*Article Submitted to ***Journal title****

WHEN MODIFIED GRAM-SCHMIDT GENERATES A WELL-CONDITIONED SET OF VECTORS

L. GIRAUD[1] AND J. LANGOU[*2]

[1]CERFACS, *42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France.*
Luc.Giraud@cerfacs.fr

[2]CERFACS, *42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France.*
Julien.Langou@cerfacs.fr

Abstract

Orthogonalization methods play a key role in many iterative methods. In this paper, we establish new properties for the modified Gram-Schmidt algorithm. We show why the modified Gram-Schmidt algorithm generates a well-conditioned set of vectors. This result holds under the assumption that the initial matrix is not "too ill-conditioned" in a way that is quantified. As a consequence we show that if two iterations of the algorithm are performed, the resulting algorithm produces a matrix whose columns are orthogonal up to machine precision. Finally we illustrate through a numerical experiment the sharpness of our result.

1. Introduction

Orthogonalization methods play a key role in many iterative methods. In this paper we study the condition number of the set of vectors generated by the Modified Gram-Schmidt (MGS) algorithm in floating-point arithmetic. After a quick review, in Section 2, of the fundamental results that we use, we devote Section 3 to our main theorem. Through this central theorem we give an upper bound close to one for the condition number of the set of vectors produced by MGS. This theorem applies to matrices that are not "too ill-conditioned". In Section 4, we combine our theorem with a well-known result from Björck to obtain that two iterations of MGS are indeed enough to get a matrix whose columns are orthogonal up to machine precision. We conclude Section 4 by exhibiting a counter example matrix. This matrix shows that if we relax the constraint on the condition number of the studied matrices, no pertinent information on the

[*]The work of this author was supported by EADS, Corporate Research Center, Toulouse.

upper bound of the condition number of the set of vectors generated by MGS can be gained.

2. Previous results and notations

We consider the MGS algorithm applied to a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with full rank $n \leq m$ and singular values : $\sigma_1 \geq \ldots \geq \sigma_n > 0$, we define the condition number of $\mathbf{A}$ as $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$.
Using results from Björck (1967) and Björck and Paige (1992), we know that, in floating-point arithmetic, MGS computes $\bar{\mathbf{Q}} \in \mathbb{R}^{m\times n}$ and $\bar{\mathbf{R}} \in \mathbb{R}^{n\times n}$ so that there exists $\bar{\mathbf{E}} \in \mathbb{R}^{m\times n}$, $\hat{\mathbf{E}} \in \mathbb{R}^{m\times n}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{m\times n}$, where

$$\mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}} \quad \text{and} \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \tag{2.1}$$

$$\|\mathbf{I} - \bar{\mathbf{Q}}^T\bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \tag{2.2}$$

$$\mathbf{A} + \hat{\mathbf{E}} = \hat{\mathbf{Q}}\bar{\mathbf{R}} \quad , \quad \hat{\mathbf{Q}}^T\hat{\mathbf{Q}} = \mathbf{I} \quad \text{and} \quad \|\hat{\mathbf{E}}\|_2 \leq c u \|\mathbf{A}\|_2, \tag{2.3}$$

$\bar{c}_i$ and c are constants depending on m, n and the details of the arithmetic, and $u = 2^{-t}$ is the unit round off.
Result (2.1) shows that $\bar{\mathbf{Q}}\bar{\mathbf{R}}$ is a backward-stable factorization of $\mathbf{A}$, that is the product $\bar{\mathbf{Q}}\bar{\mathbf{R}}$ represents accurately $\mathbf{A}$ up to machine precision.
Equation (2.3) says that $\bar{\mathbf{R}}$ solves the QR-factorization problem in a backward stable sense; that is, there exists an exact orthonormal matrix $\hat{\mathbf{Q}}$ so that $\hat{\mathbf{Q}}\bar{\mathbf{R}}$ is a QR factorization of a slight perturbation of $\mathbf{A}$.
We notice that results (2.1) from Björck (1967) and (2.3) from Björck and Paige (1992) are proved under assumptions :

$$2.12 \cdot (m+1)u < 0.01, \tag{2.4}$$

$$cu\kappa(\mathbf{A}) < 1. \tag{2.5}$$

For clarity, it is important to explicitly define the constants that are involved in the upper bounds of the inequalities. Complying with assumptions (2.4) and (2.5) we can set the constant c and $\bar{c}_1$ to

$$c = 18.53 \cdot n^{\frac{3}{2}} \quad \text{and} \quad \bar{c}_1 = 1.853 \cdot n^{\frac{3}{2}} = 0.1 \cdot c. \tag{2.6}$$

It is worth noticing that the value of c does only depend on n, the number of vectors to be orthogonalized, but not on m, the size of the vectors since (2.4) holds.
Assumption (2.5) prevents $\bar{\mathbf{R}}$ to be singular. Under this assumption and defining

$$\eta = \frac{1}{1 - cu\kappa(\mathbf{A})}, \tag{2.7}$$

Björck and Paige (1992) obtain an upper bound for $\|\bar{\mathbf{R}}^{-1}\|_2$ as

$$\|\mathbf{A}\|_2 \|\bar{\mathbf{R}}^{-1}\|_2 \leq \eta\kappa(\mathbf{A}). \tag{2.8}$$

Assuming (2.5), we note that (2.1) and (2.3) are independent of $\kappa(\mathbf{A})$. This is not the case for inequality (2.2) : the level of orthogonality in $\bar{\mathbf{Q}}$ is dependent on $\kappa(\mathbf{A})$. If $\mathbf{A}$ is well-conditioned then $\bar{\mathbf{Q}}$ is orthogonal to machine precision. But for an ill-conditioned matrix $\mathbf{A}$, the set of vectors $\bar{\mathbf{Q}}$ may lose orthogonality. An important question that arises then is whether MGS manages to preserve the full rank of $\bar{\mathbf{Q}}$ or not. In order to investigate this, we study in the next section the condition number of $\bar{\mathbf{Q}}$. For this purpose, we define the singular values of $\bar{\mathbf{Q}}$, $\sigma_1(\bar{\mathbf{Q}}) \geq \ldots \geq \sigma_n(\bar{\mathbf{Q}})$. When $\bar{\mathbf{Q}}$ is non singular, $\sigma_n(\bar{\mathbf{Q}}) > 0$, we also define the condition number $\kappa(\bar{\mathbf{Q}}) = \sigma_1(\bar{\mathbf{Q}})/\sigma_n(\bar{\mathbf{Q}})$.

3. Conditioning of the set of vectors $\bar{\mathbf{Q}}$

This section is fully devoted to the key theorem of this paper and to its proof. For sake of completeness, we establish a similar result using different arguments in the next section. The central theorem is the following.

THEOREM 3.1:

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with full rank $n \leq m$ and condition number $\kappa(\mathbf{A})$ such as

$$\begin{array}{rc} & 2.12 \cdot (m+1)u < 0.01 \\ \textit{and} & cu\kappa(\mathbf{A}) \leq 0.1, \end{array} \tag{3.9}$$

where $c = 18.53 \cdot n^{\frac{3}{2}}$ and u is the unit roundoff.
Then MGS in floating-point arithmetic computes $\bar{\mathbf{Q}} \in \mathbb{R}^{m\times n}$ so as

$$\boxed{\kappa(\bar{\mathbf{Q}}) \leq 1.3.} \tag{3.10}$$

We notice that assumption (3.9) is just slightly stronger than assumption (2.5) made by Björck and Paige (1992).

Proof :

On one hand, MGS computes $\bar{\mathbf{Q}}$, on the other hand, the matrix $\hat{\mathbf{Q}}$ has exactly orthonormal columns. It seems natural to study the distance between $\bar{\mathbf{Q}}$ and $\hat{\mathbf{Q}}$. In that respect we define $\mathbf{F}$ as

$$\mathbf{F} = \bar{\mathbf{Q}} - \hat{\mathbf{Q}}, \tag{3.11}$$

and look at its 2-norm. For this purpose, we subtract (2.3) from (2.1) to get

$$\begin{aligned} (\bar{\mathbf{Q}} - \hat{\mathbf{Q}})\bar{\mathbf{R}} &= \mathbf{A} + \bar{\mathbf{E}} - \mathbf{A} - \hat{\mathbf{E}}, \\ \mathbf{F}\bar{\mathbf{R}} &= \bar{\mathbf{E}} - \hat{\mathbf{E}}. \end{aligned}$$

Assuming $cu\kappa(\mathbf{A}) < 1$, $\bar{\mathbf{R}}$ is nonsingular and we can write

$$\mathbf{F} = (\bar{\mathbf{E}} - \hat{\mathbf{E}})\bar{\mathbf{R}}^{-1}.$$

We bound, in terms of norms, this equality

$$\|\mathbf{F}\|_2 \leq (\|\bar{\mathbf{E}}\|_2 + \|\hat{\mathbf{E}}\|_2)\|\bar{\mathbf{R}}^{-1}\|_2.$$

Using inequality (2.1) on $\|\bar{\mathbf{E}}\|_2$ and inequality (2.3) on $\|\hat{\mathbf{E}}\|_2$, we obtain

$$\|\mathbf{F}\|_2 \leq (c + \bar{c}_1)u\|\mathbf{A}\|_2\|\bar{\mathbf{R}}^{-1}\|_2.$$

Using inequality (2.8) on $\|\mathbf{A}\|_2\|\bar{\mathbf{R}}^{-1}\|_2$ and (2.6), we have

$$\boxed{\|\mathbf{F}\|_2 \leq 1.1 \cdot cu\eta\kappa(\mathbf{A}).} \tag{3.12}$$

This is the desired bound on $\|\mathbf{F}\|_2$.
Since we are interested in an upper bound on $\kappa(\bar{\mathbf{Q}})$, the condition number of $\bar{\mathbf{Q}}$, we then look for an upper bound for the largest singular value of $\bar{\mathbf{Q}}$ and a lower bound for its smallest singular value.
From Golub and Van Loan (1983, p. 449), we know that (3.11) implies :

$$\sigma_1(\bar{\mathbf{Q}}) \leq \sigma_1(\hat{\mathbf{Q}}) + \|\mathbf{F}\|_2 \quad \text{and} \quad \sigma_n(\bar{\mathbf{Q}}) \geq \sigma_n(\hat{\mathbf{Q}}) - \|\mathbf{F}\|_2.$$

Since $\hat{\mathbf{Q}}$ has exactly orthonormal columns, we have $\sigma_1(\hat{\mathbf{Q}}) = \sigma_n(\hat{\mathbf{Q}}) = 1$. Using the bound (3.12) on $\|\mathbf{F}\|_2$, we get

$$\sigma_1(\bar{\mathbf{Q}}) \leq 1 + 1.1 \cdot cu\eta\kappa(\mathbf{A}) \quad \text{and} \quad \sigma_n(\bar{\mathbf{Q}}) \geq 1 - 1.1 \cdot cu\eta\kappa(\mathbf{A}).$$

With (2.7), these inequalities can be written as

$$\sigma_1(\bar{\mathbf{Q}}) \leq \eta(1 - cu\kappa(\mathbf{A}) + 1.1 \cdot cu\kappa(\mathbf{A})) = \eta(1 + 0.1 \cdot cu\kappa(\mathbf{A}))$$

and

$$\sigma_n(\bar{\mathbf{Q}}) \geq \eta(1 - cu\kappa(\mathbf{A}) - 1.1 \cdot cu\kappa(\mathbf{A})) = \eta(1 - 2.1 \cdot cu\kappa(\mathbf{A})).$$

If we assume

$$2.1 \cdot cu\kappa(\mathbf{A}) < 1, \tag{3.13}$$

$\sigma_n(\bar{\mathbf{Q}}) > 0$ so $\bar{\mathbf{Q}}$ is nonsingular.
Under this assumption, we have :

$$\boxed{\kappa(\bar{\mathbf{Q}}) \leq \tfrac{1+0.1\cdot cu\kappa(\mathbf{A})}{1-2.1\cdot cu\kappa(\mathbf{A})}.} \tag{3.14}$$

To illustrate the behaviour of the upper bound of $\kappa(\bar{\mathbf{Q}})$, we plot in Figure 1 the upper bound as a function of $c\kappa(\mathbf{A})$. We fix $u = 1.12e - 16$.
It can be seen that this upper bound explodes when $2.1 \cdot cu\kappa(\mathbf{A}) \lesssim 1$ but in the main part of the domain where $2.1 \cdot cu\kappa(\mathbf{A}) < 1$ it is small and very close to one. For instance, if we slightly increase the constraint (2.5) used by Björck and Paige (1992) and assume that $cu\kappa(\mathbf{A}) < 0.1$ then $\kappa(\bar{\mathbf{Q}}) < 1.3$. ■

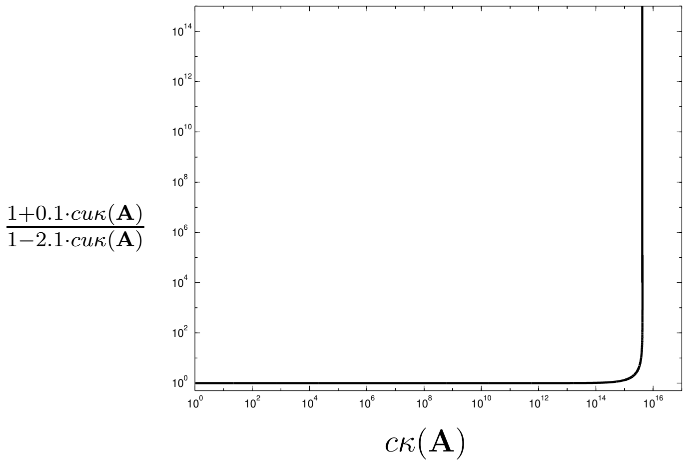

Figure 1: Behaviour of the upper bound on $\kappa(\bar{\mathbf{Q}})$ as a function of $c\kappa(\mathbf{A})$.

4. Some remarks

4.1. Iterative Modified Gram-Schmidt

If the assumption (3.9) on the condition number of $\mathbf{A}$ holds, then we obtain, after a first sweep of MGS, $\bar{\mathbf{Q}}_1$ satisfying (3.14). If we run MGS a second time on $\bar{\mathbf{Q}}_1$ to obtain $\bar{\mathbf{Q}}_2$, we deduce using (2.2) that $\bar{\mathbf{Q}}_2$ is such that :

$$\|\mathbf{I} - \bar{\mathbf{Q}}_2^T \bar{\mathbf{Q}}_2\|_2 \leq 1.71 \cdot c\kappa(\bar{\mathbf{Q}}_1)u,$$

so we get

$$\boxed{\|\mathbf{I} - \bar{\mathbf{Q}}_2^T \bar{\mathbf{Q}}_2\|_2 < 40.52 \cdot un^{\frac{3}{2}},} \tag{4.15}$$

meaning that $\bar{\mathbf{Q}}_2$ has columns orthonormal to machine precision. Two MGS sweeps are indeed enough to have an orthonormal set of vectors $\mathbf{Q}$.
We recover, in a slightly different framework, the famous sentence of Kahan

"Twice is enough."

Based on unpublished notes of Kahan, Parlett (1980) shows that an iterative Gram-Schmidt process on two vectors with a selective criterion (optional) produces two vectors orthonormal up to machine precision. In this paper, inequality (4.15) show that *twice is enough* for n vectors under assumptions (2.4) and (3.9) with MGS and a complete a posteriori re-orthogonalization (i.e. no selective criterion).

4.2. What can be said on $\kappa(\bar{\mathbf{Q}})$ when $cu\kappa(\mathbf{A}) > 0.1$

For $2.1 \cdot cu\kappa(\mathbf{A}) < 1$, the bound (3.14) on $\kappa(\bar{\mathbf{Q}})$ is well defined but when $cu\kappa(\mathbf{A}) > 0.1$, this bound explodes and very quickly nothing interesting can be said about the condition number of $\bar{\mathbf{Q}}$. For $2.1 \cdot cu\kappa(\mathbf{A}) > 1$, we even do not have any bound.

In this part, we ask the question whether there can exist or not an interesting upper bound on $\bar{\mathbf{Q}}$ when $cu\kappa(\mathbf{A}) > 0.1$. In order to answer this problem, we consider[†] the matrix *CERFACS* $\in \mathbb{R}^{3\times 3}$.

When we run MGS with Matlab on *CERFACS*, we obtain with $u = 1.12e - 16$

$$\kappa(\mathbf{A}) = 3e + 15, \quad cu\kappa(\mathbf{A}) = 37 \quad \text{and} \quad \kappa(\bar{\mathbf{Q}}) = 2e + 14 \ .$$

Matrix *CERFACS* generates a very ill-conditioned set of vectors $\bar{\mathbf{Q}}$ with $cu\kappa(\mathbf{A})$ not too far from 0.1.

If we are looking for an upper bound of $\kappa(\bar{\mathbf{Q}})$, we can take the value 1.3 up to $cu\kappa(\mathbf{A}) = 0.1$ and then this upper bound has to be greater than 2e+14 for $cu\kappa(\mathbf{A}) = 37$.

Matrix *CERFACS* proves that it is not possible to increase by much the domain of validity (i.e. $cu\kappa(\mathbf{A}) < 0.1$) of Theorem (3.1) in order to get a more interesting result.

One can also remark that with *CERFACS* two MGS sweeps are no longer enough since

$$\|\mathbf{I} - \bar{\mathbf{Q}}_2^T \bar{\mathbf{Q}}_2\|_2 = 2e - 03.$$

Acknowledgment

We would like to thank Miroslav Rozložník for fruitful discussions on the Modified Gram-Schmidt algorithm and in particular for having highlighted that the sentence *twice is enough* required the assumption of a not “too ill-conditioned” matrix **A**.

References

Wilkinson J. H., (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, Walton Street, Oxford OX2 6DP. ISBN 0-19-853403-5.

Björck Å., (1967), *Solving linear least squares problems using Gram-Schmidt orthogonalization*, BIT, **7**, 1–21.

Parlett B. N., (1980), *The Symmetric Eigenvalue Problem*, Prentice-Hall, Inc.,Englewood Cliffs,N.J.07632. ISBN 0-13-880047-2.

Golub G. H. and Van Loan C. F., (1983), *Matrix Computations*, Johns Hopkins, 2715 North Charles Street, Baltimore, Maryland 21218-4319. ISBN 0-8018-5413-X.

Björck Å. and Paige C. C., (1992), *Loss and recapture of orthogonality in the modified Gram-Schmidt Algorithm*, SIAM J. Matrix Analysis and Applications, **13**, 176–190.

[†]see Appendix

Higham N. J., (1994), *The Matrix Sign Decomposition and Its Relation to the Polar Decomposition*, Linear Algebra and its Applications, **212/213**, 3–20.

Appendix : matrix *CERFACS*

We developed a Matlab code that generates as many as desired matrices with relatively small $cu\kappa(\mathbf{A})$ and large $\kappa(\bar{\mathbf{Q}})$. *CERFACS* is one of these.

$$CERFACS = \begin{pmatrix} 0.12100300219993308 & 2.09408775152625060 & 1.26139640819301024 \\ -0.10439395064078592 & -1.80665016070527140 & -1.08825526624380808 \\ 0.21661355806776747 & 0.49451660567698374 & -0.84174336538575500 \end{pmatrix}.$$