

A reorthogonalization procedure for modified Gram–Schmidt algorithm based on a rank- k update

Luc Giraud * Serge Gratton * Julien Langou *

March 20, 2003

CERFACS Technical Report TR/PA/03/11

Abstract

The modified Gram–Schmidt algorithm is a well-known and widely used procedure to orthogonalize the column vectors of a given matrix. When applied to ill-conditioned matrices in floating point arithmetic, the orthogonality among the computed vectors may be lost. In this work, we propose an *a posteriori* reorthogonalization technique based on a rank- k update of the computed vectors. The level of orthogonality of the set of vectors built gets better when k increases and finally reaches the machine precision level for a large enough k . The rank of the update can be tuned in advance to monitor the orthogonality quality. We illustrate the efficiency of this approach in the framework of the Seed–GMRES technique for the solution of an unsymmetric linear system with multiple right-hand sides. In particular, we report experiments on numerical simulations in electromagnetic applications where a rank-one update is sufficient to recover a set of vectors orthogonal to machine precision level.

1 Introduction

Let A be an $m \times n$ real matrix, $m \geq n$ of full rank n . In exact arithmetic, the Modified Gram–Schmidt algorithm (MGS) computes an $m \times n$ matrix Q with orthonormal columns and an $n \times n$ upper triangular matrix R such that $A = QR$. The framework of this paper is the study of the MGS algorithm in the presence of rounding errors. We call *computed quantities* quantities that are computed using a well-designed floating point arithmetic [1]. We denote by \bar{Q} and \bar{R} the computed quantities obtained by running MGS in the presence of rounding errors.

*CERFACS, 42 av. Gaspard Coriolis, 31057 Toulouse Cedex, France.

In [2], Björck and Paige show that \bar{R} is as good as the triangular factor obtained using backward stable transformations such as Givens rotations or Householder reflections. This property of MGS explains why this algorithm can be safely used in applications where only the factor \bar{R} is needed. This is namely the case in the solution of linear least squares problems where the R-factor of the QR-factorization of $[A, b]$ is needed [1, 2]. Another important feature of MGS is that the number of operations required to explicitly compute the Q-factor (problem known as the orthogonal basis problem) is approximatively half that of the methods using Givens rotations or Householder reflections [6, p. 232]. However the computed factor \bar{Q} has less satisfactory properties, since for an ill-conditioned matrix A , it may exhibit a very poor orthogonality as measured by the quantity $\|\bar{Q}^T \bar{Q} - I_n\|$, where $\|\cdot\|$ denotes the spectral 2-norm, and I_n the identity matrix of order n [14]. This has stimulated significant work on modifications of MGS that enhance the orthogonality of \bar{Q} at low computational cost. One of those strategies performs reorthogonalizations during the algorithm when a prescribed criterion is satisfied. This has given rise to the family of iterated modified Gram-Schmidt algorithms, which differ in the criterion they use to enforce the reorthogonalization (see e.g. [3, 9, 15]). An alternative way to compensate for the lack of orthogonality in \bar{Q} is derived in [2] for a wide class of problems, including the linear least squares problem and computation of the minimum 2-norm solution of an underdetermined linear system and the projection of a vector onto a subspace. A careful use of \bar{Q} and \bar{R} , based on an equivalence of MGS on A and Householder QR on an augmented matrix obtained by putting a matrix of zeros on top of A , leads to a backward stable algorithm. Such a strategy implies – in general – that the use of \bar{Q} is computationally more expensive than would be the use of a Q-factor with orthonormal columns.

The error analyses related to the loss of orthogonality, that are used to derive the successful methods mentioned above, are based on the study of the quantity $\|\bar{Q}^T \bar{Q} - I_n\|$. We propose here to adopt a different approach by inspecting not only the largest singular value, as actually done in the related literature, but each singular value of the matrices involved in MGS. We denote by $\sigma_i, i = 1, \dots, n$ the singular values of A , $\sigma_1 \geq \dots \geq \sigma_n > 0$, by $\kappa = \sigma_1/\sigma_n$ the spectral condition number of A . Also we define the reduced condition numbers via

Definition 1.1 *Let κ_i , the reduced condition number, be defined by $\kappa_i = \sigma_1/\sigma_{n-i+1}, i = 1 \dots n$.*

Finally let \tilde{Q} be the matrix obtained from \bar{Q} by normalizing its columns. In

this paper, we exhibit a series of low rank matrices F_k , $k = 0, \dots, n-1$ that enables us to update the factor \tilde{Q} such that

- $\text{rank}(F_k) \leq k$,
- the columns of $\tilde{Q} + F_k$ are orthonormal up to machine precision times κ_k , if $k = n-1$, then the columns of $\tilde{Q} + F_{n-1}$ are exactly orthonormal,
- $(\tilde{Q} + F_k)\tilde{R}$ represents A up to machine precision.

In the case $k = 0$, $F_0 = 0$ so $(\tilde{Q} + F_0) = \tilde{Q}$ and the results obtained are of the same essence as the ones by Björck (1967) [1]. Namely MGS generates a Q-factor such that the columns of \tilde{Q} are orthonormal up to machine precision times $\kappa = \kappa_0$ and $\tilde{Q}\tilde{R}$ represents A up to machine precision. In the case $k = n-1$, $(\tilde{Q} + F_{n-1})$ is indeed the same matrix as \hat{Q} , the matrix exhibited by Björck and Paige [2]. That is \hat{Q} has orthonormal columns and $\hat{Q}\tilde{R}$ represents A up to machine precision. Our result can be seen as a theoretical bridge that links the result of Björck (1967) [1] to the result of Björck and Paige (1992) [2]. An algorithm to compute F_k , $k = 0, \dots, n-1$, is also derived. In our experiments this algorithm behaves well in the presence of rounding errors. For example when κ_k is close to one, the update of \tilde{Q} with F_k produces a Q-factor with columns orthonormal up to machine precision. The complexity of this algorithm increases with k . For small k , its complexity is competitive with other standard reorthogonalization techniques. We conclude our study with an application of this algorithm in the framework of the solution of unsymmetric linear systems with multiple right-hand sides where a seed-variant of GMRES can be successfully used.

In the remainder of this paper, for any $m \times n$ matrix X , we denote by $\sigma_i(X)$, $i = 1, \dots, n$ the singular values of X ordered such that $\sigma_1(X) \geq \dots \geq \sigma_n(X)$. We note that the work of this paper can be extended to complex arithmetic as well.

2 Rank considerations related to the loss of orthogonality in MGS

2.1 Introduction

A rigorous measure of the orthogonality of an $m \times n$ matrix \bar{Q} can be defined to be the distance, in the spectral 2-norm, to the set $\mathcal{O}(m, n)$ of $m \times n$ matrices with orthonormal columns

$$\min_{V \in \mathcal{O}(m, n)} \|\bar{Q} - V\|.$$

Fan and Hoffman in [4] for the case $m = n$, and Higham in [8] for the general case $n \leq m$ proved that the minimum is attained for V being the unitary polar factor of \bar{Q} . The easily computed quantity $\|I_n - \bar{Q}^T \bar{Q}\|$ is often preferred to measure the orthogonality, because, as shown in Lemma 2.1, it has the same order of magnitude as $\min_{V \in \mathcal{O}(m,n)} \|\bar{Q} - V\|$ when $\|\bar{Q}\|$ is close to one.

Lemma 2.1 [8] *Let $\bar{Q} \in \mathbb{R}^{m \times n}$, $n \leq m$,*

$$\frac{\|I_n - \bar{Q}^T \bar{Q}\|}{1 + \|\bar{Q}\|} \leq \min_{V \in \mathcal{O}(m,n)} \|\bar{Q} - V\| \leq \|I_n - \bar{Q}^T \bar{Q}\|.$$

Lemma 2.1 can be easily generalized into Lemma 2.2.

Lemma 2.2 *Let $\bar{Q} \in \mathbb{R}^{m \times n}$, $n \leq m$,*

$$\frac{\sigma_i(\bar{Q}^T \bar{Q} - I_n)}{1 + \|\bar{Q}\|} \leq \sigma_i(\bar{Q} - U) \leq \sigma_i(\bar{Q}^T \bar{Q} - I_n),$$

where $i = 1, \dots, n$ and U is the unitary polar factor associated with \bar{Q}

Proof: Let $\bar{Q} = UH$ be the polar decomposition of \bar{Q} . $U \in \mathbb{R}^{m \times n}$ has orthonormal columns and $H \in \mathbb{R}^{n \times n}$ is symmetric positive definite. From $U^T U = I_n$ and $H^T = H$, it follows that

$$(\bar{Q} - U)^T (\bar{Q} + U) = \bar{Q}^T \bar{Q} - U^T \bar{Q} + \bar{Q}^T U - U^T U = \bar{Q}^T \bar{Q} - I_n. \quad (1)$$

We also have $\bar{Q} + U = U(I_n + H)$ and H being symmetric positive definite, $I_n + H$ has full rank n , and

$$(\bar{Q} - U)^T U = (\bar{Q}^T \bar{Q} - I_n)(I_n + H)^{-1}.$$

Let us consider U_\perp such that $\begin{pmatrix} U & U_\perp \end{pmatrix}$ is unitary then

$$(\bar{Q} - U)^T \begin{pmatrix} U & U_\perp \end{pmatrix} = (\bar{Q}^T \bar{Q} - I_n)(I_n + H)^{-1} \begin{pmatrix} I_{m,n} & 0_{m-n,n} \end{pmatrix},$$

so we get

$$(\bar{Q} - U)^T = (\bar{Q}^T \bar{Q} - I_n)(I_n + H)^{-1} \begin{pmatrix} I_{m,n} & 0_{m-n,n} \end{pmatrix} \begin{pmatrix} U^T \\ U_\perp^T \end{pmatrix} = (\bar{Q}^T \bar{Q} - I_n)(I_n + H)^{-1} U^T.$$

Taking singular values and using the property $\sigma_i(XY) \leq \sigma_i(X)\|Y\|$ (see e.g. [10, p. 423]), that holds for any matrices X and Y such that the product XY exists, yields

$$\sigma_i(\bar{Q} - U) \leq \sigma_i(\bar{Q}^T \bar{Q} - I_n) \|(I_n + H)^{-1}\| \|U\|.$$

The right inequality of the lemma is a consequence of $\|(I_n + H)^{-1}\| \leq 1$ and $\|U\| = 1$.

For the left inequality of the lemma, taking singular values in Equality (1), leads to

$$\sigma_i(\bar{Q}^T \bar{Q} - I_n) \leq \sigma_i(\bar{Q} - U) \|\bar{Q} + U\|$$

and the conclusion readily follows from $\|\bar{Q} + U\| \leq \|\bar{Q}\| + 1$.

□

An important consequence of Lemma 2.2 is that if \bar{Q} does not have orthonormal columns, but if $\bar{Q}^T \bar{Q} - I_n$ has only k nonzero singular values, \bar{Q} is at most a rank- k modification of a matrix with orthonormal columns (namely U).

In Section 2.3, we derive a result for MGS that is similar in essence to Lemma 2.2. However, for any $k \leq n$, the MGS context will enable us to find explicitly a rank- k matrix F_k such that $\bar{Q} + F_k$ has an improved orthogonality compared with \bar{Q} and such that the product $(\bar{Q} + F_k)\bar{R}$ still accurately represents A .

2.2 Some useful background related to MGS in floating point arithmetic

A key result to understand the loss of orthogonality in MGS in floating point arithmetic, is that MGS on A can be interpreted as an Householder QR-factorization on $A_{aug} = \begin{bmatrix} O_n \\ A \end{bmatrix}$, where O_n is the square zero matrix of order n [2]. Since we elaborate our work on results and techniques presented in [2] we briefly outline them below.

The use of Wilkinson's analysis of Householder transformations [18, pp. 153–162] on A_{aug} enables Björck and Paige [2, Eq.(3.3)] to give an orthogonal transformation \tilde{P} such that

$$\begin{pmatrix} E_1 \\ A + E_2 \end{pmatrix} = \tilde{P} \begin{pmatrix} \bar{R} \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{P}_{11} \\ \tilde{P}_{21} \end{pmatrix} \bar{R}, \quad (2)$$

$$\|E_i\| \leq \bar{c}_i u \|A\|, \quad i = 1, 2,$$

where \bar{c}_i are constant depending on m, n and the detail of the arithmetic, and u is the unit roundoff. Here \tilde{P}_{11} is strictly upper triangular, see [2, §4 & (4.1)].

Let $\tilde{Q} = [\tilde{q}_1, \dots, \tilde{q}_n]$ be the matrix obtained from $\bar{Q} = [\bar{q}_1, \dots, \bar{q}_n]$ by normalizing its columns ($\tilde{q}_i = \bar{q}_i / \|\bar{q}_i\|$). The equality $\tilde{P}_{21} = \tilde{Q}(I_n - \tilde{P}_{11})$ holds [2, Eq.(4.5)] and the residual error of the polar factor \hat{Q} of \tilde{P}_{21} can be bounded as follows,

$$\|A - \hat{Q}\bar{R}\| \leq \bar{c}u\|A\|, \quad (3)$$

where $\bar{c} = \bar{c}_1 + \bar{c}_2$, provided that $\bar{c}u\kappa < 1$ [2, Eq.(3.7)]. Finally, let $\bar{\sigma}_1 \geq \dots \geq \bar{\sigma}_n$ be the singular values of \bar{R} . The singular values of \bar{R} approximate those of A in the following sense $|\bar{\sigma}_i - \sigma_i| \leq \bar{c}u\sigma_1$ [2, Eq.(3.8)]. This implies that under the assumption $\bar{c}u\kappa < 1$, \bar{R} has full rank n .

2.3 Recapture of the orthogonality in MGS

As $\begin{pmatrix} \tilde{P}_{11} \\ \tilde{P}_{21} \end{pmatrix}$ has orthonormal columns and $n \leq m$, we consider its CS decomposition [6, p. 77] defined by

$$\begin{aligned} \tilde{P}_{11} &= UCW^T, \\ \tilde{P}_{21} &= VSW^T, \end{aligned} \quad (4)$$

where C is singular since \tilde{P}_{11} is strictly upper triangular, the entries of S are in increasing order ($0 \leq s_1 \leq \dots \leq s_n = 1$), the entries of C are in decreasing order ($1 \geq c_1 \geq \dots \geq c_n = 0$) and $C^2 + S^2 = I_n$. and the three matrices U , V , W have orthonormal columns. C , S , U and W are $n \times n$, V is $m \times n$. Similarly as in [2], we suppose that A is not too ill-conditioned, by assuming that $(\bar{c}_1 + \bar{c})u\kappa < 1$ or equivalently (since this implies both $\bar{c}u\kappa < 1$ and $\bar{c}_1u\kappa < 1 - \bar{c}u\kappa$)

$$\bar{c}_1u\eta\kappa < 1, \quad (5)$$

where $\eta = (1 - \bar{c}u\kappa)^{-1}$. This has the following consequence. Since the leading element of C is (using (2)) $c_1 = \|\tilde{P}_{11}\| = \|E_1\bar{R}^{-1}\| \leq \bar{c}_1u\sigma_1/\bar{\sigma}_n$, and since from (3) $|\sigma_n - \bar{\sigma}_n| \leq \bar{c}u\sigma_1$, we see that $\bar{\sigma}_n \geq \sigma_n - \bar{c}u\sigma_1 = \sigma_n\eta$, and it follows $c_1 \leq \bar{c}_1u\eta\kappa < 1$ (see [2, Eq.(3.11)]), all the s_i 's are non zero, thus S is nonsingular.

Our goal is to improve the orthogonality of the Q-factor while maintaining the residual error, $\|A - Q\bar{R}\|/\|A\|$, at the level of the machine precision. Since \hat{Q} has orthonormal columns and (3) holds, \hat{Q} answers our question. Therefore a straightforward but expensive way to achieve our goal would be to compute \hat{Q} with $\hat{Q} = VW^T$ ([6, p. 149]). Let us evaluate $F = \hat{Q} - \tilde{Q}$ to find matrices that approximate the difference between \hat{Q} and \tilde{Q} at low

computational cost. Since $\hat{Q} = VW^T$, using $\tilde{P}_{21} = \tilde{Q}(I_n - \tilde{P}_{11})$ (see Section 2.3), the CS decomposition (4) and the fact that S is nonsingular, we get

$$\begin{aligned} F &= \tilde{Q} \left((I_n - \tilde{P}_{11})WS^{-1}(I_n - S)W^T - \tilde{P}_{11} \right), \\ &= \tilde{Q} (W(S^{-1} - I_n) - UCS^{-1})W^T. \end{aligned} \quad (6)$$

We define the truncated matrices U_k , V_k and W_k by retaining the first k columns in their counterparts U , V and W . In Matlab style notation, it reads $U_k = U(:, 1 : k)$. We also denote by C_k (resp. S_k) the diagonal matrix of order k whose diagonal entries are the $c_i, i = 1 \dots k$ (resp. $s_i, i = 1 \dots k$).

We define the matrix F_k obtained by setting the c_l and the $s_l, l > k$, to zero and one respectively in (6), this gives

$$F_k = \tilde{Q}(W_k(S_k^{-1} - I_k) - U_k C_k S_k^{-1})W_k^T. \quad (7)$$

so that $F_0 = 0, F_{n-1} = F_n = F$, since $s_n = 1$ and $c_n = 0$. The matrix $\tilde{Q} + F$ has orthonormal columns and accurately represents A when multiplied on the right by \tilde{R} . Theorem 2.3 shows how these properties are modified when the matrix $\tilde{Q} + F_k$ is considered instead. The matrices Q_k are then a sequence of matrices going from the matrix of normalized vectors from MGS $Q_0 = \tilde{Q}$, to the matrix of orthogonal vectors $Q_{n-1} = \hat{Q}$.

Theorem 2.3 *Assume that $\bar{c}_1 u \eta \kappa < 1$, for $k = 0, \dots, n-1$, the matrix Q_k defined by*

$$Q_k = \tilde{Q} + F_k \quad (8)$$

enjoys the following properties

a)

$$\text{rank}(Q_k - \tilde{Q}) \leq k,$$

b)

$$\|A - Q_k \tilde{R}\| \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1 + \bar{c}_1 u \eta \kappa)}{(1 - \bar{c}_1 u \eta \kappa)^2} \right) u \|A\|,$$

c) for $k = 0, \dots, n-2$,

$$\|I_n - Q_k^T Q_k\| \leq \left(2\bar{c}_1 \eta \frac{(1 + \bar{c}_1 u \eta \kappa)^2}{(1 - \bar{c}_1 u \eta \kappa)^3} \right) u \kappa_{k+1}.$$

for $k = n-1, Q_{n-1} = \hat{Q}$, and so $\|I_n - Q_{n-1}^T Q_{n-1}\| = 0$.

Proof: Part a) is a consequence of the definition (7) of F_k . We then establish part b) of this theorem. From (2), $\tilde{P}_{11}\bar{R} = E_1$, and multiplying to the left by U_k^T implies that $U_k^T U C W^T \bar{R} = U_k^T E_1$. Using the definition of the truncated matrices C_k and W_k , one gets $C_k W_k^T \bar{R} = U_k^T E_1$, and, taking norms, $\|C_k W_k^T \bar{R}\| = \|U_k^T E_1\| \leq \|E_1\|$. From (2), $\|E_1\| \leq \bar{c}_1 u \|A\|$, we obtain a first intermediate result

$$\|C_k W_k^T \bar{R}\| \leq \bar{c}_1 u \|A\|. \quad (9)$$

Let us bound the residual error $\|A - Q_k \bar{R}\|$. Using the triangular inequality, yields

$$\|A - Q_k \bar{R}\| \leq \|A - \tilde{Q} \bar{R}\| + \|F_k \bar{R}\|. \quad (10)$$

The first term of the right-hand side can be bounded using Lemma A.2. We study the second term of the right-hand side: $\|F_k \bar{R}\|$. By definition (7) of F_k ,

$$F_k \bar{R} = \tilde{Q} (W_k (S_k^{-1} - I_k) - U_k C_k S_k^{-1}) (W_k^T \bar{R}).$$

Applying the result of Lemma A.3 to $(S_k^{-1} - I_k) W_k^T \bar{R}$ and noticing that, from $c_i^2 + s_i^2 = 1$, we have

$$s_i^{-1} - 1 = (1 - c_i^2)^{-1/2} - 1 = \frac{c_i^2}{\sqrt{1 - c_i^2}(1 + \sqrt{1 - c_i^2})} \leq \frac{c_i^2}{2(1 - c_i^2)}, \quad (11)$$

so $s_i^{-1} - 1 \leq c_i \frac{c_i}{2(1 - c_i^2)}$, we get $\|(S_k^{-1} - I_k) W_k^T \bar{R}\| \leq \frac{\|C_k\|}{2(1 - \|C_k\|)} \|C_k W_k^T \bar{R}\|$, from which follows that

$$\|F_k \bar{R}\| \leq \|\tilde{Q}\| \left(\frac{\|C_k\|}{2(1 - \|C_k\|)} + \|S_k^{-1}\| \right) \|C_k W_k^T \bar{R}\|, \quad (12)$$

where we have used the fact that the two matrices C_k and S_k^{-1} being diagonal, they commute. We recall after (5) that $\|C_k\| \leq \bar{c}_1 u \eta \kappa < 1$ and therefore $\|S_k^{-1}\| \leq (1 - (\bar{c}_1 u \eta \kappa)^2)^{-1/2} \leq (1 - \bar{c}_1 u \eta \kappa)^{-1}$. Using Lemma A.1,

$$\|F_k \bar{R}\| \leq \frac{(1 + \bar{c}_1 u \eta \kappa)^2}{(1 - \bar{c}_1 u \eta \kappa)^2} \bar{c}_1 u \|A\|.$$

This proves Part b).

With Lemma A.2, this proves Part b).

We now prove part c) of the Theorem. We define the matrices $U_{\bar{k}}$, $V_{\bar{k}}$, $W_{\bar{k}}$, so that $U = [U_k, U_{\bar{k}}]$, and similarly for V and W . In Matlab style

notation, $U_{\bar{k}} = U(:, k+1 : n)$. We also define the matrices $C_{\bar{k}}$ (resp. $S_{\bar{k}}$) the diagonal matrix of order $n - k + 1$ whose diagonal elements are the c_i , $i = k + 1, \dots, n$ (resp. s_i $i = k + 1, \dots, n$). One has

$$\hat{Q} - Q_k = F - F_k, \quad (13)$$

$$\hat{Q} - Q_k = \tilde{Q} \left(W_{\bar{k}}(S_{\bar{k}}^{-1} - I_{n-k+1}) - U_{\bar{k}}C_{\bar{k}}S_{\bar{k}}^{-1} \right) W_{\bar{k}}^T, \quad (14)$$

$$\|\hat{Q} - Q_k\| \leq \|\tilde{Q}\| \left(\|S_{\bar{k}}^{-1} - I_{n-k+1}\| + \|C_{\bar{k}}S_{\bar{k}}^{-1}\| \right). \quad (15)$$

Since both the s_i 's and c_i 's belong to $[0, 1]$, and the c_i (resp. the s_i) are sorted in decreasing (resp. increasing) order, one obtains

$$\|\hat{Q} - Q_k\| \leq \|\tilde{Q}\| \left((s_{k+1}^{-1} - 1) + s_{k+1}^{-1}c_{k+1} \right),$$

which yields, using (11),

$$\|\hat{Q} - Q_k\| \leq \|\tilde{Q}\| c_{k+1} \frac{1 + c_{k+1}}{1 - c_{k+1}}.$$

From Lemma A.1 and using the fact that $c_{k+1} \leq c_1 \leq \bar{c}_1 u \eta \kappa$, we get

$$\|\hat{Q} - Q_k\| \leq \frac{(1 + \bar{c}_1 u \eta \kappa)^2}{(1 - \bar{c}_1 u \eta \kappa)^2} c_{k+1}. \quad (16)$$

Since $\tilde{P}_{11} = E_1 \bar{R}^{-1}$, and c_{k+1} is the $k+1$ th singular value of \tilde{P}_{11} , one has $c_{k+1} \leq \|E_1\| \sigma_{k+1}(\bar{R}^{-1})$. From $\sigma_{k+1}(\bar{R}^{-1}) = 1/\bar{\sigma}_{n-k}$, $\|E_1\| \leq \bar{c}_1 u \|A\|$, $|\bar{\sigma}_{k+1} - \sigma_{k+1}| \leq \bar{c} u \sigma_1$, $\kappa_{k+1} = \frac{\|A\|}{\sigma_{n-k}}$ and $\eta = (1 - \bar{c} u \kappa)^{-1}$, we obtain $c_{k+1} \leq \bar{c}_1 u \frac{\|A\|}{\sigma_{n-k}} \leq \bar{c}_1 u \eta \frac{\|A\|}{\sigma_{n-k}} = \bar{c}_1 u \eta \kappa_{k+1}$ and the conclusion for the case $k = 0, \dots, n-2$ follows using Lemma 2.1. For the case $k = n-1$, this bound on c_n gives $c_n \leq \bar{c}_1 u \eta$ which is not satisfactory. Since \tilde{P}_{11} is strictly upper triangular, a better bound is $c_n = 0$ from which we recover with Equation (16) that $\hat{Q} = Q_{n-1}$.

□

Several remarks can be made. First consistency, $\|A - Q_k \bar{R}\|/\|A\|$, is maintained close to machine precision independently of the rank- k of the update. In the introduction, we explain that in the case $k = 0$ and $k = n-1$, we recover the result of Björck [1] for $\tilde{Q} = Q_0$ and Björck and Paige [2] for $\hat{Q} = Q_{n-1}$ respectively. A consequence of this unified framework is that the bounds given are larger than the original ones but remain very close. In Table 1, we summarize the relations to be compared. Note that the results of Björck [1] have been replaced by analogous results of Björck and Paige [2] in order to compare the same quantities.

Theorem 2.3 Part b) $k = 0$ $\ A - \tilde{Q}\tilde{R}\ \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1+\bar{c}_1 u\eta\kappa)}{(1-\bar{c}_1 u\eta\kappa)^2}\right) u\ A\ $	Lemma A.2 derived from Björck and Paige (1992) [2] $\ A - \tilde{Q}\tilde{R}\ \leq \left(\bar{c}_2 + \bar{c}_1 \frac{1+\bar{c}_1 u\eta\kappa}{1-\bar{c}_1 u\eta\kappa}\right) u\ A\ $
Theorem 2.3 Part b) $k = n - 1$ $\ A - \hat{Q}\hat{R}\ \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1+\bar{c}_1 u\eta\kappa)}{(1-\bar{c}_1 u\eta\kappa)^2}\right) u\ A\ $	Björck and Paige (1992) [2, Eq.(3.7)] $\ A - \hat{Q}\hat{R}\ \leq (c_1 + c_2)u\ A\ $
Theorem 2.3 Part c) $k = 0$ $\ I_n - \tilde{Q}^T\tilde{Q}\ \leq \left(2\bar{c}_1\eta \frac{(1+\bar{c}_1 u\eta\kappa)^2}{(1-\bar{c}_1 u\eta\kappa)^3}\right) u\kappa$	Björck and Paige (1992) [2, Eq.(5.3)] $\ I_n - \tilde{Q}^T\tilde{Q}\ \leq \frac{2c_1}{1-(c+c_1)u\kappa} u\kappa$
Theorem 2.3 Part c) $k = n - 1$ $\ I_n - \hat{Q}^T\hat{Q}\ = 0$	Björck and Paige (1992) [2, Eq.(3.7)] $\ I_n - \hat{Q}^T\hat{Q}\ = 0$

Table 1: Correspondence between the bounds in Theorem 2.3 and the results of Björck and Paige[2].

3 Numerical illustrations and examples of application

3.1 Numerical illustrations of the bounds in Theorem 2.3

The aims of this section are twofold. First, we give an algorithm to compute the approximations \bar{F}_k (resp. \bar{Q}_k) of the matrices F_k (resp. Q_k), then we verify numerically that Theorem 2.3 is satisfied with these computed quantities up to machine precision.

In order to make sure that the rank- k property of the $m \times n$ matrix F_k is inherited by the computed matrix \bar{F}_k , we define \bar{F}_k as the product of the $m \times k$ computed quantities $\bar{Q}(W_k(S_k^{-1} - I_k) - U_k C_k S_k^{-1})$ times the $k \times n$ rectangular matrix W_k^T . Then by construction, the first statement a) of Theorem 2.3 is satisfied and we can now focus on the last two statements and show that the bounds are sharp.

In the following, the notation F_k (resp. Q_k) stands for the the computed quantity \bar{F}_k (resp. \bar{Q}_k). For the experiments, we proceed as follows. Starting from an initial matrix A , we run MGS to obtain \bar{Q} and \bar{R} . Then for each k from $k = 0$ to $n - 1$, we compute the associated matrix Q_k using formulae (7) and (8). In that respect, we need to compute \tilde{P}_{11} . In [2, Eq.(4.1)], Björck and Paige show that \tilde{P}_{11} is strictly upper triangular with element (i, j) equal to $\tilde{q}_i^T (I_m - \tilde{q}_1 \tilde{q}_1^T) \dots (I_m - \tilde{q}_{j-1} \tilde{q}_{j-1}^T) \tilde{q}_j$ for $i < j$. We define \tilde{T} such that \tilde{T} is strictly upper triangular with element (i, j) , $\tilde{q}_i^T \tilde{q}_j$, ($i < j$). Since $\|\tilde{q}_i\| = 1$ for

1. run MGS on A to obtain \bar{Q} and \bar{R}
2. compute \bar{T} , the strictly upper triangular matrix with entry (i, j) , $\bar{q}_i^T \bar{q}_j$, $(i < j)$ then form $\bar{P}_{11} = (I_n + \bar{T})^{-1} \bar{T}$
3. compute the k largest singular values of \bar{P}_{11} , c_i , $i = 1, \dots, n$, and the associated k right (resp. left) singular vectors U_k (resp. W_k) finally form $s_i = \sqrt{1 - c_i^2}$, $i = 1, \dots, k$. The matrix C_k (resp. S_k) is the $k \times k$ diagonal matrix with entry (i, i) equal to c_i (resp. s_i).
4. Form $Q_k = \bar{Q} + \bar{Q}(W_k(S_k^{-1} - I_k) - U_k C_k S_k^{-1})W_k^T$

Table 2: Algorithm 1 : MGS with an a-posteriori reorthogonalization by a rank- k update

all i , one may notice that $(I_n + \tilde{T})(I_n - \tilde{P}_{11}) = I_n$, that can also be written

$$\tilde{P}_{11} = (I_n + \tilde{T})^{-1} \tilde{T}. \quad (17)$$

The matrix \tilde{P}_{11} is also closely related to the T -factor of the YTY-representation of the matrix \tilde{P} (see Schreiber and Van Loan [11]). Calling $\tilde{Y} = \begin{pmatrix} -I_n \\ \tilde{Q} \end{pmatrix}$ and $\tilde{T}_{YTY} = -I_n + \tilde{P}_{11}$, we have

$$\tilde{P} = I + \tilde{Y} \tilde{T}_{YTY} \tilde{Y}^T,$$

and developing this expression leads directly to [2, Eq.(4.2)].

Note that in practice the mathematical quantities \tilde{q}_i are replaced by the computed quantities \bar{q}_i . Equation (17) is preferred to the original equation of Björck and Paige [2, Eq.(4.1)] since it enables us to compute \tilde{P}_{11} with significantly less flops when m is large compared to n . We summarize the corresponding algorithm in Table 2.

In this section, the numerical experiments are run with MATLAB 6 where the unit roundoff is $u \approx 1.1 \cdot 10^{-16}$. We consider two test matrices, that are the matrices $P(1500, 500, 1, 5)$ from Paige and Saunders [12] and GRE_216B from Matrix Market¹. The first one is a 1500×500 matrix with condition number 10^{16} while the latter is a 216×216 matrix with condition number

¹<http://math.nist.gov/MatrixMarket/>

$6 \cdot 10^{14}$. On those two matrices we investigate how sharp the bounds b) and c) in Theorem 2.3 are.

In order to quantify the orthogonality quality of the columns of different matrices, we define the level of orthogonality of Q as the quantity $\|I_n - Q^T Q\|$. In Figure 1 a), we plot the “recovered orthogonality” with \circ . For $k = 0$, we have $Q_0 = \bar{Q}$ therefore we simply plot the level of orthogonality obtained after the run of MGS on $P(1500, 500, 1, 5)$. For $k = 1$, we correct \bar{Q} by the rank-one update F_1 to obtain Q_1 and then plot the level of orthogonality of Q_1 . While k increases, we observe the benefit of adding F_k to \bar{Q} on the orthogonality quality. We stop the plot at $k = 100$. At this step, the matrix Q_{100} has nearly reached its final level of orthogonality ($1.44 \cdot 10^{-14}$ for $k = n - 1$). With \triangle , we plot the corresponding $u\kappa_{k+1}$, $k = 0, \dots, n - 1$. The theorem predicts that for each k , $\|I_n - Q_k^T Q_k\|$ is bounded above by $u\kappa_{k+1}$ times a constant. In this experiment we observe that both curves fit well. This indicates that the constant can be taken close to one for these experiments and that the bound c) of Theorem 2.3 is tight. In Figure 1 b), we illustrate that Property b) of Theorem 2.3 holds. In this case $\|A - Q_k \bar{R}\|$ is smaller than $u\|A\|$ times a constant where the constant is small.

Similar experiments are reported in Figure 2 for the matrix GRE_216B that also illustrates the tightness of the bounds.

Given the singular value distribution of A and the machine precision, Theorem 2.3 gives us a set of k for which all the associated matrices Q_k satisfy a prescribed level of orthogonality. Since the amount of work of Algorithm 1 increases with k , we can choose the lowest k of this set and update \bar{Q} with the rank- k matrix F_k . Therefore an interesting feature of Algorithm 1 is that it is able to adapt its amount of work with respect to the level of orthogonality expected. For example, if the level of orthogonality required for the Q-factor of matrix GRE_216B is 10^{-9} , with both Theorem 2.3 and the knowledge of $u\kappa_{k+1}$, we can choose $k = 10$. Meanwhile, if the level of orthogonality required is 10^{-14} , we can estimate the value a-priori $k = 37$. A-posteriori we observe in Figure 2 and curve $\|I_n - Q_k^T Q_k\|$ that these two choices are correct.

3.2 An application of choice: Seed-GMRES

A practical framework where our algorithm fits perfectly is the Seed-GMRES method for solving a sequence of linear systems with the same coefficient matrix but for a sequence of different right-hand sides. Roughly speaking one solves the linear system for one right-hand side at a time but uses the

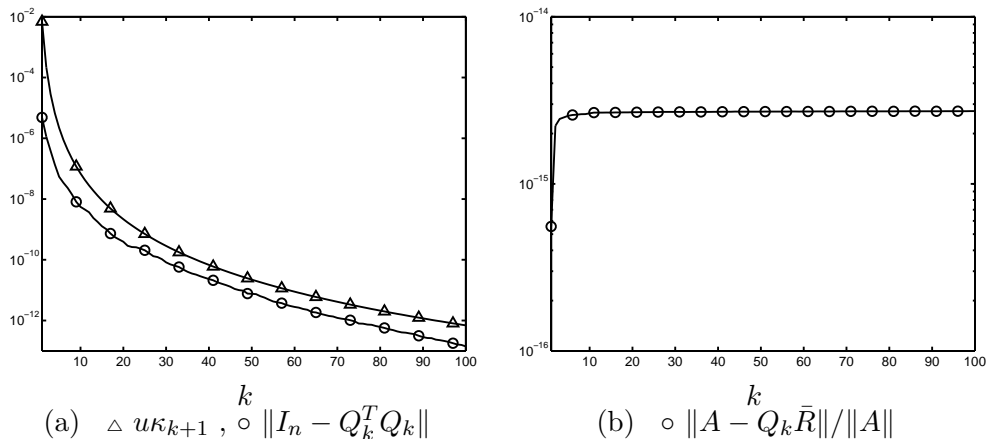


Figure 1: Illustrations of bounds (b) and (c) of Theorem 2.3 for matrix $P(1500, 500, 1, 5)$

Krylov space associated with the current right-hand side to compute a good initial guess for the next ones.

Let us now briefly describe the Seed-GMRES method and the various alternatives we consider to compare with our algorithm. Let Z be a square matrix of order m with full rank. We want to solve the linear systems $Zx^{(i)} = b^{(i)}$ for $i = 0, \dots, p$ by using Seed-GMRES with MGS (see e.g. [13, 16]). For the sake of clarity, but without loss of generality, we describe the method assuming that the initial guesses for all the right-hand sides are zeros, and we only illustrate it when the first right-hand side has converged. For the next ones, the same algorithm applies but the initial guesses are no longer zero making the notation more complicated for a purpose that is out of the scope of this paper.

We first run GMRES with MGS to solve the linear system $Zx^{(0)} = b^{(0)}$. This amounts to solving the linear least squares problem

$$\min_{y \in \mathbb{R}^{n-1}} \|b^{(0)} - ZV_{n-1}^{(0)}y\|,$$

where $V_{n-1}^{(0)}$ is a set of $n - 1$ vectors built with an Arnoldi process on Z using the starting vector $b^{(0)}$ and orthogonalization scheme MGS. In most applications, the computational burden lies in the matrix-vector products and the scalar products required to solve this linear least squares problem. In Seed-GMRES, the subsequent right-hand sides benefit from this work. An effective initial guess $x^{(i)} = V_{n-1}^{(0)}y^{(i)}$ for the system i is obtained by

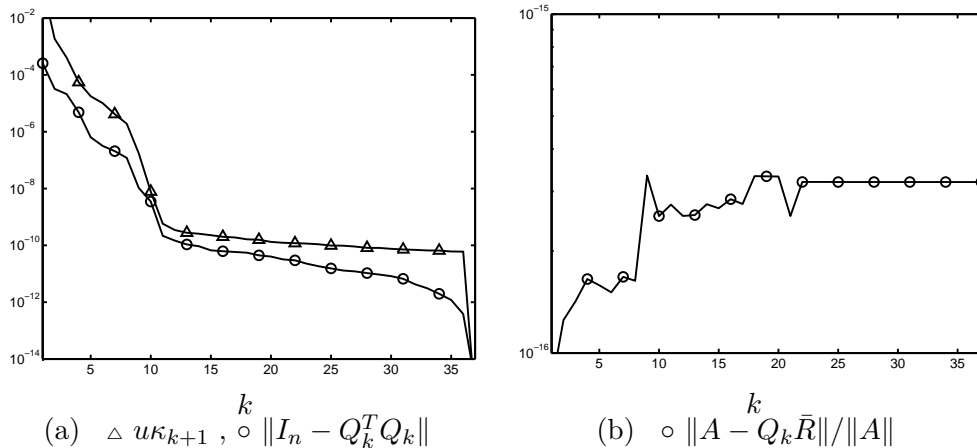


Figure 2: Illustrations of bounds (b) and (c) of Theorem 2.3 for matrix GRE_216B

solving the same linear least squares problem but with another right-hand side, namely

$$\min_{y \in \mathbb{R}^{n-1}} \|b^{(i)} - ZV_{n-1}^{(0)}y\|.$$

We first compare four approaches to solve this problem. In the first part, we present two standard algorithms and compare them in terms of floating point operations (flops) with an approach implementing Algorithm 1. In the second part, one aspect of our problem is examined in more details to show that – under reasonable assumptions – a rank-one update is enough to recover with Algorithm 1 a good level of orthogonality. In this particular case, a second algorithm is also derived based on an heuristic that enables us to save substantially computational work. Finally we illustrate the effectiveness of our approach when embedded in large electromagnetism applications.

In the sequel, the superscript (0) is omitted and the matrix A denotes the computed matrix $(b^{(0)}, ZV_{n-1}^{(0)})$ similarly as in the first section, MGS is run on A of size $m \times n$ in a well designed floating point arithmetic to obtain \bar{Q} and \bar{R} . Indeed, the Arnoldi process gives $\bar{Q} = V_n^{(0)}$ but for the sake of generality this property is not taken into account.

3.2.1 The three approaches

Since we already have computed the QR factorization of (b, ZV_{n-1}) via MGS, an efficient way to solve the linear least squares with $b^{(i)}$ is to compute the R-factor of the QR factorization of $(b, ZV_{n-1}, b^{(i)})$ via MGS. In practice it remains to compute the last column of this R-factor that is $c_{\text{MGS}}^{(i)}$ such that

$$c_{\text{MGS}}^{(i)} = \begin{pmatrix} \bar{q}_1^T b^{(i)} \\ \vdots \\ \bar{q}_n^T (I_m - \bar{q}_{n-1} \bar{q}_{n-1}^T) \cdots (I_m - \bar{q}_1 \bar{q}_1^T) b^{(i)} \end{pmatrix}. \quad (18)$$

From a complexity point of view the MGS algorithm applied to A requires $2mn^2$ flops while the $(p-1)$ projections (18) for the remaining right-hand sides require $4mnp$ flops.

A second way is to reorthogonalize \bar{Q} , the Q-factor from MGS, before performing the set of projections. We reorthogonalize \bar{Q} to obtain Q_k using formula (7), with Algorithm 1. The value of k is chosen large enough so that Q_k has columns orthonormal up to machine precision. Then we project the $(p-1)$ remaining right-hand sides with classical Gram-Schmidt type projections, that is

$$c_{\text{CGS}}^{(i)} = \begin{pmatrix} \hat{q}_1^T b^{(i)} \\ \vdots \\ \hat{q}_n^T b^{(i)} \end{pmatrix}. \quad (19)$$

This latter approach still requires $2mn^2$ flops to get the QR-factorization of A but only $2mnp$ flops for the $(p-1)$ projections. However, we have to add the cost of the reorthogonalization that is mainly governed by the construction of T , that is mn^2 flops, plus the assembly of Q_k with Equation (7), that is $4mnk$ flops.

A third approach consists in not using MGS as orthogonalization scheme in GMRES but instead iterated modified Gram-Schmidt with a criterion denoted by MGS2(K) [9]. The extra costs compared with MGS comes from the reorthogonalizations. We call ν the quantity so that the cost of MGS2(K) is $2mn^2\nu$; ν ranges from 1 (if no reorthogonalization is performed) to 2 (if one reorthogonalization per column is performed). The parameter K defines the criterion used to decide whether the reorthogonalization has to be performed or not. According to [3], we choose the value $K = \sqrt{2}$ and justify this choice later through numerical experiments. The aim here is to obtain directly an orthogonal basis to machine precision and then use Equation (19) with the Q-factor obtained with MGS2(K).

MGS and $(p - 1)$ projections (18)	$2mn^2 + 4mnp$
Algorithm 1 and $(p - 1)$ projections (19)	$2mn^2 + 2mnp + mn(n + 4k)$
MGS2(K) and $(p - 1)$ projections (19)	$2mn^2\nu + 2mnp$

Table 3: Flops required for the three orthogonalization schemes and associated projection considering m large over k , n and p .

We summarize the costs in flops of these three approaches in Table 3. From Table 3, for rather small p a good approach in term of flops seems to be MGS. However our interest is in large p . For large p , Algorithm 1 is interesting over MGS2(K) when $\frac{3}{2} + \frac{2k}{n} \leq \nu$. We have seen that the parameter k is determined a-priori by the level of orthogonality required by the user. In the sequel, we consider k small compared to n , the critical value is then $\nu = 1.50$. A larger value for ν would make our approach more efficient than MGS2(K) – and vice versa – since the construction of T which requires mn^2 is the main cost of Algorithm 1, therefore we compare $3mn^2$ (Algorithm 1) to $2mn^2\nu$ (MGS2(K)).

3.2.2 Special feature of $A = (b, ZV_{n-1})$

Greenbaum, Rozložník and Strakoš [7] have shown that for GMRES with orthogonalization schemes MGS, the quantity $\sigma_n((b, ZV_{n-1}))$ is of the order of the residual of GMRES obtained at step $n - 1$. When the residual is small, we expect $A = (b, ZV_{n-1})$ to be ill-conditioned and so an important loss of orthogonality is expected with MGS.

Since $\sigma_{n-1}((b, ZV_{n-1})) \geq \sigma_{n-1}(ZV_{n-1}) \geq \sigma_{n-1}(Z)\sigma_n(V_{n-1})$, if we assume Z and V_{n-1} well-conditioned, we get that κ_2 is close to one. We note that if the matrix (b, ZV_{n-1}) is numerically nonsingular then in [5], it is stated that \bar{Q} ($= V_n$) is well-conditioned and we only restrict our study to well-conditioned matrix Z . From this analysis, the value $k = 1$ is enough for the reorthogonalization of \bar{Q} with Algorithm 1 to obtain a Q -factor orthogonal up to machine precision. In the experimental part, we illustrate that $k = 1$ is indeed necessary and sufficient in the Seed-GMRES context.

For small k compared to n , the cost of the a-posteriori reorthogonalization procedure of MGS performed with Algorithm 1 is mainly governed by the computation of the $n(n + 1)/2$ entries of the matrix \bar{T} (Section 3.2.1). We degrade Algorithm 1 to get a second algorithm, this algorithm relies mainly on an heuristic that attempts to avoid the complete computation of \bar{T} . First of all we consider that the rank of \bar{P}_{11} is one, – this is justified by

the special feature of the problem: κ large and κ_2 close to one – and since P_{11} is strictly upper triangular therefore nilpotent (i.e. $P_{11}^n = 0$), we have $P_{11}^2 = 0$ and so Equation (17) reduces to $P_{11} = T$. Therefore in practice we just compute T and use it as P_{11} . But computing all the entries of a rank–one matrix may be considered as a waste of time. In theory, it is enough to build a row i and a column j so that the entry (i, j) is nonzero. With rounding errors, the best choice is to build the row i and the column j such that the entry (i, j) is the largest in magnitude. In practice, if the entry (i, j) is not the largest but of the order of the largest entry of \bar{T} , the procedure is still reliable. A good candidate to be of the order of the largest entry of \bar{T} is $|\bar{q}_1^T \bar{q}_n|$ since the orthogonality given by MGS of \bar{q}_n over \bar{q}_1 assumes in theory the orthogonality of all the previous vectors ; in practice, we expect the loss of orthogonality in V to be maximal between \bar{q}_n and \bar{q}_1 . This defines our heuristic:

Heuristic

$|\bar{q}_1^T \bar{q}_n|$ is of the order of the largest entry in magnitude of \bar{T} .

Thanks to this heuristic only the first row and the last column of \bar{T} are computed.

Algorithm 2 uses the reorthogonalization based on this heuristic, it is described in Table 4. The fourth approach to compute the orthogonalization and the projections in Seed–GMRES is to use Algorithm 2 and then project the $(p - 1)$ other right–hand sides with Equation (19). The whole algorithm is very cheap and only requires $2mn^2 + 2mnp + 8mn$ flops in which $8mn$ flops are necessary for the reorthogonalization. For comparison, $8mn$ corresponds to the extra cost of the reorthogonalization of about 4 columns. Finally, let us remark that if $p < n$, then it is worth to use the factorized form of Q_1 instead of computing it explicitly as suggested by line 4 of Algorithm 2.

3.2.3 Numerical experiments in a large electromagnetism calculation

Our case study arises from large calculations in electromagnetism. The boundary element method is used to discretize the 3D Maxwell’s equations on the surface of an object. The formulation relies on the combined field integral equations and the preconditioner used is a sparse approximate inverse [17], this means that in practice the preconditioned matrix Z is well–conditioned. Moreover one can notice that the matrix Z is not explicitly known and is accessed through matrix–vector product done via the fast multipole method. All the calculations are performed using double precision

1. run MGS on $A = (b, ZV_{n-1})$ to obtain \bar{Q} and \bar{R} ,
2. compute $u^T = (\bar{q}_n^T \bar{q}_1, \dots, \bar{q}_n^T \bar{q}_{n-1}, 0)$, $c = u(1)$
and $w^T = (0, \bar{q}_1^T \bar{q}_2, \dots, \bar{q}_1^T \bar{q}_n)$,
3. $c = u(1)$, $u = u/\|u\|$, $w = w/\|w\|$, $c = c/u(1)/w(n)$, $s = \sqrt{1 - c^2}$,
4. compute $Q_1 = \bar{Q} + \bar{Q}(w(s^{-1} - 1) - ucs^{-1})w^T$.

Table 4: Algorithm 2: MGS with an a-posteriori reorthogonalization by a rank-one update using the heuristic.

arithmetic. There are several linear systems $Zx^{(i)} = b^{(i)}$ to be solved, for this typical calculation we have $p = 180$ but this value might be much larger if several radar cross sections have to be computed, as is often the case in engineering applications. For each right-hand side, GMRES is stopped at iteration l if the approximate solution $x_l^{(i)}$ verifies $\|b^{(i)} - Ax_l^{(i)}\|/\|b^{(i)}\| \leq 10^{-14}$. We remark that the problem is defined in complex arithmetic, however in order to be consistent with the whole paper the real notation is maintained.

Four geometries are considered, they represent standard test-cases for electromagnetism calculations, namely a cetaf, an Airbus airplane, a sphere and an almond [17]. In Table 5, we give the characteristics of the matrices (b, ZV_{n-1}) obtained by a GMRES-MGS run on these matrices. The values obtained with GMRES-MGS2(K) are the same. For more information on the method and the test-case, we refer to [17].

In Table 5, ($\#$ iter) represents the number of GMRES iterations required to converge. The number of columns of the matrix $A = (b, ZV_{n-1})$ is $n = \#$ iter + 1, the number of rows is m . As expected (see Section 3.2.2), the condition number κ is such that $\kappa \cdot 10^{-14}$ is close to one, while κ_2 is of order $\mathcal{O}(1)$.

The fourth column of Table 5 corresponds to the average number of reorthogonalizations obtained with MGS2($\sqrt{2}$). In this cases, MGS2($\sqrt{2}$) systematically performs an extra reorthogonalization per matrix-vector product, which explains the constant value ($\nu = 2.00$).

In Table 6, we illustrate that all the residual errors $\|A - \bar{Q}\bar{R}\|$ – where \bar{Q} and \bar{R} designed the QR-factor given by one the four algorithms – are of the order of the machine precision. In Table 7, the different levels of

	m	# iter	κ	κ_2	ν
Cetaf	5391	31	$9.7 \cdot 10^{14}$	27	2.00
Airbus	23676	104	$3.6 \cdot 10^{14}$	14	2.00
Sphere	40368	59	$3.9 \cdot 10^{14}$	6.4	2.00
Almond	104973	71	$5.1 \cdot 10^{14}$	5.9	2.00

Table 5: Characteristics of $A = (b, ZV_{n-1})$

	MGS	Algorithm 1	MGS2($\sqrt{2}$)	Algorithm 2
Cetaf	$2.8 \cdot 10^{-17}$	$2.8 \cdot 10^{-16}$	$1.8 \cdot 10^{-16}$	$2.9 \cdot 10^{-16}$
Airbus	$4.0 \cdot 10^{-17}$	$4.4 \cdot 10^{-16}$	$2.7 \cdot 10^{-16}$	$4.4 \cdot 10^{-16}$
Sphere	$5.8 \cdot 10^{-17}$	$2.7 \cdot 10^{-16}$	$1.6 \cdot 10^{-16}$	$2.7 \cdot 10^{-16}$
Almond	$3.9 \cdot 10^{-17}$	$3.9 \cdot 10^{-16}$	$3.9 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$

Table 6: Residual errors for the four case-test and the different algorithms.

orthogonality characterized with $\|I_n - \bar{Q}^T \bar{Q}\|$ are given. As expected, MGS completely loses the orthogonality while the three other approaches give a set of vectors orthogonal up to machine precision. In the context of Seed-GMRES, this enables us to use confidently Equation (19) to project the $(p - 1)$ remaining right-hand sides.

A conclusion drawn from Table 7 is that in the case of GMRES-MGS applied to a not too ill-conditioned matrix the value $k = 1$ is satisfactory (Algorithm 1 with $k = 1$). Moreover from Table 6 and Table 7, we observe that Algorithm 2 relying on the heuristic works fine in practice.

One might question about the relevance of the choice $K = \sqrt{2}$ and its possible artificial high cost. In Table 8 we report on the sensitivity of the orthogonality quality with respect to the choice of the threshold. These experiments assess the choice of $K = \sqrt{2}$ for MGS2(K). This value gives a

	MGS	Algorithm 1	MGS2($\sqrt{2}$)	Algorithm 2
Cetaf	$1.6 \cdot 10^{-02}$	$1.9 \cdot 10^{-15}$	$2.8 \cdot 10^{-16}$	$2.4 \cdot 10^{-15}$
Airbus	$1.8 \cdot 10^{-02}$	$1.5 \cdot 10^{-15}$	$3.7 \cdot 10^{-16}$	$1.6 \cdot 10^{-15}$
Sphere	$3.9 \cdot 10^{-02}$	$5.4 \cdot 10^{-16}$	$3.0 \cdot 10^{-16}$	$7.8 \cdot 10^{-16}$
Almond	$4.1 \cdot 10^{-02}$	$6.8 \cdot 10^{-16}$	$2.8 \cdot 10^{-16}$	$7.9 \cdot 10^{-16}$

Table 7: $\|I_n - \bar{Q}^T \bar{Q}\|$ for the four case-test and the different algorithms.

	MGS2($\sqrt{2}$)	MGS2(2)	MGS2($\sqrt{5}$)
Cetaf	$2.8 \cdot 10^{-16}$ ($\nu = 2.00$)	$6.3 \cdot 10^{-16}$ ($\nu = 1.90$)	$1.2 \cdot 10^{-15}$ ($\nu = 1.87$)
Airbus	$3.7 \cdot 10^{-16}$ ($\nu = 2.00$)	$3.9 \cdot 10^{-03}$ ($\nu = 1.02$)	$8.8 \cdot 10^{-03}$ ($\nu = 1.01$)
Sphere	$3.0 \cdot 10^{-16}$ ($\nu = 2.00$)	$7.5 \cdot 10^{-15}$ ($\nu = 1.52$)	$4.9 \cdot 10^{-04}$ ($\nu = 1.07$)
Almond	$2.8 \cdot 10^{-16}$ ($\nu = 2.00$)	$1.7 \cdot 10^{-03}$ ($\nu = 1.06$)	$5.2 \cdot 10^{-03}$ ($\nu = 1.03$)

Table 8: $\|I_n - \bar{Q}^T \bar{Q}\|$.

good orthogonality level for all the examples while the others tested ($K = 2$ and $K = \sqrt{5}$) fail. However $K = \sqrt{2}$ implies in these cases $\nu = 2.00$ meaning that the criterion is unable to save any reorthogonalization. This result is not satisfactory and highlights a weakness of the MGS2(K) procedure. Even if κ_2 is close to one, improving noticeably the condition number κ cannot be obtained in the general case by removing only a column of (b, ZV_{n-1}) , it is a global phenomenon that needs a global treatment (e.g. to add the singular vector associated to the smallest singular value to all the columns). In the same way, the loss of orthogonality is global and affects all the columns of \bar{Q} . An algorithm like MGS2(K) that acts locally on each column performs poorly in this case, whereas Algorithm 1 and 2 represent appealing strategies since the reorthogonalization – that has to be global – is expressed as a rank-one update.

Finally, there exists other examples where the value of $k > 1$ can be given a priori. Still for the solution of linear systems with multiple right-hand sides, we mention for instance the Block(k)–Seed–GMRES–MGS algorithm; that is one run Block GMRES on k vectors, when the convergence is observed, a rank- k update is performed to recover an orthogonal set of vector, that we use to project the $p - k$ right-hand sides as in Seed–GMRES.

4 Conclusion

In this paper we propose an *a posteriori* reorthogonalization technique based on a rank- k update to reorthogonalize a set of vectors built by the modified Gram–Schmidt algorithm. We show that for large enough k , we can fully recover the orthogonality. We illustrate the effectiveness of our technique in the framework of the iterative solution of linear systems based on the GMRES algorithm. On a set of industrial test problems we demonstrate that our algorithm is efficient and outperforms classical approaches that also permit to remedy the loss of orthogonality observed when GMRES has

converged.

Acknowledgments

The authors would like to thank C. C. Paige for his careful reading of an earlier version of this report paper and for his fruitful comments. The work of J. Langou was supported by EADS, Corporate Research Centre, Toulouse.

References

- [1] Åke Björck. Solving linear least squares problems by Gram–Schmidt orthogonalization. *BIT*, 7:1–21, 1967.
- [2] Åke Björck and Chris C. Paige. Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, 13(1):176–190, 1992.
- [3] J. W. Daniel, W. B. Gragg, L. Kaufman and G. W. Stewart. Re-orthogonalization and Stable Algorithms for Updating the Gram–Schmidt QR Factorization. *Math. Comp.*, 30(136):772–795, 1976.
- [4] Ky Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proc. Amer. Math.*, 6:111–116, 1955.
- [5] Luc Giraud and Julien Langou. When modified Gram–Schmidt generates a well–conditioned set of vectors. *IMA Journal of Numerical Analysis*, 22(4):521–528, 2002.
- [6] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Third Edition, Johns Hopkins University Press, Baltimore, MD, USA, 1996.
- [7] Anne Greenbaum, Miroslav Rozložník and Zdenek Strakoš. Numerical behaviour of the modified Gram–Schmidt GMRES implementation. *BIT*, 37:706–719, 1997.
- [8] Nicholas J. Higham. Matrix nearness problems and applications. In *Applications of Matrix Theory*, Oxford University Press, Oxford, UK, 1989.
- [9] W. Hoffmann. Iterative algorithms for Gram–Schmidt orthogonalization. *Computing*, 41:335–348, 1989.

- [10] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.
- [11] Robert S. Schreiber and Charles F. Van Loan. A storage-efficient WY representations for products of Householder transformations. *SIAM J. Sci. Stat. Computing*, 10:53–57, 1989.
- [12] Chris C. Paige and Michael A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM TOMS*, 8(1):43–71, 1982.
- [13] Beresford N. Parlett. A new look at the Lanczos algorithm for solving systems of linear equations. *Linear Algebra Appl.*, 29:323–346, 1980.
- [14] John R. Rice. Experiments on Gram–Schmidt orthogonalization. *Math. Comp.*, 20:325–328, 1966.
- [15] Heinz Rutishauser. Description of Algol 60. In *Handbook for Automatic Computation*, Volume 1, Part a, Springer Verlag, New York Inc., 1967.
- [16] Yousef Saad. On the Lanczos method for solving symmetric linear systems with several right–hand sides. *Math. Comp.*, 48:651–662, 1987.
- [17] Guillaume Sylvand. *La méthode multipôle rapide en électromagnétisme : performances, parallélisation et applications*. Ph.D. Thesis manuscript from Ecole Nationale des Ponts et Chaussées, 2002.
- [18] James H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, Oxford, UK, 1965.

A Three technical Lemmas

In this appendix, we prove three technical lemmas that are used in the proof of Theorem 2.3. Lemma A.1 relates the norm of the computed \tilde{Q} to the condition number of A . We prove that for a well–conditioned matrix A , $\|\tilde{Q}\|$ is close to one. Lemma A.2 gives an upper bound for the quantity $\|A - Q\tilde{R}\|$. Similar residual bounds have been derived in [1, 2], but they involve the computed \tilde{Q} , instead of \tilde{Q} . In these two lemmas, notations directly refers to the ones used in the article. Lemma A.3 is a general theorem of matrix theory.

Lemma A.1 *Suppose that $\bar{c}_1 u \eta \kappa < 1$, then*

$$\|\tilde{Q}\| \leq \frac{1 + \bar{c}_1 u \eta \kappa}{1 - \bar{c}_1 u \eta \kappa}.$$

Proof: Since $\tilde{P}_{21} = \tilde{Q}(I_n - \tilde{P}_{11})$, we obtain

$$\|\tilde{Q}\| \leq \frac{\|P_{21}\|}{\sigma_n(I_n - P_{11})},$$

where we have used that matrix $I_n - \tilde{P}_{11}$ is nonsingular. Indeed \tilde{P}_{11} is strictly upper triangular so $I_n - \tilde{P}_{11}$ is unit upper triangular and therefore nonsingular. From the variational characterization of the singular values, it follows that, for any matrices X and Y , $\sigma_i(X + Y) \geq \sigma_i(X) - \|Y\|$ (see e.g. [10]), therefore,

$$\|\tilde{Q}\| \leq \frac{\|\tilde{P}_{21}\|}{1 - \|\tilde{P}_{11}\|}. \quad (20)$$

Using (2), we obtain that $\|\tilde{P}_{11}\| = \|E_1 \bar{R}^{-1}\| \leq \|E_1\| \|\bar{R}^{-1}\| \leq \bar{c}_1 u \|A\| / \bar{\sigma}_n$, which yields, using $\bar{\sigma}_n \geq \sigma_n - \bar{c} u \sigma_1$,

$$\|\tilde{P}_{11}\| \leq \bar{c}_1 u \sigma_1 / (\sigma_n - \bar{c} u \sigma_1) \leq \bar{c}_1 u \kappa / (1 - \bar{c} u \kappa) \leq \bar{c}_1 \eta \kappa u. \quad (21)$$

Since \tilde{P} has orthonormal columns, this implies that $\tilde{P}_{11}^T \tilde{P}_{11} + \tilde{P}_{21}^T \tilde{P}_{21} = I_n$, hence, $\|\tilde{P}_{21}\|^2 \leq 1 + \|\tilde{P}_{11}\|^2 \leq (1 + \|\tilde{P}_{11}\|)^2$, that is

$$\|\tilde{P}_{21}\| \leq 1 + \|\tilde{P}_{11}\|, \quad (22)$$

and the conclusion follows readily from (20), (21) and (22). □

Lemma A.2 *Suppose that $\bar{c}_1 u \eta \kappa < 1$, then*

$$\|A - \tilde{Q} \bar{R}\| \leq \left(\bar{c}_2 + \bar{c}_1 \frac{1 + \bar{c}_1 u \eta \kappa}{1 - \bar{c}_1 u \eta \kappa} \right) u \|A\|.$$

Proof: Since $-E_2 = A - \tilde{P}_{21} \bar{R}$, $E_1 = \tilde{P}_{11} \bar{R}$ and $\tilde{P}_{21} = \tilde{Q}(I_n - \tilde{P}_{11})$,

$$A - \tilde{Q} \bar{R} = -E_2 - \tilde{Q} E_1.$$

The conclusion follows Lemma A.1 and $\|E_i\| \leq \bar{c}_i u \|A\|$, $i = 1, 2$.

Lemma A.3 *Let consider A an $n \times n$ matrix and C (resp. D diagonal matrix of order n with diagonal entry i equal to c_i (resp. d_i), $i = 1, \dots, n$ such that*

$$\forall k = 1, \dots, n, \quad |d_k| \leq |c_k| \eta,$$

where η is a real. Then

$$\|DA\| \leq \|CA\| \eta.$$

Proof: First of all, we remark that $\eta \geq 0$. Then if $\eta = 0$, this implies $d_k = 0$ and the lemma is true. Therefore the non-trivial case is $\eta > 0$. In this case, we remark that $c_k = 0 \Rightarrow d_k = 0$. Let us define

$$J_0 = \{i \text{ such that } c_i = 0\} \quad \text{and} \quad J_1 = \{i \text{ such that } c_i \neq 0\},$$

we consider the truncated matrix A_1 by retaining the lines i such that $i \in J_1$ in its counterpart A . In Matlab style notation, $A_1 = A(J_1, :)$. We also denote by C_1 (resp. D_1) the diagonal matrix whose diagonal elements are the $c_i, i \in J_1$ (resp. $d_i, i \in J_1$).

From this we have

$$\|DA\| = \|D_1 A_1\| = \|(D_1 C_1^{-1})(C_1 A_1)\|.$$

In term of norms,

$$\|DA\| \leq \|D_1 C_1^{-1}\| \|C_1 A_1\| = \|D_1 C_1^{-1}\| \|CA\|.$$

We conclude by noticing that $\|D_1 C_1^{-1}\| \leq \eta$.