

# HOW (UN-)STABLE IS THE GRAM-SCHMIDT PROCESS?

**Miro Rozložník**

Institute of Computer Science,  
Czech Academy of Sciences,  
Prague, Czech Republic and  
Technical University of Liberec,  
email: miro@cs.cas.cz

joint results with  
**Luc Giraud, Jasper van den Eshof and  
Julien Langou**

5th ERCIM WS, Praha, August 27-29, 2004

# GRAM-SCHMIDT ORTHOGONALIZATION

$$A = (a_1, \dots, a_n) \in \mathcal{R}^{m,n}$$

$$m \geq \text{rank}(A) = n$$

orthogonal basis  $Q$  of  $\text{span}(A)$

$$Q = (q_1, \dots, q_n) \in \mathcal{R}^{m,n}, \quad Q^T Q = I_n$$

$$A = QR, \quad R \text{ upper triangular}$$

# CLASSICAL/MODIFIED GRAM-SCHMIDT ALGORITHMS

**classical** Gram-Schmidt (CGS) process

Schmidt, 1907,1908

**modified** Gram-Schmidt (MGS) process

Laplace, 1816, Cauchy, 1837

**classical** and **modified** Gram-Schmidt are mathematically equivalent, but they have "**different**" numerical properties

# CLASSICAL/MODIFIED GRAM-SCHMIDT ALGORITHMS

**classical (CGS)**

for  $j = 1, \dots, n$

$$u_j = a_j$$

for  $k = 1, \dots, j - 1$

$$u_j = u_j - (a_j, q_k)q_k$$

end

$$q_j = u_j / \|u_j\|$$

end

**modified (MGS)**

for  $j = 1, \dots, n$

$$u_j = a_j$$

for  $k = 1, \dots, j - 1$

$$u_j = u_j - (u_j, q_k)q_k$$

end

$$q_j = u_j / \|u_j\|$$

end

# CLASSICAL/MODIFIED GRAM-SCHMIDT ALGORITHMS

finite precision arithmetic:

$$\bar{Q} = (\bar{q}_1, \dots, \bar{q}_n), \quad \bar{Q}^T \bar{Q} \neq I_n,$$

$$\|I - \bar{Q}^T \bar{Q}\| \leq ?$$

$$A \neq \bar{Q} \bar{R}, \quad \|A - \bar{Q} \bar{R}\| \leq ?$$

$$\bar{R}?, \quad \text{cond}(\bar{R}) \leq ?$$

but are they really so different?

# ILLUSTRATION (BJÖRCK, 1967)

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & \sigma \end{pmatrix}$$

Läuchli, 1961

$$\kappa(A) = \sigma^{-1}(3 + \sigma^2)^{1/2} \approx \sigma^{-1}\sqrt{3}, \quad \sigma \ll 1$$

$$\sigma_{\min}(A) = \sigma, \quad \|A\| = \sqrt{3 + \sigma^2}$$

assume first that  $\sigma^2 \leq \varepsilon$ , so  $\text{fl}(1 + \sigma^2) = 1$

# ILLUSTRATION

if no other rounding errors are made, the matrices computed in CGS and MGS have the following form:

$$\begin{pmatrix} 1 & 0 & 0 \\ \sigma & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ \sigma & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix}$$

# ILLUSTRATION

$$\text{CGS: } (\bar{q}_3, \bar{q}_1) = -\sigma/\sqrt{2}, \quad (\bar{q}_3, \bar{q}_2) = 1/2,$$

$$\text{MGS: } (\bar{q}_3, \bar{q}_1) = -\sigma/\sqrt{6}, \quad (\bar{q}_3, \bar{q}_2) = 0$$

complete loss of orthogonality (  $\iff$  loss of lin. independence, loss of (numerical) rank ):

$$\sigma^2 \leq \varepsilon \text{ (CGS)}, \quad \sigma \leq \varepsilon \text{ (MGS)}$$

MGS: numerical full rank of  $A$ ,

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

CGS: numerical nonsingularity of  $A^T A$ ,

$$c(m, n)\varepsilon\kappa^2(A) < 1$$

# GRAM-SCHMIDT PROCESS VERSUS ROUNDING ERRORS

- **modified** Gram-Schmidt (MGS):

assuming  $\hat{c}_1 \varepsilon \kappa(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{\hat{c}_2 \varepsilon \kappa(A)}{1 - \hat{c}_1 \varepsilon \kappa(A)}$$

Björck, 1967 , Björck, Paige, 1992

- **classical** Gram-Schmidt (CGS)?

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{\tilde{c}_2 \varepsilon \kappa^{n-1}(A)}{1 - \tilde{c}_1 \varepsilon \kappa^{n-1}(A)}?$$

Kielbasinski, Schwettlik, 1994

Polish version of the book, 2nd edition

# CLASSICAL GRAM-SCHMIDT ORTHOGONALIZATION

$$\begin{aligned}u_j &= a_j - \sum_{k=1}^{j-1} r_{k,j} q_k \\ &= (I - Q_{j-1} Q_{j-1}^T) a_j\end{aligned}$$

$$r_{j,j} = \|u_j\|, \quad q_j = u_j / r_{j,j}$$

$$\begin{aligned}\sigma_{\min}(A) &= \sigma_{\min}(R) \leq |r_{j,j}| = \\ &\|u_j\| \leq \|a_j\| \leq \|A\| = \|R\|\end{aligned}$$

# CLASSICAL GRAM-SCHMIDT ORTHOGONALIZATION

$$\bar{u}_j = a_j - \sum_{k=1}^{j-1} \bar{r}_{k,j} \bar{q}_k + \delta u_j$$

$$\|\delta u_j\| \leq c_0 \varepsilon \|a_j\|$$

$$\bar{q}_j = \bar{u}_j / \bar{r}_{j,j} + \delta q_j$$

$$\|\delta q_j\| \leq (m + 4) \varepsilon$$

$$\bar{r}_{j,j} = \|\bar{u}_j\| + \delta r_{j,j}$$

$$|\delta r_{j,j}| \leq (m + 2) / 2 \varepsilon \|\bar{u}_j\|$$

$$A + \delta A = \bar{Q}\bar{R}$$

$$\|\delta A\| \leq c_1 \varepsilon \|A\|$$

$$A^T A + E = \bar{R}^T \bar{R}$$

$$\|E\| \leq c_2 \varepsilon \|A\|^2$$

# CLASSICAL GRAM-SCHMIDT ORTHOGONALIZATION

$$\bar{r}_{i,j} = (a_j, \bar{q}_i) + \delta r_{i,j}, \quad |\delta r_{i,j}| \leq m\varepsilon \|\bar{q}_i\| \|a_j\|$$

$$\bar{q}_i = \bar{u}_i / \bar{r}_{i,i} + \delta q_i, \quad \|\delta q_i\| \leq (m + 4)\varepsilon$$

$$\bar{u}_i = a_i - \sum_{k=1}^{i-1} \bar{r}_{k,i} \bar{q}_k + \delta u_i,$$

$$\|\delta u_i\| \leq c_0 \varepsilon \|a_i\|$$

$$\bar{r}_{i,i} \bar{r}_{i,j} = (a_j, a_i) - \sum_{k=1}^{i-1} \bar{r}_{k,i} \bar{r}_{k,j} + \dots$$

COMPUTED UPPER TRIANGULAR  
FACTOR IN CLASSICAL  
GRAM-SCHMIDT

$$A^T A + E = \bar{R}^T \bar{R}$$

$$\|E\| \leq c_2 \varepsilon \|A\|^2$$

assuming  $c_2 \varepsilon \kappa^2(A) < 1$

$$\|\bar{R}^{-1}\| \leq \frac{1}{\sigma_{\min}(A) [1 - c_2 \varepsilon \kappa^2(A)]^{1/2}}$$

# THE LOSS OF ORTHOGONALITY IN CLASSICAL GRAM-SCHMIDT

assuming  $c_2 \varepsilon \kappa^2(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{c_3 \varepsilon \kappa^2(A)}{1 - c_2 \varepsilon \kappa^2(A)}$$

$$A + \delta A = \bar{Q} \bar{R}, \quad A^T A + E = \bar{R}^T \bar{R}$$

$$\begin{aligned} & \bar{R}^T (I - \bar{Q}^T \bar{Q}) \bar{R} = \\ & -(\delta A)^T A - A^T \delta A - (\delta A)^T \delta A + E \end{aligned}$$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \left( 2\|\delta A\| \|A\| + \|\delta A\|^2 + \|E\| \right) \|\bar{R}^{-1}\|^2$$

# NUMERICAL EXAMPLE

$$A = \begin{pmatrix} 1 & \dots & 1 \\ \sigma & & \\ & \dots & \\ & & \sigma \end{pmatrix}$$

Läuchli, 1961

$$j \quad \|I - \bar{Q}_j^T \bar{Q}_j\| \quad \|I - (A\bar{L}^{-T})^T A\bar{L}^{-T}\|$$

$\sigma = 10^{-7}$ :

1	0.0000e+00	0.0000e+00
2	1.6266e-09	1.2045e-02
3	1.3280e-02	1.6991e-02
4	1.6491e-02	2.0117e-02

$\sigma = 10^{-4}$ :

1	0.0000e+00	0.0000e+00
2	2.7747e-13	1.0775e-09
3	2.2646e-09	6.0774e-09
4	2.9616e-09	6.0774e-09

# GRAM-SCHMIDT PROCESS VERSUS ROUNDING ERRORS

- modified Gram-Schmidt (MGS): assuming  $\hat{c}_1 \varepsilon \kappa(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{\hat{c}_2 \varepsilon \kappa(A)}{1 - \hat{c}_1 \varepsilon \kappa(A)}$$

Björck, 1967, Björck, Paige, 1992

- **classical Gram-Schmidt (CGS)**: assuming  $c_2 \varepsilon \kappa^2(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{c_3 \varepsilon \kappa^2(A)}{1 - c_2 \varepsilon \kappa^2(A)}!$$

Giraud, Van den Eshof, Langou, R, 2004