

COMMENTS

The slides named "COMMENTS" were not present during the talk, it is just to help the reader to understand.

A reorthogonalization procedure for MGS applied to a low rank deficient matrix.

Julien Langou

CERFACS

joint work with [Luc Giraud](#),
and [Serge Gratton](#) .

Gram-Schmidt algorithm

Starting from $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n) \in \mathbb{R}^{m \times n}$ with full rank,
Gram-Schmidt algorithm computes \mathbf{Q} and \mathbf{R} so as

- $\mathbf{A} = \mathbf{QR}$,
- $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$,
- $\mathbf{R} \in \mathbb{R}^{n \times n}$, \mathbf{R} is upper triangular with positive elements on the diagonal .

Gram-Schmidt algorithm

Starting from $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n) \in \mathbb{R}^{m \times n}$ with full rank,
Gram-Schmidt algorithm computes \mathbf{Q} and \mathbf{R} so as

- $\mathbf{A} = \mathbf{QR}$,
- $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$,
- $\mathbf{R} \in \mathbb{R}^{n \times n}$, \mathbf{R} is upper triangular with positive elements on the diagonal .

Classical Gram-Schmidt (CGS)

for $j = 1, n$ do

$$\mathbf{w} = \mathbf{a}_j$$

$$\mathbf{w} = (\mathbf{I} - \mathbf{Q}_{j-1} \mathbf{Q}_{j-1}^T) \mathbf{a}_j$$

$$\mathbf{q}_j = \mathbf{w} / \|\mathbf{w}\|_2$$

end for

Gram-Schmidt algorithm

Starting from $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n) \in \mathbb{R}^{m \times n}$ with full rank,
Gram-Schmidt algorithm computes \mathbf{Q} and \mathbf{R} so as

- $\mathbf{A} = \mathbf{QR}$,
- $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$,
- $\mathbf{R} \in \mathbb{R}^{n \times n}$, \mathbf{R} is upper triangular with positive elements on the diagonal .

Classical Gram-Schmidt (CGS)

for $j = 1, n$ do

$$\mathbf{w} = \mathbf{a}_j$$

$$\mathbf{w} = (\mathbf{I} - \mathbf{Q}_{j-1} \mathbf{Q}_{j-1}^T) \mathbf{a}_j$$

$$\mathbf{q}_j = \mathbf{w} / \|\mathbf{w}\|_2$$

end for

Modified Gram-Schmidt (MGS)

for $j = 1, n$ do

$$\mathbf{w} = \mathbf{a}_j$$

$$\mathbf{w} = (\mathbf{I} - \mathbf{q}_{j-1} \mathbf{q}_{j-1}^T) \dots (\mathbf{I} - \mathbf{q}_1 \mathbf{q}_1^T) \mathbf{a}_j$$

$$\mathbf{q}_j = \mathbf{w} / \|\mathbf{w}\|_2$$

end for

Gram-Schmidt algorithm

Starting from $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n) \in \mathbb{R}^{m \times n}$ with full rank,

- $\mathbf{A} = \mathbf{QR}$,
- $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$,
- $\mathbf{R} \in \mathbb{R}^{n \times n}$, \mathbf{R} is upper triangular with positive elements on the diagonal .

Classical Gram-Schmidt (CGS)

```

for  $j = 1, n$  do
   $\mathbf{w} = \mathbf{a}_j$ 
  for  $i = 1, j - 1$  do
     $r_{ij} = \mathbf{q}_i^T \mathbf{a}_j$ 
     $\mathbf{w} = \mathbf{w} - \mathbf{q}_i r_{ij}$ 
  end for
   $r_{jj} = \|\mathbf{w}\|_2$ 
   $\mathbf{q}_j = \mathbf{w} / r_{jj}$ 
end for

```

Modified Gram-Schmidt (MGS)

```

for  $j = 1, n$  do
   $\mathbf{w} = \mathbf{a}_j$ 
  for  $i = 1, j - 1$  do
     $r_{ij} = \mathbf{q}_i^T \mathbf{w}$ 
     $\mathbf{w} = \mathbf{w} - \mathbf{q}_i r_{ij}$ 
  end for
   $r_{jj} = \|\mathbf{w}\|_2$ 
   $\mathbf{q}_j = \mathbf{w} / r_{jj}$ 
end for

```

Outline

- Some previous results on modified Gram-Schmidt algorithm.
- Rank considerations in the modified Gram-Schmidt algorithm.
- A reorthogonalization procedure for modified Gram-schmidt applied to low rank matrices.
- Numerical experiments.

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,

MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR},$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR}, \quad \mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR},$$

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

[bjor:67]

$$\mathbf{I} - \mathbf{Q}^T \mathbf{Q} = \mathbf{0},$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR},$$

$$\mathbf{I} - \mathbf{Q}^T \mathbf{Q} = \mathbf{0},$$

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

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

[bjor:67]

[bjor:67]

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR}, \quad \mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

$$\mathbf{I} - \mathbf{Q}^T \mathbf{Q} = \mathbf{0}, \quad \|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \quad [\text{bjor:67}]$$

$$\mathbf{A} = \hat{\mathbf{Q}}\mathbf{R},$$

$$\mathbf{I} - \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{0},$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} = \mathbf{QR}, \quad \mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

$$\mathbf{I} - \mathbf{Q}^T \mathbf{Q} = \mathbf{0}, \quad \|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \quad [\text{bjor:67}]$$

$$\mathbf{A} = \hat{\mathbf{Q}}\mathbf{R}, \quad \mathbf{A} + \hat{\mathbf{E}} = \hat{\mathbf{Q}}\bar{\mathbf{R}}, \quad \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{I} \quad \text{and} \quad \|\hat{\mathbf{E}}\|_2 \leq cu \|\mathbf{A}\|_2, \quad [\text{bjpa:92}]$$

$$\mathbf{I} - \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{0},$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

$$\|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \quad [\text{bjor:67}]$$

$$\mathbf{A} + \hat{\mathbf{E}} = \hat{\mathbf{Q}}\bar{\mathbf{R}}, \quad \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{I} \quad \text{and} \quad \|\hat{\mathbf{E}}\|_2 \leq c u \|\mathbf{A}\|_2, \quad [\text{bjpa:92}]$$

where \bar{c}_i and c are constants depending on m , n and the details of the arithmetic,
and u is the unit round off.

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

$$\|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \quad [\text{bjor:67}]$$

$$\mathbf{A} + \hat{\mathbf{E}} = \hat{\mathbf{Q}}\bar{\mathbf{R}}, \quad \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{I} \quad \text{and} \quad \|\hat{\mathbf{E}}\|_2 \leq c u \|\mathbf{A}\|_2, \quad [\text{bjpa:92}]$$

where \bar{c}_i and c are constants depending on m , n and the details of the arithmetic,
and u is the unit round off.

$$c = 18.53n^{\frac{3}{2}}, \quad \bar{c}_1 = 1.5n^{\frac{3}{2}} \quad \text{and} \quad \bar{c}_2 = 31.6863n^{\frac{3}{2}}.$$

Historical Framework

$\mathbf{A} \in \mathbb{R}^{m \times n}$ with full rank $n \leq m$, with singular values : $\sigma_1 \geq \dots \geq \sigma_n > 0$, $\kappa(\mathbf{A}) = \sigma_1/\sigma_n$,
MGS computes in floating point arithmetic $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ so as

$$\mathbf{A} + \bar{\mathbf{E}} = \bar{\mathbf{Q}}\bar{\mathbf{R}}, \quad \|\bar{\mathbf{E}}\|_2 \leq \bar{c}_1 u \|\mathbf{A}\|_2, \quad [\text{bjor:67}]$$

$$\|\mathbf{I} - \bar{\mathbf{Q}}^T \bar{\mathbf{Q}}\|_2 \leq \bar{c}_2 \kappa(\mathbf{A}) u, \quad [\text{bjor:67}]$$

$$\mathbf{A} + \hat{\mathbf{E}} = \hat{\mathbf{Q}}\bar{\mathbf{R}}, \quad \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} = \mathbf{I} \quad \text{and} \quad \|\hat{\mathbf{E}}\|_2 \leq cu \|\mathbf{A}\|_2, \quad [\text{bjpa:92}]$$

where \bar{c}_i and c are constants depending on m , n and the details of the arithmetic,
and u is the unit round off.

$$c = 18.53n^{\frac{3}{2}}, \quad \bar{c}_1 = 1.5n^{\frac{3}{2}} \quad \text{and} \quad \bar{c}_2 = 31.6863n^{\frac{3}{2}}.$$

This results holds under the assumptions :

$$2.12 \cdot (m + 1) \cdot u < 0.01 \quad \text{and} \quad \bar{c}_2 u \kappa(\mathbf{A}) < 1.$$

Singular values of T .

$$(Id - \tilde{Q}^T \tilde{Q}) = - \begin{pmatrix} 0 & \tilde{q}_1^T \tilde{q}_2 & \dots & \tilde{q}_1^T \tilde{q}_n \\ \tilde{q}_2^T \tilde{q}_1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{q}_{n-1}^T \tilde{q}_n \\ \tilde{q}_n^T \tilde{q}_1 & \dots & \tilde{q}_n^T \tilde{q}_{n-1} & 0 \end{pmatrix}.$$

Singular values of T .

$$\tilde{T} = \mathbf{triu}(Id - \tilde{Q}^T \tilde{Q}) = - \begin{pmatrix} 0 & \tilde{q}_1^T \tilde{q}_2 & \dots & \tilde{q}_1^T \tilde{q}_n \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & \tilde{q}_{n-1}^T \tilde{q}_n \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Singular values of T .

$$\tilde{T} = \mathbf{triu}(Id - \tilde{Q}^T \tilde{Q}) = - \begin{pmatrix} 0 & \tilde{q}_1^T \tilde{q}_2 & \dots & \tilde{q}_1^T \tilde{q}_n \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & \tilde{q}_{n-1}^T \tilde{q}_n \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Under the assumption that $(c + c_1)u\kappa < 1$

$$[\text{bjpa:92}] \quad \|\tilde{T}\|_2 \leq c\xi u\kappa, \quad \kappa = \frac{\sigma_1}{\sigma_n}.$$

Singular values of T .

$$\tilde{T} = \mathbf{triu}(Id - \tilde{Q}^T \tilde{Q}) = - \begin{pmatrix} 0 & \tilde{q}_1^T \tilde{q}_2 & \dots & \tilde{q}_1^T \tilde{q}_n \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & \tilde{q}_{n-1}^T \tilde{q}_n \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Under the assumption that $(c + c_1)u\kappa < 1$

$$[\text{bjpa:92}] \quad \|\tilde{T}\|_2 \leq c\xi u\kappa, \quad \kappa = \frac{\sigma_1}{\sigma_n}.$$

$$i = 1, \dots, n, \quad \sigma_i(\tilde{T}) \leq c\xi u\kappa_i, \quad \kappa_i = \frac{\sigma_1}{\sigma_{n-i+1}}.$$

Proof.

Follow Björck and Paige (1992) and instead of doing

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2,$$

do

$$\sigma_i(AB) \leq \sigma_i(A) \|B\|_2, \quad i = 1, \dots, n.$$

Singular values of F .

$$[\text{bjpa:92}] \quad A + \hat{E} = \hat{Q}\bar{R} \quad , \quad \hat{Q}^T \hat{Q} = Id \quad \text{and} \quad \|\hat{E}\|_2 \leq cu\|A\|_2,$$

Singular values of F .

$$[\text{bjpa:92}] \quad A + \hat{E} = \hat{Q}\bar{R} \quad , \quad \hat{Q}^T \hat{Q} = Id \quad \text{and} \quad \|\hat{E}\|_2 \leq cu\|A\|_2,$$

$$F = \tilde{Q} - \hat{Q}.$$

Singular values of F .

$$[\text{bjpa:92}] \quad A + \hat{E} = \hat{Q}\bar{R} \quad , \quad \hat{Q}^T \hat{Q} = Id \quad \text{and} \quad \|\hat{E}\|_2 \leq cu\|A\|_2,$$

$$F = \tilde{Q} - \hat{Q}.$$

under the assumption $cu\kappa < 1$, with $\eta = (1 - cu\kappa)^{-1}$,

$$\sigma_i(F) \leq 2c\eta u\kappa_i, \quad i = 1, \dots, n.$$

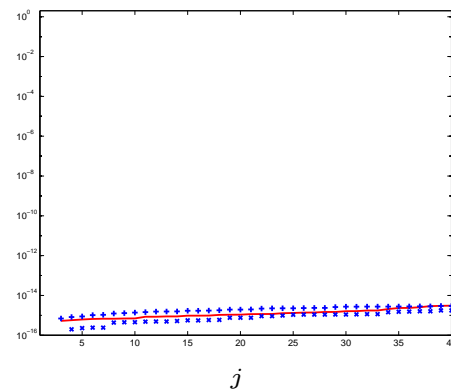
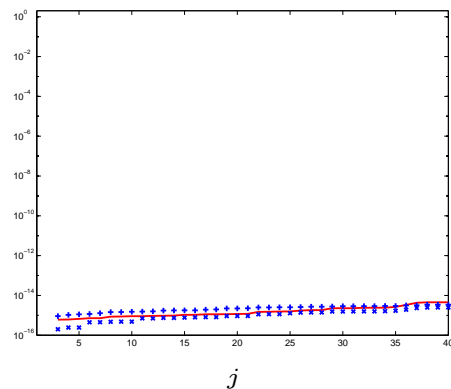
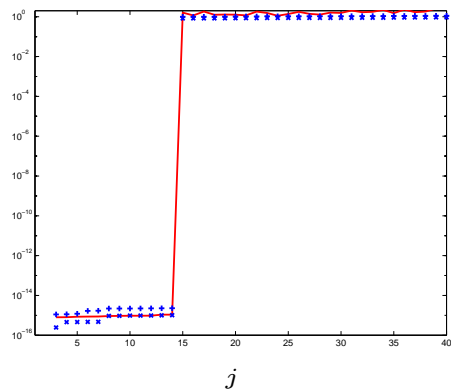
Tightness of the bounds(1)

— $u\kappa(A(1:j))$
 × $\sigma_1(T(1:j))$
 + $\sigma_1(F(1:j))$

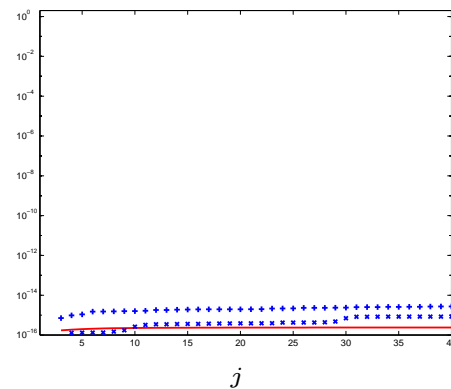
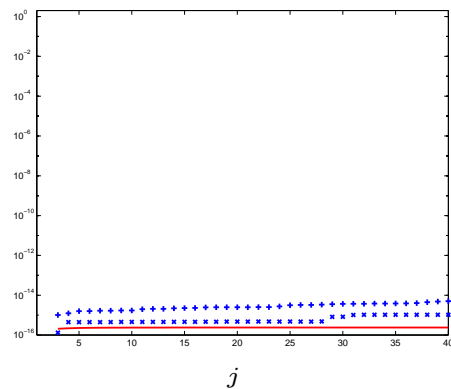
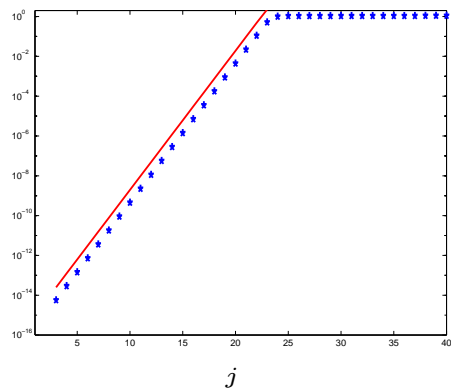
— $u\kappa_2(A(1:j))$
 × $\sigma_2(T(1:j))$
 + $\sigma_2(F(1:j))$

— $u\kappa_3(A(1:j))$
 × $\sigma_3(T(1:j))$
 + $\sigma_3(F(1:j))$

A_1



A_2



$A \in \mathbb{R}^{50 \times 40}$

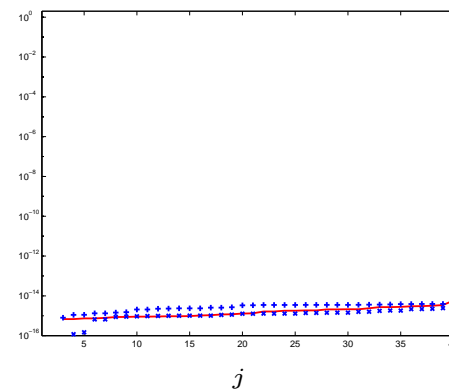
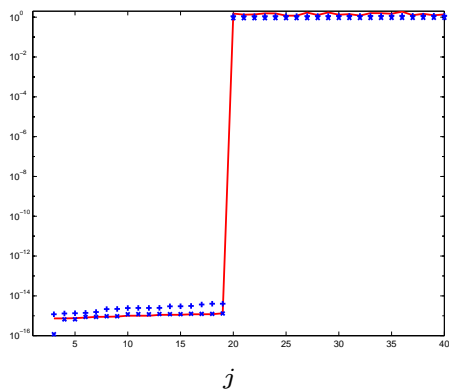
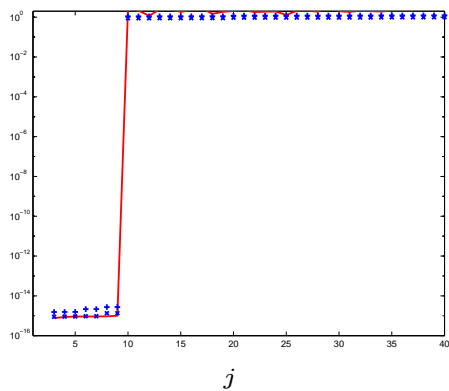
Tightness of the bounds(2)

- $u\kappa(A(1:j))$
- × $\sigma_1(T(1:j))$
- + $\sigma_1(F(1:j))$

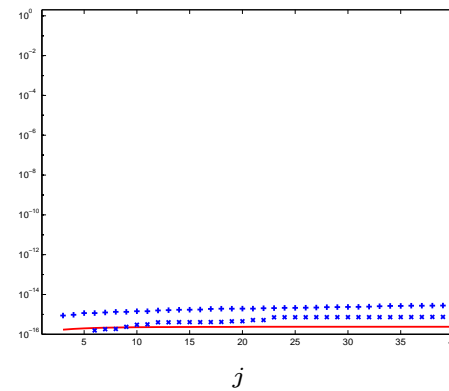
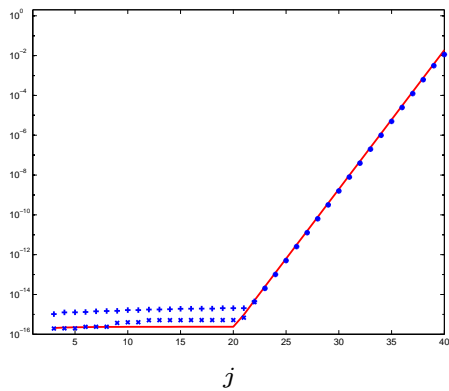
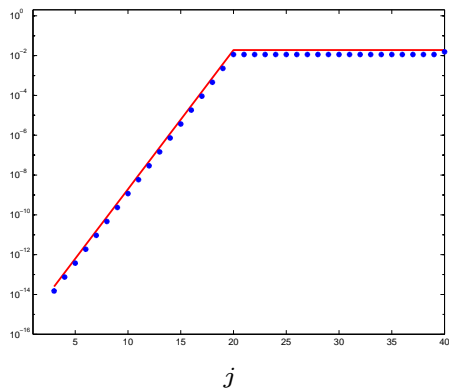
- $u\kappa_2(A(1:j))$
- × $\sigma_2(T(1:j))$
- + $\sigma_2(F(1:j))$

- $u\kappa_3(A(1:j))$
- × $\sigma_3(T(1:j))$
- + $\sigma_3(F(1:j))$

A_3



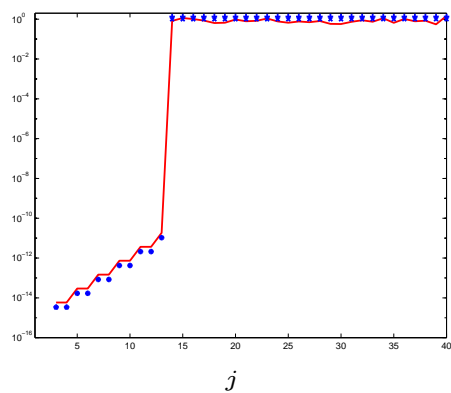
A_4



$$A \in \mathbb{R}^{50 \times 40}$$

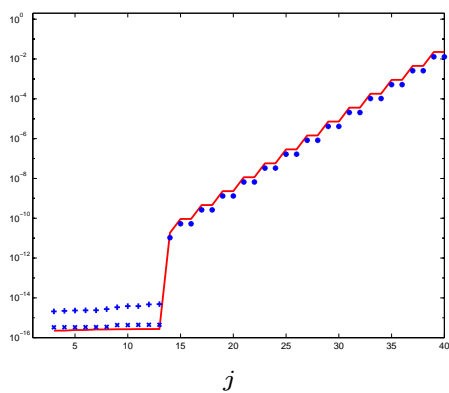
Tightness of the bounds(3)

— $u\kappa(A(1:j))$
 × $\sigma_1(T(1:j))$
 + $\sigma_1(F(1:j))$

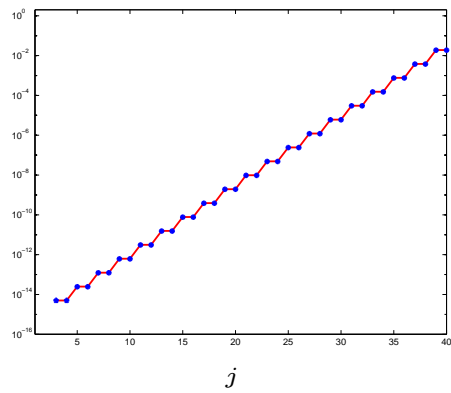
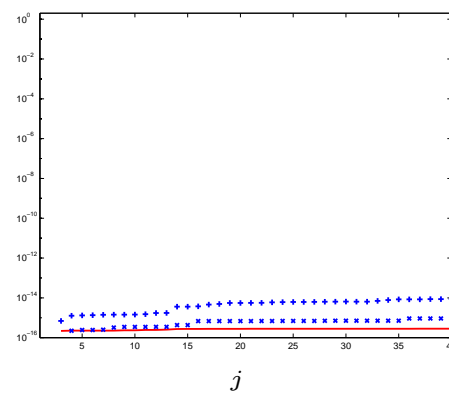


A_5

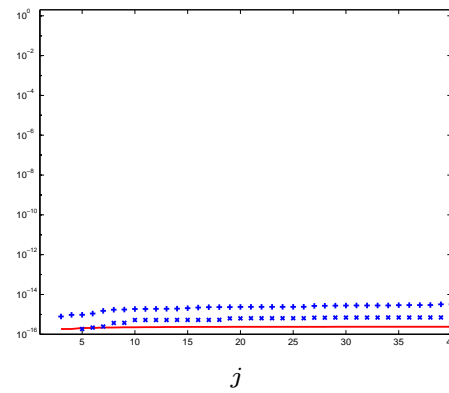
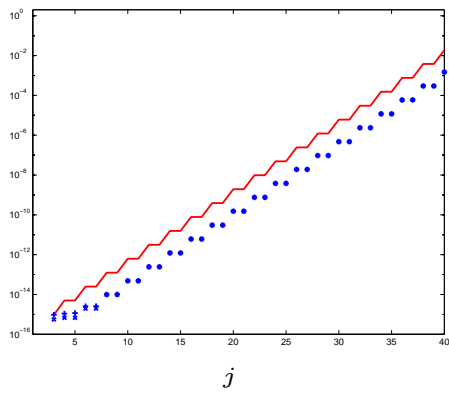
— $u\kappa_2(A(1:j))$
 × $\sigma_2(T(1:j))$
 + $\sigma_2(F(1:j))$



— $u\kappa_3(A(1:j))$
 × $\sigma_3(T(1:j))$
 + $\sigma_3(F(1:j))$



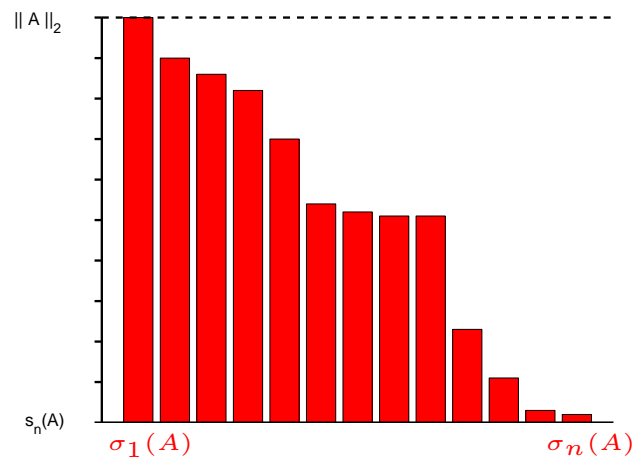
A_6



$A \in \mathbb{R}^{50 \times 40}$

Interpretation.

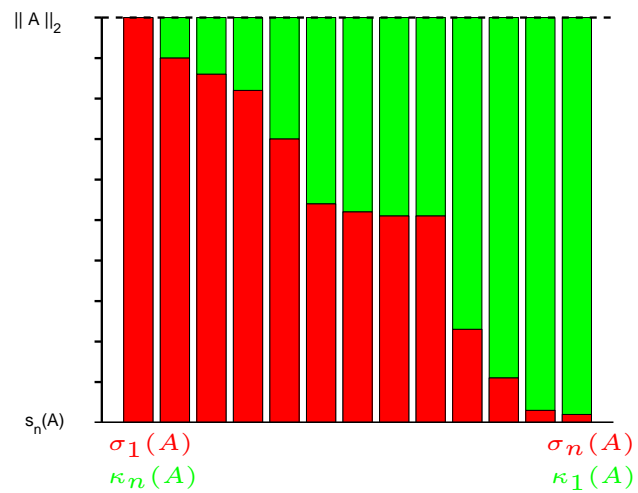
$$\sigma_i(F) \leq 2c\omega\eta\kappa_i(A) = 2c\omega\eta\frac{\|A\|_2}{\sigma_{n-i+1}(A)}, \quad i = 1, \dots, n.$$



Singular values of A .

Interpretation.

$$\sigma_i(F) \leq 2cu\eta\kappa_i(A), \quad i = 1, \dots, n.$$

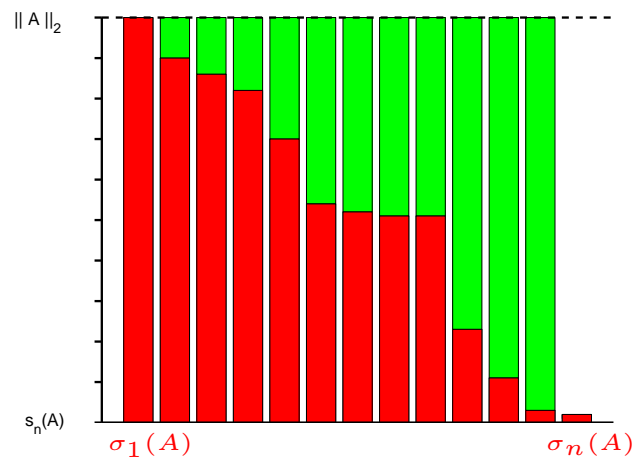


Singular values of A .

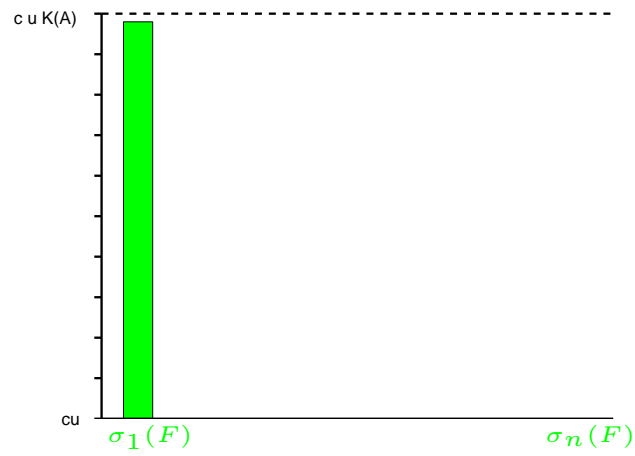
Condition numbers $\kappa_i(A)$ of A .

Interpretation.

$$\sigma_i(F) \leq 2cu\eta\kappa_i(A), \quad i = 1, \dots, n.$$



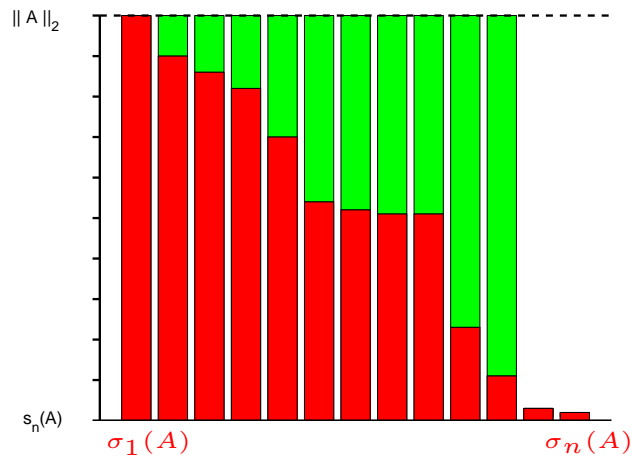
Singular values of A .



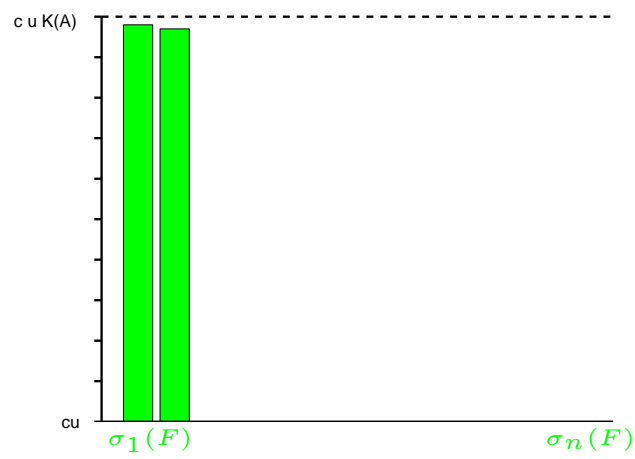
Singular values of F .

Interpretation.

$$\sigma_i(F) \leq 2cu\eta\kappa_i(A), \quad i = 1, \dots, n.$$



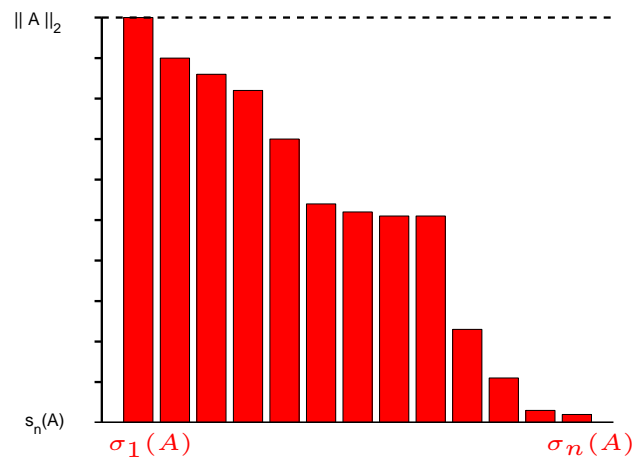
Singular values of A .



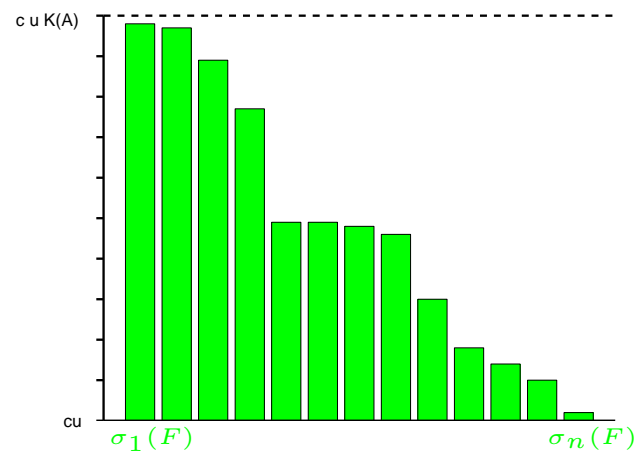
Singular values of F .

Interpretation.

$$\sigma_i(F) \leq 2cu\eta\kappa_i(A), \quad i = 1, \dots, n.$$

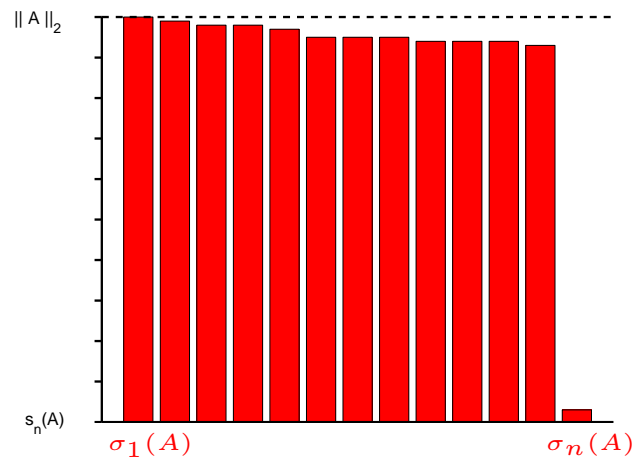


Singular values of A .

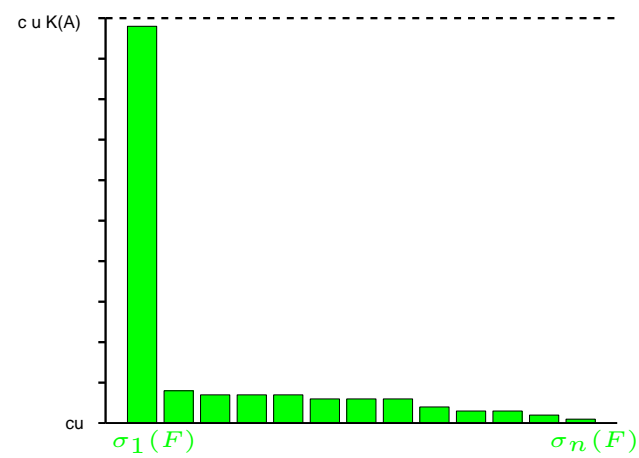


Singular values of F .

Interpretation in term of rank.



Singular values of A .



Singular values of F .

If A has (numerical) rank $n - 1$, F has (numerical) rank 1.

Interpretation in term of rank.

$$F = \tilde{Q} - \hat{Q}$$

we say that

the difference from \tilde{Q} , the Q -factor computed by modified Gram-Schmidt,
and \hat{Q} , an exact orthogonal matrix such that $A = \hat{Q}R + E$,
is F , a low rank matrix.

Background - Derivation of \hat{Q} .

-

$$\begin{array}{ccc}
 \text{MGS} & & \text{Householder} \\
 A = QR & \longleftrightarrow & \begin{pmatrix} 0 \\ A \end{pmatrix} = \begin{pmatrix} P_{11} \\ P_{21} \end{pmatrix} R
 \end{array}$$

with

$$= \begin{pmatrix} P_{11} \\ P_{21} \\ 0 & q_1^T q_2 & q_1^T (Id - q_2 q_2^T) q_3 & \dots & q_1^T (Id - q_2 q_2^T) \dots (Id - q_{n-1} q_{n-1}^T) q_n \\ 0 & 0 & q_2^T q_3 & \dots & q_2^T (Id - q_3 q_3^T) \dots (Id - q_{n-1} q_{n-1}^T) q_n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & q_{n-1}^T q_n \\ 0 & 0 & 0 & \dots & 0 \\ q_1 & (Id - q_1 q_1^T) q_2 & (Id - q_1 q_1^T) (Id - q_2 q_2^T) q_3 & \dots & (Id - q_1 q_1^T) \dots (Id - q_{n-1} q_{n-1}^T) q_n \end{pmatrix} .$$

- some useful formulae

$$P_{21} = Q(Id - P_{11}) \quad \tilde{T} = P_{11}(P_{11} - Id)^{-1}$$

Background - Derivation of \hat{Q} .

As $\begin{pmatrix} P_{11} \\ P_{21} \end{pmatrix}$ has orthonormal columns, we can do its CS-decomposition.

$$\begin{aligned} P_{11} &= U_1 C W^T & \text{with} & & U_1^T U_1 &= I \\ P_{21} &= U_2 S W^T & & & U_2^T U_2 &= I \\ & & & & C^2 + S^2 &= I, C \text{ and } S \text{ diagonal.} \end{aligned}$$

Björck and Paige (1992) define \hat{Q} as

$$\hat{Q} = U_2 W^T.$$

How to find F ?

$$\begin{aligned}
 F &= Q - \hat{Q}, \\
 &= P_{21}(I - T) - \hat{Q}, & Q &= P_{21}(I - T) \\
 &= U_2(S - I)W^T - P_{21}T, & P_{21} &= U_2SW^T \\
 &\sim U_{2k}(S_k - I)W_k^T - P_{21}(U_{Tk}\Sigma_{Tk}W_{Tk}^T).
 \end{aligned}$$

We have expressed F as a sum of to low rank matrix of rank k . The choice of k is a compromise between **accuracy** and **efficiency**.

How to find F ?

$$\begin{aligned}
 F &= Q - \hat{Q}, \\
 &= P_{21}(I - T) - \hat{Q}, & Q &= P_{21}(I - T) \\
 &= U_2(S - I)W^T - P_{21}T, & P_{21} &= U_2SW^T \\
 &\sim U_{2k}(S_k - I)W_k^T - P_{21}(U_{Tk}\Sigma_{Tk}W_{Tk}^T).
 \end{aligned}$$

We have expressed F as a sum of to low rank matrix of rank k . The choice of k is a compromise between **accuracy** and **efficiency**.

We avoid the computation of P_{11} and P_{21} using $P_{11} = (T - I)^{-1}T$ and $P_{21} = Q(I - T)^{-1}$

$$F = (Q(I - T)^{-1}) W_k S_k^{-1} (S_k - I_k) W_k^T - Q(I - T)^{-1} (U_{Tk} \Sigma_{Tk} W_{Tk}^T),$$

$$F = Q \left((I - U_{Tk} \Sigma_{Tk} W_{Tk}^T)^{-1} (W_k (I_k - S_k^{-1}) W_k^T - U_{Tk} \Sigma_{Tk} W_{Tk}^T) \right).$$

How to find F ? - the rank 1 case

- Run MGS to have Q and R .
- Form T and find its singular value decomposition such that $T = u_T \sigma_T w_T^T$.
- Compute $c = \sigma + \sigma^2(w_T^T u_T)$ and $s = \sqrt{1 - c^2}$.
- Form F :

$$F = \left(\left(Q \left(w_T (1 - s^{-1}) + u_T (1 - s^{-1} - c) \right) \right) w_T^T \right) .$$

Things to retain.

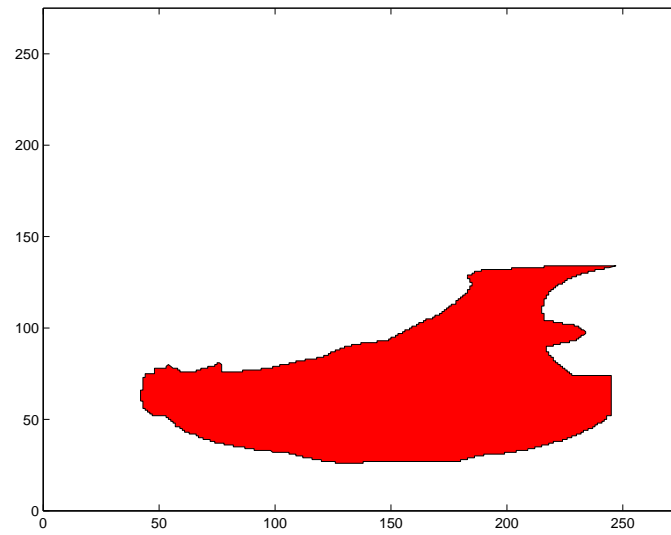
- We need to fix the orthogonality $\|I - Q^T Q\|_2$ that we want to reach. In a way doing this we fix the *numerical rank* of A .
- The algorithm costs $\mathcal{O}(mn^2)$. (MGS costs $\mathcal{O}(2mn^2)$).
The main cost is governed by the construction of $T = \mathbf{triu}(I - Q^T Q)$.
- For both of these reasons, it is particularly well-suited when A has a **neat low** rank deficiency.

Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)



GOLDORAK ($n = 715$, $\kappa(A) = 546$).

Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)



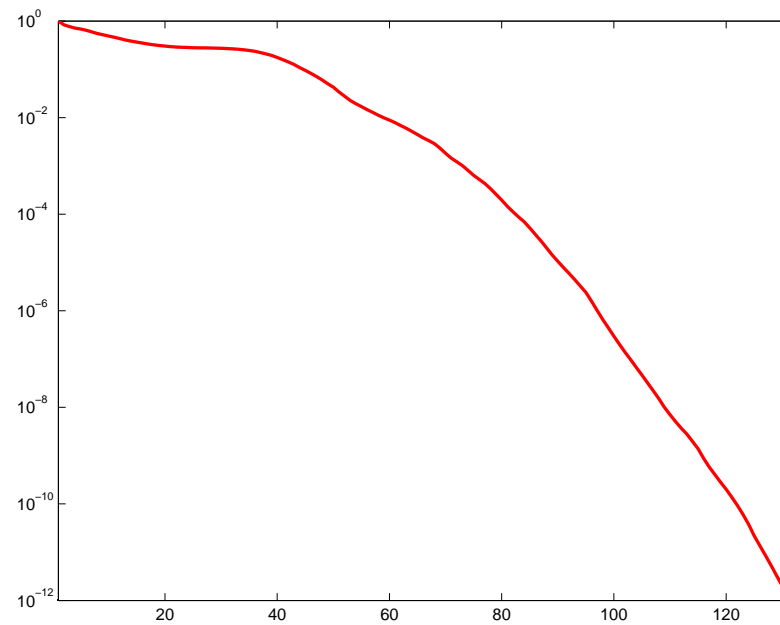
GOLDORAK ($n = 715$, $\kappa(A) = 546$).

Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)

We run full GMRES with MGS without preconditionner and a right-hand side corresponding to an incident wave at 0° .

The tolerance is fixed to 10^{-12} .

$$- \|r_m\|_2 / \|b\|_2$$



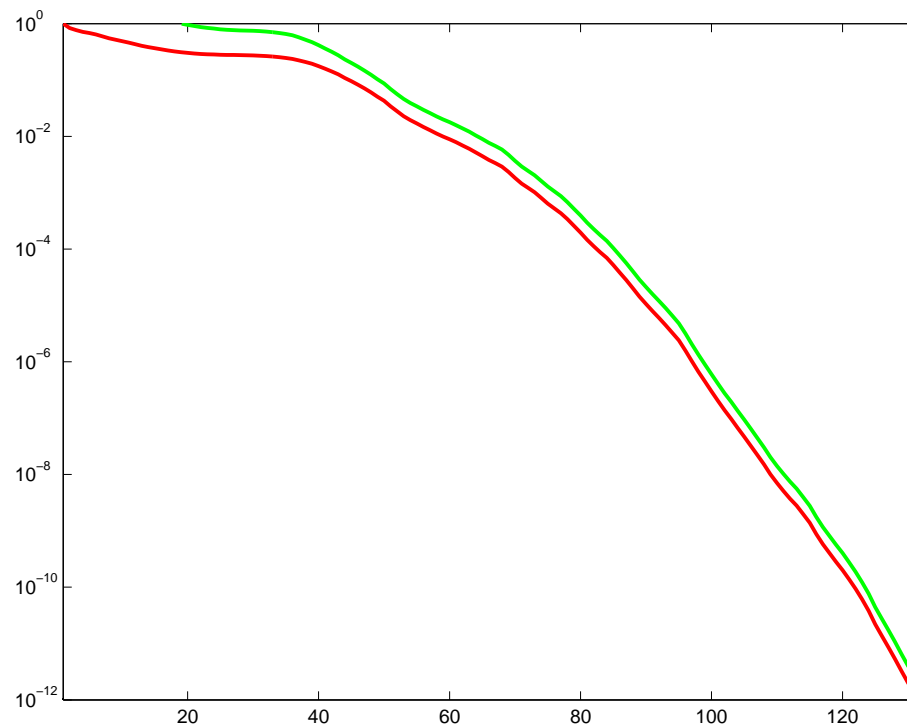
Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)

From Greenbaum, Rozložník and Strakoš (1997) we know that

$$\|r_m\|_2 \sim \sigma_{\min}(v_0, AV_m)$$

$$- \|r_m\|_2 / \|b\|_2$$

$$- \frac{\sigma_{\min}(v_0, AV_m)}{\sigma_{\min}(A)}$$



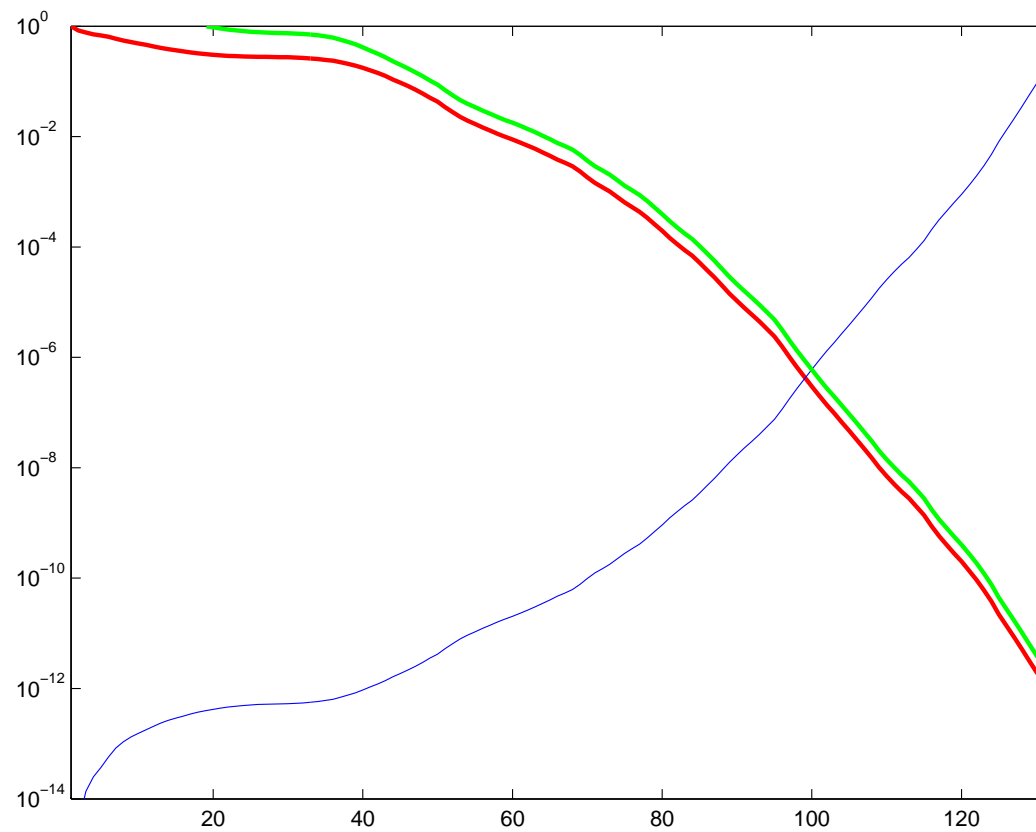
Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)

so from Björck (1967) we can check that the loss of orthogonality is proportionnal to $1/\|r_m\|_2$.

- $\|r_m\|_2/\|b\|_2$

- $\frac{\sigma_{\min}(v_0, AV_m)}{\sigma_{\min}(A)}$

- $\|I - V_m^T V_m\|_2$



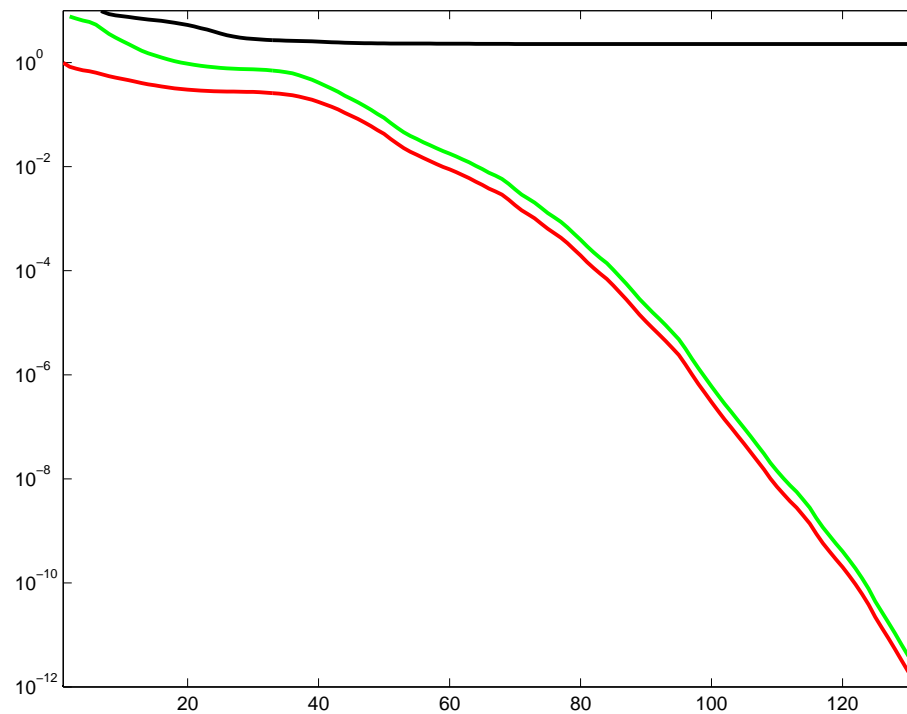
Experimental results - the case-test.(toolbox proved by Abderahmanne Bendali)

As $AV_m = V_m \bar{H}_m$, V_m and A are well-conditioned, we deduced that the second smallest singular value of (v_0, AV_m) is reasonably big.

$$- \|r_m\|_2 / \|b\|_2$$

$$- \frac{\sigma_{m+1}(v_0, AV_m)}{\sigma_{\min}(A)}$$

$$- \frac{\sigma_m(v_0, AV_m)}{\sigma_{\min}(A)}$$



EXPLANATION

$$\sigma_i(A) = \max_{E, \dim(E)=i} \min_{x \in E} \frac{\|Ax\|_2}{\|x\|_2}.$$

therefore

$$\sigma_{n-1}(v_1, AV) = \max_{E, \dim(E)=n-1} \min_{x \in E} \frac{\|Ax\|_2}{\|x\|_2}.$$

let consider $E = \begin{pmatrix} 0 \\ \mathbb{R}^{n-1} \end{pmatrix}$, $\dim(E) = n - 1$ and

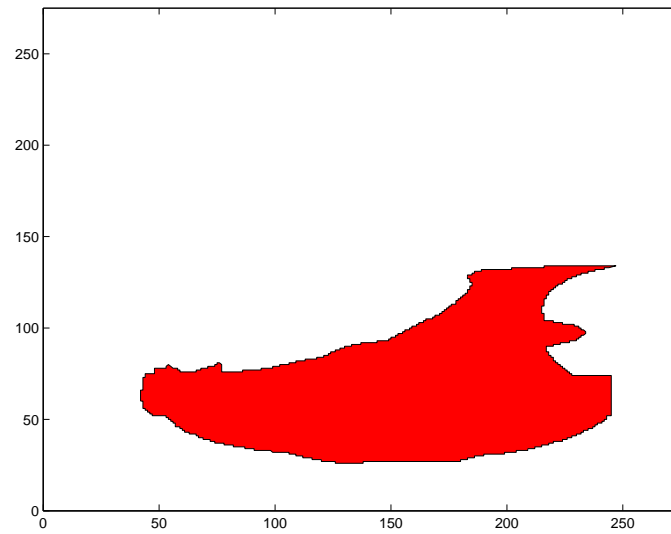
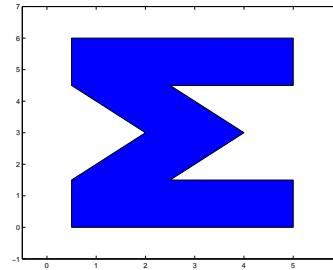
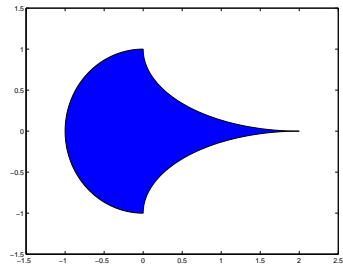
$$\min_{x \in E} \frac{\|(v_0, AV)x\|_2}{\|x\|_2} = \min_{y \in \mathbb{R}^{n-1}} \frac{\|(v_0, AV) \begin{pmatrix} 0 \\ y \end{pmatrix}\|_2}{\left\| \begin{pmatrix} 0 \\ y \end{pmatrix} \right\|_2} = \min_{y \in \mathbb{R}^{n-1}} \frac{\|AVy\|_2}{\|y\|_2} = \sigma_{\min}(AV)$$

so

$$\sigma_{n-1}(v_1, AV) = \max_{E, \dim(E)=n-1} \min_{x \in E} \frac{\|Ax\|_2}{\|x\|_2} \geq \min_{x \in E} \frac{\|(v_0, AV)x\|_2}{\|x\|_2} \geq \sigma_{\min}(A)\sigma_{\min}(V)$$

While $\sigma_n(v_1, AV)$ is controlled by the convergence of GMRES, $\sigma_{n-1}(v_1, AV)$ is controlled by A . We deduced that the matrix (v_1, AV) can be considered as a rank $n-1$ matrix. When it is orthogonalize with MGS, it is therefore easy to reorthogonalize.

Experimental results - the case-test.

GOLDORAK ($n = 715$, $\kappa(A) = 546$).CNSPH ($n = 310$, $\kappa(A) = 3,977$). CERFACS ($n = 312$, $\kappa(A) = 38$).

COMMENTS

- We may note that the matrix studied are complex whereas the study has been done only in the real case.
- We are going to run one GMRES with MGS for the first right-hand side (corresponding to an incident wave of 0°) then we project the 359 other right-hand sides (corresponding to incident waves from 1° to 359°) on the Krylov Space (V) given by the GMRES run.

When using MGS the right way to project is :

$$\begin{pmatrix} v_1^T b^{(i)} \\ v_2^T (I - v_1 v_1^T) b^{(i)} \\ \vdots \\ v_n^T (I - v_{n-1} v_{n-1}^T) \dots (I - v_1 v_1^T) b^{(i)} \end{pmatrix} \quad i = 1, \dots, 359$$

When using an orthonormal basis Q the right way to project is :

$$\begin{pmatrix} q_1^T b^{(i)} \\ q_2^T b^{(i)} \\ \vdots \\ q_n^T b^{(i)} \end{pmatrix} \quad i = 1, \dots, 359$$

Therefore we will see that the cost of the reorthogonalization (obtaining Q from V) is largely amortized.

- We may also remark that the relation $AV = VH$ is changed in the relation $AV = QH$ then the reorthogonalization does not matter the solves.

Experimental results - 1.

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-12}$

	GOLDORAK	CNSPH	CERFACS
m	715	310	312
$\kappa(A)$	516	3,977	38
n	123	101	68
$\ I - V^T V\ _2$ (MGS)	$1.33 \cdot 10^{-1}$	$1.82 \cdot 10^{-1}$	$5.39 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$1.35 \cdot 10^{-13}$	$1.43 \cdot 10^{-13}$	$1.60 \cdot 10^{-14}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$1.74 \cdot 10^{-15}$	$5.84 \cdot 10^{-16}$	$6.18 \cdot 10^{-16}$
time for projecting with MGS (s.)	36.07	7.59	4.49
time for reorth (s.)	4.79	1.11	0.59
time for projecting with reorth (s.)	5.66	1.18	0.79

Experimental results - 1.

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-12}$

	GOLDORAK	CNSPH	CERFACS
m	715	310	312
$\kappa(A)$	516	3,977	38
n	123	101	68
$\ I - V^T V\ _2$ (MGS)	$1.33 \cdot 10^{-1}$	$1.82 \cdot 10^{-1}$	$5.39 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$1.35 \cdot 10^{-13}$	$1.43 \cdot 10^{-13}$	$1.60 \cdot 10^{-14}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$1.74 \cdot 10^{-15}$	$5.84 \cdot 10^{-16}$	$6.18 \cdot 10^{-16}$
time for projecting with MGS (s.)	36.07	7.59	4.49
time for reorth (s.)	4.79	1.11	0.59
time for projecting with reorth (s.)	5.66	1.18	0.79

Experimental results - 1.

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-12}$

	GOLDORAK	CNSPH	CERFACS
m	715	310	312
$\kappa(A)$	516	3,977	38
n	123	101	68
$\ I - V^T V\ _2$ (MGS)	$1.33 \cdot 10^{-1}$	$1.82 \cdot 10^{-1}$	$5.39 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$1.35 \cdot 10^{-13}$	$1.43 \cdot 10^{-13}$	$1.60 \cdot 10^{-14}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$1.74 \cdot 10^{-15}$	$5.84 \cdot 10^{-16}$	$6.18 \cdot 10^{-16}$
time for projecting with MGS (s.)	36.07	7.59	4.49
time for reorth (s.)	4.79	1.11	0.59
time for projecting with reorth (s.)	5.66	1.18	0.79

Experimental results - 1.

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-12}$

	GOLDORAK	CNSPH	CERFACS
m	715	310	312
$\kappa(A)$	516	3,977	38
n	123	101	68
$\ I - V^T V\ _2$ (MGS)	$1.33 \cdot 10^{-1}$	$1.82 \cdot 10^{-1}$	$5.39 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$1.35 \cdot 10^{-13}$	$1.43 \cdot 10^{-13}$	$1.60 \cdot 10^{-14}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$1.74 \cdot 10^{-15}$	$5.84 \cdot 10^{-16}$	$6.18 \cdot 10^{-16}$
time for projecting with MGS (s.)	36.07	7.59	4.49
time for reorth (s.)	4.79	1.11	0.59
time for projecting with reorth (s.)	5.66	1.18	0.79

Experimental results - 1.

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-12}$

	GOLDORAK	CNSPH	CERFACS
m	715	310	312
$\kappa(A)$	516	3,977	38
n	123	101	68
$\ I - V^T V\ _2$ (MGS)	$1.33 \cdot 10^{-1}$	$1.82 \cdot 10^{-1}$	$5.39 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$1.35 \cdot 10^{-13}$	$1.43 \cdot 10^{-13}$	$1.60 \cdot 10^{-14}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$1.74 \cdot 10^{-15}$	$5.84 \cdot 10^{-16}$	$6.18 \cdot 10^{-16}$
time for projecting with MGS (s.)	36.07	7.59	4.49
time for reorth (s.)	4.79	1.11	0.59
time for projecting with reorth (s.)	5.66	1.18	0.79

Experimental results - 2 .

We run full GMRES without preconditioning and stop when $\frac{\|r_n\|_2}{\|b\|_2} < 10^{-9}$

	IMPCOLA	STEAM2	STEAM1
m	225	600	240
$\kappa(A)$	$9.2 \cdot 10^6$	$3.5 \cdot 10^6$	$3.0 \cdot 10^7$
n	210	159	199
$\ I - V^T V\ _2$ (MGS)	$5.93 \cdot 10^{-2}$	$2.48 \cdot 10^{-1}$	$2.26 \cdot 10^{-3}$
$\ I - Q^T Q\ _2$ (reorth)	$4.18 \cdot 10^{-11}$	$2.26 \cdot 10^{-10}$	$1.36 \cdot 10^{-10}$
$\ AV - Q\bar{H}\ _2$ (reorth)	$2.36 \cdot 10^{-16}$	$9.09 \cdot 10^{-16}$	$3.53 \cdot 10^{-16}$

COMMENTS

This application may seem very promising however we present them just as some test case for the reorthogonalization issue. Not like real interesting from the computational point of view. For two reasons :

- First of all, we remarked in our applications that two CGS runs faster than a single MGS and the basis obtained after 2 CGS is orthogonal. This is due to the big size of our matrix in this case the parallelism and BLAS3 operation in CGS counts a lot.

- If we want to compute $Q^T b$ in a good way from V of MGS the only solution is the long formula already presented. However $Q^T b$ is not a goal in itself in our application, we want y such as $y = \arg \min_{y \in \mathbb{R}^n} \|Q^T b - \bar{H}y\|_2$ We can compute y to a *good level* with the formula $y = \arg \min_{y \in \mathbb{R}^n} \|V^T b - \bar{H}y\|_2$

EXPLANATION

First of all we have

$$\bar{H} = Q_{n+1}^T AV \Rightarrow \kappa(H) \leq \kappa(V)\kappa(A),$$

we consider H well-conditioned.

We have $AV = VH + E_V = QH + E_Q$.

Let assume that the least square problem on H is solved exactly.

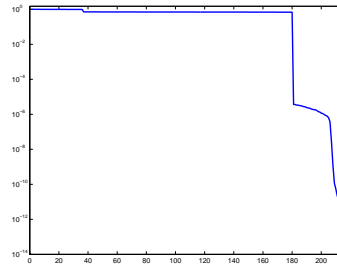
We can write

$$\begin{aligned} y^{(i)} &= (\bar{H}^T \bar{H})^{-1} \bar{H}^T (Q^T b^{(i)}) \\ &= (\bar{H}^T \bar{H})^{-1} (Q\bar{H})^T b^{(i)} \\ &= (\bar{H}^T \bar{H})^{-1} (V\bar{H} + E_V - E_Q)^T b^{(i)} \\ &= (\bar{H}^T \bar{H})^{-1} \bar{H}^T (V^T b^{(i)}) + (\bar{H}^T \bar{H})^{-1} (E_V - E_Q)^T b^{(i)} \end{aligned}$$

Meaning that the difference between the *good* y obtained from $Q^T b^{(i)}$ and the *bad* y obtained from $V^T b^{(i)}$ is of the order of $u\kappa(A)$. It is not backward stable, but it is as good as one have when he reorthogonalize MGS in just one direction.

Experimental results - 4 .

We take $A = \text{gre_216b}$ with $m = n = 216$ and $\kappa(A) = 8.10 \cdot 10^{14}$.



Singular values of gre_216b.

We run MGS on A and observe that

$$\|I - \bar{Q}^T \bar{Q}\|_2 = 1.54 \cdot 10^{-4}.$$

We fix the tolerance of the reorthogonalization $\text{tol} = 10^{-7}$, 10^{-10} and 10^{-13} and observe

tolerance	k	$\ I - \hat{Q}^T \hat{Q}\ _2$	$\ A - \hat{Q} \bar{R}\ _2$
10^{-7}	7	$7.32 \cdot 10^{-8}$	$1.79 \cdot 10^{-16}$
10^{-10}	11	$6.93 \cdot 10^{-11}$	$1.59 \cdot 10^{-16}$
10^{-13}	36	$2.44 \cdot 10^{-15}$	$1.59 \cdot 10^{-16}$

k is fixed by the tolerance wanted and the s .

Conclusion :

- During modified Gram-Schmidt algorithm, information from the distribution of the singular values of the initial matrix A is transferred to the constructed matrix Q ,
- this remark can be used for reorthogonalize Q .
- the reorthogonalization process is particularly efficient when A has a **near low** rank-deficiency.

References

- [1] BJÖRCK Å., (1967), Solving linear least squares problems using gram-schmidt orthogonalization. *BIT*, 7:1-21.
- [5] 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(1):176-190.

contact: `julien.langou@cerfacs.fr`

homepage: <http://www.cerfacs.fr/~langou/>

technical reports: <http://www.cerfacs.fr/algor/reports.html>