CERFACS **Technical Report No. TR/PA/02/33**

Luc Giraud · Julien Langou · Miroslav Rozložník

On the round-off error analysis of the Gram-Schmidt algorithm with reorthogonalization

Abstract. In this paper we analyse the numerical behavior of the Gram-Schmidt orthogonalization process with reorthogonalization. Assuming numerical nonsingularity of the matrix we prove that two steps of iterative Gram-Schmidt process are enough for preserving the orthogonality of computed vectors close to the machine precision level. We give a rounding error analysis of classical reorthogonalization and modified Gram-Schmidt algorithm with (exactly one) reorthogonalization and relate our results to the approach used in the Kahan-Parlett "twice is enough" algorithm as well as to results shown by Abdelmalek, Daniel et al, Hoffmann and others.

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 or Givens rotations (see e.g. [3, 11, 13, 23]). In this paper we focus on the Gram-Schmidt (GS) orthogonalization procedure [22] which also produces a QR factorization of the matrix A. Several computation variants of the Gram-Schmidt algorithm exist - each formulation leading to a different numerical behavior of the associated scheme.

One of the first methods for successive orthogonalization of the columns of A is the classical Gram-Schmidt algorithm (CGS) [3]. It is well known, and confirmed by extensive numerical experiments, that this technique may produce a set of vectors which is far from orthogonal and sometimes the orthogonality can be completely lost [2, 18]. Rearrangement of the CGS scheme leads to the modified Gram-Schmidt algorithm (MGS) (see e.g. [3, 23]) with much better numerical properties [18, 24]. Björck [2] showed that for a numerically nonsingular matrix A the loss of orthogonality in MGS occurs in a predictable way and it can be bounded by a term proportional to the condition number $\kappa(A)$ times the machine precision ε. Therefore, the loss of orthogonality of computed vectors is close to machine precision level only for well-conditioned matrices, while for ill-conditioned matrices it can be much larger. The results on MGS were reinforced and extended by Paige and Björck [4] who showed that despite the loss of orthogonality in Q the computed upper triangular factor R in MGS is numerically as good as the computed R from the Householder or Givens QR factorization [3, 13]. This result also indicates that the possible loss of orthogonality in Q does not necessarily imply that the orthogonalization techniques like Householder or Givens QR orthogonalization, which deliver almost orthogonal set of vectors are superior to MGS when applied for some problems. E.g. it was shown by Björck (see e.g. [3, 13]) that the MGS method can be used to solve least squares problems and that the algorithm is backward-stable. Another example of this type is the GMRES method when used together with the MGS Arnoldi implementation. Surprisingly, it was shown in [12] that the

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

Julien Langou: CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France. e-mail: Julien.Langou@cerfacs.fr

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

linear independence of the computed (Arnoldi) basis vectors, not their orthogonality, is important for the convergence of GMRES and that the MGS GMRES performs as well as GMRES with the Householder implementation of the Arnoldi process [12]. On the other hand, in some other applications it is important to produce vectors Q so that their orthogonality is kept close to the machine precision. An example of such problem can be the eigenvalue problem solved by the Arnoldi method, where the orthogonality is essential to get an accurate projection onto the space spanned by the computed vectors [21].

The problem of computing Q accurately is sometimes called as the orthogonal basis problem [3]. Then it is necessary to iterate GS process twice or several times. Depending on the number of iteration steps, the name of the resulting scheme changes. GS with reorthogonalization (GS2) is used for the two-step algorithms (i.e. the algorithms where the orthogonalization of a current vector against previously computed set is performed exactly twice), while the algorithms with more steps are called iterative GS (GSI). First experiments with iterative versions of GS were performed by Rice [18] who, in order to get sufficiently orthogonal vectors, reports experiments from 3 up-to 6 iteration steps for CGSI, while he needed just 2 iterations for MGSI when applied on rank-deficient problems. A first rounding error analysis for the CGS2 algorithm was given by Abdelmalek [1] who considered exactly two iteration steps. For showing that the scheme produces sufficiently orthogonal set of vectors he needed an assumption on diagonal elements of the computed upper triangular factors at each orthogonalization step.

In some other algorithms, one can find a selective reorthogonalization criterion. The reorthogonalization is not systematic, the decision whether to reorthogonalize or not is made upon a criterion which describes the quality of the computed basis. In these algorithms, the maximal number of loops can also be fixed to 2, 3 or either ∞; ∞ means that GS is iterated until the criterion is satisfied. This suggests that the criterion has to be reliable, a bad criterion could be satisfied in a case where reorthogonalization might be necessary. The most commonly used criterion for checking the "a priori" quality of the computed basis is certainly the one introduced by Rutishauser [20]. This criterion is dependent of a single parameter σ. We call the resulting algorithms CGSI(σ) or MGSI(σ) and we only focus here on such selective reorthogonalization algorithms. Rutishauser [20] showed an example of two vectors where one reorthogonalization step was enough for full orthogonality and recommended some choices of the parameter σ. Daniel, Gragg, Kaufman and Stewart [6] presented a rounding error analysis of the CGSI(σ) algorithm essentially showing that under certain technical assumptions either the iterative 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 latter essentially implies the case of rank-deficient (singular) problem. Daniel et al. also illustrated experimentally on rank-deficient problems that the iterative Gram-Schmidt algorithm may need more than two iteration steps when the condition number of the matrix A exceeds $1/\varepsilon$. Theoretical analysis of the situation with two vectors, attributed to Kahan, was given by Parlett in his book [16]. He analysed the so called "twice is enough" algorithm. Parlett showed for the case of two vectors that two steps are enough for obtaining the orthogonality level prescribed by σ (Cf. Section 2). CGSI(σ) and MGSI(σ) algorithms were further thoroughly analysed by Hoffmann [14] who, taking into account experimental results, conjectured a bound on the orthogonality of the computed vectors by CGSI(σ) in terms of the parameter σ chosen in the criterion (as Kahan and Parlett for 2 vectors). From his extensive experiments he concluded that a third iteration never occurred for both CGSI(σ) and MGSI(σ) algorithms but theoretical proof of this fact remained still an open question. We note that the problems he used in experiments were all (numerically) nonsingular.

A slightly different variant of MGS2 has recently been analysed in [10]. This variant (called as TMGS) corresponds to the MGS2 scheme with the exception in the opposite order of two loops. In MGS2, for each vector, we orthogonalize this vector against the previous set of vectors then immediately reorthogonalize the obtained vector against the (same) previous set of vectors. In TMGS, we orthogonalize all the vectors with MGS then reorthogonalize the obtained set a second time with MGS. Using the results of Björck [2] it was shown in [10] that for numerically nonsingular matrices the second application of the MGS orthogonalization scheme onto resulting set of vectors in the first application of MGS leads to very accurate orthogonal

basis. When the matrix A is not known in advance (e.g. in the context of Krylov subspace methods), restarting the MGS algorithm from the "scratch" requires to keep all vectors computed in both two loops. This is not the case with GS2. As a consequence, we see that the memory cost of TMGS is around twice as much as the memory cost of GS2. This is one of the reason to prefer GS2.

Similarly to Rice [18] or Abdelmalek [1], in this paper we analyse the GS2 algorithm, where each orthogonalization step is performed exactly twice. The orthogonal matrix Q is constructed successively column-by-column so that for each $j = 1,\ldots,n$ we have $Q_j = (q_1,\ldots q_j)$ and $span(q_1,\ldots,q_j) = span(a_1,\ldots,a_j)$. For the description we use the following notation. We start with the matrix $A_j = (a_1,\ldots,a_j)$ and for $i = 1,2$ we generate successively the vectors $A_j^{(i)} = (a_1^{(i)},\ldots,a_j^{(i)})$ such that

$$a_j^{(1)} = (I - Q_{j-1}Q_{j-1}^T)a_j, \tag{1.1}$$

$$a_j^{(2)} = (I - Q_{j-1}Q_{j-1}^T)a_j^{(1)}. \tag{1.2}$$

The new vector q_j is then just a result of normalization of the vector $a_j^{(2)}$ and it is given as $q_j = a_j^{(2)}/\|a_j^{(2)}\|$. If the elements of the triangular factor $R_j = (r_1,\ldots,r_j)$ are wanted too, then for each $j = 1,\ldots,n$ we start with the initial column $r_j^{(1)} = Q_{j-1}^T a_j$ and by using $r_j^{(2)} = r_j^{(1)} + Q_{j-1}^T a_j^{(1)}$ we end up with $r_j = ((r_j^{(2)})^T, \|a_j^{(2)}\|)^T$.

The organization of the paper is as follows. In Section 2 we recall Kahan's analysis for two vectors and discuss the relation of the behavior of the "twice is enough" algorithm to the linear independence of the initial vectors. In particular we show that the Case III in the algorithm corresponds to the rank-deficient problem. Section 3 is devoted to our main result on the CGS2 method. Assuming numerically nonsingular matrix A we show that two iterations are enough to achieve the orthogonality of computed vectors on the level of small multiple of the machine precision. The main key here is to show that the norm of the vector resulting from the first projection step (1.1) cannot be infinitely small, indeed it is linked to the minimal singular value of the matrix A. Based on this observation and using the results of Hoffmann [14] we conclude our analysis with the bound on the loss of orthogonality for the computed vectors. The implications of the main result and related questions are discussed in Section 4. We discuss several issues, including the difference between the CGS and the MGS variants, the applicability of our analysis and its relation to other works and approaches.

Throughout the paper, $\|X\|$ denotes the 2-norm, $\sigma_{min}(X)$ the minimal singular value and $\|X\|_F$ the Frobenius norm of the matrix X; $\|x\|$ denotes the Euclidean norm of a vector x. The condition number of X is denoted by $\kappa(X)$. For distinction, we denote quantities computed in finite precision arithmetic using an upper-bar. We assume that usual rules of a well designed floating-point arithmetic hold, and use the notation $fl(\cdot)$ for the computed result of some expression. As it was already mentioned before, the machine precision is denoted by ε. Constant factors ξ_j, $j = 1,2,\ldots$ are introduced in several places. They are constants, independent of the problem parameters like m, n or $\kappa(A)$ and the machine precision ε, but they do depend on the details of computer arithmetic.

2. Twice-is-enough algorithm of Kahan and Parlett

A complete analysis of the situation with two vectors was given by Parlett [16, pp. 105-109] who described the "twice is enough" algorithm and presented the analysis due to W. Kahan. Their algorithm is not GS2, there is a selective reorthogonalization criterion. This point does not represent any difficulty and the results of Kahan and Parlett can be translated also for the case of GS2 with two vectors.

We consider two given vectors a_1 and a_2, where the second vector a_2 is orthogonalize against the normalized first vector $q_1 = a_1/\|a_1\|$ in two steps

$$a_2^{(1)} = (I - q_1 q_1^T)a_2, \tag{2.1}$$

$$a_2^{(2)} = (I - q_1 q_1^T)a_2^{(1)}. \tag{2.2}$$

The result of (2.2) is then normalized as $q_2 = a_2^{(2)}/\|a_2^{(2)}\|$. It is clear that in exact arithmetic it follows that $a_2^{(2)} = a_2^{(1)}$. However, in finite precision arithmetic, one computes only approximations to (2.1) and (2.2), which we denote as $\bar{q}_2^{(1)} = fl((I - q_1 q_1^T)a_2)$ and $\bar{q}_2^{(2)} = fl((I - q_1 q_1^T)\bar{q}_2^{(1)})$ and which satisfy

$$\bar{q}_2^{(1)} = (I - q_1 q_1^T)a_2 + \eta_2^{(1)}, \quad \|\eta_2^{(1)}\| \leq \phi(m,n)\varepsilon\|a_2\| \tag{2.3}$$

$$\bar{q}_2^{(2)} = (I - q_1 q_1^T)\bar{q}_2^{(1)} + \eta_2^{(2)}, \quad \|\eta_2^{(2)}\| \leq \phi(m,n)\varepsilon\|\bar{q}_2^{(1)}\|, \tag{2.4}$$

where $\phi(m,n)$ is a low degree polynomial in m and n. This polynomial is universally used throughout paper and it is specified in Section 3. Of course, in this Section, we have $n = 2$. Furthermore, the normalization of vectors is not performed exactly in finite precision arithmetic. Since these errors do not bring any additional problems in our analysis, so throughout this paper, for the sake of simplicity and brevity, we assume that all vectors are exactly normalized and we do not distinguish between vectors $fl(\bar{q}_j^{(2)}/\|\bar{q}_j^{(2)}\|)$ and $\bar{q}_j^{(2)}/\|\bar{q}_j^{(2)}\|$ and everywhere we use the notation $q_j = \bar{q}_j^{(2)}/\|\bar{q}_j^{(2)}\|$ for $j = 1,\ldots,n$.

The "twice is enough" algorithm due to Kahan and Parlett can be summarized as follows:

"twice is enough" algorithm of Kahan and Parlett:
$q_1 = a_1/\|a_1\|$
$\bar{q}_2^{(1)} = fl(a_2 - (a_2, q_1)q_1)$
if $\frac{\|a_2\|}{\|\bar{q}_2^{(1)}\|} \leq \sigma$ then
  Case I: $q_2 = \bar{q}_2^{(1)}/\|\bar{q}_2^{(1)}\|$
else
  $\bar{q}_2^{(2)} = fl(\bar{q}_2^{(1)} - (\bar{q}_2^{(1)}, q_1)q_1)$
  if $\frac{\|\bar{q}_2^{(1)}\|}{\|\bar{q}_2^{(2)}\|} \leq \sigma$ then
    Case II: $q_2 = \bar{q}_2^{(2)}/\|\bar{q}_2^{(2)}\|$
  else
    Case III: $\bar{q}_2^{(3)} = 0$ and $q_2 = 0$
  end if
end if

The parameter σ is a fixed real number chosen in the interval $1/(0.83-\varepsilon) \leq \sigma \leq 0.83/\varepsilon$ which describes the criterion on reorthogonalization. On the one hand, it is clear that larger σ implies that reorthogonalization will be rarer. On the other hand, it was shown in [16] that it determines the quality of orthogonality of computed vectors. Indeed, the vector q_2 computed by the "twice is enough" algorithm satisfies the relation

$$|q_1^T q_2| \leq \sigma\phi(m,n)\varepsilon. \tag{2.5}$$

The bound on orthogonality is thus for larger values of σ less good. We note that the bound (2.5) is valid for all three Cases (Case I-III) of the algorithm. For Case II, the equation (2.5) can be easily transferred to GS2, since the scheme is numerically equivalent to (2.3) - (2.4). In GS2, there is no criterion. The result (2.5) holds for all σ; the best value to choose σ is certainly the smallest one, therefore we can take $\sigma = 1/(0.83-\varepsilon)$.

Moreover, Kahan and Parlett have shown that for all three cases we have

$$\bar{q}_2^{(i)} = (I - q_1 q_1^T)a_2 + \delta_2^{(i)}, \quad \|\delta_2^{(i)}\| \leq (1 + 1/\sigma)\phi(m,n)\varepsilon\|a_2\|, \quad i = 1,2,3. \tag{2.6}$$

We conclude that if Case III happens in "twice is enough" algorithm, $\bar{q}_2^{(3)} = 0$ implies $\delta_2^{(3)} = -a_2 + (a_1^T a_2)/(a_1^T a_1)a_1$ and therefore

$$\sigma_{min}(A) \leq \left\| \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \frac{(a_1^T a_2)}{(a_1^T a_1)} \\ -1 \end{pmatrix} \right\| \leq \|\delta_2^{(3)}\| \leq (1 + 1/\sigma)\phi(m,n)\varepsilon\|a_2\|. \tag{2.7}$$

Since $\|a_2\| \leq \|A\|$ and $1/\sigma \leq 0.83 - \varepsilon$ it is clear that $1.83\phi(m,n)\varepsilon\kappa(A) \geq 1$. This inequality clearly implies that the Case III occurs only in the numerically rank deficient case, when the vectors a_1 and a_2 are (almost) linearly dependent. In other words, if we assume

$$\phi(m,n)\varepsilon\kappa(A) < 1$$

then Case III never happens and Kahan and Parlett's results can be interpreted as

$$\|I - Q^T Q\| \leq \sigma\phi(m,n)\varepsilon. \tag{2.8}$$

In our paper we generalize this result for n column vectors of the matrix A. We will show that under assumption of numerical nonsingularity of the matrix $A = (a_1, \ldots, a_n)$ which can be written in the form

$$\zeta(m,n)\varepsilon\kappa(A) < 1, \tag{2.9}$$

the orthogonality of vectors $Q = (q_1, \ldots, q_n)$ computed by the CGS2 algorithm can be bounded as

$$\|I - Q^T Q\| \leq \alpha(m,n)\varepsilon, \tag{2.10}$$

where $\zeta(m,n)$ and $\alpha(m,n)$ are some sufficiently chosen low degree polynomials in m and n. Indeed we show that two iterations steps are enough for ensuring the orthogonality on the level of a small multiple of machine precision, when we apply the CGS (or MGS) process on a set of numerically nonsingular vectors.

3. Numerical nonsingularity and orthogonality of computed vectors

We first recall well known results obtained by Daniel et al. [6] (see also [3]) who analysed the elementary projections (1.1) and (1.2) and gave the formulae for the computed projections $\bar{q}_j^{(1)}$ and $\bar{q}_j^{(2)}$ in the form

$$\bar{q}_j^{(1)} = (I - Q_{j-1}Q_{j-1}^T)a_j + \eta_j^{(1)}, \quad \|\eta_j^{(1)}\| \leq \phi(m,n)\varepsilon\|a_j\|, \tag{3.1}$$

$$\bar{q}_j^{(2)} = (I - Q_{j-1}Q_{j-1}^T)\,\bar{q}_j^{(1)} + \eta_j^{(2)}, \quad \|\eta_j^{(2)}\| \leq \phi(m,n)\varepsilon\|\bar{q}_j^{(1)}\|, \tag{3.2}$$

where $\phi(m,n) = \xi_1 mn$. We point out that the bounds (3.1) and (3.2) could be refined for every j in terms of polynomials in m and j ($j \leq n \leq m$) but like for the case of two vectors (see Section 2) we use for all local errors a general bound with polynomial $\phi(m,n)$ only in terms of m and n.

In the following we will use an incremental approach similar to the approach used by Hoffmann [14]. We assume that at step $j-1$

$$\|I - Q_{j-1}^T Q_{j-1}\| \leq \alpha_{j-1}(m,n)\varepsilon, \tag{3.3}$$

with

$$\alpha_{j-1}(m,n) = \sqrt{2(j-1)}\,\omega(m,n) \quad \text{and} \quad \omega(m,n) = 2\sqrt{n}\,\phi(m,n). \tag{3.4}$$

Our goal is to show that the vector $q_j = \bar{q}_j^{(2)}/\|\bar{q}_j^{(2)}\|$ computed at the j-th step of the CGS2 algorithm satisfies the bound

$$\|Q_{j-1}^T q_j\| \leq \omega(m,n)\varepsilon. \tag{3.5}$$

The idea is that if (3.5) is verified then we can apply Theorem 3 of Hoffmann [14, p. 343] showing that the loss of orthogonality of vectors Q_j can be at the step j bounded as

$$\begin{aligned} \|I - Q_j^T Q_j\| &\leq \frac{1}{2}\left(\alpha_{j-1}(m,n) + \sqrt{\alpha_{j-1}(m,n)^2 + 4\omega(m,n)^2}\right)\varepsilon \\ &\leq \sqrt{2j}\,\omega(m,n)\varepsilon = \alpha_j(m,n)\varepsilon, \end{aligned} \tag{3.6}$$

which proves the induction statement at step j.

First of all, it is easy to see that for $j = 1$ the statement of (3.6) holds with $\alpha_0(m,n) = 0$. The proof is divided into two parts: in the first part we analyse the orthogonality of the vector $\bar{q}_j^{(1)}$ with respect to the column space of the matrix Q_{j-1} and give a bound for $\omega_j^{(1)} = \frac{\|Q_{j-1}^T \bar{a}_j^{(1)}\|}{\|\bar{a}_j^{(1)}\|}$. Based on that we give in the second part of the proof a bound for the quotient $\omega_j^{(2)} = \frac{\|Q_{j-1}^T \bar{a}_j^{(2)}\|}{\|\bar{a}_j^{(2)}\|}$. In our analysis we will make sure that $\|\bar{q}_j^{(1)}\| \neq 0$ and $\|\bar{q}_j^{(2)}\| \neq 0$ for every step $j = 1,\ldots,n$. Important role in the analysis will be played by the factors $\sigma_j^{(1)} = \frac{\|a_j\|}{\|\bar{a}_j^{(1)}\|}$ and $\sigma_j^{(2)} = \frac{\|a_j^{(1)}\|}{\|\bar{a}_j^{(2)}\|}$ which appear also in the "twice is enough" algorithm of Kahan and Parlett where their magnitude determines decision on further reorthogonalization. Here we show that, assuming numerical nonsingularity of initial column vectors in the form (2.9), the factor $\sigma_j^{(1)}$ cannot be infinitely large and using this fact we prove that $\sigma_j^{(2)}$ must be necessarily very close to 1. This is the main reason why two iterations of the CGS process are enough for preserving the orthogonality of computed vectors on the level close to the machine precision.

Multiplication of the expression (3.1) from left by Q_{j-1}^T leads to the identity

$$Q_{j-1}^T \bar{q}_j^{(1)} = (I - Q_{j-1}^T Q_{j-1}) Q_{j-1}^T a_j + Q_{j-1}^T \eta_j^{(1)}.$$

Dividing its norm by the norm of the computed vector $\bar{q}_j^{(1)}$ we obtain the inequality

$$\frac{\|Q_{j-1}^T \bar{q}_j^{(1)}\|}{\|\bar{q}_j^{(1)}\|} \leq \|I - Q_{j-1}^T Q_{j-1}\| \frac{\|Q_{j-1}^T a_j\|}{\|\bar{q}_j^{(1)}\|} + \frac{\|Q_{j-1}^T \eta_j^{(1)}\|}{\|a_j\|} \frac{\|a_j\|}{\|\bar{q}_j^{(1)}\|}.$$

The quotient $\omega_j^{(1)}$ can be then bounded as

$$\omega_j^{(1)} \leq \left[\|I - Q_{j-1}^T Q_{j-1}\| + \frac{\|\eta_j^{(1)}\|}{\|a_j\|} \right] \|Q_{j-1}\| \sigma_j^{(1)}. \tag{3.7}$$

Since the columns of the matrix Q_{j-1} are exactly normalized its norm can be bounded by $\|Q_{j-1}\| \leq \sqrt{j-1}$ for every $j = 1,\ldots,n$. Therefore for the term $\|Q_{j-1}\|$, wherever it appears, we will use the bound $\sqrt{n}$. Considering the induction assumption (3.3) and the bound (3.1), the inequality (3.7) can be using the fact that $\phi(m,n) \leq \omega(m,n)$ further rewritten as

$$\omega_j^{(1)} \leq \left[\sqrt{2(j-1)}\omega(m,n) + \phi(m,n) \right] \sqrt{n}\varepsilon\sigma_j^{(1)} \leq 2n\omega(m,n)\varepsilon\sigma_j^{(1)}. \tag{3.8}$$

The inequality (3.8) is easy to interpret. It is well known and described in many papers that if $\sigma_j^{(1)}$ is large then we can observe a severe loss of orthogonality after first orthogonalization step (1.1) in the algorithm. As it was already pointed out in previous Section, its actual value forms a criterion for further reorthogonalization in several algorithms. Our approach is different. Instead of using a criterion on reorthogonalization and an uncertain number of iterations we perform exactly two iterations, give an explicit upper bound for the factor $\sigma_j^{(1)}$ and use this bound for showing that the second iteration step is already enough for preserving the orthogonality on the level close to machine precision.

From the rounding error analysis of the CGS process (see e.g. Daniel et al. [6]) it follows that the vectors Q_{j-1}, the upper triangular matrix $\bar{R}_{j-1}$ and the vector $\bar{q}_j^{(1)}$ computed after j steps of the CGS2

algorithm satisfy

$$A_{j-1}+E_{j-1}=Q_{j-1}\bar{R}_{j-1},\quad \|E_{j-1}\|_F\le\psi(m,n)\varepsilon\|A_{j-1}\|_F, \tag{3.9}$$

$$a_j+\delta_j^{(1)}-\bar{q}^{(1)}=Q_{j-1}\bar{r}_j^{(1)},\quad \|\delta_j^{(1)}\|\le\psi(m,n)\varepsilon\|a_j\|, \tag{3.10}$$

where $\psi(m,n)=\xi_2 n^{\frac{3}{2}}$. Summarizing (3.9) and (3.10) we can write these relations in a matrix form and obtain the formula

$$A_j+\Delta_j=Q_{j-1}\bar{R}_{j-1,j}, \tag{3.11}$$

where $\bar{R}_{j-1,j}$ is a $(j-1)\times j$ matrix with $[\bar{R}_{j-1},\bar{r}_j^{(1)}]$ and the perturbation matrix is defined as $\Delta_j=[E_{j-1},\delta_j^{(1)}-\bar{q}^{(1)}]$. We remark that the matrix $Q_{j-1}\bar{R}_{j-1,j}$ is $(j-1)$-rank matrix. Therefore the matrix $A_j+\Delta_j$ is singular, whereas we have assumed that the matrix A_j is (numerically) nonsingular. The distance to singularity for a matrix A_j can be related to its minimal singular value. Theorem on relative distance to singularity can be found in many books (e.g. [13, p. 123] or [11, 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 $\sigma_{min}(A_j)\le\|\Delta_j\|$ and so using the bounds (3.9) and (3.10) on norms of the matrix E_{j-1} and the vector $\delta_j^{(1)}$ we can write

$$\sigma_{min}(A_j)\le\|\Delta_j\|\le\sqrt{\|E_{j-1}\|_F^2+\|\delta_j^{(1)}\|^2+\|\bar{q}^{(1)}\|^2}\le\sqrt{n\psi^2(m,n)\varepsilon^2\|A\|^2+\|\bar{q}^{(1)}\|^2}.$$

Since $\sigma_{min}(A)\le\sigma_{min}(A_j)$ and assuming

$$\sqrt{n}\psi(m,n)\varepsilon\kappa(A)<1, \tag{3.12}$$

we can give a lower bound for the norm of the vector $\bar{q}^{(1)}$ as follows

$$\|\bar{q}^{(1)}\|\ge\sigma_{min}(A)\sqrt{1-n\psi^2(m,n)\varepsilon^2\kappa^2(A)}. \tag{3.13}$$

The bound (3.13) shows that under assumption on numerical nonsingularity (3.12) the norm $\|\bar{q}^{(1)}\|$ cannot be arbitrarily small and it is essentially bounded by the minimal singular value of A. We note that the result (3.13) corresponds well to the bound on $\|a_j^{(1)}\|$ in exact arithmetic, $\|a_j^{(1)}\|\ge\sigma_{min}(A)$. Trivially, using $\|a_j\|\le\|A\|$, the ratio $\|a_j\|/\|a_j^{(1)}\|$ can be bounded from above by the condition number of the matrix A, which is certainly an interesting observation. The desired bound on the factor $\sigma_j^{(1)}$ can be then obtained from (3.13) considering that

$$\sigma_j^{(1)}=\frac{\|a_j\|}{\|\bar{q}^{(1)}\|}\le\frac{\kappa(A)}{\sqrt{1-n\psi^2(m,n)\varepsilon^2\kappa^2(A)}}. \tag{3.14}$$

Consequently, the quotient $\omega_j^{(1)}$ which describes orthogonality between the vector $\bar{q}^{(1)}$ computed in the first iteration step and the column vectors in Q_{j-1} can be combining (3.8) and (3.14) bounded by

$$\omega_j^{(1)}\le\frac{2n\omega(m,n)\varepsilon\kappa(A)}{\sqrt{1-n\psi^2(m,n)\varepsilon^2\kappa^2(A)}}. \tag{3.15}$$

In the second part of the proof we proceed similarly as in the first part. Using the derived bound (3.15) we study the orthogonality of the vector $\bar{q}^{(2)}$ computed in the second iteration step with respect to the

column vectors Q_{j-1} and give a bound for the quotient $\omega_j^{(2)}$. From (3.2) it follows that

$$\begin{aligned} Q_{j-1}^T \bar{q}^{(2)} &= (I - Q_{j-1}^T Q_{j-1}) Q_{j-1}^T \bar{q}^{(1)} + Q_{j-1}^T \eta_j^{(2)}, \\ \frac{\|Q_{j-1}^T \bar{q}^{(2)}\|}{\|\bar{q}^{(2)}\|} &\leq \left[\|I - Q_{j-1}^T Q_{j-1}\| \frac{\|Q_{j-1}^T \bar{q}^{(1)}\|}{\|\bar{q}^{(1)}\|} + \frac{\|Q_{j-1}^T \eta_j^{(2)}\|}{\|\bar{q}^{(1)}\|} \right] \frac{\|\bar{q}^{(1)}\|}{\|\bar{q}^{(2)}\|}, \end{aligned}$$

which can be using (3.2) and (3.3) further rewritten to the following simpler form

$$\begin{aligned} \omega_j^{(2)} &\leq \left[\|I - Q_{j-1}^T Q_{j-1}\| \omega_j^{(1)} + \|Q_{j-1}\| \frac{\|\eta_j^{(2)}\|}{\|\bar{q}^{(1)}\|} \right] \sigma_j^{(2)}, \\ \omega_j^{(2)} &\leq \left[\sqrt{2n}\omega(m,n)\omega_j^{(1)} + \sqrt{n}\phi(m,n) \right] \varepsilon \sigma_j^{(2)}. \end{aligned} \tag{3.16}$$

The factor $\sigma_j^{(2)} = \frac{\|\bar{a}_j^{(1)}\|}{\|\bar{a}_j^{(2)}\|}$ can be using the relation for the local error in the second iteration step (3.2) estimated as follows

$$\frac{\|\bar{q}^{(2)}\|}{\|\bar{q}^{(1)}\|} \geq \frac{\|\bar{q}^{(1)}\|}{\|\bar{q}^{(1)}\|} - \|Q_{j-1}\| \frac{\|Q_{j-1}^T \bar{q}^{(1)}\|}{\|\bar{q}^{(1)}\|} - \frac{\|\eta_j^{(2)}\|}{\|\bar{q}^{(1)}\|} \geq 1 - \|Q_{j-1}\| \omega_j^{(1)} - \frac{\|\eta_j^{(2)}\|}{\|\bar{q}^{(1)}\|}.$$

Considering the bound for the local error in the second step (3.2), the bound on $\omega_j^{(1)}$ from (3.15) and under assumption

$$\frac{3n^{3/2}\omega(m,n)\varepsilon\kappa(A)}{\sqrt{1 - n\psi^2(m,n)\varepsilon^2\kappa^2(A)}} < 1, \tag{3.17}$$

we obtain the final bound for the factor $\sigma_j^{(2)}$ as follows

$$\sigma_j^{(2)} \leq \left[1 - \frac{3n^{3/2}\omega(m,n)\varepsilon\kappa(A)}{\sqrt{1 - n\psi^2(m,n)\varepsilon^2\kappa^2(A)}} \right]^{-1}. \tag{3.18}$$

The upper bound (3.18) shows that if we strenghten a bit assumption (3.17), the factor $\sigma_j^{(2)}$ becomes very close to 1, which essentially means that $\|\bar{q}^{(2)}\|$ is not significantly smaller than $\|\bar{q}^{(1)}\|$. Due to (3.16) this implies that there cannot be a severe loss of orthogonality of the vector $\bar{q}^{(2)}$ with respect to the column vectors in Q_{j-1} and thus the second iteration step is already enough for preserving the orthogonality of computed vectors. We note that in exact arithmetic we have $a_j^{(2)} = a_j^{(1)}$ giving thus $\|a_j^{(1)}\|/\|a_j^{(2)}\| = 1$. We also note here that the approach used in this paper is very similar to the approach used in [1]. The important difference to results of Abdelmalek is, that his analysis is based on the criterion $(j-2)^2\|Q_{j-1}^T \bar{q}^{(1)}\|/\|\bar{q}^{(2)}\| \leq 1$ and this criterion is expected to hold in most practical cases. Indeed, this criterion can be rewritten as $(j-2)^2\omega_j^{(1)}\sigma_j^{(2)} \leq 1$ and it is clearly seen from (3.15) and (3.18) that it is met under certain assumption on numerical nonsingularity of A. Consequently, using (3.8), (3.16) and (3.18) we obtain the final form for the bound on the quotient $\omega_j^{(2)}$

$$\omega_j^{(2)} \leq \left[\frac{2\sqrt{2}n^{3/2}\omega^2(m,n)\varepsilon^2\kappa(A)}{\sqrt{1 - n\psi^2(m,n)\varepsilon^2\kappa^2(A)}} + \sqrt{n}\phi(m,n)\varepsilon \right] \left[1 - \frac{3n^{3/2}\omega(m,n)\varepsilon\kappa(A)}{\sqrt{1 - n\psi^2(m,n)\varepsilon^2\kappa^2(A)}} \right]^{-1}. \tag{3.19}$$

Now we must fix a polynomial $\zeta(m,n)$ so that the right-hand side of (3.19) can be bounded by $\omega(m,n)\varepsilon$ in order to have (3.5). E.g., if we set $\zeta(m,n) = 25\max(\sqrt{n}\psi(m,n), n^2\phi(m,n))$ and assume

$$\zeta(m,n)\varepsilon\kappa(A) < 1, \tag{3.20}$$

then (3.12) and (3.17) are true and this proves (3.5) which implies (3.6).

If the induction assumption is true at step $j-1$, under assumption (3.20), the statement is true at step j. Then it follows for the last step $j = n$ that

$$\|I - Q^T Q\| \leq 2\sqrt{2}n\phi(m,n)\varepsilon. \tag{3.21}$$

We note that the form of polynomials $\omega(m,n)$ and $\zeta(m,n)$ can be optimized with respect to their degree in n. Indeed, throughout the paper we have used the bounds $\|Q_{j-1}\| \leq \sqrt{j-1}$ which can be improved using the assumption (3.3).

4. Concluding notes and remarks

In this paper we have shown that the orthogonality of vectors Q computed by the CGS2 algorithm is of order of a small multiple of machine precision. Indeed, we have shown that exactly two iteration-steps are already enough when full precision is wanted in the orthogonal basis problem and when the algorithm is applied to (numerically) nonsingular initial set of column vectors. Our results fulfill the theoretical gap in understanding the GS2 process and extend the results of Kahan and Parlett [16] from 2 to n vectors or the results of Abdelmalek [1].

We have seen that the computed basis Q and the computed upper triangular factor $\bar{R}$ satisfy the relation

$$A + E = Q\bar{R}, \quad \|E\| \leq \sqrt{n}\psi(m,n)\varepsilon\|A\|. \tag{4.1}$$

Furthermore, considering our main result (3.21), and using the polar decomposition of the matrix Q (see e.g. [13, p. 389]) one can show that there exists an exactly orthogonal matrix $\hat{Q}$ with $\hat{Q}^T\hat{Q} = I$ such that the computed matrix $\bar{R}$ is an upper triangular factor in the QR factorization of a perturbed matrix $A + \hat{E}$ satisfying

$$A + \hat{E} = \hat{Q}\bar{R}, \quad \|\hat{E}\| \leq \varphi(m,n)\varepsilon\|A\|, \tag{4.2}$$

where $\varphi(m,n)$ stands for some low degree polynomial in dimensions m and n. Clearly, this result shows that the CGS2 algorithmn has similar numerical properties as Householder or Givens orthogonalizations [11, 13]. The operation count for the CGS with reorthogonalization (CGS2) algorithm is $2mn^2$ flops, this is comparable to the operation count for the Householder orthogonalization process which requires $2(mn^2 - n^3/3)$ flops.

Our results are conformal with the practical use of reorthogonalization criterion in the Kahan-Parlett and related algorithms, that we have denoted CGSI(σ). All these algorithms use the criterion introduced by Rutishauser [20]: for a fixed threshold σ, if the criterion $\sigma_j^{(1)} \leq \sigma$ is satisfied after the first orthogonalization loop, then no reorthogonalization is performed. Many experiments (see e.g. Hoffmann [14]) proved that in the most practical case this criterion enables to keep the orthogonality while saving useless reorthogonalization steps. Experiments of Daniel et al. and others [6,7] who observed more iterations of the GS algorithm for rank-deficient problems indicate that this criterion is also suitable to ill-conditioned problems. This gives rise to theoretical question on the quality of the CGSI(σ) algorithm and the validity of this criterion for $n > 2$ vectors. One answer that we can draw from our paper is that if in the CGSI(σ) algorithm we use $\sigma \geq \kappa(A)$, then, due to 3.14), no reorthogonalization is performed and the scheme reduces to CGS. In addition, it is clear that due to (3.18) the criterion $\sigma_j^{(2)} \leq \sigma$ will be satisfied in the second step of CGSI(σ) (if it occurs) for every σ (sufficiently) larger than 1, this means that, at most, two CGS-loops are performed.

Another question that is not adressed in this paper is when A is numerically singular, then CGS2 is clearly unable to perform well. The application of column pivoting in the GS algorithm has been considered by Golub and Businger (see e.g. [11]) and by Chan [5] (see also Dax [7]) who proposed versions of the GS algorithm which can handle the rank-deficient problems and estimate the numerical rank of the matrix A.

In our paper we have considered the CGS2 method and not the corresponding MGS2 algorithm which can be also used to solve the orthogonal basis problem. The same result (3.21) can be shown also for MGS version of the algorithm. Indeed the results similar to (3.1), (3.2) and (3.9), (3.10) can be shown also for MGS process and these are the only parts of the proof, which depend on the implementation of the orthogonalization process.

From a practical point of view it is clear that the CGS2 algorithm is a better candidate than MGS2 algorithm for parallel implementation and this aspect could not be overlooked in certain computing environments. We point out once again that the orthogonal basis problem is of the primary interest in this paper. The situation could be completely different in many applications, where the orthogonality of computed vectors does not play a crucial role and where the one-step MGS algorithm is a successful and practically used technique. As examples we have already mentioned the solution of the least squares problems using MGS [3] or the MGS-GMRES method [12]. However a new trend is emerging nowadays, several experiments are reporting that even if performing twice as many operations as MGS, CGS2 may be faster in some cases because it takes advantage of BLAS 2 or parallel facilities (e.g. in a GMRES context [8], in a GCR context [9], in an eigenvalue context [15]). This gives a superiority of CGS2 over MGS even when MGS is performing well.

5. Acknowledgements

The authors would like to thank for fruitful discussion to Å. Björck and Z. Strakoš. The work of M. R. was supported by the GA CR under the grant No. 201/02/0595 and by the GA AS CR under grant No. A1030103. The work of J. L. was supported by EADS, Corporate Research Centre, Toulouse.

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. Å. Björck and C. Paige. Loss and Recapture of Orthogonality in the Modified Gram-Schmidt Algorithm. *SIAM J. Matrix Anal. Appl. 13(1) (1992), 176-190.*
5. T.F. Chan. Rank revealing QR factorizations. *Linear Algebra and its Appl. 88/89 (1987), 67-82.*
6. 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.*
7. A. Dax. A modified Gram-Schmidt algorithm with iterative orthogonalization and pivoting. *Linear Alg. and its Appl. 310 (2000), 25-42.*
8. V. Frayssé, L. Giraud and H. Kharraz-Aroussi. On the influence of the orthogonalization scheme on the parallel performance of GMRES. *EUROPAR'98 Parallel Processing, Springer 1470,(1998), 751-762.*
9. J. Frank and C. Vuik. Parallel implementation of a multiblock method with approximate subdomain solution. *Appl. Num. Math. 30 (1999), 403-423.*
10. L. Giraud and J. Langou. When modified Gram-Schmidt generates a well-conditioned set of vectors. *Technical Report TR/PA/01/17, CERFACS, Toulouse, France, 2001. To appear in IMAJNA.*

11. G. H. Golub and C. F. Van Loan. Matrix Computations, 3rd ed. *John Hopkins University Press, Baltimore, MD, 1996.*
12. A. Greenbaum, M. Rozložník and Z. Strakoš. Numerical behaviour of the modified Gram-Schmidt GMRES implementation. *BIT 37:3 (1997), 706-719.*
13. N. Higham. Accuracy and Stability of Numerical Algorithms. *SIAM, Philadelphia, PA, 1996.*
14. W. Hoffmann. Iterative Algorithms for Gram-Schmidt Orthogonalization. *Computing 41 (1989), 335-348.*
15. R. B. Lehoucq and A. G. Salinger. Large-Scale Eigenvalue Calculations for Stability Analysis of Steady Flows on Massively Parallel Computers. *Int. J. Numerical Methods in Fluids 36 (2001), 309-327.*
16. B. N. Parlett. The Symmetric Eigenvalue Problem. *Englewood Cliffs, N.J., Prentice-Hall, 1980.*
17. L. Reichel and W. B. Gragg. FORTRAN subroutines for updating the QR decomposition, *ACM Trans. Math. Software, 16 (1990), 369-377.*
18. J. R. Rice. Experiments on Gram-Schmidt Orthogonalization. *Math. Comp. 20 (1966), 325-328.*
19. A. Ruhe. Numerical Aspects of Gram-Schmidt Orthogonalization of Vectors. *Linear Algebra and its Applications 52/53 (1983), 591-601.*
20. H. Rutishauser. Description of Algol 60. Handbook for Automatic Computation, Vol. 1a. *Springer Verlag, Berlin, 1967.*
21. Y. Saad. Numerical Methods for Large Eigenvalue Problems. *Halstead Press, NY, 1992.*
22. E. Schmidt. Über die Auflösung linearer Gleichungen mit unendlich vielen Ubbekannten. *Rend. Circ. Mat. Palermo. Ser. 1, 25 (1908), 53-77.*
23. G. W. Stewart. Matrix Algorithms. Volume I: Basic Decompositions. *SIAM, Philadelphia, PA, 1998.*
24. J. H. Wilkinson. Modern error analysis. *SIAM rev., 13:4, (1971), 548-569.*