as that of $\mathbf{T}_A(n,\alpha)$ and can be bounded by

$$(2.4) \qquad \kappa(\mathbf{A}(n,\alpha)) \geq \alpha^{-n}\sqrt{\frac{1-\alpha^4}{1-\alpha^{2n}}}.$$

For a given $L > 1$, the parameter α is set by using (2.3) so that $\alpha < 1/L < 1$ (Property 2). Using (2.4), we increase n, the size of the matrix $\mathbf{A}(n,\alpha)$, to have a sufficiently ill-conditioned matrix to comply with Property 1.

We have performed some numerical experiments with these matrices using MAT-LAB. The machine precision is $\varepsilon = 1.12 \cdot 10^{-16}$. We set $\alpha = 0.98$ and $n = 1500$ with a random unitary matrix $\mathbf{V}$ to obtain $\mathbf{A}(n,\alpha)$. The condition number of the matrix is $\kappa(\mathbf{A}(n,\alpha)) = 7.28 \cdot 10^{14}$. We should point out that even though the theoretical result was proved for the square root–free MGS algorithm, we consider in our experiments the classical implementation that involves the square root calculation. In Table 1, we display the numerical experiments. When $L = 1.03$, a few reorthogonalizations are performed and the algorithm is in fact very close to MGS applied to $\mathbf{A}(n,\alpha)$. $\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ is far from machine precision. When $L = 0.99$, the criterion permits all the reorthogonalizations; the algorithm is in fact exactly MGS2 and gives rise to a matrix $\mathbf{Q}$ that is orthogonal up to machine precision.

TABLE 1
$\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ for $\mathbf{Q}$ obtained by MGS2(L) for different values of L applied on $\mathbf{A}(n = 1500, \alpha = 0.98)$.

MGS2($L = 1.03$)	$5.44 \cdot 10^{-1}$
MGS2($L = 0.99$)	$4.57 \cdot 10^{-14}$

We show how to construct matrices such that the L-criterion with $L > 1$ fails. This strategy permits us to construct matrices such that $L = 1.03$ is not a good criterion. We have been limited by the size of the matrices used and we conjectured that increasing the size of the matrices would enable us to decrease the value of L. Furthermore, we remark that in our experiments we do not observe the influence of the terms in n and m in either assumption (1.2) on $\mathbf{A}$ or the final orthogonality given by inequality (1.3).

3. Lack of robustness of the K-criterion. Assuming that $\sum_{k=1}^{j-1}(r_{kj}^{(1)})^2 + \|\mathbf{a}_j^{(1)}\|_2^2 = \|\mathbf{a}_j\|_2^2$ (which corresponds to the theorem of Pythagoras if $\mathbf{Q}_{j-1}$ has orthogonal columns), we can rewrite the K-criterion as

$$(3.1) \qquad \frac{\sqrt{\sum_{k=1}^{j-1}(r_{kj}^{(1)})^2}}{\|\mathbf{a}_j^{(1)}\|_2} \leq \sqrt{K^2 - 1}.$$

Formula (3.1) means that the K-criterion compares the 2-norm of the nondiagonal entries $r_{kj}^{(1)}$, $k < j$, to the diagonal entry $\|\mathbf{a}_j^{(1)}\|_2$. We recall that the L-criterion consists of comparing the 1-norm of the nondiagonal entries $r_{kj}^{(1)}$, $k < j$, to the diagonal entry $\|\mathbf{a}_j^{(1)}\|_2$.

By analogy with inequality (1.43) we call a *diagonally dominant matrix by columns in the* 2-*norm* a matrix $\mathbf{A}$ such that for all j,

$$(3.2) \qquad |\mathbf{a}_{jj}| > \sqrt{\sum_{i \neq j} \mathbf{a}_{ij}^2}.$$

The value $L = 1$ for the L-criterion, which means that the matrix is diagonally dominant by columns, can be related to the value $K = \sqrt{2}$ for the K-criterion, which means that the matrix is diagonally dominant by columns in the 2-norm. Therefore, our point of view is that the K-criterion forced $\mathbf{R}$ to be diagonally dominant by columns in the 2-norm, whereas the L-criterion forced $\mathbf{R}$ to be diagonally dominant by columns.

We also notice that, if the K-criterion is satisfied, we have

$$\frac{\|\mathbf{a}_j\|_2}{\|\mathbf{a}_j^{(1)}\|_2} < K$$

$$\Leftrightarrow \frac{\sqrt{\|\mathbf{a}_j^{(1)}\|_2^2 + \sum_{k=1}^{j-1} r_{kj}^{(1)2}}}{\|\mathbf{a}_j^{(1)}\|_2} < K$$

$$\Leftrightarrow \frac{\sqrt{\sum_{k=1}^{j-1} r_{kj}^{(1)2}}}{\|\mathbf{a}_j^{(1)}\|_2} < \sqrt{K^2 - 1}$$

$$\Leftarrow \frac{\sum_{k=1}^{j-1} |r_{kj}^{(1)}|}{\|\mathbf{a}_j^{(1)}\|_2} < \sqrt{K^2 - 1}.$$

Thus if the L-criterion is satisfied with $L = 1$, this implies that the K-criterion with $K = \sqrt{2}$ is also satisfied. In other words, MGS2($L = 1$) reorthogonalizes more often than MGS2($K = \sqrt{2}$). In terms of diagonal dominance, we get that a matrix that is diagonally dominant by columns is diagonally dominant by columns in 2-norm.

We have compared MGS2($K = \sqrt{2}$) and MGS2($L = 1$) on several numerically nonsingular matrices from Matrix Market (http://math.nist.gov/MatrixMarket/) and also on the set of matrices of Hoffmann [12]. From our experiments, it appears that the K-criterion with $K = \sqrt{2}$ gives us results as good as the L-criterion with $L = 1$ in terms of orthogonality on all these matrices. However, the L-criterion with $L = 1$ may perform a few extra useless reorthogonalizations. Therefore, on these cases, the K-criterion is to be preferred.

TABLE 2

$\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ *for* $\mathbf{Q}$ *obtained MGS2(K) applied on* $\mathbf{A}(n = 1500, \alpha = 0.98)$.

MGS2($K = 1.43$)	$1.82 \cdot 10^0$

In this section, we look for matrices such that the K-criterion performs poorly. An initial idea is to simply take the matrix $\mathbf{A}(n, \alpha)$, $\alpha < 1$. For those matrices, in exact arithmetic, MGS2(K) does not perform any reorthogonalization for any

$$K \geq \sqrt{1 + \left(\frac{1}{\alpha}\right)^2}.$$

If we consider $\mathbf{A}(n = 1500, \alpha = 0.98)$, MGS2($K = 1.43$) performs no reorthogonalization and therefore reduces to MGS (cf. Table 2). With the $\mathbf{A}(n, \alpha)$ matrices, the smallest value of K for which MGS2(K) may fail is $K = \sqrt{2}$.

However, we can find better counterexample matrices by considering the matrices $\mathbf{B}(n, \alpha) \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{B}(n, \alpha) = \mathbf{VT}(n, \alpha) = \mathbf{V} \begin{pmatrix} 1 & -\alpha & -\alpha/\sqrt{2} & -\alpha/\sqrt{3} & & -\alpha/\sqrt{n-1} \\ & 1 & -\alpha/\sqrt{2} & -\alpha/\sqrt{3} & & -\alpha/\sqrt{n-1} \\ & & 1 & -\alpha/\sqrt{3} & & -\alpha/\sqrt{n-1} \\ & & & 1 & & \vdots \\ & & & & \ddots & \vdots \\ & & & & & -\alpha/\sqrt{n-1} \\ & & & & & 1 \end{pmatrix},$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ is such that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

For $\alpha < 1$, the unit triangular matrix $\mathbf{T}(n, \alpha)$ is a *diagonally dominant matrix by columns in the* 2-*norm* but is not a *diagonally dominant matrix* in the usual sense. For the reorthogonalization criterion, this means that if we apply MGS2($K \geq \sqrt{1 + \alpha^2}$) to $\mathbf{B}(n, \alpha)$, no reorthogonalization is performed, whereas for MGS2($L = 1$) nearly all the reorthogonalizations are performed. With $\alpha < 1$ and matrix $\mathbf{B}(n, \alpha)$, Property 2 is verified for MGS2($K \geq \sqrt{1 + \alpha^2}$).

Moreover, for $\alpha < 1$ the numerical experiments show that when n increases, $\mathbf{T}(n, \alpha)$ becomes ill-conditioned. Property 1 is also verified. It seems therefore that matrices $\mathbf{B}(n, \alpha)$ are good counterexamples for the K-criterion.

The experimental results are in Table 3. We run different versions of Gram–Schmidt with reorthogonalization on a set of matrices $\mathbf{B}(n, \alpha)$. The experiments are carried out using MATLAB. With $\mathbf{B}(n = 2500, \alpha = 0.30)$, the MGS2($K = 1.05$) algorithm gives a matrix $\mathbf{Q}$ that is far from orthogonal. This means that to guarantee good accuracy K has to be set to a value lower than 1.05. We recall that the value $K = 1$ implies that the algorithm reduces to MGS2. By diminishing α and increasing n, we expect that it is possible to exhibit smaller K than 1.05. We notice that the algorithm MGS2($L = 0.99$) behaves well.

4. What about CGS? The main focus of this paper is the MGS algorithm and its selective reorthogonalization variant. A natural question is whether the results extend to the CGS variant CGS2(L). In [17], the behavior of CGS2 is analyzed. However, to our knowledge, no study exists for either the CGS2(K) algorithm or the CGS2(L) algorithm. For that latter variant, we notice that the proof proposed

Table 3

$\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ for $\mathbf{Q}$ obtained with the MGS2(L) and MGS2(K) algorithms applied to four matrices $\mathbf{B}(n, \alpha)$.

(L, K)	$L = 0.99 \quad K = 1.40$	$L = 0.99 \quad K = 1.30$
matrix $\mathbf{B}$	$\mathbf{B}(n = 400, \alpha = 0.97)$	$\mathbf{B}(n = 500, \alpha = 0.82)$
$\kappa(\mathbf{B})$	$3.4 \cdot 10^{15}$	$8.6 \cdot 10^{14}$
MGS2(K)	$7.2 \cdot 10^{-1}$	$1.1 \cdot 10^{0}$
MGS2(L)	$1.5 \cdot 10^{-14}$	$1.9 \cdot 10^{-14}$

(L, K)	$L = 0.99 \quad K = 1.17$	$L = 0.99 \quad K = 1.05$
matrix $\mathbf{B}$	$\mathbf{B}(n = 1000, \alpha = 0.50)$	$\mathbf{B}(n = 2500, \alpha = 0.30)$
$\kappa(\mathbf{B})$	$1.8 \cdot 10^{13}$	$5.9 \cdot 10^{12}$
MGS2(K)	$1.0 \cdot 10^{-2}$	$7.6 \cdot 10^{-3}$
MGS2(L)	$3.5 \cdot 10^{-14}$	$8.0 \cdot 10^{-14}$

Table 4

$\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ for $\mathbf{Q}$ obtained by CGS2(L) and CGS2(K) for different values of L and K applied on $\mathbf{A}(n = 1500, \alpha = 0.98)$.

CGS2(L = 1.03)	$6.67 \cdot 10^{0}$
CGS2(L = 0.99)	$3.56 \cdot 10^{-14}$
CGS2(K = 1.43)	$1.82 \cdot 10^{0}$

in this paper for MGS2(L) does not apply. Even though the theoretical behavior is still an open question, we want to present some numerical experiments that tend to indicate that a similar behavior might exist for CGS2(L). In Table 4, we display the orthogonality quality produced by CGS2(L) and CGS2(K) on the same test matrix used in Tables 1 and 2. We observe that, on that matrix, CGS2(L) with $L = 1.03$ does not produce an orthogonal matrix while, for $L = 0.99$, the computed $\mathbf{Q}$ factor is orthogonal to machine precision. Similarly to MGS2(K), CGS2(K) for K slightly larger than $\sqrt{2}$ cannot compute an orthogonal set of vectors.

Similar experiments to those displayed in Table 3 are reported in Table 5, and thus similar comments can be made. That is, the CGS2($K = 1.05$) algorithm gives a matrix $\mathbf{Q}$ that is far from orthogonal. This means that, to guarantee good accuracy, K has to be set to a value lower than 1.05. We recall that the value $K = 1$ implies that the algorithm reduces to CGS2. On the other hand, the algorithm CGS2($L = 0.99$) behaves well. This is a clue suggesting that a theoretical analysis might be done to show that CGS2(L) with $L < 1$ generates an orthogonal set of vectors. This latter study might be the focus of future work, which would require developing a completely different proof from the one exposed in this paper, which does not apply.

Conclusion. In this paper, we give a new reorthogonalization criterion for the MGS algorithm with selective reorthogonalization, which is referred to as the L-criterion. This criterion depends on a single parameter L. When L is chosen smaller than 1 (e.g., $L = 0.99$), for numerically nonsingular matrices, this criterion is able to realize the compromise between saving useless reorthogonalizations and giving a set of vectors $\mathbf{Q}$ orthogonal up to machine precision level. On the other hand if we set $L > 1$, we exhibit some matrices for which the MGS algorithm with selective reorthogonalization based on the L-criterion (MGS2(L)) performs very poorly. The condition $L < 1$ is therefore necessary to ensure the robustness of MGS2(L).

To justify the need for a new criterion, we also show counterexample matrices for which a standard criterion, the K-criterion, gives a final set of vectors which are

TABLE 5

$\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_2$ *for* $\mathbf{Q}$ *obtained by different CGS algorithms applied to four matrices* $\mathbf{B}(n, \alpha)$.

(L, K)	$L = 0.99 \quad K = 1.40$	$L = 0.99 \quad K = 1.30$
matrix $\mathbf{B}$	$\mathbf{B}(n = 400, \alpha = 0.97)$	$\mathbf{B}(n = 500, \alpha = 0.82)$
$\kappa(\mathbf{B})$	$3.4 \cdot 10^{15}$	$8.6 \cdot 10^{14}$
CGS2(K)	$1.6 \cdot 10^0$	$1.6 \cdot 10^0$
CGS2(L)	$1.2 \cdot 10^{-14}$	$1.5 \cdot 10^{-14}$

(L, K)	$L = 0.99 \quad K = 1.17$	$L = 0.99 \quad K = 1.05$
matrix $\mathbf{B}$	$\mathbf{B}(n = 1000, \alpha = 0.50)$	$\mathbf{B}(n = 2500, \alpha = 0.30)$
$\kappa(\mathbf{B})$	$1.8 \cdot 10^{13}$	$5.9 \cdot 10^{12}$
CGS2(K)	$1.6 \cdot 10^0$	$1.6 \cdot 10^0$
CGS2(L)	$2.8 \cdot 10^{-14}$	$6.0 \cdot 10^{-14}$

far from orthogonal for any value of the parameter K. On all these counterexample matrices, we have verified the theory and observe that MGS2($L < 1$) behaves well.

Moreover, we have compared the K-criterion with $K = \sqrt{2}$ and the L-criterion with $L = 1$ on a wide class of standard test matrices. It appears that the K-criterion with $K = \sqrt{2}$ works fine in terms of orthogonality of the computed set of vectors for all these matrices, but it also saves more reorthogonalizations than the L-criterion with $L = 1$. Note that both criteria save reorthogonalizations on standard test matrices. Therefore in many cases, the K-criterion with $K = \sqrt{2}$ may nevertheless be preferred over the L-criterion with $L = 1$.

Finally, even though no theory yet exists, we give some numerical evidence that a similar analysis might exist for the CGS algorithm with selective orthogonalization based on the L-criterion. Furthermore, these numerical experiments show that neither MGS2(K) nor CGS2(K) succeeds in generating a set of orthogonal vectors. This also illustrates the lack of robustness of this criterion when implementing a CGS algorithm with selective reorthogonalization.

REFERENCES

[1] E. SCHMIDT, *Uber die Auflösung linearer Gleichungen mit unendlich vielen Unbekannten,* Rend. Circ. Mat. Palermo Ser. 1, 25 (1908), pp. 53–77.

[2] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice–Hall, Englewood Cliffs, NJ, 1963.

[3] J. R. RICE, *Experiments on Gram–Schmidt orthogonalization,* Math. Comp., 20 (1966), pp. 325–328.

[4] A. BJÖRCK, *Solving linear least squares problems by Gram–Schmidt orthogonalization,* BIT, 7 (1967), pp. 1–21.

[5] H. RUTISHAUSER, *Description of ALGOL* 60, in Handbook for Automatic Computation, Vol. 1, Part a, F. L. Bauer, A. S. Householder, F. W. J. Olver, H. Rutishauser, K. Samelson, and E. Stiefel, eds., Springer-Verlag, New York, 1967, pp. 220–221.

[6] J. M. VARAH, *A lower bound for the smallest singular value of a matrix,* Linear Algebra Appl., 11 (1975), pp. 3–5.

[7] J. W. DANIEL, W. B. GRAGG, L. KAUFMAN, AND G. W. STEWART, *Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization,* Math. Comp., 30 (1976), pp. 772–795.

[8] W. GANDER, L. MOLINARI, AND H. ŠVECOVÁ, *Numerische Prozeduren aus Nachlass und Lehre von Prof. Heinz Rutishauser,* Internat. Ser. Numer. Math. 33, Birkhäuser-Verlag, Basel, Stuttgart, 1977.

[9] W. Gander, *Algorithms for the QR Decomposition,* Research report 80-02, Eidgenössische Technische Hochschule, Zürich, 1980.

[10] B. N. Parlett, *The Symmetric Eigenvalue Problem,* Prentice–Hall, Englewood Cliffs, NJ, 1980.

[11] A. Ruhe, *Numerical aspects of Gram–Schmidt orthogonalization of vectors,* Linear Algebra Appl., 52/53 (1983), pp. 591–601.

[12] W. Hoffmann, *Iterative algorithms for Gram–Schmidt orthogonalization,* Computing, 41 (1989), pp. 335–348.

[13] L. Reichel and W. B. Gragg, *FORTRAN subroutines for updating the QR decomposition,* ACM Trans. Math. Software, 16 (1990), pp. 369–377.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[15] N. Higham, *Accuracy and Stability of Numerical Algorithms,* 2nd ed., SIAM, Philadelphia, 2002.

[16] V. Frayssé, L. Giraud, and S. Gratton, *A Set of GMRES Routines for Real and Complex Arithmetics,* Tech. report TR/PA/97/49, CERFACS, Toulouse, France, 1997.

[17] L. Giraud, J. Langou, and M. Rozložník, *On the Round-off Error Analysis of the Gram–Schmidt Algorithm with Reorthogonalization,* Tech. report TR/PA/02/33, CERFACS, Toulouse, France, 2002.