ROUNDING ERROR ANALYSIS OF THE CLASSICAL 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, Julien Langou and Jasper van den Eshof

Householder Symposium XVI, Seven Springs Mountain Resort,
Champion, Pennsylvania, USA, May 23-27, 2005

GRAM-SCHMIDT PROCESS AS QR ORTHOGONALIZATION

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

orthogonal basis Q of $span(A)$

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

$A = QR$, R upper triangular ($A^T A = R^T R$)

CLASSICAL AND 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 Gram-Schmidt can be **"quite unstable", can "quickly" lose** all semblance of **orthogonality**

GRAM-SCHMIDT PROCESS VERSUS ROUNDING ERRORS

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

  assuming $\widehat{c}_1 u\kappa(A) < 1$

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

  Björck, 1967 , Björck, Paige, 1992

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

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

  Kielbasinski, Schwettlik, 1994

  Polish version of the book, 2nd edition

TRIANGULAR FACTOR FROM CLASSICAL GRAM-SCHMIDT VS. CHOLESKY FACTOR OF THE CROSS-PRODUCT MATRIX

exact arithmetic:

$$r_{i,j} = (a_j, q_i) = \left(a_j, \frac{a_i - \sum_{k=1}^{i-1} r_{k,i} q_k}{r_{i,i}} \right)$$

$$= \frac{(a_j, a_i) - \sum_{k=1}^{i-1} r_{k,i} r_{k,j}}{r_{i,i}}$$

The computation of R in the classical Gram-Schmidt is closely related to the left-looking Cholesky factorization of the cross-product matrix $A^T A = R^T R$

$$
\begin{aligned}
\bar{r}_{i,j} &= fl(a_j, \bar{q}_i) = (a_j, \bar{q}_i) + \Delta e_{i,j}^{(1)} \\
&= \left(a_j, \frac{fl(a_i - \sum_{k=1}^{i-1} \bar{q}_k \bar{r}_{k,i})}{\bar{r}_{i,i}} + \Delta e_i^{(2)} \right) + \Delta e_{i,j}^{(1)}
\end{aligned}
$$

$$
\begin{aligned}
\bar{r}_{i,i}\bar{r}_{i,j} &= \left(a_j, a_i - \sum_{k=1}^{i-1} \bar{q}_k \bar{r}_{k,i} + \Delta e_i^{(3)} \right) \\
&+ \bar{r}_{i,i} \left[(a_j, \Delta e_i^{(2)}) + \Delta e_{i,j}^{(1)} \right] \\
&= (a_i, a_j) - \sum_{k=1}^{i-1} \bar{r}_{k,i} [\bar{r}_{k,j} - \Delta e_{k,j}^{(1)}] \\
&+ (a_j, \Delta e_i^{(3)}) + \bar{r}_{i,i} \left[(a_j, \Delta e_i^{(2)}) + \Delta e_{i,j}^{(1)} \right]
\end{aligned}
$$

CLASSICAL GRAM-SCHMIDT PROCESS: COMPUTED TRIANGULAR FACTOR

$$\Sigma_{k=1}^{i} \bar{r}_{k,i}\bar{r}_{k,j} = (a_i, a_j) + \Delta e_{i,j}$$

$$A^T A + \Delta E_1 = \bar{R}^T \bar{R}!$$

$$\|\Delta E_1\| \leq c_1 u \|A\|^2$$

The CGS process is another way how to compute **a backward stable Cholesky factor of the cross-product matrix** $A^T A$!

CLASSICAL GRAM-SCHMIDT PROCESS: COMPUTED TRIANGULAR FACTOR

$$A^T A + \Delta E_1 = \bar{R}^T \bar{R} \ , \ \|\Delta E_1\| \leq c_1 \varepsilon \|A\|^2$$

assuming $c_1 \varepsilon \kappa^2(A) < 1$,

$$\|\bar{R}^{-1}\| \leq \frac{1}{\sigma_{min}(A)\left[1 - c_1 \varepsilon \kappa^2(A)\right]^{1/2}}, \ \|\bar{R}\| \leq \|A\| \left[1 + c_1 \varepsilon \kappa^2(A)\right]^{1/2}$$

$$A + \Delta E_2 = \bar{Q}\bar{R}, \ \|\Delta E_2\| \leq c_2 \varepsilon \|A\|$$

CLASSICAL GRAM-SCHMIDT PROCESS: THE LOSS OF ORTHOGONALITY

$$A^T A + \Delta E_1 = \bar{R}^T \bar{R},\ A + \Delta E_2 = \bar{Q}\bar{R}$$

$$\bar{R}^T (I - \bar{Q}^T \bar{Q})\bar{R} =$$
$$-(\Delta E_2)^T A - A^T \Delta E_2 - (\Delta E_2)^T \Delta E_2 + \Delta E_1$$

assuming $c_1 \varepsilon \kappa^2(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{c_3 u \kappa^2(A)}{1 - c_1 u \kappa^2(A)}$$

GRAM-SCHMIDT PROCESS VERSUS ROUNDING ERRORS

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

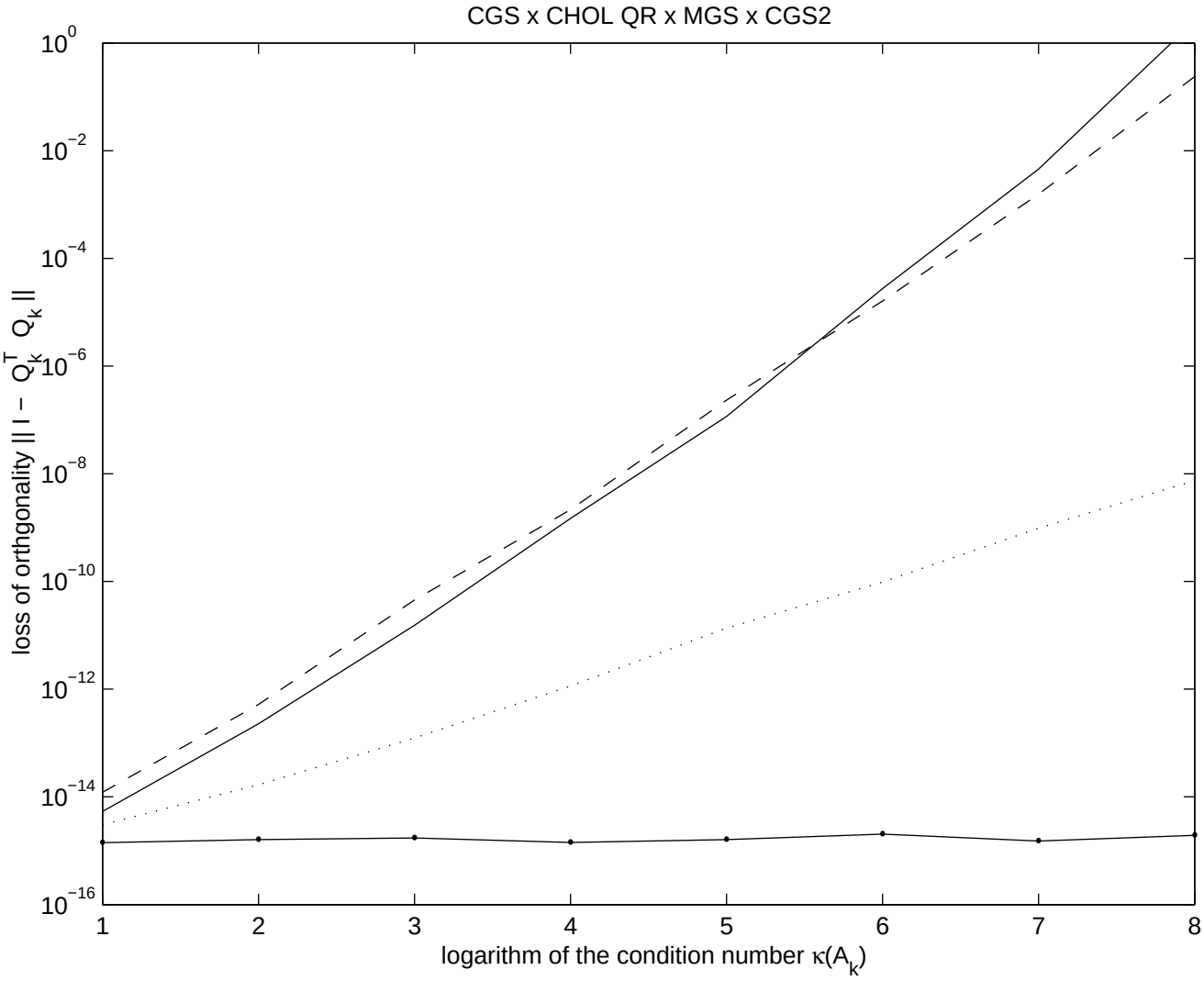
$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{\widehat{c}_2 u\kappa(A)}{1-\widehat{c}_1 u\kappa(A)}$$

Björck, 1967, Björck, Paige, 1992

- **classical Gram-Schmidt (CGS):** assuming $c_1 u\kappa^2(A) < 1$

$$\|I - \bar{Q}^T \bar{Q}\| \leq \frac{c_3 u\kappa^2(A)}{1-c_1 u\kappa^2(A)}!$$

Giraud, Van den Eshof, Langou, R, 2004



Stewart, "Matrix algorithms" book, p. 284, 1998

LEAST SQUARES PROBLEM WITH CLASSICAL GRAM-SCHMIDT

$$\|b - Ax\| = \min_u \|b - Au\|,\ r = b - Ax$$

$$A^T A x = A^T b$$

$$\bar{r} = (I - \bar{Q}\bar{Q}^T)b + \Delta e_1$$

$$(\bar{R} + \Delta E_3)\bar{x} = \bar{Q}^T b + \Delta e_2$$

$$\|\Delta e_1\|, \|\Delta e_2\| \leq c_0 u \|b\|,\ \|\Delta E_3\| \leq c_0 u \|\bar{R}\|$$

LEAST SQUARES PROBLEM WITH CLASSICAL GRAM-SCHMIDT

$$\bar{R}^T(\bar{R} + \Delta E_3)\bar{x} = (\bar{Q}\bar{R})^T b + \bar{R}^T \Delta e_2$$

$$(A^T A + \Delta E_1 + \bar{R}^T \Delta E_3)\bar{x} = (A + \Delta E_2)^T b + \bar{R}^T \Delta e_2$$

$$(A^T A + \Delta E)\bar{x} = A^T b + \Delta e$$

$$\|\Delta E\| \leq c_4 u \|A\|^2, \ \|\Delta e\| \leq c_4 u \|A\| \|b\|$$

LEAST SQUARES PROBLEM WITH CLASSICAL GRAM-SCHMIDT

$$\frac{\|\bar{r}-r\|}{\|b\|} \leq \kappa(A)(2\kappa(A)+1)\frac{c_5 u}{[1-c_1)u\kappa^2(A)]^{1/2}}$$

$$\frac{\|\bar{x}-x\|}{\|x\|} \leq \kappa^2(A)\left(2+\frac{\|r\|}{\|A\|\|x\|}\right)\frac{c_5 u}{1-c_1 u\kappa^2(A)}$$

The least square solution with classical Gram-Schmidt has the same forward error bound as the normal equation method:
$\bar{R}-\bar{Q}^T A = \bar{R}-\bar{R}^{-T}(A+\Delta E_2)^T A = -\bar{R}^{-T}[\Delta E_1 + (\Delta E_2)^T A]$

Björck, 1967

THE ARNOLDI PROCESS AND THE GMRES METHOD WITH THE CLASSICAL GRAM-SCHMIDT PROCESS

$$V_n = [v_1, v_2, \ldots, v_n]$$

$$[r_0, AV_n] = V_{n+1}[\|r_0\|e_1, H_{n+1,n}]$$

$H_{n+1,n}$ is an upper Hessenberg matrix

Arnoldi process is a (recursive) column-oriented QR decomposition of the (special) matrix $[r_0, AV_n]$!

$x_n = x_0 + V_n y_n$, $\|\|r_0\|e_1 - H_{n+1,n}y_n\| = \min_y \|\|r_0\|e_1 - H_{n+1,n}y\|$

THE GRAM-SCHMIDT PROCESS IN THE ARNOLDI CONTEXT: LOSS OF ORTHOGONALITY

- modified Gram-Schmidt (MGS):

  $\|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\| \leq \bar{c}_1 u \kappa([\bar{v}_1, A\bar{V}_n])$

  Björck, Paige 1967, 1992

- classical Gram-Schmidt (CGS):
  $\|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\| \leq \bar{c}_2 u \kappa^2([\bar{v}_1, A\bar{V}_n])$

  Giraud, Langou, R, Van den Eshof 2004

CONDITION NUMBER IN ARNOLDI VERSUS RESIDUAL NORM IN GMRES

The loss of orthogonality in Arnoldi is controlled by the convergence of the residual norm in GMRES:

$$\|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\| \leq \bar{c}_\alpha u \kappa^\alpha([\bar{v}_1, A\bar{V}_n]), \qquad \alpha = 1, 2$$

Björck 1967, Björck and Paige , 1992

Giraud, Langou, R, Van den Eshof 2003

$$\kappa([\bar{v}_1, A\bar{V}_n]) \leq \frac{\|[\bar{v}_1, A\bar{V}_n]\|}{\frac{\|\hat{r}_n\|}{\|\bar{r}_0\|}[1+\frac{\|\hat{y}_n\|^2}{1-\delta_n^2}]^{1/2}}$$

$$\frac{\|\hat{r}_n\|}{\|\bar{r}_0\|} = \|\bar{v}_1 - A\bar{V}_n\hat{y}_n\| = min_y\|\bar{v}_1 - A\bar{V}_n y\| \ , \ \delta_n = \frac{\sigma_{n+1}([\bar{v}_1, A\bar{V}_n])}{\sigma_n(A\bar{V}_n)} < 1$$

Paige, Strakoš, 2000-2002

Greenbaum, R, Strakoš, 1997

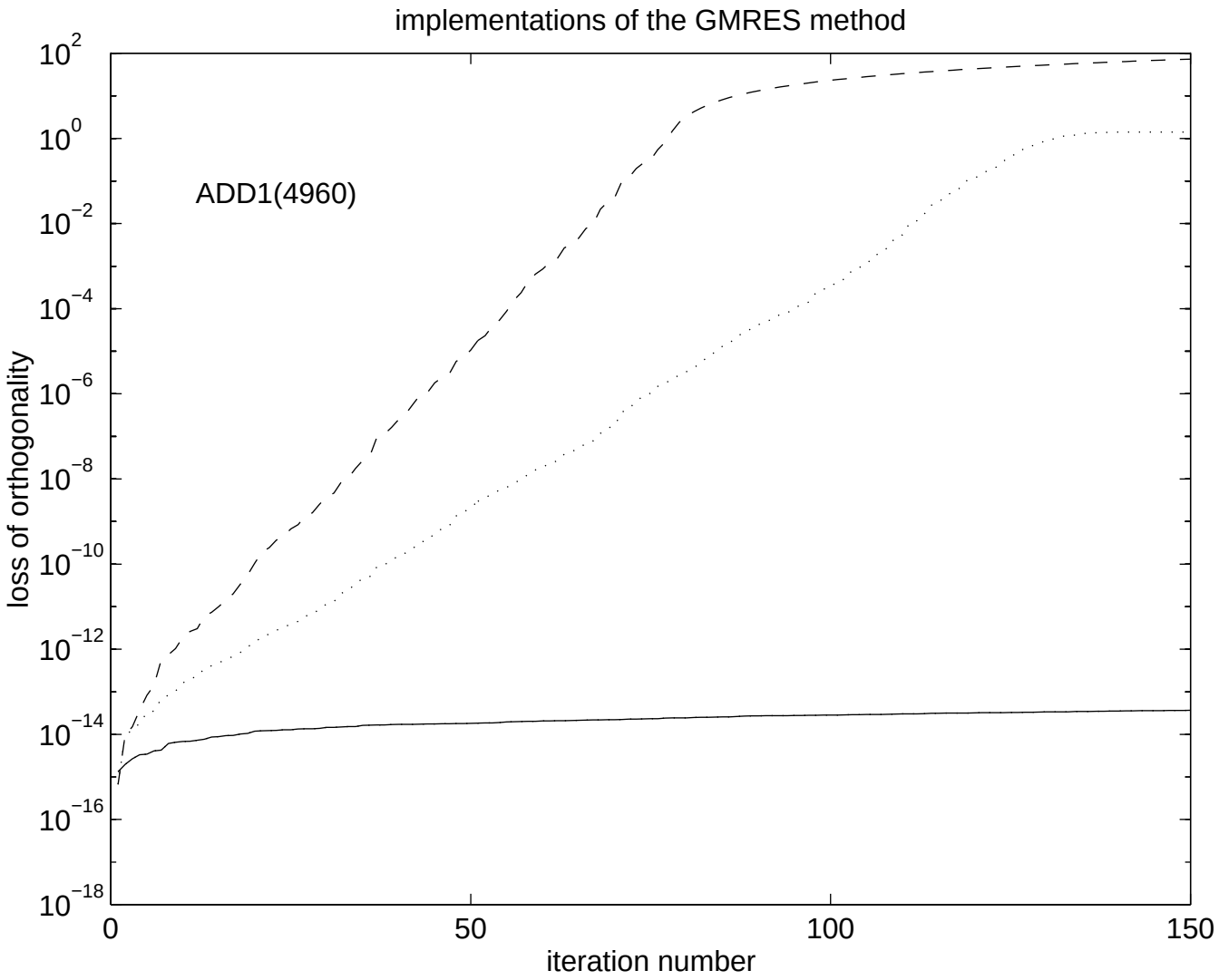
THE GMRES METHOD WITH THE GRAM-SCHMIDT PROCESS

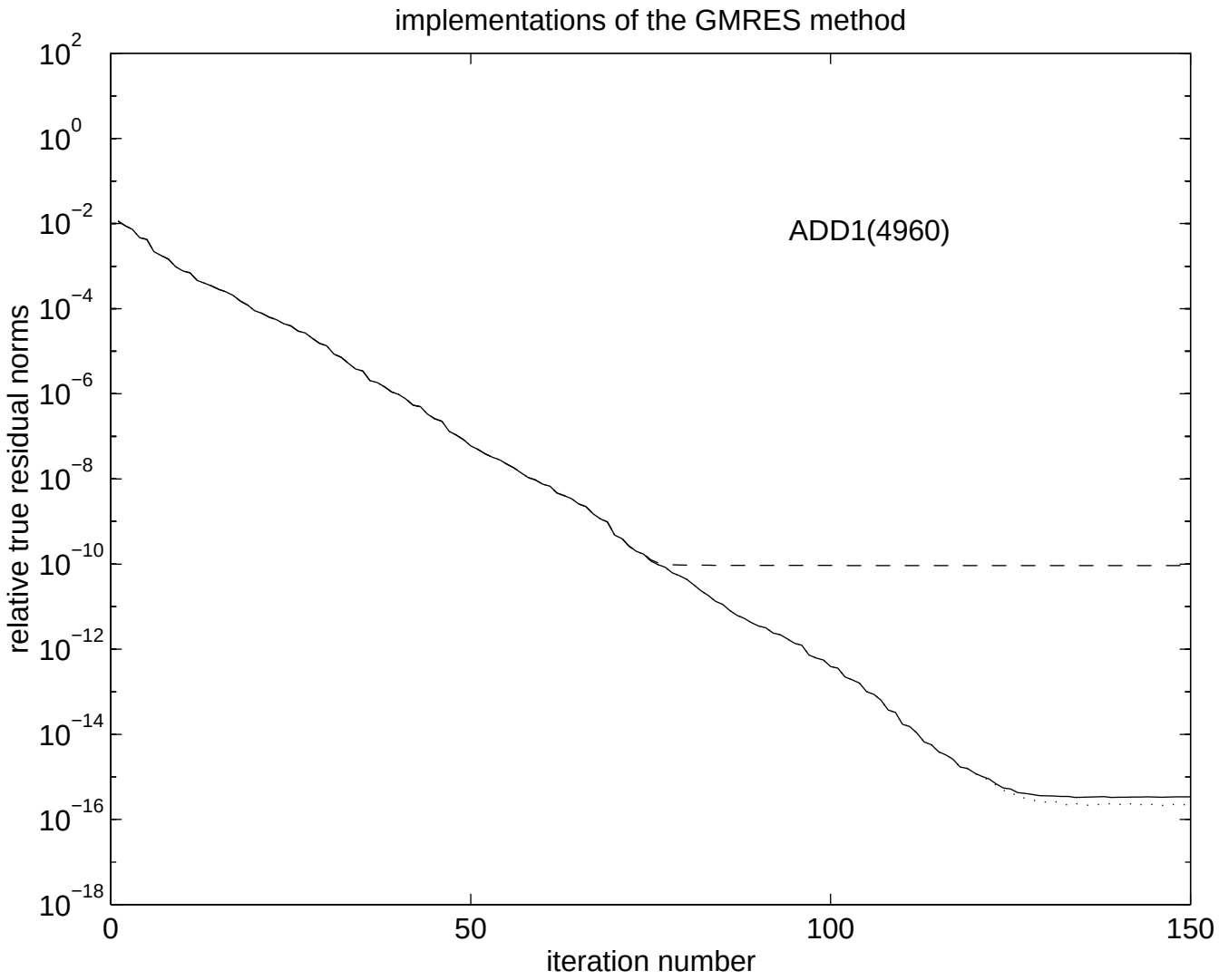
The total loss of orthogonality (rank-deficiency) in the Arnoldi process with Gram-Schmidt can occur **only after** GMRES reaches its final accuracy level:

- modified Gram-Schmidt (MGS):
$$\frac{\|\widehat{r}_n\|\|}{\|\bar{r}_0\|[1+\frac{\|\widehat{y}_n\|^2}{1-\delta_n^2}]^{1/2}} \approx \bar{c}_1[\bar{v}_1, A\bar{V}_n]\|u$$

see the talk of Chris Paige

- classical Gram-Schmidt (CGS):
$$\frac{\|\widehat{r}_n\|}{\|\bar{r}_0\|[1+\frac{\|\widehat{y}_n\|^2}{1-\delta_n^2}]^{1/2}} \approx [\bar{c}_2\|[\bar{v}_1, A\bar{V}_n]\|u]^{1/2}$$


implementations of the GMRES method
ADD1(4960)
loss of orthogonality
iteration number
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{-18}
0
50
100
150


implementations of the GMRES method
ADD1(4960)
relative true residual norms
iteration number
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{-18}
0
50
100
150

THE FLOATING POINT GMRES METHOD WITH THE GRAM-SCHMIDT PROCESS AS INEXACT KRYLOV METHOD

$$[\bar{v}_1, A\bar{V}_n] = \hat{V}_{n+1}[e_1, \bar{H}_{n+1,n}] + [\Delta f_1, \Delta F_n]$$

$$\hat{V}_{n+1}^T \hat{V}_{n+1} = I, \ \frac{\|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\|}{1 + \|\bar{V}_{n+1}\|} \leq \|\bar{V}_{n+1} - \hat{V}_{n+1}\| \leq \|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\|$$

$$(A + \mathcal{E}_n)\hat{V}_{n+1} = \hat{V}_{n+1}\bar{H}_{n+1,n}$$

$$\mathcal{E}_n = \left[A(\bar{V}_n - \hat{V}_n) - \Delta F_n\right] \hat{V}_n^T$$

Simoncini, Szyld, 2003

van den Eshof, Sleijpen, 2004

Giraud, Gratton, Langou, 2004

THE FLOATING POINT GMRES METHOD WITH THE GRAM-SCHMIDT PROCESS AS INEXACT KRYLOV METHOD

The **computed** upper Hessenberg matrix $\bar{H}_{n+1,n}$ satisfies the **exact** Arnoldi recurrence for the perturbed matrix $A + \mathcal{E}_n$ and initial vector $\widehat{v}_1$

$$\|I - \bar{V}_{n+1}^T \bar{V}_{n+1}\| \approx \|\mathcal{E}_n\|/\|A\| \nearrow c_\alpha(m,n) u \kappa^\alpha([\bar{v}_1, A\bar{V}_n]), \quad \alpha = 1,\, 2$$

$$\kappa([\bar{v}_1, A\bar{V}_n]) \leq \frac{\|[\bar{v}_1, A\bar{V}_n]\|}{\frac{\|\widehat{r}_n\|}{\|\bar{r}_0\|}[1+\frac{\|\widehat{y}_n\|^2}{1-\delta_n^2}]^{1/2}}$$

Paige, Strakoš, 2000-2002