Numer. Math. (2005) 101: 87–100
DOI 10.1007/s00211-005-0615-4

Numerische
Mathematik

Luc Giraud · Julien Langou · Miroslav Rozložník · Jasper van den Eshof

Rounding error analysis of the classical Gram-Schmidt orthogonalization process

Received: 10 August 2004 / Revised: 12 January 2005 / Published online: 25 May 2005

Abstract 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 of the initial vectors. The second result states that, provided the initial set of vectors has numerical full rank, two

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 project 1ET400300415 within the National Program of Research “Information Society” and by the GA AS CR under grant No. IAA1030405.

Luc Giraud
CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France
E-mail: giraud@cerfacs.fr

Julien Langou (✉)
Department of Computer Science, The University of Tennessee, 1122 Volunteer Blvd.,
Knoxville,TN 37996-3450, USA
E-mail: langou@cs.utk.edu

Miroslav Rozložník
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8,Czech Republic
E-mail: miro@cs.cas.cz

Jasper van den Eshof
Heinrich-Heine-Universität, Mathematisches Institut, Universitätsstrasse 1, D-40225 Düsseldorf, Germany
E-mail: eshof@am.uni-duesseldorf.de

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.

Mathematics Subject Classification (2000) : 65F25, 65G50, 15A23

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 orthonormal 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. [3,10,12,23]). In this paper, we focus on the Gram-Schmidt (GS) orthogonalization process [22] which numerical properties are certainly less understood than those of 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. [3,23]). Although those two variants are mathematically equivalent, due to roundoff errors the set of vectors produced by either of these two methods can be far from orthogonal and sometimes the orthogonality can even be completely absent [2,20]. Generally it is agreed that the MGS algorithm has much better numerical properties than the CGS algorithm [20,24].

When the columns of A are generated one at a time, one has to use the column variant of these algorithms. In this case, even though CGS and MGS algorithms have the same operation count and memory requirement, CGS uses higher level BLAS operations enabling a higher performance rate on modern computing platforms. This performance aspect can not be overlooked in certain computing environments (in particular parallel distributed). In addition, there are applications, where the orthogonality of computed vectors does not play a crucial role and where the CGS algorithm can be a successful alternative to the MGS algorithm (for example the implementation of the GMRES method [11]).

Up to now there was no bound available for the loss of orthogonality of vectors computed by the CGS process (with the exception of a loose bound given in Kiełbasiński and Schwetlick, see [16, p.284] or [17, p.299]). In the first part of this paper, we derive a new bound for the loss of orthogonality in the CGS process. Provided that the matrix $A^T A$ is numerically nonsingular, we show that the loss orthogonality of the vectors computed by the CGS algorithm can be bounded by a term proportional to the square of the condition number $\kappa^2(A)$ times the unit roundoff. We illustrate through a numerical experiment that this bound is tight.

In some other applications it may be important to produce a set of basis vectors whose orthogonality is close to the machine precision. In this case the orthogonality of the vectors computed by a Gram-Schmidt process can be improved by reorthogonalization, where the orthogonalization step (either in the CGS or MGS algorithm) is iterated twice or several times. Extensive experiments with iterative versions of GS were performed by Rice [20]. Various schemes and aspects of reorthogonalization have been analyzed by several authors, including Abdelmalek [1], Daniel, Gragg, Kaufman and Stewart [5] and Hoffmann [13]. In particular, Hoffmann [13], taking into account experimental results on (numerically) nonsingular problems, observed that a third iteration never occurred for both iterated CGS and MGS algorithms. He conjectured that two steps are enough for obtaining orthogonality on the level of machine precision but a theoretical foundation for this observation still remained an open question.

The importance of having such a result is even more profound nowadays since many numerical experiments have shown that CGS with reorthogonalization may be faster than the MGS method despite the fact that it performs twice as many floating-point operations.For examples in Krylov solvers on parallel distributed computers see [7,8,18].

In the second part of this paper, we analyze the CGS algorithm with reorthogonalization, where each orthogonalization step is performed exactly twice (it is frequently denoted as the CGS2 algorithm). The main result of this section is a proof of the fact that, assuming full rank of the matrix A, two iterations are sufficient to guarantee that the level of orthogonality of the computed vectors is close to the unit roundoff level.

In both algorithms (CGS and CGS with reorthogonalization), the matrix with orthonormal columns $Q = (q_1, \dots, q_n)$ is assumed to be constructed column-by-column so that for each index $j = 1, \dots, n$ we have $span(q_1, \dots, q_j) = span(a_1, \dots, a_j)$. In the course of this paper we will not make a distinction in our notation between these two mathematically equivalent algorithms and we will use the same notation for the 'plain' CGS algorithm as well as 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 $j = 2, \dots, n$, it successively generates

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

where $q_j = v_j/\|v_j\|$. The corresponding column r_j in the upper triangular factor $R = (r_1, \dots, r_n)$ 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 $j = 2, \dots, n$, we successively compute the vectors

$$v_j = [I - Q_{j-1}Q_{j-1}^T]a_j, \tag{2}$$

$$w_j = [I - Q_{j-1}Q_{j-1}^T]v_j. \tag{3}$$

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 $(r_j, 0)^T + s_j = (Q_{j-1}^T a_j, 0)^T + (Q_{j-1}^T v_j, \|w_j\|)^T$. Note that the above CGS2 algorithm is different from the algorithm that consists of applying consecutively the CGS algorithm twice. That is when one first applies the 'plain' CGS, as defined by 1, followed by another sweep of CGS on the results from the first run as $\tilde{v}_j = [I - \tilde{Q}_{j-1}\tilde{Q}_{j-1}^T]q_j$, for $j = 2, \dots, n$, where $\tilde{q}_j = \tilde{v}_j/\|\tilde{v}_j\|$ (and $\tilde{q}_1 = q_1$). For details and thorough analysis of the MGS version of such algorithm we refer to [9].

Throughout the paper, $\|X\|$ denotes the 2-norm of the matrix X. Similarly, $\sigma_{min}(X)$ is the minimal singular value and $\kappa(X)$ refers to the condition number of the matrix; $\|x\|$ denotes the Euclidean norm of a vector x. For distinction with their exact arithmetic counterparts, we denote quantities computed in finite precision arithmetic using an extra upper-bar. We assume the standard model for floating-point computations, and use the notation $fl(\cdot)$ for the computed results of some expressions (see e.g. [12]). The unit roundoff is denoted by u. The terms $c_k(m,n)$, $k = 1, 2, \ldots$ 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 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 below 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 in the proof is Lemma 1 stating that the R-factor computed by CGS is a backward stable Cholesky factor for the matrix $A^T A$. We note that the factor obtained from the Cholesky factorization of $A^T A$ is another backward stable Cholesky factor, in the presence of rounding error they are 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 (4), (5), (6) and (7) (to be fully correct, throughout the whole paper we assume that $mu \ll 1$). The vector $\bar{v}_j$ computed in (1) satisfies

$$\bar{v}_j = a_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{r}_{k,j} + \delta v_j, \quad \|\delta v_j\| \le c_0(m,n) u \|a_j\|, \tag{4}$$

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

$$\bar{q}_j = \bar{v}_j / \|\bar{v}_j\| + \delta q_j, \quad \|\delta q_j\| \le (m+4)u, \quad \|\bar{q}_j\|^2 \le 1 + (m+4)u. \tag{5}$$

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:

$$\bar{r}_{i,j} = \bar{q}_i^T a_j + \delta r_{i,j}, \quad |\delta r_{i,j}| \le mu \|\bar{q}_i\| \, \|a_j\|, \tag{6}$$

$$\bar{r}_{j,j} = \|\bar{v}_j\| + \delta r_{j,j}, \quad |\delta r_{j,j}| \le mu \|\bar{v}_j\|. \tag{7}$$

Summarizing (4) together with (6) and (7) for steps $j = 1, \ldots, n$ in matrix notation (for details we also refer to Daniel et al [5]), the basis vectors $\bar{Q}$ and the upper triangular factor $\bar{R}$ computed by the CGS algorithm satisfy the recurrence relation

$$A + \delta A = \bar{Q}\bar{R}, \quad \|\delta A\| \le c_1(m,n) u \|A\|, \tag{8}$$

where $c_1(m,n) = \mathcal{O}(mn^{3/2})$. The key point in the analysis of the CGS algorithm is 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 $A^T A$ perturbed by a matrix of relatively 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 $A^T A$.

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

$$\bar{R}^T\bar{R} = A^TA + E, \quad \|E\| \le c_2(m,n)u\|A\|^2, \tag{9}$$

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

Proof We begin with the formula (6) for the orthogonalization coefficient $\bar{r}_{i,j}$ in the form

$$\bar{r}_{i,j} = \bar{q}_i^T a_j + \delta r_{i,j} = (\bar{v}_i/\|\bar{v}_i\| + \delta q_i)^T a_j + \delta r_{i,j}. \tag{10}$$

Multiplying both sides of (10) by $\|\bar{v}_i\|$ and substituting $\bar{r}_{i,i}$ on the left-hand side using (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 (4) and using the identities (6) for $\bar{r}_{k,i}$, we obtain, after some manipulations, the identity

$$\begin{aligned}\bar{r}_{i,i}\bar{r}_{i,j} &= (a_i - \sum_{k=1}^{i-1}\bar{q}_k\bar{r}_{k,i} + \delta v_i)^T a_j + \Big((\delta q_i)^T a_j + \delta r_{i,j}\Big)\|\bar{v}_i\| + \bar{r}_{i,j}\delta r_{i,i} \\ &= a_i^T a_j - \sum_{k=1}^{i-1}\bar{r}_{k,i}(\bar{r}_{k,j} - \delta r_{k,j}) + (\delta v_i)^T a_j + ((\delta q_i)^T a_j + \delta r_{i,j})\|\bar{v}_i\| \\ &\quad + \bar{r}_{i,j}\delta r_{i,i}.\end{aligned} \tag{11}$$

Thus we can immediately write

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

which gives rise to the expression for the (i, j)-element in the matrix equation $\bar{R}^T\bar{R} = A^TA + E$. 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 (7), the bound on $<\|\delta v_i\|$ from (4), the bound on $\|\delta q_i < \|<$ from (5) and considering that

$$\begin{aligned}|\bar{r}_{k,i}| &\le \|\bar{q}_k\|\,\|a_i\| + |\delta r_{k,i}| \le [1 + 2(m+4)u]\,\|a_i\|, \\ \|\bar{v}_i\| &\le \|a_i\| + \sum_{k=1}^{i-1}|\bar{r}_{k,i}|\,\|\bar{q}_k\| + \|\delta v_i\| \le [n + 2c_0(m,n)u]\,\|a_i\|.\end{aligned} \tag{12}$$

Note that a much smaller bound on $\|\bar{v}_i\|$ than the one given by (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 $c_2(m,n)u\|A\|^2$ for a properly chosen polynomial $c_2(m,n)$. □

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

$$\|\bar{R}^{-1}\| \leq \frac{1}{\sigma_{min}(A)\left[1 - c_2(m,n)u\kappa^2(A)\right]^{1/2}}. \tag{13}$$

The analogy of this corollary in exact arithmetic is the fact 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 $c_2(m,n)u\kappa^2(A) < 1$, the loss of orthogonality of the vectors $\bar{Q}$ computed by the CGS algorithm is bounded by*

$$\|I - \bar{Q}^T\bar{Q}\| \leq \frac{c_3(m,n)u\kappa^2(A)}{1 - c_2(m,n)u\kappa^2(A)}, \tag{14}$$

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

Proof It follows from (8) that $(A + \delta A)^T(A + \delta A) = \bar{R}^T\bar{Q}^T\bar{Q}\bar{R}$. Substituting $A^T A$ from (9), we have

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

Assuming $c_2(m,n)u\kappa^2(A) < 1$, 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

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

Using the bounds on $\|\delta A\|$, $\|E\|$ and $\|\bar{R}^{-1}\|$ in Equations (8), (9) and (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 $A^T A$ is numerically nonsingular, i.e. $c_2(m,n)u\kappa^2(A) < 1$. In contrast, Björck [2] proved that the loss of orthogonality in MGS depends only linearly on $\kappa(A)$. For this, he assumed the numerical full rank of the matrix A, i.e. he assumed that $c(m,n)u\kappa(A) < 1$. As far as we could check, there was only one attempt to give a bound for the CGS algorithm by Kiełbasiński and Schwetlick (see [16, p.284] or [17, p.299])

$$\|I - \bar{Q}^T\bar{Q}\| \leq \min\left\{\tilde{c}(m,n)u\left[\kappa(A)\right]^{n-1}, n+1\right\}, \tag{15}$$

where the presence of the first term in the right-hand side is explained only intuitively (see also another paper of Kiełbasiński [15]), while the second term is a trivial consequence of the inequalities $\|I - \bar{Q}^T\bar{Q}\| \leq 1 + \|\bar{Q}\|_F^2 \leq 1 + n[1 + (m+4)u]$.

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. First 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 (34) 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 [5]. 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. [5] is the same as for the paper of Abdelmalek [1]. We clearly define what happens for numerically nonsingular matrices and give a Theorem 2 which proves the conjecture of Hoffmann [13]. The main motivation here is to have a proof as self-contained, modern and short as possible. 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 $c_4(m,n)u\kappa(A) < 1$ 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 $c_4(m,n)u\kappa(A) < 1$, the loss of orthogonality of the vectors $\bar{Q}$ computed by the CGS2 algorithm can be bounded as*

$$\|I - \bar{Q}^T\bar{Q}\| \le c_5(m,n)u. \tag{16}$$

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

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

$$\|\bar{Q}_{i-1}^T\bar{q}_i\| \le c_6(m,n)u, \quad i = 1,\ldots,j-1, \tag{17}$$

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 (17) is also true at step j; that is to say we want to prove that $\|\bar{Q}_{j-1}^T\bar{q}_j\| \le c_6(m,n)u$. Of particular importance for us is the result proved by Hoffmann [13, p. 343-4]. He proved that if $\|\bar{Q}_{i-1}^T\bar{q}_i\| \le c_6(m,n)u$ for $i = 1,\ldots,j$ then

$$\|I - \bar{Q}_j^T\bar{Q}_j\| \le \max_{i=1,\ldots,j}\left\{\|\bar{q}_i\|^2 - 1 + \|\bar{Q}_{i-1}^T\bar{q}_i\|\sqrt{2j}\right\} \le c_5(m,n)u, \tag{18}$$

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 (16). Note that (18) also implies that $\|\bar{Q}_{j-1}\| \leq [1+c_5(m,n)u]^{1/2}$. Similarly to (4), we first recall the results for the elementary projections (2) and (3)

$$\bar{v}_j = a_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{r}_{k,j} + \delta v_j, \quad \|\delta v_j\| \leq c_0(m,n)u\|a_j\|, \tag{19}$$

$$\bar{w}_j = \bar{v}_j - \sum_{k=1}^{j-1} \bar{q}_k \bar{s}_{k,j} + \delta w_j, \quad \|\delta w_j\| \leq c_0(m,n)u\|\bar{v}_j\|, \tag{20}$$

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

$$\bar{r}_{k,j} = \bar{q}_k^T a_j + \delta r_{k,j}, \quad \bar{s}_{k,j} = \bar{q}_k^T \bar{v}_j + \delta s_{k,j}, \quad \bar{s}_{j,j} = \|\bar{w}_j\| + \delta s_{j,j}, \tag{21}$$

$$|\delta r_{k,j}| \leq mu\|\bar{q}_k\|\,\|a_j\|, \quad |\delta s_{k,j}| \leq mu\|\bar{q}_k\|\,\|\bar{v}_j\|, \quad |\delta s_{j,j}| \leq mu\|\bar{w}_j\|. \tag{22}$$

The vector $\bar{q}_j$ comes from the normalization of the vector $\bar{w}_j$. Analogously to (5), we have

$$\bar{q}_j = \bar{w}_j/\|\bar{w}_j\| + \delta q_j, \quad \|\delta q_j\| \leq (m+4)u, \quad \|\bar{q}_j\|^2 \leq 1+(m+4)u. \tag{23}$$

The relations (19) and (20) can be added to give

$$a_j + \delta v_j + \delta w_j = \sum_{k=1}^{j-1} (\bar{r}_{k,j} + \bar{s}_{k,j})\bar{q}_k + \bar{w}_j. \tag{24}$$

Taking also into account the errors (22) and (23), the recurrence (24) for $j = 1,\dots,n$ can be rewritten into the matrix relation

$$A + \delta V + \delta W = \bar{Q}(\bar{R} + \bar{S}), \tag{25}$$

where $\delta V = (\delta v_1,\dots,\delta v_n)$ and $\delta W = (\delta w_1,\dots,\delta w_n)$. 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 (8). Actually, the possible differences can be hidden into definition of the polynomial $c_1(m,n)$. In order to prove that $\|\bar{Q}_{j-1}^T \bar{q}_j\| \leq c_6(m,n)u$, 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 $\|\bar{Q}_{j-1}^T \bar{v}_j\|/\|\bar{v}_j\|$. In the second part of the proof, a bound for the quotient $\|\bar{Q}_{j-1}^T \bar{w}_j\|/\|\bar{w}_j\|$ is given. The factors $\|a_j\|/\|\bar{v}_j\|$ and $\|\bar{v}_j\|/\|\bar{w}_j\|$ play a significant role in the proof. Assuming that A has numerical full rank, we prove a lower bound for the factor $\|a_j\|/\|\bar{v}_j\|$ proportional to the minimum singular value of A. Using this bound, we prove that the factor $\|\bar{v}_j\|/\|\bar{w}_j\|$ 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 (19) 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 (21) and (22), the quotient $\|\bar{Q}_{j-1}^T \bar{v}_j\| / \|\bar{v}_j\|$ can be bounded as

$$\frac{\|\bar{Q}_{j-1}^T \bar{v}_j\|}{\|\bar{v}_j\|} \leq [c_5(m,n) + mn(1 + (m+4)u) + c_0(m,n)]\,(1 + c_5(m,n)u)^{1/2} u \times \frac{\|a_j\|}{\|\bar{v}_j\|}. \tag{26}$$

The inequality (26) is easy to interpret. It is well known and described in many papers (e.g. [6,13,21]) that the loss of orthogonality after the first orthogonalization step (2) is proportional to the quantity $\|a_j\|/\|\bar{v}_j\|$. The next lemma provides us some control on this quantity.

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

$$\frac{\|a_j\|}{\|\bar{v}_j\|} \leq \kappa(A)\,[1 - c_7(m,n)u\kappa(A)]^{-1}, \tag{27}$$

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

Proof We consider the matrix recurrence (25) for the first $j-1$ orthogonalization steps

$$A_{j-1} + \delta V_{j-1} + \delta W_{j-1} = \bar{Q}_{j-1}(\bar{R}_{j-1} + \bar{S}_{j-1}). \tag{28}$$

Summarizing (28) with (19), we can rewrite these two relations into the matrix relation

$$A_j + \left[\delta V_{j-1} + \delta W_{j-1}, \delta v_j - \bar{v}_j\right] = \bar{Q}_{j-1}\left[\bar{R}_{j-1} + \bar{S}_{j-1}, \bar{r}_j\right], \tag{29}$$

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. [12, p. 123] or [10, 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

$$\sigma_{min}(A) \leq \sigma_{min}(A_j) \leq \|\Delta_j\| \leq \sqrt{\|\delta V_{j-1}\|^2 + \|\delta W_{j-1}\|^2 + \|\delta v_j\|^2 + \|\bar{v}_j\|^2}. \tag{30}$$

We are now going to use the bounds on the norms of the matrices δV_{j-1}, δW_{j-1}, the bound (19) on the vector δv_j and an argumentation similar to (12). Assuming $c_7(m,n)u\kappa(A) < 1$ (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

$$\|\bar{v}_j\| \geq \sigma_{min}(A)(1 - c_7(m,n)u\kappa(A)). \tag{31}$$

□

The bound (31) shows that, under the assumption that A has numerical full rank (i.e. assuming $c_7(m,n)u\kappa(A) < 1$), the norm $\|\bar{v}_j\|$ is bounded by the minimal singular value of A. We note that the result (31) is analogous to the exact arithmetic bound $\|v_j\| \geq \sigma_{min}(A)$. 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 (26) and (27), bounded by

$$\begin{aligned}
&\frac{\|\bar{Q}_{j-1}^T\bar{v}_j\|}{\|\bar{v}_j\|}\\
&\quad\leq \frac{[c_5(m,n) + mn(1+(m+4)u) + c_0(m,n)]\,(1+c_5(m,n)u)^{1/2}u\kappa(A)}{1 - c_7(m,n)u\kappa(A)}\\
&\quad= c_8(m,n)u\kappa(A).
\end{aligned} \tag{32}$$

We are now ready to start the second step to prove that $\|\bar{Q}_{j-1}^T\bar{q}_j\| \leq c_6(m,n)u$. We proceed similarly as in the first part. Using the derived bound (32), 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 $\|\bar{v}_j\|/\|\bar{w}_j\|$. Using the relation for the local error in the second iteration step (20), it can be bounded as follows

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

Thus, under the assumption that

$$\begin{aligned}
c_9(m,n)u\kappa(A) = &\left[c_8(m,n)\kappa(A)(1+c_5(m,n)u)^{1/2} + mn(1+(m+4)u)\right.\\
&\left.+ c_0(m,n)\right]u < 1,
\end{aligned} \tag{33}$$

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

$$\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|} \leq [1 - c_9(m,n)u\kappa(A)]^{-1}. \tag{34}$$

The upper bound (34) shows that if we slightly strengthen the assumption (33), the factor $\|\bar{v}_j\|/\|\bar{w}_j\|$ becomes very close to 1, which 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 (34). In his analysis, Abdelmalek needs that $(j-2)^2\|\bar{Q}_{j-1}^T\bar{v}_j\|/\|\bar{w}_j\| \le 1$, a statement that he expects to hold in most practical cases. Indeed, this criterion can be rewritten as $(j-2)^2(\|\bar{Q}_{j-1}^T\bar{v}_j\|/\|\bar{v}_j\|)(\|\bar{v}_j\|/\|\bar{w}_j\|) \le 1$ and it can been seen from (32) and (34) that Abdelmalek's assumption is met under a clear assumption on the numerical rank of A. From (20), 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 (22) and (32) leads to

$$\frac{\|\bar{Q}_{j-1}^T\bar{w}_j\|}{\|\bar{w}_j\|} \le [c_5(m,n)c_8(m,n)u\kappa(A) + mn(1+(m+4)u) + c_0(m,n)]\,u \times(1+(m+4)u)^{1/2}\frac{\|\bar{v}_j\|}{\|\bar{w}_j\|}. \tag{35}$$

Consequently, using (26), (35) and (34), and remarking that $\|\bar{Q}_{j-1}^T\bar{q}_j\| \le \|\bar{Q}_{j-1}^T \bar{w}_j\|/\|\bar{w}_j\| + \|\bar{Q}_{j-1}^T\delta q_j\|$, we can write

$$\|\bar{Q}_{j-1}^T\bar{q}_j\| \le [c_5(m,n)c_8(m,n)u\kappa(A) + mn(1+(m+4)u) + c_0(m,n)] \times\frac{[1+(m+4)u]^{1/2}}{1-c_9u\kappa(A)}u + (m+4)u[1+c_5(m,n)u]^{1/2}. \tag{36}$$

Now, let us assume that $[1-c_9(m,n)u\kappa(A)]^{-1} \le 2$, $mn^{3/2}u \ll 1$ and $c_5(m,n)$ $c_8(m,n)u\kappa(A) \le 1$. Then $1+(m+4)u \le 2$, $1+c_5(m,n)u \le 2$ and we have

$$\|\bar{Q}_{j-1}^T\bar{q}_j\| \le 2\sqrt{2}\,[1+2mn+c_0(m,n)]\,u + \sqrt{2}(m+4)u = c_6(m,n)u, \tag{37}$$

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

$$c_4(m,n)u\kappa(A) < 1, \tag{38}$$

where $c_4(m,n) = \mathcal{O}(m^2n^3)$. We are now able to conclude the proof by induction. If the induction assumption (17) is true at the step $j-1$, under assumption (38), the statement is true at the step j. Consequently we have at the step j the bound (18). For the last step $j = n$, it follows that

$$\|I - \bar{Q}^T\bar{Q}\| \le c_5(m,n)u, \tag{39}$$

which proves Theorem 2. □

Finally we illustrate our theoretical results. We consider a 200 × 100 matrices A_k generated by computing $A_k = U\Sigma_k V^T$, where U and V are randomly chosen

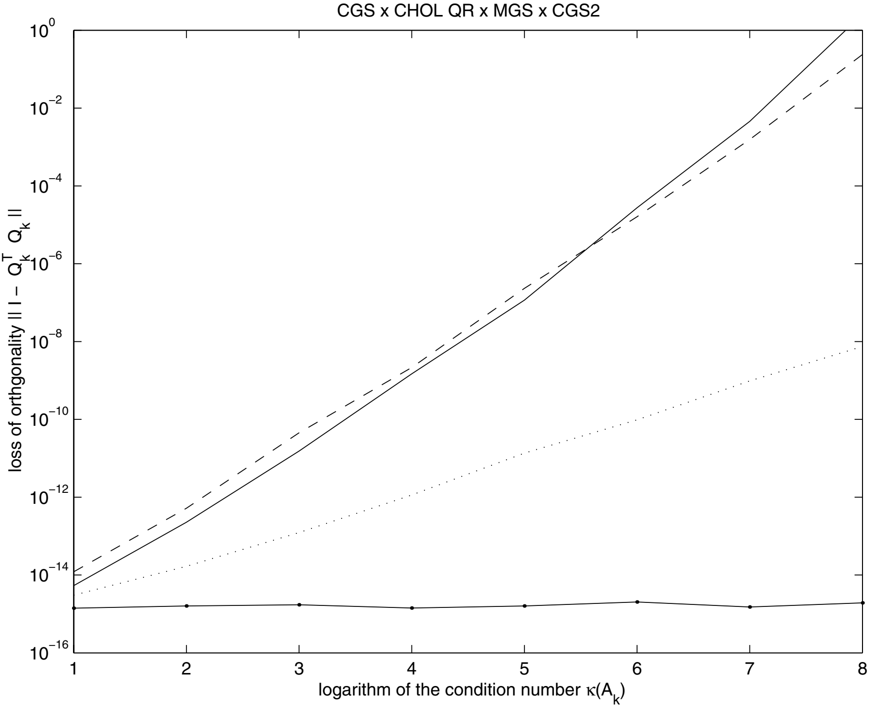

Fig. 1 The loss of orthogonality in the CGS (solid line), MGS (dotted line), Cholesky QR (dashed line) and CGS2 (solid line with dots) algorithms (measured by the corresponding $\|I - \bar{Q}_k^T \bar{Q}_k\|$) with respect to the decadic logarithm of the condition number $\kappa(A_k) = 10^k$

orthonormal matrices and Σ_k contains the singular values of A_k uniformly distributed between 1 and 10^{-k} for $k = 1, \ldots, 8$. For details we refer to Example 1.15 in [23]. In Figure 1, we have plotted the loss of orthogonality (measured by $\|I - \bar{Q}_k^T \bar{Q}_k\|$) between the vectors $\bar{Q}_k$ computed by CGS (solid line), MGS (dotted line), Cholesky QR (dashed line) and the CGS2 (solid line with dots) versus the logarithm of the condition number $\kappa(A_k) = 10^k$, where $k = 1, \ldots, 8$. All experiments are performed with MATLAB with $u = 2.2204e - 16$. It is clear from Figure 1 that the loss of orthogonality in CGS is proportional to the square of the condition number of A_k (while the dependence is only linear for MGS), and thus the bound (14) is meaningful. In addition the CGS algorithm and the Cholesky QR algorithm (where Q is given as $Q = A_k R^{-1}$ and R is the computed Cholesky factor of the matrix $A_k^T A_k$) deliver qualitatively the same results. It is also clear from Figure 1 that two iteration steps of the CGS2 algorithm are enough for preserving the orthogonality of the computed vectors close to the machine level; agreeing with the theoretical results developed in this section.

4 Conclusions and remarks

In this paper, we give a bound on the loss of orthogonality of the CGS algorithm. We prove 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. The bound is similar to that for MGS except that the loss of orthogonality in MGS depends linearly on $\kappa(A)$ and the assumption depends on the numerical full rank 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 prove that the orthogonality of the vectors computed by the CGS2 algorithm is close to the machine precision level. That is, exactly two iteration steps are already enough when full orthogonality is required requested and when the algorithm is applied to a (numerically) independent initial set of column vectors. This result extends the ones of Abdelmalek [1], Daniel et al [5], Kahan and Parlett [19] and Hoffmann [13].

Acknowledgements The authors would like to thank for the fruitful discussions, useful comments and help to Å. Björck, A. Kiełbasiński, A. Smoktunowicz, P. Tichý and K. Ziętak as well as the anonymous referees.

References

1. Abdelmalek, N.: Round off error analysis for Gram-Schmidt method and solution of linear least squares problems. BIT **11**, 345–368 (1971)
2. Björck, Å.: Solving linear least squares problems by Gram-Schmidt orthogonalization. BIT **7**, 1–21 (1967)
3. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia, PA, 1996
4. Björck, Å., Paige, C.: Loss and Recapture of Orthogonality in the Modified Gram-Schmidt Algorithm. SIAM J. Matrix Anal. Appl **13**(1), 176–190 (1992)
5. Daniel, J.W., Gragg, W.B., Kaufman, L., Stewart, G.W.: Reorthogonalization and Stable Algorithms for Updating the Gram-Schmidt QR Factorization. Math. Comp **30**, 772–795 (1976)
6. Dax, A.: A modified Gram-Schmidt algorithm with iterative orthogonalization and pivoting. Linear Alg. and its Appl **310**, 25–42 (2000)
7. Frayssé, V., Giraud, L., Kharraz-Aroussi, H.: On the influence of the orthogonalization scheme on the parallel performance of GMRES. EUROPAR'98 Parallel Processing, Springer **1470**, 751–762 (1998)
8. Frank, J., Vuik, C.: Parallel implementation of a multiblock method with approximate subdomain solution. Appl. Num. Math **30**, 403–423 (1999)
9. Giraud, L., Langou, J.: When modified Gram-Schmidt generates a well-conditioned set of vectors. IMA Journal of Numerical Analysis, **22**(4), 521–528
10. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd ed. John Hopkins University Press, Baltimore, MD, 1996
11. Greenbaum, A., Rozložník, M., Strakoš, Z.: Numerical behaviour of the modified Gram-Schmidt GMRES implementation. BIT **37**(3), 706–719 (1997)
12. Higham, N.: Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia, PA, 2nd ed., 2002
13. Hoffmann, W.: Iterative Algorithms for Gram-Schmidt Orthogonalization. Computing **41**, 335–348 (1989)
14. Jalby, W., Philippe, B.: Stability analysis and improvement of the block Gram-Schmidt algorithm. SIAM J. Sci. Stat. Comput **12**, 1058–1073 (1991)
15. Kiełbasiński, A.: Numerical analysis of the Gram-Schmidt orthogonalization algorithm (Analiza numeryczna algorytmu ortogonalizacji Grama-Schmidta). (In Polish). Roczniki Polskiego Towarzystwa Matematycznego, Seria III: Matematyka Stosowana II, 15–35 (1974)
16. Kiełbasiński, A., Schwetlick, H.: Numerische lineare Algebra. Eine computerorientierte Einführung. (In German). Mathematik für Naturwissenschaft und Technik 18. Deutscher Verlag der Wissenschaften, Berlin, 1988

17. Kiełbasiński, A., Schwetlick, H.: Numeryczna algebra liniowa. (In Polish). Second edition. Wydawnictwo Naukowo-Techniczne, Warszawa, 1994
18. Lehoucq,R.B., Salinger, A.G.: Large-Scale Eigenvalue Calculations for Stability Analysis of Steady Flows on Massively Parallel Computers. Int. J. Numerical Methods in Fluids **36**, 309–327 (2001)
19. Parlett, B.N.: The Symmetric Eigenvalue Problem. Englewood Cliffs, N.J., Prentice-Hall, 1980
20. Rice, J.R.: Experiments on Gram-Schmidt Orthogonalization. Math. Comp. **20**, 325–328 (1966)
21. Rutishauser, H.: Description of Algol 60. Handbook for Automatic Computation, Vol. 1a. Springer Verlag, Berlin, 1967
22. Schmidt, E.: Über die Auflösung linearer Gleichungen mit unendlich vielen Unbekannten. (In German.) Rend. Circ. Mat. Palermo. Ser. 1, **25**, 53–77 (1908)
23. Stewart, G.W.: Matrix Algorithms. Volume I: Basic Decompositions. SIAM, Philadelphia, PA, 1998
24. Wilkinson, J.H.: Modern error analysis. SIAM rev., **13**(4), 548–569 (1971)