Cerfacs Technical report TR-PA-04-77
submitted to **Numerische Mathematik manuscript No. 5271**

Rounding error analysis of the classical Gram-Schmidt orthogonalization process

Luc Giraud[1], Julien Langou[2⋆], Miroslav Rozložník[3⋆⋆], Jasper van den Eshof[4]

[1] CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France (giraud@cerfacs.fr).
[2] The University of Tennessee, Department of Computer Science, 1122 Volunteer Blvd., Knoxville, TN 37996-3450, USA (langou@cs.utk.edu).
[3] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic (miro@cs.cas.cz).
[4] Heinrich-Heine-Universität, Mathematisches Institut, Universitätsstrasse 1, D-40225 Düsseldorf, Germany (eshof@am.uni-duesseldorf.de).

Submitted: August, 2004

Summary This paper provides two results on the numerical behavior of the classical Gram-Schmidt algorithm. The first result states that, provided the normal equations associated with the initial vectors are numerically nonsingular, the loss of orthogonality of the vectors computed by the classical Gram-Schmidt algorithm depends quadratically on the condition number. The second result states that, provided the initial set of vectors has numerical full rank, two iterations of the classical Gram-Schmidt algorithm are enough for ensuring the orthogonality of the computed vectors to be close to the unit roundoff level.

1 Introduction

Let $A = (a_1, \ldots, a_n)$ be a real $m \times n$ matrix ($m \geq n$) with full column rank ($rank(A) = n$). In many applications it is important to compute an orthogonal basis $Q = (q_1, \ldots, q_n)$ of $span(A)$ such that $A = QR$, where R is upper triangular matrix of order n. For this purpose, many orthogonalization algorithms and techniques have been proposed and are widely used, including those based on Householder transformations and Givens rotations (see e.g. [3,6,7,15]). In this paper we focus on the Gram-Schmidt (GS) orthogonaliza-

⋆ The work of the second author was supported in part by the US Department of Energy, Office of Basic Energy Science under LAB03-17 initiative, DOE contract No. DE-FG02-03ER25584, and in part by the TeraScale Optimal PDE Simulations (TOPS) SciDAC, DoE Contract No. DE-FC02-01ER25480
⋆⋆ The work of the third author was supported by the GA CR under the grant No. 201/02/0595 and by the GA AS CR under grant No. IAA1030405.

tion process [14] which numerical properties are certainly less understood than the two previously mentioned techniques.

The Gram-Schmidt process has two basic computational variants: the classical Gram-Schmidt (CGS) algorithm and the modified Gram-Schmidt (MGS) algorithm (see e.g. [3, 15]). From a numerical point of view, both of these techniques may produce a set of vectors which is far from orthogonal and sometimes the orthogonality can be completely lost [2,12]. A generally agreed opinion is that the MGS algorithm has much better numerical properties than the CGS algorithm [12,16]. Björck [2] has proved that, for a matrix A with numerical full rank, the loss of orthogonality in MGS occurs in a predictable way and it can be bounded by a term proportional to the condition number $\kappa(A)$ and to the unit roundoff u. Many textbooks (see e.g. [3,7,15]) give examples where the orthogonality of the vectors computed by CGS is lost completely, but the connection to the (ill-)conditioning of the problem has not been analyzed yet. As far as we could check, there was only one attempt to give a bound for the CGS algorithm by Kiełbasiński and Schwetlick (but unfortunately only in the Polish version of the 2nd edition of the book [10], p. 299). In the first part of the paper, we give Theorem 1 that provides a new bound for the loss of orthogonality in the CGS process. More precisely, Theorem 1 states that, provided that the matrix A^TA is numerically nonsingular, the loss orthogonality of the vectors $\bar{Q} = (\bar{q}_1, \ldots, \bar{q}_n)$ computed by the CGS algorithm (measured by the norm of the matrix $I - \bar{Q}^T\bar{Q}$) can be bounded by a term proportional to the square of the condition number $\kappa^2(A)$ times the unit roundoff. This result is based on the fact that the upper triangular factor $\bar{R}$ computed by CGS is the exact Cholesky factor of a matrix relatively close to the matrix A^TA, i.e. there exists a perturbation matrix E of relative small norm such that $A^TA + E = \bar{R}^T\bar{R}$.

On the other hand, in some other applications, it may be important to produce vectors Q so that their orthogonality is kept close to the machine precision. The orthogonality of the vectors computed by Gram-Schmidt process can then be improved by reorthogonalization: the orthogonalization step is iterated twice or several times. Section 3 is devoted to the analysis of this algorithm and basically consists in the proof of Theorem 2. This theorem states that, assuming numerical full rank of the matrix A, two iterations are enough to provide that the level of orthogonality of the computed vectors is close to the unit roundoff level. A rounding error analysis for this algorithm has already been given by Abdelmalek [1] who considered exactly two iteration steps. To prove that the scheme produces a set of vectors sufficiently orthogonal, Abdelmalek needed to assume that the diagonal elements of the computed upper triangular factor are large enough. The main contribution of this paper with respect to the work of Abdelmalek is to provide Lemma 2 and formula (3.33) which explain how these diagonal elements are controlled by the condition number of the matrix A. With the results presented in this paper, we are able to state that the results of Abdelmalek are indeed true for any set of initial vectors with numerical full rank. A second rounding error analysis for the iterated classical Gram-Schmidt algorithm has been done by Daniel, Gragg, Kaufman and Stewart [4]. Under

certain assumptions they proved that either the algorithm converges (theoretically in an infinite number of steps but in practice rapidly) to a sufficient level of orthogonality or the termination criterion they use may continually fail to be satisfied. The contribution of this paper with respect to the paper of Daniel et al. [4] is the same as for the paper of Abdelmalek [1]. We clearly define what happens for numerically nonsingular matrices. The iterated Gram-Schmidt algorithm was further thoroughly analyzed by Hoffmann [8]. His paper provided some nice tools for an elegant proof by induction. The proof of Theorem 2 uses Hoffmann's analysis (in contrast to [1,4]), the main motivation here is to have a proof as self-contained, modern and short as possible.

In this paper we analyze the CGS algorithm and its version with reorthogonalization, where each orthogonalization step is performed exactly twice (it is frequently denoted as the CGS2 algorithm). The matrix with orthonormal columns $Q = (q_1, \ldots, q_n)$ in both algorithms is constructed successively column-by-column so that for each $j = 1, \ldots, n$ we have $span(q_1, \ldots, q_j) = span(a_1, \ldots, a_j)$. For description we use the following notation. We will not distinguish between these two mathematically equivalent algorithms and we will use the same notation for the 'plain' CGS algorithm and for the CGS2 algorithm. The actual meaning of some quantity will be clear from the context of the section. The CGS algorithm starts with $q_1 = a_1/\|a_1\|$ and, for $j = 2, \ldots, n$, it successively generates

$$v_j = [I - Q_{j-1}Q_{j-1}^T]a_j, \tag{1.1}$$

where $q_j = v_j/\|v_j\|$. The corresponding column r_j in the upper triangular factor $R = (r_1, \ldots, r_n)$ is then given as $r_j = (Q_{j-1}^T a_j, \|v_j\|)^T$. In the CGS2 algorithm, we start with $q_1 = a_1/\|a_1\|$ and, for $j = 2, \ldots, n$, we successively compute the vectors

$$v_j = [I - Q_{j-1}Q_{j-1}^T]a_j, \tag{1.2}$$

$$w_j = [I - Q_{j-1}Q_{j-1}^T]v_j. \tag{1.3}$$

The vector q_j is the result of the normalization of w_j and it is given as $q_j = w_j/\|w_j\|$. The elements of the triangular factor are given by $(r_j, 0)^T + s_j = (Q_{j-1}^T a_j, 0)^T + (Q_{j-1}^T v_j, \|w_j\|)^T$.

Throughout the paper, $\|X\|$ denotes the 2-norm, $\sigma_{min}(X)$ is the minimal singular value and $\kappa(X)$ is the condition number of the matrix X; $\|x\|$ denotes the Euclidean norm of a vector x. For distinction, we denote quantities computed in finite precision arithmetic using an upper-bar. We assume the standard model for floating-point computations, and use the notation $fl(\cdot)$ for the computed result of some expression (see e.g. [7]). The unit roundoff is denoted by u. The terms $c_k = c_k(m, n)$, $k = 1, 2, \ldots$ are low-degree polynomials in the problem dimensions m and n, they are independent of the condition number $\kappa(A)$ and the unit roundoff u, but they do depend on some of the details of the computer arithmetic.

2 Loss of orthogonality in the classical Gram-Schmidt algorithm

This section is devoted to the analysis of the CGS algorithm. Theorem 1 states that the bound on the loss of orthogonality of the vectors computed by CGS depends on the

square of the condition number $\kappa(A)$. The key point of the proof is the Lemma 1 stating that the R-factor computed by CGS is a backward stable Cholesky factor for the matrix A^TA. We note that the factor obtained from the Cholesky factorization of A^TA is another backward stable Cholesky factor, in practice different by a factor of $u\kappa(A)$ (while they are equal in exact arithmetic).

Our analysis is based on standard results from the rounding error analysis of an elementary orthogonalization step (see, e.g., Björck [2,3]). These are first consecutively recalled in relations (2.4), (2.5), (2.6) and (2.7) (To be fully correct, throughout the whole paper we assume that $mu \ll 1$). The vector $\bar{v}_j$ computed in (1.1) satisfies

$$\bar{v}_j = a_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{r}_{k,j} + \delta v_j, \quad \|\delta v_j\| \leq c_0 u \|a_j\|, \tag{2.4}$$

where $c_0(m,n) = \mathcal{O}(mn)$. The vector $\bar{q}_j$ results from the normalization of the vector $\bar{v}_j$ and it is given as

$$\bar{q}_j = \bar{v}_j/\|\bar{v}_j\| + \delta q_j, \quad \|\delta q_j\| \leq (m+4)u, \quad \|\bar{q}_j\|^2 \leq 1 + (m+4)u. \tag{2.5}$$

The standard rounding error analysis for computing the orthogonalization coefficients $\bar{r}_{i,j}$, $i = 1, \ldots, j-1$, and the diagonal element $\bar{r}_{j,j}$ leads to the following error bounds:

$$\bar{r}_{i,j} = \bar{q}_i^T a_j + \delta r_{i,j}, \quad |\delta r_{i,j}| \leq mu \|\bar{q}_i\| \, \|a_j\|, \tag{2.6}$$

$$\bar{r}_{j,j} = \|\bar{v}_j\| + \delta r_{j,j}, \quad |\delta r_{j,j}| \leq mu \|\bar{v}_j\|. \tag{2.7}$$

Summarizing (2.4) together with (2.6) and (2.7) for steps $j = 1, \ldots, n$ into matrix notation (for details we also refer to Daniel et al [4]), the basis vectors $\bar{Q}$ and the upper triangular factor $\bar{R}$ computed by the CGS algorithm satisfy the recurrence relation

$$A + \delta A = \bar{Q}\bar{R}, \quad \|\delta A\| \leq c_1 u \|A\|, \tag{2.8}$$

where $c_1(m,n) = \mathcal{O}(mn^{3/2})$. The key point in the analysis of the CGS algorithm consists in understanding the numerical properties of the computed upper triangular factor $\bar{R}$. In the next lemma, we prove that it is an exact Cholesky factor of A^TA perturbed by a matrix of relative small norm. An interpretation of this lemma is that the CGS algorithm on the matrix A is a backward stable algorithm for the computation of the Cholesky decomposition of the matrix A^TA.

Lemma 1 *The upper triangular factor $\bar{R}$ computed by the CGS algorithm is such that*

$$\bar{R}^T\bar{R} = A^TA + E, \quad \|E\| \leq c_2 u \|A\|^2, \tag{2.9}$$

where $c_2(m,n) = \mathcal{O}(mn^2)$.

Proof We begin with the formula (2.6) for the orthogonalization coefficient $\bar{r}_{i,j}$ in the form

$$\bar{r}_{i,j} = \bar{q}_i^T a_j + \delta r_{i,j} = (\bar{v}_i/\|\bar{v}_i\| + \delta q_i)^T a_j + \delta r_{i,j}. \tag{2.10}$$

Multiplying both sides of (2.10) by $\|\bar{v}_i\|$ and substituting $\bar{r}_{i,i}$ on the left-hand side using (2.7), we get the relation

$$\bar{r}_{i,j}(\bar{r}_{i,i} - \delta r_{i,i}) = \bar{v}_i^T a_j + ((\delta q_i)^T a_j + \delta r_{i,j})\|\bar{v}_i\|.$$

Substituting for the computed vector $\bar{v}_i$ from (2.4) and using the identities (2.6) for $\bar{r}_{k,i}$, we obtain, after some manipulations, the identity

$$\bar{r}_{i,i}\bar{r}_{i,j} = (a_i - \sum_{k=1}^{i-1} \bar{q}_k \bar{r}_{k,i} + \delta v_i)^T a_j + \left((\delta q_i)^T a_j + \delta r_{i,j}\right)\|\bar{v}_i\| + \bar{r}_{i,j}\delta r_{i,i} \quad (2.11)$$

$$= a_i^T a_j - \sum_{k=1}^{i-1} \bar{r}_{k,i}(\bar{r}_{k,j} - \delta r_{k,j}) + (\delta v_i)^T a_j + ((\delta q_i)^T a_j + \delta r_{i,j})\|\bar{v}_i\| + \bar{r}_{i,j}\delta r_{i,i}.$$

Thus we can immediately write

$$\sum_{k=1}^{i} \bar{r}_{k,i}\bar{r}_{k,j} = a_i^T a_j + \sum_{k=1}^{i-1} \bar{r}_{k,i}\delta r_{k,j} + (\delta v_i)^T a_j + ((\delta q_i)^T a_j + \delta r_{i,j})\|\bar{v}_i\| + \bar{r}_{i,j}\delta r_{i,i},$$

which gives rise to the expression for the (i,j)-element in the matrix equation $\bar{R}^T\bar{R} = A^T A + E$. The bound for the norm of the matrix E can be obtained using the bounds on $\delta r_{k,i}$ and $\delta r_{i,i}$ from (2.7), the bound on δv_i from (2.4), the bound on δq_i from (2.5) and considering that

$$|\bar{r}_{k,i}| \le \|\bar{q}_k\|\,\|a_i\| + |\delta r_{k,i}| \le [1 + 2(m+4)u]\,\|a_i\|,$$
$$\|\bar{v}_i\| \le \|a_i\| + \sum_{k=1}^{i-1} |\bar{r}_{k,i}|\,\|\bar{q}_k\| + \|\delta v_i\| \le [n + 2c_0 u]\,\|a_i\|. \quad (2.12)$$

Note that a much smaller bound on $\|\bar{v}_i\|$ than the one given by (2.12) can be derived, but this one is small enough to get a bound in $\mathcal{O}(mn)u\|a_i\|\,\|a_j\|$ for all the entries of E. The norm of the error matrix E can then be bounded by $c_2 u\|A\|^2$ for a properly chosen c_2.

Corollary 1 *Under assumption on numerical nonsingularity of the matrix $A^T A$, i.e. assuming $c_2 u\kappa^2(A) < 1$, the upper triangular factor $\bar{R}$ computed by the CGS algorithm is nonsingular and we have*

$$\|\bar{R}^{-1}\| \le \frac{1}{\sigma_{min}(A)\,[1 - c_2 u\kappa^2(A)]^{1/2}}. \quad (2.13)$$

The analogy of this corollary in exact arithmetic is that if A is nonsingular, then R is nonsingular and $\|R^{-1}\| = 1/\sigma_{min}(A)$. We are now ready to prove the main result of this section.

Theorem 1 *Assuming $c_2 u\kappa^2(A) < 1$, the loss of orthogonality of the vectors $\bar{Q}$ computed by the CGS algorithm is bounded by*

$$\|I - \bar{Q}^T\bar{Q}\| \le \frac{c_3 u\kappa^2(A)}{1 - c_2 u\kappa^2(A)}, \quad (2.14)$$

where $c_3(m,n) = \mathcal{O}(mn^2)$.

	$\ell = 10^{-4}$, $\kappa(A) = 2.0000 \times 10^4$	$\ell = 10^{-7}$, $\kappa(A) = 2.0000 \times 10^7$
j	$\|I - \bar{Q}_j^T \bar{Q}_j\|$	$\|I - \bar{Q}_j^T \bar{Q}_j\|$
2	2.7747e-13	1.6266e-09
3	2.2646e-09	1.3280e-02
4	2.9616e-09	1.6491e-02

Table 1. The loss of orthogonality in CGS measured by $\|I-\bar{Q}_j^T\bar{Q}_j\|$ with respect to the orthogonalization step j (Experiments performed with MATLAB, where $u = 2.2204e-16$).

Proof It follows from (2.8) that $(A+\delta A)^T(A+\delta A) = \bar{R}^T\bar{Q}^T\bar{Q}\bar{R}$. Substituting A^TA from (2.9), we have

$$\bar{R}^T(I - \bar{Q}^T\bar{Q})\bar{R} = -(\delta A)^TA - A^T(\delta A) - (\delta A)^T(\delta A) + E.$$

Assuming $c_2u\kappa^2(A) < 1$, we can pre-multiply this identity from the left (resp. from the right) by $\bar{R}^{-T}$ (resp. by $\bar{R}^{-1}$). The loss of orthogonality $I - \bar{Q}^T\bar{Q}$ can then be bounded as

$$\|I - \bar{Q}^T\bar{Q}\| \leq \Big(2\|\delta A\|\,\|A\| + \|\delta A\|^2 + \|E\|\Big)\,\|\bar{R}^{-1}\|^2.$$

Using the bounds on $\|\delta A\|$, $\|E\|$ and $\|\bar{R}^{-1}\|$ in Equations (2.8), (2.9) and (2.13), we obtain the statement of the theorem.

We have proved that for CGS the loss of orthogonality can be bounded in terms of the square of the condition number $\kappa(A)$. This is true for every matrix A such that A^TA is numerically nonsingular, i.e. $c_2u\kappa^2(A) < 1$. In contrast, Björck [2] proved that the loss of orthogonality in MGS depends only linearly on $\kappa(A)$. For this, he has assumed the numerical full rank of the matrix A, i.e. $cu\kappa(A) < 1$.

Let us illustrate how tight the bound (2.14) is. We consider the famous Läuchli matrix of order $n+1 \times n$ (for details see [2] or [9]) with nonzero elements defined as $A_{1,j} = 1$ and $A_{j+1,j} = \ell \ll 1$ $(j = 1, \ldots, n)$. This matrix has the following properties: $\sigma_{min}(A) = \ell$, $\|A\| = (n + \ell^2)^{1/2} \approx \|a_j\| = (1 + \ell^2)^{1/2}$, $j = 1, \ldots, n$ and $\kappa(A) = \ell^{-1}(n + \ell^2)^{1/2} \approx n^{1/2}\ell^{-1}$. It is particularly interesting to study the Gram-Schmidt algorithm with this matrix since, at every step $j \geq 2$, we have $r_{j,j} = \|v_j\| \approx \sigma_{min}(A)$. This significantly affects the numerical behavior of the CGS algorithm in finite precision arithmetic. In Table 1, we report the loss of orthogonality between the vectors computed by CGS (measured by $\|I - \bar{Q}_j^T\bar{Q}_j\|$) for $\ell = 10^{-4}$ and $\ell = 10^{-7}$. It is clear from Table 1 that the loss of orthogonality depends quadratically on the condition number of A (except for the step $j = 2$, where CGS coincides with MGS), and thus the bound (2.14) is justified.

3 Loss of orthogonality in the Gram-Schmidt algorithm with reorthogonalization

In this section we analyze the CGS2 algorithm, where the orthogonalization of the current vector a_j against the previously computed set is performed exactly twice. In contrast to

the CGS algorithm, we use a standard assumption on the numerical full rank of the initial set of vectors in the form $c_4 u\kappa(A) < 1$ and prove that two steps are enough for preserving the orthogonality of computed vectors close to the machine precision level. Indeed, the main result of this section is formulated in the following theorem.

Theorem 2 *Assuming $c_4 u\kappa(A) < 1$, the loss of orthogonality of the vectors $\bar{Q}$ computed by the CGS2 algorithm can be bounded as*

$$\|I - \bar{Q}^T \bar{Q}\| \leq c_5 u. \tag{3.15}$$

where $c_4 = \mathcal{O}(m^2 n^3)$ and $c_5 = \mathcal{O}(mn^{3/2})$.

The proof of Theorem 2 is done using induction. We assume that, at step $j-1$, we have

$$\|\bar{Q}_{i-1}^T \bar{q}_i\| \leq c_6 u, \quad i = 1, \ldots, j-1, \tag{3.16}$$

where $c_6(m,n) = \mathcal{O}(mn)$ (note that this is trivially true at step 1). The goal is to prove that the statement (3.16) is also true at step j; that is to say we want to prove that $\|\bar{Q}_{j-1}^T \bar{q}_j\| \leq c_6 u$. Of particular importance for us is the result proved by Hoffmann [8, p. 343-4]. He proved that if $\|\bar{Q}_{i-1}^T \bar{q}_i\| \leq c_6 u$ for $i = 1, \ldots, j$ then

$$\|I - \bar{Q}_j^T \bar{Q}_j\| \leq \max_{i=1,\ldots,j} \left\{ \|\bar{q}_i\|^2 - 1 + \|\bar{Q}_{i-1}^T \bar{q}_i\| \sqrt{2j} \right\} \leq c_5 u, \tag{3.17}$$

where $c_5(m,n) = (1 + (m+4)u)\sqrt{2} n^{1/2} c_6(m,n) + m + 4 = \mathcal{O}(mn^{3/2})$. This will finally give the statement (3.15). Note that (3.17) also implies that $\|\bar{Q}_{j-1}\| \leq [1 + c_5 u]^{1/2}$.

Similarly to (2.4), we first recall the results for the elementary projections (1.2) and (1.3)

$$\bar{v}_j = a_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{r}_{k,j} + \delta v_j, \quad \|\delta v_j\| \leq c_0 u \|a_j\|, \tag{3.18}$$

$$\bar{w}_j = \bar{v}_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{s}_{k,j} + \delta w_j, \quad \|\delta w_j\| \leq c_0 u \|\bar{v}_j\|, \tag{3.19}$$

where $c_0(m,n) = \mathcal{O}(mn)$. The orthogonalization coefficients $\bar{r}_{k,j}$ and $\bar{s}_{k,j}$, $k = 1, \ldots, j-1$ and the diagonal elements $\bar{s}_{j,j}$ (note that the normalization of the vector is performed only after the second iteration) satisfy

$$\bar{r}_{k,j} = \bar{q}_k^T a_j + \delta r_{k,j}, \quad \bar{s}_{k,j} = \bar{q}_k^T \bar{v}_j + \delta s_{k,j}, \quad \bar{s}_{j,j} = \|\bar{w}_j\| + \delta s_{j,j}, \tag{3.20}$$

$$|\delta r_{k,j}| \leq mu \|\bar{q}_k\| \, \|a_j\|, \quad |\delta s_{k,j}| \leq mu \|\bar{q}_k\| \, \|\bar{v}_j\|, \quad |\delta s_{j,j}| \leq mu \|\bar{w}_j\|. \tag{3.21}$$

The vector $\bar{q}_j$ comes from the normalization of the vector $\bar{w}_j$. Analogously to (2.5), we have

$$\bar{q}_j = \bar{w}_j / \|\bar{w}_j\| + \delta q_j, \quad \|\delta q_j\| \leq (m+4)u, \quad \|\bar{q}_j\|^2 \leq 1 + (m+4)u. \tag{3.22}$$

The relations (3.18) and (3.19) can be added to give

$$a_j + \delta v_j + \delta w_j = \sum_{k=1}^{j-1} (\bar{r}_{k,j} + \bar{s}_{k,j}) \bar{q}_k + \bar{w}_j. \tag{3.23}$$

Taking also into account the errors (3.21) and (3.22), the recurrence (3.23) for $j = 1, \ldots, n$ can be rewritten into the matrix relation

$$A + \delta V + \delta W = \bar{Q}(\bar{R} + \bar{S}), \tag{3.24}$$

where $\delta V = (\delta v_1, \ldots, \delta v_n)$ and $\delta W = (\delta w_1, \ldots, \delta w_n)$. For simplicity, we will assume a bound for the perturbation matrices δV and δW in the same form as the one for the perturbation matrix δA in (2.8). Actually, the possible differences can be hidden into definition of the constant c_1.

In order to prove that $\|\bar{Q}_{j-1}^T \bar{q}_j\| \leq c_6 u$, we proceed in two steps. In the first step, we analyze the orthogonality of the vector $\bar{v}_j$ with respect to the column space of the matrix $\bar{Q}_{j-1}$. We give a bound for $\frac{\|\bar{Q}_{j-1}^T \bar{v}_j\|}{\|\bar{v}_j\|}$. In the second part of the proof, a bound for the quotient $\frac{\|\bar{Q}_{j-1}^T \bar{w}_j\|}{\|\bar{w}_j\|}$ is given. The factors $\frac{\|a_j\|}{\|\bar{v}_j\|}$ and $\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}$ play a significant role in the proof. Assuming that A has numerical full rank, we prove a lower bound for the factor $\frac{\|a_j\|}{\|\bar{v}_j\|}$ proportional to the minimum singular value of A. Using this bound, we prove that the factor $\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}$ is necessarily close to 1. This last statement is the main reason why two iterations of the CGS process are enough for preserving the orthogonality of the computed vectors close to the level of the unit roundoff.

Let us start now with the analysis of the first step. Multiplication of the expression (3.18) from the left by $\bar{Q}_{j-1}^T$ leads to the identity

$$\bar{Q}_{j-1}^T \bar{v}_j = (I - \bar{Q}_{j-1}^T \bar{Q}_{j-1}) \bar{Q}_{j-1}^T a_j + \bar{Q}_{j-1}^T (-\sum_{k=1}^{j-1} \bar{q}_k \delta r_{k,j} + \delta v_j).$$

Taking the norm of this expression, dividing by the norm of $\bar{v}_j$ and using (3.20) and (3.21), the quotient $\|\bar{Q}_{j-1}^T \bar{v}_j\| / \|\bar{v}_j\|$ can be bounded as

$$\frac{\|\bar{Q}_{j-1}^T \bar{v}_j\|}{\|\bar{v}_j\|} \leq [c_5 + mn(1 + (m+4)u) + c_0] \, (1 + c_5 u)^{1/2} u \frac{\|a_j\|}{\|\bar{v}_j\|}. \tag{3.25}$$

The inequality (3.25) is easy to interpret. It is well known and described in many papers (e.g. [5,8,13]) that the loss of orthogonality after the first orthogonalization step (1.2) is proportional to the quantity $\frac{\|a_j\|}{\|\bar{v}_j\|}$. The next lemma provides us some control on this quantity.

Lemma 2 *Assuming $c_7 u \kappa(A) < 1$, the norms of the vectors $\bar{v}_j$ computed by the first iteration of the CGS2 algorithm satisfy the inequalities*

$$\frac{\|a_j\|}{\|\bar{v}_j\|} \leq \kappa(A) \, [1 - c_7 u \kappa(A)]^{-1}, \tag{3.26}$$

where $c_7(m,n) = \mathcal{O}(mn^{3/2})$.

Proof We consider the matrix recurrence (3.24) for the first $j-1$ orthogonalization steps

$$A_{j-1} + \delta V_{j-1} + \delta W_{j-1} = \bar{Q}_{j-1}(\bar{R}_{j-1} + \bar{S}_{j-1}). \tag{3.27}$$

Summarizing (3.27) with (3.18), we can rewrite these two relations into the matrix relation

$$A_j + [\delta V_{j-1} + \delta W_{j-1}, \delta v_j - \bar{v}_j] = \bar{Q}_{j-1} \left[\bar{R}_{j-1} + \bar{S}_{j-1}, \bar{r}_j\right], \tag{3.28}$$

where $[\bar{R}_{j-1}+\bar{S}_{j-1}, \bar{r}_j]$ is a $(j-1)\times j$ matrix. Let us define $\Delta_j = [\delta V_{j-1}+\delta W_{j-1}, \delta v_j - \bar{v}_j]$ and remark that the matrix $\bar{Q}_{j-1}[\bar{R}_{j-1}+\bar{S}_{j-1}, \bar{r}_j]$ is of rank $(j-1)$. Therefore the matrix $A_j + \Delta_j$ has rank $j-1$ whereas we have assumed that the matrix A_j has full rank j. This means that the distance from A_j to the set of matrices of rank $j-1$ is less than the norm of Δ_j. The distance to singularity for a square matrix can be related to its minimal singular value. Theorems on relative distance to singularity can be found in many books (e.g. [7, p. 123] or [6, p. 73]). Although the textbooks usually assume the case of square matrix, the statement is valid also for rectangular matrices. Indeed, in our case the minimal singular value of A_j can be then bounded by the norm of the perturbation matrix Δ_j that is to say $\sigma_{min}(A_j) \leq \|\Delta_j\|$ and so we can write

$$\sigma_{min}(A) \leq \sigma_{min}(A_j) \leq \|\Delta_j\| \leq \sqrt{\|\delta V_{j-1}\|^2 + \|\delta W_{j-1}\|^2 + \|\delta v_j\|^2 + \|\bar{v}_j\|^2}. \tag{3.29}$$

We are now going to use the bounds on the norms of the matrices δV_{j-1}, δW_{j-1}, the bound (3.18) on the vector δv_j and an argumentation similar to (2.12). Assuming $c_7 u\kappa(A) < 1$ (with a properly chosen polynomial $c_7(m,n) = \mathcal{O}(mn^{3/2})$ with the same degree as the one of $c_1(m,n)$), a lower bound for the norm of the vector $\bar{u}_j$ can be given in the form

$$\|\bar{v}_j\| \geq \sigma_{min}(A)(1 - c_7 u\kappa(A)). \tag{3.30}$$

The bound (3.30) shows that, under the assumption that A has numerical full rank (i.e. assuming $c_7 u\kappa(A) < 1$), the norm $\|\bar{v}_j\|$ is essentially bounded by the minimal singular value of A. We note that the result (3.30) corresponds well to the bound in exact arithmetic $\|v_j\| \geq \sigma_{min}(A)$. Consequently, the quotient $\|\bar{Q}_{j-1}^T \bar{v}_j\|/\|\bar{v}_j\|$, which describes the orthogonality between the vector $\bar{v}_j$ computed by the first iteration step and the column vectors of $\bar{Q}_{j-1}$, can be, combining (3.25) and (3.26), bounded by

$$\frac{\|\bar{Q}_{j-1}^T \bar{v}_j\|}{\|\bar{v}_j\|} \leq \frac{[c_5 + mn(1+(m+4)u) + c_0]\,(1+c_5 u)^{1/2} u\kappa(A)}{1 - c_7 u\kappa(A)} = c_8 u\kappa(A). \tag{3.31}$$

We are now ready to start the second step to prove that $\|\bar{Q}_{j-1}^T \bar{q}_j\| \leq c_6 u$. We proceed similarly as in the first part. Using the derived bound (3.31), we study the orthogonality of the vector $\bar{w}_j$ computed by the second iteration step with respect to the column vectors $\bar{Q}_{j-1}$ and finally give a bound for the quotient $\|\bar{Q}_{j-1}^T \bar{q}_j\|$. Let us concentrate first on $\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}$. Using the relation for the local error in the second iteration step (3.19), it can be bounded as follows

$$\begin{aligned}\frac{\|\bar{w}_j\|}{\|\bar{v}_j\|} &\geq \frac{\|\bar{v}_j\|}{\|\bar{v}_j\|} - \|\bar{Q}_{j-1}\| \frac{\|\bar{Q}_{j-1}^T \bar{v}_j\|}{\|\bar{v}_j\|} - \frac{\|\sum_{k=1}^{j-1} \bar{q}_k \delta s_{k,j}\| + \|\delta v_j\|}{\|\bar{v}_j\|} \\ &\geq 1 - \left[c_8\kappa(A)(1+c_5 u)^{1/2} + mn(1+(m+4)u) + c_0\right] u.\end{aligned}$$

Thus, under the assumption that

$$c_9 u\kappa(A) = \left[c_8\kappa(A)(1+c_5 u)^{1/2} + mn(1+(m+4)u) + c_0\right] u < 1, \tag{3.32}$$

we obtain the final bound for the factor $\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}$ as follows

$$\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|} \leq [1 - c_9 u\kappa(A)]^{-1}. \tag{3.33}$$

The upper bound (3.33) shows that if we slightly strengthen the assumption (3.32), the factor $\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}$ becomes very close to 1, which essentially means that $\|\bar{w}_j\|$ is not significantly smaller than $\|\bar{v}_j\|$. We note that, in exact arithmetic, we have $w_j = v_j$ implying $\|v_j\|/\|w_j\| = 1$. Finally, we also note that the main contribution of this section with respect to the results of Abdelmalek is Equation (3.33). In his analysis, Abdelmaleck needs that $(j-2)^2\|\bar{Q}_{j-1}^T\bar{v}_j\|/\|\bar{w}_j\| \leq 1$, a statement that he expects to hold in most practical cases. Indeed, this criterion can be rewritten as $(j-2)^2\|\bar{Q}_{j-1}^T\bar{v}_j\|/\|\bar{v}_j\|\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|} \leq 1$ and it can been seen from (3.31) and (3.33) that Abdelmaleck's assumption is met under a clear assumption on the numerical rank of A.

From (3.19), it follows that

$$\bar{Q}_{j-1}^T\bar{w}_j = (I - \bar{Q}_{j-1}^T\bar{Q}_{j-1})\bar{Q}_{j-1}^T\bar{v}_j + \bar{Q}_{j-1}^T(-\sum_{k=1}^{j-1}\bar{q}_k\delta s_{k,j} + \delta w_j).$$

Taking the norm of this expression and using (3.21) and (3.31) leads to

$$\frac{\|\bar{Q}_{j-1}^T\bar{w}_j\|}{\|\bar{w}_j\|} \leq [c_5c_8u\kappa(A) + mn(1+(m+4)u) + c_0]\,u(1+(m+4)u)^{1/2}\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}. \tag{3.34}$$

Consequently, using (3.25), (3.34) and (3.33), and remarking that $\|\bar{Q}_{j-1}^T\bar{q}_j\| \leq \|\bar{Q}_{j-1}^T\bar{w}_j\|/\|\bar{w}_j\| + \|\bar{Q}_{j-1}^T\delta q_j\|$, we can write

$$\begin{aligned}\|\bar{Q}_{j-1}^T\bar{q}_j\| \leq\ & [c_5c_8u\kappa(A) + mn(1+(m+4)u) + c_0]\,[1 - c_9u\kappa(A)]^{-1}\,[1+(m+4)u]^{1/2}u \\ & + (m+4)u[1+c_5u]^{1/2}.\end{aligned} \tag{3.35}$$

Now, let us assume that $[1 - c_9u\kappa(A)]^{-1} \leq 2$, $mn^{3/2}u \ll 1$ and $c_5c_8u\kappa(A) \leq 1$. Then $1+(m+4)u \leq 2$, $1+c_5u \leq 2$ and we have

$$\|\bar{Q}_{j-1}^T\bar{q}_j\| \leq 2\sqrt{2}\,[1+2mn+c_0]\,u + \sqrt{2}(m+4)u = c_6u, \tag{3.36}$$

where $c_6(m,n) = \mathcal{O}(mn)$. We can summarize all the assumptions made so far in a single one

$$c_4u\kappa(A) < 1, \tag{3.37}$$

where $c_4(m,n) = \mathcal{O}(m^2n^3)$.

We are now able to conclude the proof by induction. If the induction assumption (3.16) is true at the step $j-1$, under assumption (3.37), the statement is true at the step j. Consequently we have at the step j the bound (3.17). For the last step $j = n$, it follows that

$$\|I - \bar{Q}^T\bar{Q}\| \leq c_5u. \tag{3.38}$$

Finally we illustrate our theoretical results. We consider once again the Läuchli matrix with $\ell = 10^{-4}$ and $\ell = 10^{-7}$. In Table 2, we compare the loss of orthogonality computed

	$\ell = 10^{-4}$, $\kappa(A) = 2.0000 \times 10^4$		$\ell = 10^{-7}$, $\kappa(A) = 2.0000 \times 10^7$	
j	CGS	CGS2	CGS	CGS2
2	2.7747e-13	2.2204e-16	1.6266e-09	4.4409e-16
3	2.2646e-09	2.2888e-16	1.3280e-02	4.5719e-16
4	2.9616e-09	2.5713e-16	1.6491e-02	4.6350e-16

Table 2. The loss of orthogonality in the CGS and CGS2 algorithms (measured by corresponding $\|I - \bar{Q}_j^T \bar{Q}_j\|$) with respect to the orthogonalization step j (Experiments performed with MATLAB, where $u = 2.2204e - 16$).

by the plain CGS algorithm and the CGS2 algorithm. It is clear from Table 2 that two iteration steps are enough for preserving the orthogonality of the computed vectors close to the machine level, which is in agreement with the theoretical results developed in this section.

4 Conclusions and remarks

In this paper, we give a bound for the loss of orthogonality of the CGS algorithm. We proved that the loss of orthogonality of CGS can be bounded by a term proportional to the square of the condition number $\kappa(A)$ and to the unit roundoff u. This assumes that $A^T A$ is numerically nonsingular. Indeed, the loss of orthogonality occurs in a predictable way and our bound is tight. This is very similar to MGS up to the difference that the loss of orthogonality in MGS depends (only) linearly on $\kappa(A)$ and the assumption depends on the numerical full rank (only) of the matrix A. This result fills the theoretical gap in understanding the CGS process and agrees well with all examples used in textbooks. In addition, we have proved that the orthogonality of the vectors computed by the CGS2 algorithm is close to the machine precision level. Indeed, exactly two iteration-steps are already enough when full orthogonality is requested and when the algorithm is applied to (numerically) independent initial set of column vectors. This result extends the ones of Abdelmalek [1], Daniel et al [4], Kahan and Parlett [11] and Hoffmann [8].

5 Acknowledgements

The authors would like to thank for the fruitful discussion, useful comments and help to Å. Björck, A. Kiełbasiński, A. Smoktunowicz, P. Tichý and K. Zietak.

References

1. N. Abdelmalek. Round off error analysis for Gram-Schmidt method and solution of linear least squares problems. *BIT 11 (1971), 345-368.*
2. Å. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT 7 (1967), 1-21.*

3. Å. Björck. Numerical Methods for Least Squares Problems. *SIAM, Philadelphia, PA, 1996.*
4. 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), 772-795.*
5. A. Dax. A modified Gram-Schmidt algorithm with iterative orthogonalization and pivoting. *Linear Alg. and its Appl. 310 (2000), 25-42.*
6. G. H. Golub and C. F. Van Loan. Matrix Computations, 3rd ed. *John Hopkins University Press, Baltimore, MD, 1996.*
7. N. Higham. Accuracy and Stability of Numerical Algorithms. *SIAM, Philadelphia, PA, 2nd ed., 2002.*
8. W. Hoffmann. Iterative Algorithms for Gram-Schmidt Orthogonalization. *Computing 41 (1989), 335-348.*
9. W. Jalby and B. Philippe. Stability analysis and improvement of the block Gram-Schmidt algorithm. *SIAM J. Sci. Stat. Comput 12 (1991), 1058–1073.*
10. A. Kiełbasiński and H. Schwetlick. Numeryczna algebra liniova (in Polish). *Wydawnictwo Naukowo-Technyczne, Warszawa (1994), Second edition.*
11. B. N. Parlett. The Symmetric Eigenvalue Problem. *Englewood Cliffs, N.J., Prentice-Hall, 1980.*
12. J. R. Rice. Experiments on Gram-Schmidt Orthogonalization. *Math. Comp. 20 (1966), 325-328.*
13. H. Rutishauser. Description of Algol 60. Handbook for Automatic Computation, Vol. 1a. *Springer Verlag, Berlin, 1967.*
14. E. Schmidt. Über die Auflösung linearer Gleichungen mit unendlich vielen Ubbekannten. *Rend. Circ. Mat. Palermo. Ser. 1, 25 (1908), 53-77.*
15. G. W. Stewart. Matrix Algorithms. Volume I: Basic Decompositions. *SIAM, Philadelphia, PA, 1998.*
16. J. H. Wilkinson. Modern error analysis. *SIAM rev., 13:4, (1971), 548-569.*