

Luc Giraud and Julien Langou

ANOTHER PROOF FOR MODIFIED GRAM-SCHMIDT WITH REORTHOGONALIZATION

Abstract. In this note, we consider the modified Gram-Schmidt algorithm with reorthogonalization applied on a numerical nonsingular matrix, we explain why the resulting set of vectors is orthogonal up to the machine precision level. To establish this result, we show that a certain L -criterion is necessarily verified after the second reorthogonalization step, then we prove that this L -criterion implies the desired level of orthogonality.

If the L -criterion is verified after the first orthogonalization step, then there is no need to reorthogonalize. From this simple observation, we deduce that the L -criterion is an interesting selective reorthogonalization criterion for modified Gram-Schmidt algorithm.

AMS Subject Classification : 65F25, 65G50, 15A23.

1. Introduction

Let $A = (a_1, \dots, a_n)$ be a real $m \times n$ matrix whose columns are linearly independent and $\kappa_2(A)$ be its condition number. In this paper, we give a rounding error analysis of the MGS2 algorithm applied on A . We based our work on the rounding error analysis of MGS done by Björck in [1] therefore the algorithm used is not the *classical* version of MGS2 but the square root free version of MGS2. We assume also that single precision floating point arithmetic is used and set the machine precision to 2^{-t} .

The result established in this note is already in [3]. The originality comes from the fact that the proof is not at all the same. We may say that the proof of [3] is adapted to CGS and the one that is presented here fit only for MGS. The interest of the proof given in [3] is that it applies either for CGS or MGS. The interest of the proof presented in this note is that it is straightforward to adapt for a MGS2(L), modified Gram-Schmidt algorithm with reorthogonalization and a L -criterion.

2. Description of the algorithm MGS2

The algorithm MGS2 in [3] uses square roots when it computes $\|\mathbf{q}_j\|_2$. In order to avoid the use of square roots during the run of the algorithm, we replace the normalized vectors $\mathbf{q}_1, \dots, \mathbf{q}_n$ by the non-normalized vectors $\mathbf{q}'_1, \dots, \mathbf{q}'_n$. This gives the algorithm MGS2 without square roots, that we will assimilate to MGS2. The factorization we get is then :

$$\mathbf{A} = \mathbf{Q}'\mathbf{R}'$$

where \mathbf{R}' is a unit upper triangular triangle and $(\mathbf{Q}')^T \mathbf{Q}'$ diagonal.

Algorithm 1 MGS2 without square roots

```

for  $j = 1$  to  $n$  do
   $\mathbf{a}_j^{(1)(1)} = \mathbf{a}_j$ 
  for  $k = 1$  to  $j - 1$  do
     $r_{kj}'^{(1)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(1)} / d_k$ 
     $\mathbf{a}_j^{(k+1)(1)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k r_{kj}'^{(1)}$ 
  end for
   $\mathbf{a}_j^{(1)(2)} = \mathbf{a}_j^{(j)(1)}$ 
  for  $k = 1$  to  $j - 1$  do
     $r_{kj}'^{(2)} = \mathbf{q}_k^T \mathbf{a}_j^{(k)(2)} / d_k$ 
     $\mathbf{a}_j^{(k+1)(2)} = \mathbf{a}_j^{(k)(1)} - \mathbf{q}_k r_{kj}'^{(2)}$ 
  end for
   $\mathbf{q}_j' = \mathbf{a}_j^{(j)(2)}$ 
   $d_j = \|\mathbf{q}_j'\|_2^2$ 
   $r_{kj}' = r_{kj}'^{(1)} + r_{kj}'^{(2)}, 1 \leq k \leq j - 1$ 
   $r_{jj}' = 1$ 
end for

```

3. Basic definitions for the error analysis

We define for $j = 1, \dots, n$, the computed quantity:

$$\begin{aligned}
 \bar{k}_j^{(r)} &= \text{fl}(\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(r)} / \bar{d}_k), & \text{for } k = 1, \dots, j - 1 \text{ and } r = 1, 2, \\
 \bar{\mathbf{a}}_j^{(k+1)(r)} &= \text{fl}(\bar{\mathbf{a}}_j^{(k)(r)} - \bar{\mathbf{q}}_k \bar{k}_j^{(r)}), & \text{for } k = 1, \dots, j - 1 \text{ and } r = 1, 2, \\
 \bar{\mathbf{q}}_j' &= \bar{\mathbf{a}}_j^{(j)(2)}, \\
 \bar{d}_j &= \text{fl}(\|\bar{\mathbf{q}}_j'\|_2^2), \\
 \bar{k}_j &= \text{fl}(\bar{k}_j^{(1)} + \bar{k}_j^{(2)}), & \text{for } k = 1, \dots, j - 1, \\
 \bar{f}_j &= \text{fl}(1).
 \end{aligned}$$

For the convenience of notation we also introduce the normalized quantities :

$$\bar{\mathbf{q}}_j = d_j^{-1/2} \bar{\mathbf{q}}_j', \quad \bar{k}_j^{(r)} = d_j^{1/2} \bar{k}_j^{(r)}, \quad \bar{k}_j = \bar{k}_j^{(1)} + \bar{k}_j^{(2)}, \quad \bar{f}_j = d_j^{1/2}, \quad (3.1)$$

where

$$d_j^{1/2} = \begin{cases} \|\bar{\mathbf{q}}_j'\|_2, & \bar{\mathbf{q}}_j' \neq 0, \\ 1, & \bar{\mathbf{q}}_j' = 0. \end{cases}$$

Note that these quantities are never computed and thus (3.1) are exact relations.

4. Errors in an elementary projection

In this section, we recall results from Björck [1]. We assume

$$m \geq 2, \quad 2n(m+1)2^{-t} < 0.01 \quad \text{and} \quad n \cdot 2^{-t} \leq 0.1, \quad (4.1)$$

with $t_1 = t - \log_2(1.06)$.

If $\bar{\mathbf{q}}_k' \neq \mathbf{0}$, we define the related errors $\delta_j^{(k)(r)}$ and $\eta_j^{(k)(r)}$ by

$$\bar{\mathbf{a}}_j^{(k+1)(r)} = \bar{\mathbf{a}}_j^{(k)(r)} - \bar{\mathbf{q}}_k \bar{\kappa}_j^{-r} + \delta_j^{(k)(r)}, \quad (4.2)$$

$$\bar{\mathbf{a}}_j^{(k+1)(r)} = (\mathbf{I} - \bar{\mathbf{q}}_k \bar{\mathbf{q}}_k^T) \bar{\mathbf{a}}_j^{(k)(r)} + \eta_j^{(k)(r)}. \quad (4.3)$$

In the singular case when $\bar{\mathbf{q}}_k' = \mathbf{0}$ these relations are satisfied with

$$\bar{\mathbf{a}}_j^{(k+1)(r)} = \bar{\mathbf{a}}_j^{(k)(r)} \quad \text{and} \quad \delta_j^{(k)(r)} = \eta_j^{(k)(r)} = \mathbf{0}. \quad (4.4)$$

In the nonsingular case, Björck has shown that :

$$\|\delta_j^{(k)(r)}\|_2 \leq 1.45 \cdot 2^{-t} \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 \quad \text{and} \quad \|\eta_j^{(k)(r)}\|_2 \leq (2m+3) \cdot 2^{-t_1} \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2. \quad (4.5)$$

The error between $\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(r)}$ and the computed value $\bar{\kappa}_j^{(r)}$ is known by the formula

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(r)} - \bar{\kappa}_j^{(r)}| < ((m+1) \cdot |\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(r)}| + m \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2) 2^{-t_1} \leq (2m+1) 2^{-t_1} \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2. \quad (4.6)$$

In exact arithmetic, the norm obtained after each orthogonal projection is always smaller than the initial vector. In floating point arithmetic, due to rounding errors, we can imagine that the norm of the vectors grows however we can control this increase. After n projections Björck have shown that

$$\|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 < 1.006 \|\mathbf{a}_j^{(1)(r)}\|_2 \quad \text{and} \quad \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 < 1.013 \|\mathbf{a}_j\|_2. \quad (4.7)$$

5. Errors in the factorization

We define

$$\mathbf{E} = \bar{\mathbf{Q}} \bar{\mathbf{R}} - \mathbf{A}. \quad (5.1)$$

We shall prove the following estimate :

$$\|\mathbf{E}\|_E < 2.94(n-1) \cdot 2^{-t} \|\mathbf{A}\|_E. \quad (5.2)$$

Summing (4.2) for $k = 1, 2, \dots, j-1$ and for $r = 1, 2$ and using

$$\bar{\mathbf{a}}_j^{(1)(1)} = \mathbf{a}_j, \quad \bar{\mathbf{a}}_j^{(j)(1)} = \bar{\mathbf{a}}_j^{(1)(2)}, \quad \bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{q}}_j \bar{\kappa}_j^{-1}, \quad \bar{\kappa}_j = \bar{\kappa}_j^{(1)} + \bar{\kappa}_j^{(2)},$$

we get

$$\sum_{k=1}^j \bar{\mathbf{q}}_k \bar{\kappa}_j - \mathbf{a}_j = \sum_{k=1}^{j-1} (\delta_j^{(k)(1)} + \delta_j^{(k)(2)}) = \delta_j. \quad (5.3)$$

From (4.5) follows

$$\|\delta_j\|_2 < 1.45 \cdot 2^{-t} \sum_{r=1}^2 \sum_{k=1}^{j-1} \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.$$

Using (4.7) we guess

$$\|\delta_j\|_2 < 2.94 \cdot 2^{-t} (j-1) \|\mathbf{a}_j\|_2$$

finally we get the desired result that is to say

$$\|\mathbf{E}\|_E = \left(\sum_{j=1}^n \|\delta_j\|_2^2 \right)^{1/2} < 2.94 \cdot 2^{-t} (n-1) \left(\sum_{j=1}^n \|\mathbf{a}_j\|_2^2 \right)^{1/2} = 2.94(n-1) \cdot 2^{-t} \|\mathbf{A}\|_E.$$

6. Nonsingularity of $\bar{\mathbf{A}}$

We now derive a sufficient condition for $\bar{\mathbf{A}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}$ to have rank n . Let the exact factorization of \mathbf{A} be $\mathbf{A} = \mathbf{Q}\mathbf{R}$ then using (5.1) and following Björck, we get that $\bar{\mathbf{A}}$ has rank n if :

$$2.94(n-1) \cdot 2^{-t} \|\mathbf{A}\|_E \|\mathbf{R}^{-1}\|_2 \leq \sqrt{2} - 1. \quad (6.1)$$

We assume in the following that (6.1) is satisfied.

7. Induction assumption

The orthogonality of the computed vectors $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \dots, \bar{\mathbf{q}}_n$ can be measured by the norm of the matrix $(\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}})$. Let $\mathbf{U}_p, p = 1, \dots, n$ be the strictly upper triangular matrix of size (p, p) with elements :

$$u_{ij} = \bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_j, \quad 1 \leq i < j \leq p \quad \text{and } u_{ij} = 0, \quad 1 \leq j \leq i \leq p.$$

We call $\mathbf{U} = \mathbf{U}_n$. Now we have :

$$\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}} = -(\mathbf{U} + \mathbf{U}^T). \quad (7.1)$$

We make an induction to show that $\|\mathbf{U}\|_2$ is small. Therefore, we assume that at step $p-1$:

$$\boxed{\|\mathbf{U}_{p-1}\|_2 \leq \lambda.} \quad (7.2)$$

Our aim is to show that at step p , we have $\|\mathbf{U}_p\|_2 \leq \lambda$. The value of λ is fixed during the proof. In the following, the variables are so that

$$1 \leq j \leq p \leq n.$$

8. Theorem of Pythagorus

In this part, we just look each step separately, therefore the exponent (r) is useless. In exact arithmetic, after the j -step of MGS, we would have

$$\mathbf{a}_j = \sum_{k=1}^{j-1} (\mathbf{q}_k r_{kj}) + \mathbf{a}_j^{(j)}$$

and as the vectors $\mathbf{q}_k, k = 1, \dots, j-1$ are orthonormal

$$\sum_{k=1}^{j-1} (r_{kj})^2 + \|\mathbf{a}_j^{(j)}\|_2^2 = \|\mathbf{a}_j\|_2^2. \quad (8.1)$$

This is the theorem of Pythagorus, we are interested in its validity with rounding errors.

The main effect of the rounding errors is the lost of orthogonality of the vectors (\mathbf{q}_k) . Let us take $\mathbf{q}_k, k = 1, \dots, j-1$ such that $\|\mathbf{q}_k\|_2 = 1$ without no other assumption. Then from \mathbf{a}_j we run the step j of the MGS algorithm, in exact arithmetic, we get

$$\begin{aligned} \mathbf{a}_j^{(1)} &= (\mathbf{I} - \mathbf{q}_1 \mathbf{q}_1^T) \mathbf{a}_j, \quad \text{with } r_{1j} = \mathbf{q}_1^T \mathbf{a}_j &\Rightarrow \|\mathbf{a}_j\|_2^2 &= (r_{1j})^2 + \|\mathbf{a}_j^{(1)}\|_2^2, \\ &\vdots &&\vdots \\ \mathbf{a}_j^{(j)} &= (\mathbf{I} - \mathbf{q}_1 \mathbf{q}_1^T) \mathbf{a}_{j-1}, \quad \text{with } r_{j-1,j} = \mathbf{q}_{j-1}^T \mathbf{a}_j &\Rightarrow \|\mathbf{a}_j^{(j-1)}\|_2^2 &= (r_{j-1,j})^2 + \|\mathbf{a}_j^{(j)}\|_2^2, \\ &&&\Rightarrow \|\mathbf{a}_j\|_2^2 &= \sum_{k=1}^{j-1} (r_{kj})^2 + \|\mathbf{a}_j^{(j)}\|_2^2. \end{aligned}$$

This is once more a remarkable property of MGS. Property (8.1) is independent of the orthogonality of the vectors (\mathbf{q}_k) .

We apply the same idea with rounding errors. From (4.2),

$$\begin{aligned}\bar{\mathbf{a}}_j^{(k+1)} &= \bar{\mathbf{a}}_j^{(k)} - \bar{\mathbf{q}}_k \bar{\kappa}_j + \delta_j^{(k)}, \\ \Rightarrow \bar{\mathbf{a}}_j^{(k)} + \delta_j^{(k)} &= \bar{\mathbf{a}}_j^{(k+1)} + \bar{\mathbf{q}}_k \bar{\kappa}_j, \\ \Rightarrow \|\bar{\mathbf{a}}_j^{(k)}\|_2^2 + \alpha_j^{(k)} &= \|\bar{\mathbf{a}}_j^{(k+1)}\|_2^2 + (\bar{\kappa}_j)^2,\end{aligned}\tag{8.2}$$

where

$$\alpha_j^{(k)} = (\delta_j^{(k)})^T \delta_j^{(k)} + 2(\delta_j^{(k)})^T \bar{\mathbf{a}}_j^{(k)} - 2\bar{\kappa}_j (\bar{\mathbf{q}}_k)^T \bar{\mathbf{a}}_j^{(k+1)}.$$

It is straightforward to get

$$|\alpha_j^{(k)}| \leq \|\delta_j^{(k)}\|_2^2 + 2\|\delta_j^{(k)}\|_2 \|\bar{\mathbf{a}}_j^{(k)}\|_2 + 2|\bar{\kappa}_j| \|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k+1)}\|.$$

From (4.3) follows

$$(\bar{\mathbf{q}}_k)^T \bar{\mathbf{a}}_j^{(k+1)} = (\bar{\mathbf{q}}_k)^T \eta_j^{(k)},\tag{8.3}$$

therefore :

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k+1)}| \leq \|\eta_j^{(k)}\|_2.$$

For $|\bar{\kappa}_j|$, (4.6) gives us

$$|\bar{\kappa}_j| \leq (1 + (2m + 1)2^{-t_1}) \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2 \leq 1.01 \cdot \|\bar{\mathbf{a}}_j^{(k)(r)}\|_2.\tag{8.4}$$

Using (4.5) and (8.4) we get

$$\begin{aligned}|\alpha_j^{(k)}| &\leq 1.006 \times [1.45^2/1.06 \times 2^{-t_1} + 2 \times 1.45/1.06 + 2 \times 1.01 \times (2m + 3)] \cdot 2^{-t_1} \|\bar{\mathbf{a}}_j\|_2^2, \\ &\leq (4.07m + 8.85)2^{-t_1} \cdot \|\mathbf{a}_j\|_2^2.\end{aligned}\tag{8.5}$$

Summing (8.2) for $k = 1, \dots, j-1$

$$\|\mathbf{a}_j\|_2^2 + \sum_{k=1}^{j-1} \alpha_j^{(k)} = \|\mathbf{a}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{\kappa}_j)^2,$$

and then using (8.5)

$$\left| (\|\mathbf{a}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{\kappa}_j)^2) - \|\mathbf{a}_j\|_2^2 \right| \leq (4.07m + 8.85)(j-1)2^{-t_1} \cdot \|\mathbf{a}_j\|_2^2.$$

Equations that we can also rewrite as

$$\sqrt{\|\mathbf{a}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{\kappa}_j)^2} \leq [1 + (2.04m + 4.43)(j-1)2^{-t_1}] \cdot \|\mathbf{a}_j\|_2.\tag{8.6}$$

Let assume that

$$(2.04m + 4.43)(j-1)2^{-t_1} \leq 0.01,\tag{8.7}$$

then we get

$$\boxed{\sqrt{\|\mathbf{a}_j^{(j)}\|_2^2 + \sum_{k=1}^{j-1} (\bar{\kappa}_j)^2} \leq 1.01 \cdot \|\mathbf{a}_j\|_2.}\tag{8.8}$$

Note that Equation (8.8) (and Assumption (8.7)) is independent of λ , the theorem of Pythagorus is “respected” without any assumption on the orthogonality of the column of $\bar{\mathbf{Q}}_{j-1}$.

9. Condition number of A and maximum value of $K_1^{(j)} = \frac{\|\mathbf{a}_j\|_2}{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}$

In exact arithmetic, $K_1^{(j)}$ is always less than the condition number of A , $\kappa_2(A)$. With rounding errors, this is nearly the same.

We define

$$K_1^{(j)} = \frac{\|\mathbf{a}_j\|_2}{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2} \quad \text{and} \quad K_2^{(j)} = \frac{\bar{\mathbf{a}}_j^{(1)(2)}}{\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2}. \quad (9.1)$$

Notice that $\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2 \neq 0$ and $\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \neq 0$ because of the nonsingularity assumption (6.1).

We recall (5.3) :

$$\mathbf{a}_k = \sum_{i=1}^k \bar{\mathbf{q}}_i \cdot \bar{r}_k - \delta_k, \quad k = 1, \dots, j-1.$$

For $k = j$, we just take into account the first loop ($r = 1$), this gives

$$\mathbf{a}_j = \sum_{i=1}^j \bar{\mathbf{q}}_i \cdot \bar{r}_j^{(1)} + \bar{\mathbf{a}}_j^{(j)(1)} - \delta_j^{(1)}$$

with $\delta_j^{(1)} = \sum_{k=1}^{j-1} \delta_j^{(k)(1)}$. In matrix form, this two relations together gives

$$\mathbf{A}_j = \bar{\mathbf{Q}}_{j-1} \bar{\mathbf{R}}_{(j-1,j)} - \mathbf{E}_j$$

with $\bar{\mathbf{Q}}_{j-1} \in \mathbb{R}^{m \times j-1}$

$$\bar{\mathbf{Q}}_{j-1} = [\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_{j-1}]$$

and $\bar{\mathbf{R}}_{(j-1,j)} \in \mathbb{R}^{j-1 \times j}$, the columns 1 to $j-1$ of $\bar{\mathbf{R}}_{(j-1,j)}$ are the $\bar{r}_i, i = 1, \dots, j-1$, the j^{th} is the column of the $\bar{r}_j^{(1)}, i = 1, \dots, j-1$.

Finally $\mathbf{E}_j \in \mathbb{R}^{m \times j}$ is defined by

$$\mathbf{E}_j = [\delta_1, \dots, \delta_{j-1}, \delta_j^{(1)} - \bar{\mathbf{a}}_j^{(j)(1)}],$$

therefore

$$0 < \|\mathbf{E}_j\|_E \leq 2.94(j-1) \cdot 2^{-t} \|\mathbf{A}_j\|_E + \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2.$$

Obviously the matrix $\bar{\mathbf{Q}}_{j-1} \bar{\mathbf{R}}_{(j-1,j)}$ is singular. The minimum singular value is the distance to the singularity therefore

$$\frac{1}{\kappa_2(\mathbf{A})} \leq \frac{1}{\kappa_2(\mathbf{A}_j)} \leq \min\left\{ \frac{\|\mathbf{E}\|_2}{\|\mathbf{A}_j\|_2}, \mathbf{E} \in \mathbb{R}^{m \times j} \text{ so as } \mathbf{A}_j + \mathbf{E} \text{ is singular} \right\} \leq \frac{\|\mathbf{E}_j\|_2}{\|\mathbf{A}_j\|_2} \leq \frac{\|\mathbf{E}_j\|_E}{\|\mathbf{A}_j\|_2},$$

so :

$$\kappa_2(\mathbf{A}) \geq \frac{\|\mathbf{A}_j\|_2}{\|\mathbf{E}_j\|_E} \geq \frac{1}{2.94(n-1) \cdot 2^{-t} \frac{\|\mathbf{A}_j\|_E}{\|\mathbf{A}_j\|_2} + \frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{\|\mathbf{A}_j\|_2}} = \frac{1}{2.94(j-1) \cdot 2^{-t} \frac{\|\mathbf{A}_j\|_E}{\|\mathbf{A}_j\|_2} + \frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{\|\mathbf{a}_j\|_2} \frac{\|\mathbf{a}_j\|_2}{\|\mathbf{A}_j\|_2}}$$

but

$$\frac{\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2}{\|\mathbf{a}_j\|_2} = K_1^{(j)}, \quad \frac{\|\mathbf{a}_j\|_2}{\|\mathbf{A}_j\|_2} \leq 1 \quad \text{and} \quad \frac{\|\mathbf{A}_j\|_E}{\|\mathbf{A}_j\|_2} < j^{\frac{1}{2}}$$

therefore

$$\kappa_2(\mathbf{A}) \geq \frac{1}{2.94(j-1)j^{\frac{1}{2}} \cdot 2^{-t} + \frac{1}{\kappa_1^{(j)}}}.$$

By example let assume that

$$2.94(n-1)n^{\frac{1}{2}} \cdot 2^{-t} \cdot \kappa_2(\mathbf{A}) < 0.09, \quad (9.2)$$

we have the inequality

$$\kappa_1^{(j)} \leq \frac{1}{1 - 2.94(j-1)j^{\frac{1}{2}} 2^{-t} \cdot \kappa_2(\mathbf{A})} \kappa_2(\mathbf{A}).$$

Using assumption (9.2) we get

$$\boxed{\kappa_1^{(j)} \leq 1.1 \cdot \kappa_2(\mathbf{A})}. \quad (9.3)$$

10. Bound for $K_2^{(j)}$ $= \frac{\|\bar{\mathbf{a}}_j^{(1)(2)}\|_2}{\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2}$

We are interested in finding an upper value for $K_2^{(j)}$. The idea is that if $K_2^{(j)}$ is small then the computed vector $\bar{\mathbf{q}}_j$ will be orthogonal to the previous. We sum (4.2) for $r = 2, k = 1, 2, \dots, j-1$ we get :

$$\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{r}_{kj}^{(2)} + \sum_{k=1}^{j-1} \delta_j^{(k)(2)}. \quad (10.1)$$

Therefore we see that the modification of $\bar{\mathbf{a}}_j^{(j)(1)}$ to give $\bar{\mathbf{a}}_j^{(j)(2)}$ are due to the $\bar{r}_{kj}^{(2)}$. If the $\bar{r}_{kj}^{(2)}$ are small enough, then $K_2^{(j)} \sim 1$. The quantity that interest us is therefore $\bar{r}_{kj}^{(2)}, k = 1, \dots, j-1$.

10.1. Bound on $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}|$

First of all, let's take a look to $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}|$. This represents the orthogonality of $\bar{\mathbf{q}}_k$ and the vector $\bar{\mathbf{a}}_j^{(j)(1)}$ given by the first step of MGS ($r = 1$). Following Björck, we sum (4.2) for $r = 1, i = k+1, k+2, \dots, j-1$ we get

$$\bar{\mathbf{a}}_j^{(j)(1)} = \bar{\mathbf{a}}_j^{(k+1)(1)} - \sum_{i=k+1}^{j-1} \bar{\mathbf{q}}_i \bar{r}_{ij}^{(1)} + \sum_{i=k+1}^{j-1} \delta_j^{(i)(1)}.$$

Hence multiplying this relation by $\bar{\mathbf{q}}_k^T$ and using (8.3) we get

$$\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)} = - \sum_{i=k+1}^{j-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i) \bar{r}_{ij}^{(1)} + \bar{\mathbf{q}}_k^T (\bar{\mathbf{n}}_j^{(k)(1)}) + \sum_{i=k+1}^{j-1} \delta_j^{(i)(1)}.$$

Therefore :

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| \leq \sqrt{\sum_{i=k+1}^{j-1} (\bar{r}_{ij}^{(1)})^2} \sqrt{\sum_{i=k+1}^{j-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i)^2 + \|\bar{\mathbf{n}}_j^{(k)(1)}\|_2} + \sum_{i=k+1}^{j-1} \|\delta_j^{(i)(1)}\|_2.$$

Using (4.5) and (4.7) we have :

$$\|\bar{\eta}_j^{(k)(1)}\|_2 + \sum_{i=k+1}^{j-1} \|\delta_j^{(i)(1)}\|_2 \leq (2.14m + 3.20 + 1.46(j-k-1))2^{-t} \cdot \|\mathbf{a}_j\|_2.$$

Finally, using (7.2) and (8.8), we get

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| \leq [1.01\lambda + (2.14m + 1.46(j-k-1) + 3.20)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

10.2. Bound on $|\bar{k}_j^{(2)}|$

Now that we control the orthogonality of the first step, we study the influence in the second step by computing : $|\bar{k}_j^{(2)}|$. In the same way as preceding, we sum (4.2) for $r = 2, i = 1, 2, \dots, k-1$ we get :

$$\bar{\mathbf{a}}_j^{(k)(2)} = \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{i=1}^{k-1} \bar{\mathbf{q}}_i \bar{k}_j^{(2)} + \sum_{i=1}^{k-1} \delta_j^{(i)(2)}.$$

Hence multiplying this relation by $\bar{\mathbf{q}}_k^T$, we get

$$\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)} = \bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{i=1}^{k-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i) \bar{k}_j^{(2)} + \bar{\mathbf{q}}_k^T \sum_{i=1}^{k-1} \delta_j^{(i)(2)}.$$

We deduce

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)}| \leq |\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(j)(1)}| + \sqrt{\sum_{i=1}^{k-1} (\bar{k}_j^{(2)})^2} \sqrt{\sum_{i=1}^{k-1} (\bar{\mathbf{q}}_k^T \bar{\mathbf{q}}_i)^2} + \sum_{i=1}^{k-1} \|\delta_j^{(i)(2)}\|_2.$$

We know how to bound all the right-side quantities, we get

$$|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)}| \leq [2.02\lambda + (2.14m + 1.46(j-2) + 3.20)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

Using (4.6) and (4.7) we know that $|\bar{\mathbf{q}}_k^T \bar{\mathbf{a}}_j^{(k)(2)} - \bar{k}_j^{(2)}| \leq (2.15m + 1.08) \cdot 2^{-t} \|\mathbf{a}_j\|_2$, therefore

$$|\bar{k}_j^{(2)}| \leq [2.02\lambda + (4.29m + 1.46(j-2) + 4.28)2^{-t}] \cdot \|\mathbf{a}_j\|_2,$$

and so

$$\boxed{|\bar{k}_j^{(2)}| \leq [2.02\lambda + 5.75(m+1)2^{-t}] \cdot \|\mathbf{a}_j\|_2.} \quad (10.2)$$

10.3. Bound on $K_2^{(j)}$

Let us sum (4.2) for $r = 2, k = 1, 2, \dots, j-1$ we get :

$$\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{a}}_j^{(j)(1)} - \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)} + \sum_{k=1}^{j-1} \delta_j^{(k)(2)}.$$

In norms we obtain

$$\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \geq \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2 - \left\| \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)} \right\|_2 - \sum_{k=1}^{j-1} \|\delta_j^{(k)(2)}\|_2. \quad (10.3)$$

Let us try to have an upper bound for $\|\sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)}\|_2$, we have by applying the induction assumption (7.1)

$$\left\| \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)} \right\|_2^2 \leq (1 + 2\lambda) \cdot \left\| \begin{pmatrix} \bar{k}_j^{(2)} \\ \vdots \\ \bar{k}_{j-1,j}^{(2)} \end{pmatrix} \right\|_2^2.$$

Using (10.2) we get

$$\left\| \sum_{k=1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)} \right\|_2 \leq \sqrt{1 + 2\lambda} \cdot \sqrt{j-1} [2.02\lambda + 5.75(m+1)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

With (10.3) this gives

$$\|\bar{\mathbf{a}}_j^{(j)(2)}\|_2 \geq \|\bar{\mathbf{a}}_j^{(j)(1)}\|_2 - \sqrt{1 + 2\lambda} \cdot \sqrt{j-1} [2.02\lambda + 5.75(m+1)2^{-t}] \cdot \|\mathbf{a}_j\|_2.$$

Dividing by $\|\bar{\mathbf{a}}_j^{(j)(1)}\|_2$ we have

$$1/K_2^{(j)} \geq 1 - K_1^{(j)} \sqrt{1 + 2\lambda} \cdot \sqrt{j-1} [2.02\lambda + 5.75(m+1)2^{-t}].$$

Let assume that

$$1.1 \kappa_2(\mathbf{A}) \sqrt{1 + 2\lambda} \cdot \sqrt{n-1} [2.02\lambda + 5.75(m+1)2^{-t}] \leq 0.67 < 1, \quad (10.4)$$

we obtain

$$K_2^{(j)} \leq \frac{1}{1 - K_1^{(j)} \sqrt{1 + 2\lambda} \cdot \sqrt{j-1} [2.02\lambda + 5.75(m+1)2^{-t}]} \leq \frac{1}{0.67}.$$

We conclude by

$$\boxed{K_2^{(j)} \leq 1.5.} \quad (10.5)$$

We remark that assumption (10.4) is dependent on λ which is not known yet.

11. Bound on the orthogonality of the vectors

Summing (4.2) from $k = i+1, i+2, \dots, j-1$ and $r = 2$ we get

$$\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{a}}_j^{(i+1)(2)} - \sum_{k=i+1}^{j-1} \bar{\mathbf{q}}_k \bar{k}_j^{(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)}. \quad (11.1)$$

From (8.3) we have $\bar{\mathbf{q}}_i^T \bar{\mathbf{a}}_j^{(i+1)(2)} = \bar{\mathbf{q}}_i^T \boldsymbol{\eta}_j^{(i)(2)}$ and we also have $\bar{\mathbf{a}}_j^{(j)(2)} = \bar{\mathbf{q}}_j \bar{k}_j^{(2)}$ therefore multiplying (11.1) by $\bar{\mathbf{q}}_i^T$ we get

$$\sum_{k=i+1}^j \bar{k}_j^{(2)} (\bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_k) = \bar{\mathbf{q}}_i^T (\boldsymbol{\eta}_j^{(i)(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)}).$$

We divide by $|\bar{k}_j^{(2)}|$ (which is different from 0)

$$s_{ij} = \sum_{k=i+1}^j \frac{\bar{k}_j^{(2)}}{|\bar{k}_j^{(2)}|} (\bar{\mathbf{q}}_i^T \bar{\mathbf{q}}_k) = \frac{\bar{\mathbf{q}}_i^T (\boldsymbol{\eta}_j^{(i)(2)} + \sum_{k=i+1}^{j-1} \delta_j^{(k)(2)})}{|\bar{k}_j^{(2)}|}$$

where s_{ij} is the component (i, j) of the matrix $\mathbf{S}_p = \mathbf{U}_p \mathbf{M}_p$.
with \mathbf{M}_p , an unit upper triangular matrix such as the (k, j) entry m_{kj} is

$$\begin{cases} m_{kj} = \frac{\bar{r}_{kj}^{(2)}}{|\bar{r}_{jj}^{(2)}|}, & \text{if } k < j, \\ m_{jj} = 1, \\ m_{kj} = 0. \end{cases}$$

Björck [1] gives an upper bound for L_2 -norm of each column of \mathbf{S}_p

$$\|\mathbf{s}_j\|_2 \leq 0.87 K_2^{(j)} \cdot n^{\frac{1}{2}} (n + 1 + 2.5m) 2^{-t}.$$

It is straightforward to get using (10.5) and $0.87 \times 1.5 = 1.305$

$$\|\mathbf{S}_p\|_2 \leq 1.305 n^{\frac{1}{2}} (n + 1 + 2.5m) 2^{-t}. \quad (11.2)$$

\mathbf{M}_p is non singular therefore, we guess

$$\|\mathbf{U}_p\|_2 \leq \|\mathbf{M}_p^{-1}\|_2 \|\mathbf{S}_p\|_2. \quad (11.3)$$

The quantity that interests us is therefore $\|\mathbf{M}_p^{-1}\|_2$. From (10.2) we have

$$\sum_{k=1}^{j-1} |m_{kj}| \leq (j-1) [2.02\lambda + 5.75(m+1)2^{-t}] \frac{\|\mathbf{a}_j\|_2}{|\bar{r}_{jj}^{(2)}|},$$

therefore

$$\sum_{k=1}^{j-1} |m_{kj}| \leq (j-1) [2.02\lambda + 5.75(m+1)2^{-t}] K_j^{(1)} K_j^{(2)}.$$

Using Equations (9.3) and (10.5), we get as $1.1 \times 1.5 = 1.65$

$$\sum_{k=1}^{j-1} |m_{kj}| \leq 1.65 (j-1) [2.02\lambda + 5.75(m+1)2^{-t}] \kappa_2(A).$$

We assume that

$$1.65(n-1) [2.02\lambda + 5.75(m+1)2^{-t}] \kappa_2(A) \leq 0.5, \quad (11.4)$$

therefore

$$\sum_{k=1}^{j-1} |m_{kj}| \leq 0.5. \quad (11.5)$$

Let us define L so as

$$L = \min_k (|m_{jj}| - \sum_{k \neq j} |m_{kj}|) = \min_k (1 - \sum_{k=1}^{j-1} |m_{kj}|),$$

then with (11.5) we have $L \geq 0.5$. The matrix \mathbf{M}_p is therefore strictly diagonally dominant by columns. Using a result of Varah [2], we know that

$$\|\mathbf{M}_p^{-1}\|_1 \leq 1/L.$$

Equivalently, we can also use the decomposition $M_p = I_p + B_p$ with B_p nilpotent, and from (11.5), $\|B_p\|_1 \leq 0.5$ we have the well-known formula

$$\|M_p^{-1}\|_1 \leq (1 - \|B_p\|_1)^{-1} \leq 2. \quad (11.6)$$

With (11.2), (11.3) and (11.6), we get as $2 \times 1.305 = 2.61$

$$\|\mathbf{U}_p\|_2 \leq 2.61 \cdot n(n+1+2.5m)2^{-t}. \quad (11.7)$$

In a very natural way, we define a posteriori λ as

$$\lambda = 2.61 \cdot n(n+1+2.5m)2^{-t} = \lambda. \quad (11.8)$$

12. Assumptions on \mathbf{A}

To be clear on the assumption made on \mathbf{A} , we recall that we have assumed (4.1), (6.1), (8.7), (9.2), (10.4) and (11.4). We focus here on the main assumption that is (11.4). We replace λ by its value and get

$$2 \times 1.65(n-1)[2.02 \times 2.61(n^2+n+2.5mn) + 5.75(m+1)]2^{-t} \cdot \kappa_2(\mathbf{A}) \leq 1.$$

Equation that we replace for the sake of the simplicity by

$$50(m+2)n^2\kappa_2(\mathbf{A}) \cdot 2^{-t} < 1. \quad (12.1)$$

13. Conclusion of the proof by induction

We have shown that if we assume (12.1) and define λ with (11.8) then

$$\text{if at step } p-1, \text{ we have } \|\mathbf{U}_{p-1}\|_2 \leq \lambda \text{ then } \|\mathbf{U}_p\|_2 \leq \lambda.$$

At step $n=1$, \mathbf{U}_1 is not defined but we can define it for the proof $\|\mathbf{U}_1\|_2 = 0$ and so $\|\mathbf{U}_1\|_2 \leq \lambda$. From this, we conclude that at step n , we have

$$\|\mathbf{I} - \mathbf{Q}^T \mathbf{Q}\|_2 \leq 5.22 \cdot n(n+1+2.5m)2^{-t}.$$

14. Conclusion

MGS2 applied on a numerically nonsingular matrix \mathbf{A} gives a well-orthogonal set of vectors \mathbf{Q} in floating point arithmetic. This fact was already stated since [3].

The key point of the proof given in this paper is if

$$\sum_{k=1}^{j-1} \frac{|k_j^{-(2)}|}{\|\mathbf{a}_j^{(j)(2)}\|_2} \leq L, \quad (14.1)$$

with $L < 1$ then the matrix \mathbf{M} is well-conditioned and so $\|\mathbf{U}\|_2$ is small. We have shown that if \mathbf{A} is numerically nonsingular matrix then Equation (14.1) is verified at the second orthogonalization step. However if (14.1) is verified at the first step, that is to say

$$\sum_{k=1}^{j-1} \frac{|k_j^{-(1)}|}{\|\mathbf{a}_j^{(j)(1)}\|_2} \leq L, \quad (14.2)$$

clearly this means that no reorthogonalization is needed. Therefore we propose to check at each step whether (14.2) is verified or not. The algorithm obtained is the modified Gram-Schmidt algorithm with selective reorthogonalization using L -criterion, $\text{MGS2}(L)$. In this paper, L is set to 0.5, it is clear that any value lower than 1 is correct. Therefore this paper also state that $\text{MGS2}(L)$ with $L < 1$ gives good results.

One may remark that the matrix \mathbf{M} is not only diagonal dominant by column but also by row. This observation permits us to use a result of Varah [2] and we can directly bound $\|\mathbf{M}^{-1}\|_2$ by 2 instead of $2\sqrt{n}$. Consequently the orthogonality obtained is in $O(n^{3/2}2^{-t})$, the assumption on \mathbf{A} also get better. The proof is not done in this way because then it can not be adapted to $\text{MGS2}(L)$.

References

1. BJÖRCK Å. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT* 7 (1967), 1-21.
2. VARAH J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications* 11 (1975), 3-5.
3. GIRAUD L. & LANGOU J. & ROZLOŽNÍK M. On the round-off error analysis of the Gram-Schmidt algorithm with reorthogonalization. *CERFACS Technical Report, TR/PA/02/33, (2002)*