NUMERICAL STABILITY OF ITERATIVE METHODS

Miro Rozložník

Institute of Computer Science
Czech Academy of Sciences
CZ-182 07 Prague, Czech Republic
email: miro@cs.cas.cz

joint work with
Christopher C. Paige and Julien Langou

GAMM-SIAM Conference on Applied Linear Algebra 2006
Düsseldorf, Germany , July 23-27, 2006

ACKNOWLEDGEMENT

Besides results of many other authors mentioned during the presentation, we thank to the coauthors of joint results:

Anne Greenbaum, Washington University, Seattle
Martin Gutknecht, ETH Zürich, Switzerland
Jörg Liesen, TU Berlin, Germany
Zdeněk Strakoš, Czech Academy of Sciences, Prague

OUTLINE

1. ROUNDING ERROR EFFECTS: DELAY OF CONVERGENCE AND MAXIMUM ATTAINABLE ACCURACY

2. BACKWARD ERROR AND BACKWARD STABILITY

3. THE CONJUGATE GRADIENT METHOD AND OTHER KRYLOV SUBSPACE METHODS WITH SHORT-TERM RECURRENCES

4. LOSS OF ORTHOGONALITY AND NUMERICAL BEHAVIOR OF GMRES

ITERATIVE METHODS IN EXACT ARITHMETIC

generate approximate solutions to the solution of $Ax = b$

$$x_0, x_1, \ldots, x_n \longrightarrow x$$

with residual vectors $r_0 = b - Ax_0, \ldots, r_n = b - Ax_n \to 0$

METHODS IN FINITE PRECISION ARITHMETIC

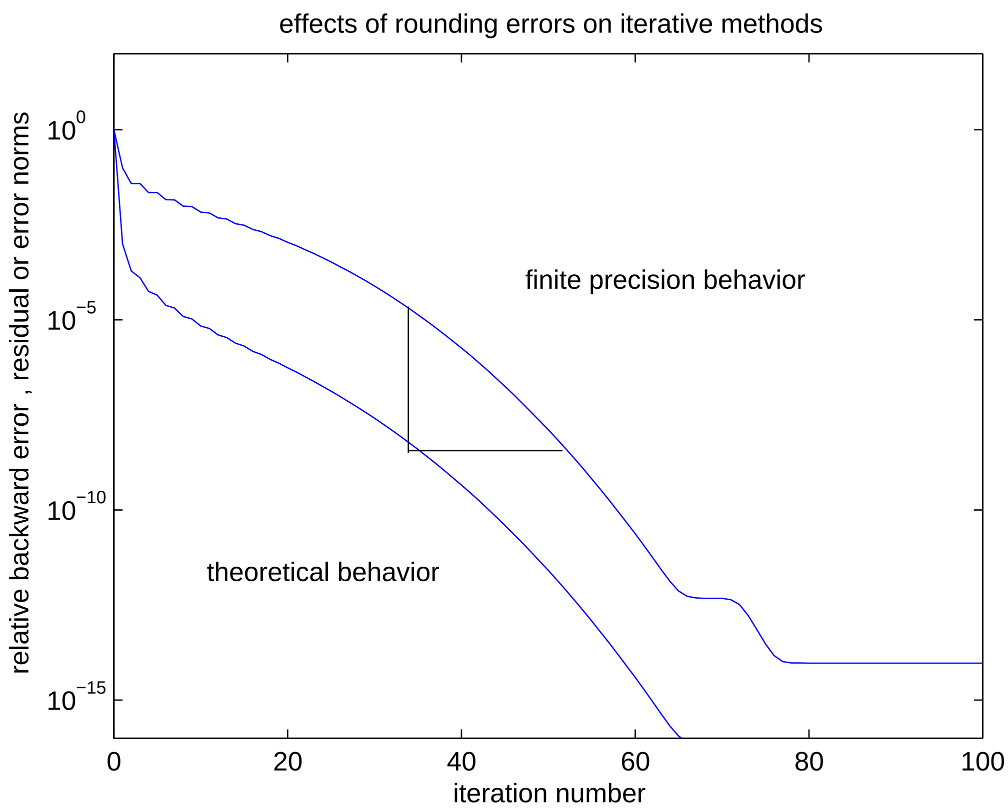
compute approximations $x_0, \bar{x}_1, \ldots, \bar{x}_n$ and updated residual vectors $\bar{r}_0, \bar{r}_1, \ldots, \bar{r}_n$ which are usually close to (but different from) the true residuals $b - A\bar{x}_n$

TWO MAIN QUESTIONS

- How good is the computed approximate solution $\bar{x}_n$? How many (extra) steps do we need to reach the same accuracy as in the exact method?

- How well the computed vector $\bar{r}_n$ approximates the (true) residual $b - A\bar{x}_n$? Is there a limitation on the accuracy of the computed approximate solution?

TWO EFFECTS OF ROUNDING ERRORS: DELAY OF CONVERGENCE AND LIMITING (MAXIMUM ATTAINABLE) ACCURACY



effects of rounding errors on iterative methods
relative backward error , residual or error norms
10^0
10^−5
10^−10
10^−15
finite precision behavior
theoretical behavior
0
20
40
60
80
100
iteration number

THE CONCEPT OF BACKWARD STABILITY

A backward stable algorithm eventually computes the exact answer to a nearby problem, i.e. the vector $\bar{x}_n$ satisfying

$$(A + \Delta A_n)\bar{x}_n = b + \Delta b_n$$

$$\|\Delta A_n\|/\|A\| \leq O(\varepsilon), \ \|\Delta b_n\|/\|b\| \leq O(\varepsilon)$$

$\Longleftrightarrow$ The normwise backward error associated with the approximate solution $\bar{x}_n$ satisfies

$$\frac{\|b - A\bar{x}_n\|}{\|b\| + \|A\|_{(F)}\|\bar{x}_n\|} \leq O(\varepsilon)$$

Prager, Oettli, 1964; Rigal, Gaches, 1967

see also Higham, 2nd ed. 2002; Stewart, Sun, 1990; Meurant 1999

THE LEVEL OF MAXIMUM ATTAINABLE ACCURACY

We are looking for the difference between the updated $\bar{r}_n$ and true residual $b - A\bar{x}_n$ (divided by $\|A\|\|\bar{x}_n\| + \|b\|$ or $\|A\|_F\|\bar{x}_n\| + \|b\|$)

$$\frac{\|b - A\bar{x}_n - \bar{r}_n\|}{\|A\|\|\bar{x}_n\| + \|b\|} \leq ?$$

$$\|\bar{r}_n\| \longrightarrow 0 \Longrightarrow \lim_{n\to\infty} \frac{\|b - A\bar{x}_n\|}{\|A\|\|\bar{x}_n\| + \|b\|} \leq ?$$

In the optimal case the bound is of $O(\varepsilon)$; then we have a backward stable solution

talk of Chris Paige, Vancouver, 2005

HISTORICAL REMARKS

- error analysis for stationary iterative methods (including the estimates for the forward and backward error for various classical schemes)

  Higham 2002, Chapter 17

- finite precision behavior of the symmetric Lanczos process

  Paige 1972, 1976, 1980

  Parlett and Scott 1979, 1980; Simon 1982, 1984

  Greenbaum, 1989; Strakoš, Greenbaum 1991, Druskin, Knizhnerman 1991

  and many other authors

- early results on maximum attainable accuracy of the conjugate gradient method not applicable to practical implementations

  Wozniakowski 1978, 1980, Bollen 1984

ITERATIVE METHODS USING TWO-TERM RECURRENCES

$$x_{n+1} = x_n + \alpha_n p_n$$

$$r_{n+1} = r_n - \alpha_n A p_n$$

Greenbaum 1994,1997

Sleijpen, Van der Vorst, Fokkema 1994

$$\|b - A\bar{x}_{n+1} - \bar{r}_{n+1}\| \leq O(\varepsilon)\|A\| \max_{k=0,\ldots,n+1} \{\|x - \bar{x}_k\|\}$$

$$\frac{\|b - A\bar{x}_{n+1} - \bar{r}_{n+1}\|}{\|A\|\|\bar{x}_{n+1}\| + \|b\|} \leq O(\varepsilon) \frac{\max_{k=0,\ldots,n+1}\{\|\bar{x}_k\|, \|x\|\}}{\|\bar{x}_{n+1}\|}$$

THE CONJUGATE GRADIENT METHOD

- classical coupled two-term recurrence (Hestenes, Stiefel) implementation, the estimation of the A-norm of the error $\|x - \bar{x}_n\|_A$ based on the relationship to Gauss quadrature

Hestenes, Stiefel 1952

Golub, Meurant 1994, 1997; Golub, Strakoš 1997

Dahlquist, Eisenstat, Golub, 1972; Dahlquist, Golub, Nash 1978

Strakoš, Tichý, 2002, 2005

- the backward stability of the Hestenes-Stiefel implementation $\frac{\|b - A\bar{x}_n - \bar{r}_n\|}{\|A\|\|\bar{x}_n\| + \|b\|} \leq O(\varepsilon)$

Greenbaum 1997

- similar result for the conjugate gradients via symmetric Lanczos process

Paige, Saunders 1976, Sleijpen, Van der Vorst, Modersitzki 1997

ITERATIVE METHODS USING TWO-TERM RECURRENCES

$$x_{n+1} = -(r_n + \alpha_n x_n + \beta_{n-1} x_{n-1})/\gamma_n$$
$$r_{n+1} = (A r_n - \alpha_n r_n - \beta_{n-1} r_{n-1})/\gamma_n$$

Stiefel 1955

Young, Jea 1980

Rutishauser, 1959

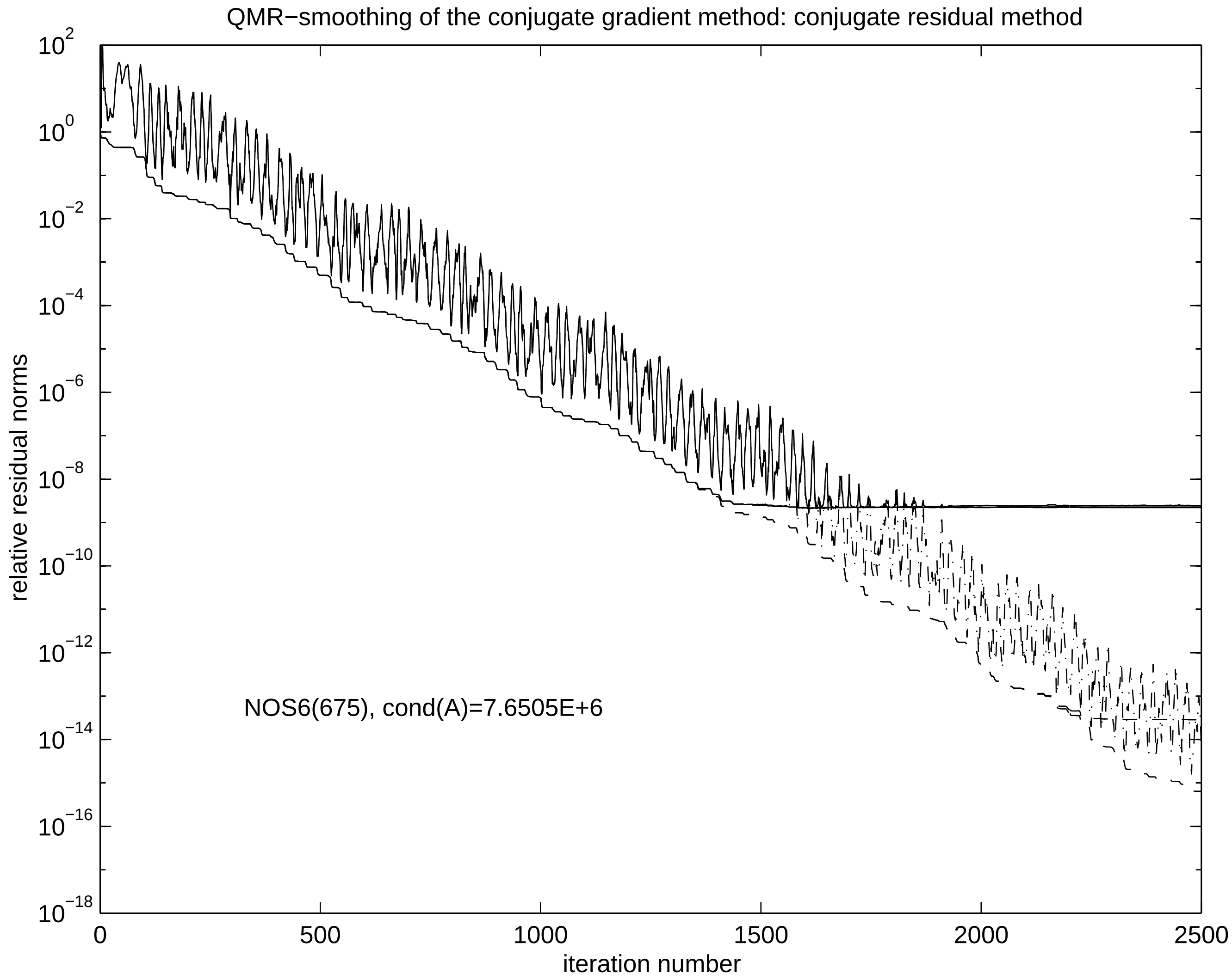
- methods based on two three term recurrences can give significantly less accurate approximate solutions than mathematically equivalent solvers implemented with coupled two-term recurrences

Gutknecht, Strakoš, 2001

RESIDUAL SMOOTHING TECHNIQUES IN FINITE PRECISION ARITHMETIC

- For the most conjugate gradient-type methods, the approximations $\bar{x}_n$ with residuals $\bar{r}_n \neq b - A\bar{x}_n$ usually exhibit an erratic convergence behavior. The residual smoothing techniques produce a better looking residual norm history plot (often nonincreasing)
- when properly implemented residual smoothing does not improve but also does not deteriorate the rate of primary (unsmoothed) method (FP analogues for the peak/ plateau property)
- smoothing does not improve the final accuracy of the primary method (it remains on the same level)

Gutknecht, R, 2001



QMR−smoothing of the conjugate gradient method: conjugate residual method
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{-18}
relative residual norms
0
500
1000
1500
2000
2500
iteration number
NOS6(675), cond(A)=7.6505E+6

THE RELATIONSHIP BETWEEN SYMMETRIC LANCZOS PROCESS AND CONJUGATE GRADIENT METHOD

THE CONCEPT OF CONVERGENCE DELAY

Greenbaum, 1989; Strakoš, Greenbaum, 1991

Paige, Strakoš, 1999

Meurant, Strakoš, Acta Numerica 2006

Strakoš, Liesen, ZAMM 2006

Delay in convergence of the conjugate gradient method (due to rounding errors) is given by the rank-deficiency of the computed Lanczos basis!

NONSYMMETRIC ARNOLDI PROCESS AND THE GMRES METHOD

Saad, Schultz 1986

THE CONCEPT OF CONVERGENCE DELAY:

Once the rank-deficiency occurs in the Arnoldi process the GMRES method stagnates on its final accuracy level

WELL-PRESERVED ORTHOGONALITY $\Rightarrow$ BACKWARD STABILITY

HOUSEHOLDER GMRES:

$$\|I - \bar{V}_N^T \bar{V}_N\| \leq O(\varepsilon)$$

Walker 1988, 1989

Greenbaum, Drkošová, R, Strakoš, 1995

$$\frac{\|b - A\bar{x}_N\|}{\|A\|\|\bar{x}_N\| + \|b\|} \leq O(\varepsilon)$$

$\bar{x}_N$ represents an exact solution to the nearby problem

$$(A + \Delta A)\bar{x}_N = b + \Delta b$$

The GRAM-SCHMIDT GMRES IMPLEMENTATION

The (modified) Gram-Schmidt version of GMRES (MGS-GMRES) is efficient, but looses orthogonality.

The rank-deficiency (total loss of orthogonality $\equiv$ loss of linear independence of computed basis vectors) in the Arnoldi process with (modified) Gram-Schmidt can occur **only after** GMRES reaches its final accuracy level!

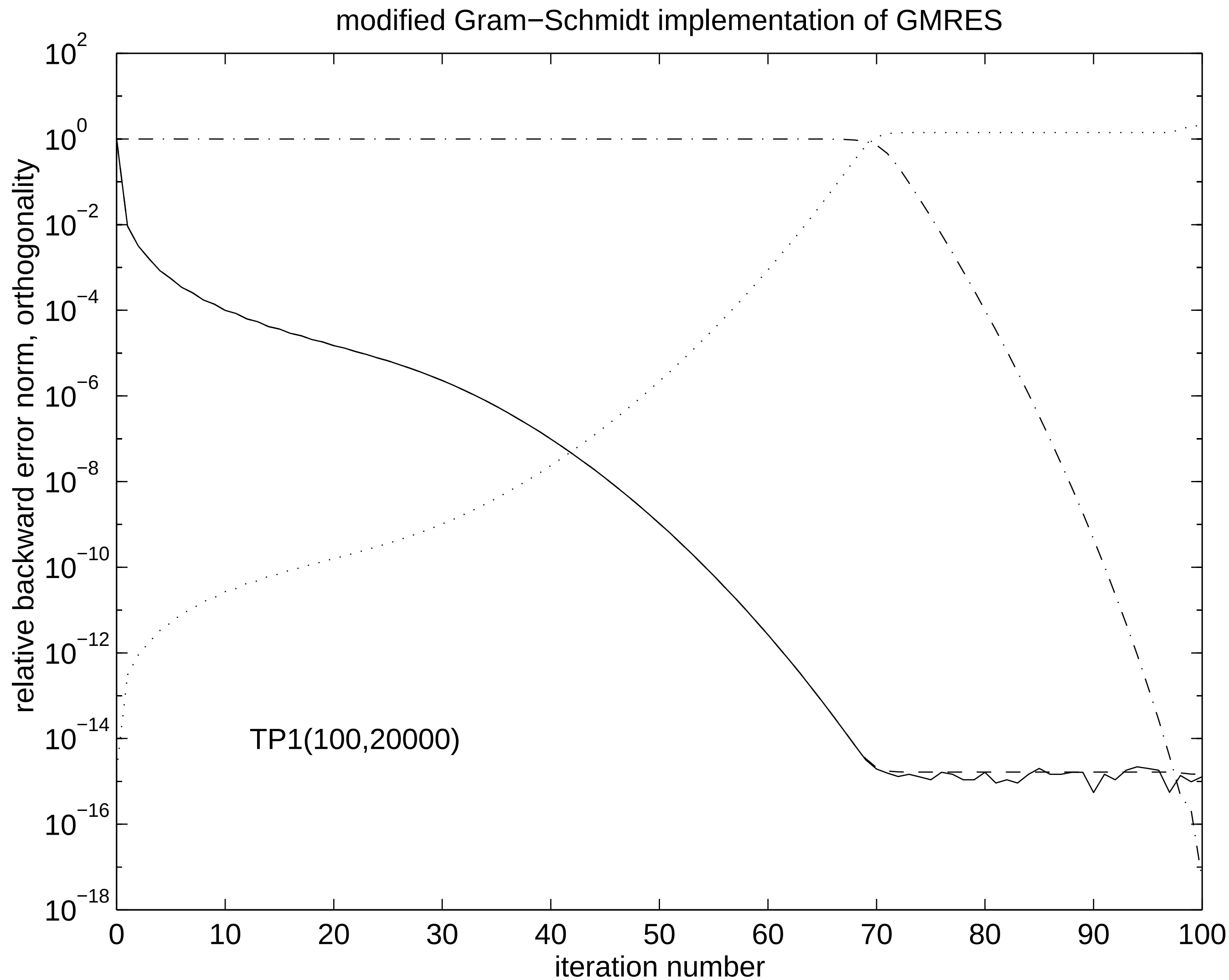
Greenbaum, R, Strakoš, 1997

Paige, R, Strakoš, 2006

GMRES WITH CGS ARNOLDI PROCESS

van den Eshof, Giraud, Langou, R, 2005

Smoktunowicz, Barlow, Langou, 2006


modified Gram−Schmidt implementation of GMRES
relative backward error norm, orthogonality
iteration number
TP1(100,20000)
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{-18}
0
10
20
30
40
50
60
70
80
90
100

GMRES WITH MGS ARNOLDI PROCESS

The MGS-GMRES implementation is a **backward stable** iterative method.

STATEMENT:

For some iteration step $n \leq N$ the computed approximate solution $\bar{x}_n$ satisfies

$$(A + \Delta A_n)\bar{x}_n = b + \Delta b_n$$

$$\|\Delta A_n\| / \|A\| \leq O(\varepsilon), \ \|\Delta b_n\| / \|b\| \leq O(\varepsilon)$$

Paige, R, Strakoš, 2006

THANK YOU FOR YOUR ATTENTION!