On numerical stability in large scale linear algebraic computations

Zdeněk Strakoš

Institute of Computer Science,
Academy of Sciences, Prague.

http://www.cs.cas.cz/~strakos

GAMM Annual Meeting,
Dresden, March 2004

Backward comments on the title

- Linear algebraic computations:
  - can be strongly nonlinear
  - are tightly related to classical analysis
- Large scale assumes but does not mean large scale computer environment. It attributes to mathematical approach
- Numerical stability is studied because things can go wrong due to rounding errors

Collaborators

Jörg Liesen, TU Berlin,

and coauthors of joint results mentioned during the presentation, in particular

Anne Greenbaum, Washington U, Seattle,
Gene H. Golub, Stanford U,
Miroslav Rozložník, ICS AS Prague,
Martin Gutknecht, ETH Zürich,
Vlastimil Pták, IM and ICS AS Prague,
Christopher C. Paige, McGill U, Montreal,
Petr Tichý, TU Berlin and ICS AS Prague,
Dianne P. O'Leary, U of Maryland, C. Park.

Outline

1. Introduction and example

2. Concept of backward stability

3. Surprising consequence: intermediate quantitites and desired accuracy

4. Results linking cost of computation and numerical stability

   (a) Lanczos method

   (b) Conjugate gradient method

   (c) GMRES method

1. Introduction

We start with a real-world problem and assume a typical case modeled by integro-differential equations

(in other cases the situation is similar)

Real-world problem

Integro-differential equations: model

- Infinite dimensional problem
- Existence and uniqueness of solution

Discretized problem: approximation

- Finite dimensional approximation
- Convergence to analytic solution

Algebraic problem: computation

- Linearization
- Matrix computation

Correspondence to reality

Errors of the model

- Caused by simplifying assumptions
- Limit applicability of results

Discretization errors

- Determined by discretization methods
- Limit relevance of numerical solutions

Computational errors

- Truncation and rounding errors
- Limit numerical accuracy

Large scale computation

- Direct and iterative means combination
- Principal advantage of the iterative part is in possibility of stopping the computation at a desired accuracy level
- Requires a meaningful stopping criterion
- Errors of the model, discretization errors and computational errors
  should be in balance

How fast can we get an acceptable approximate solution?

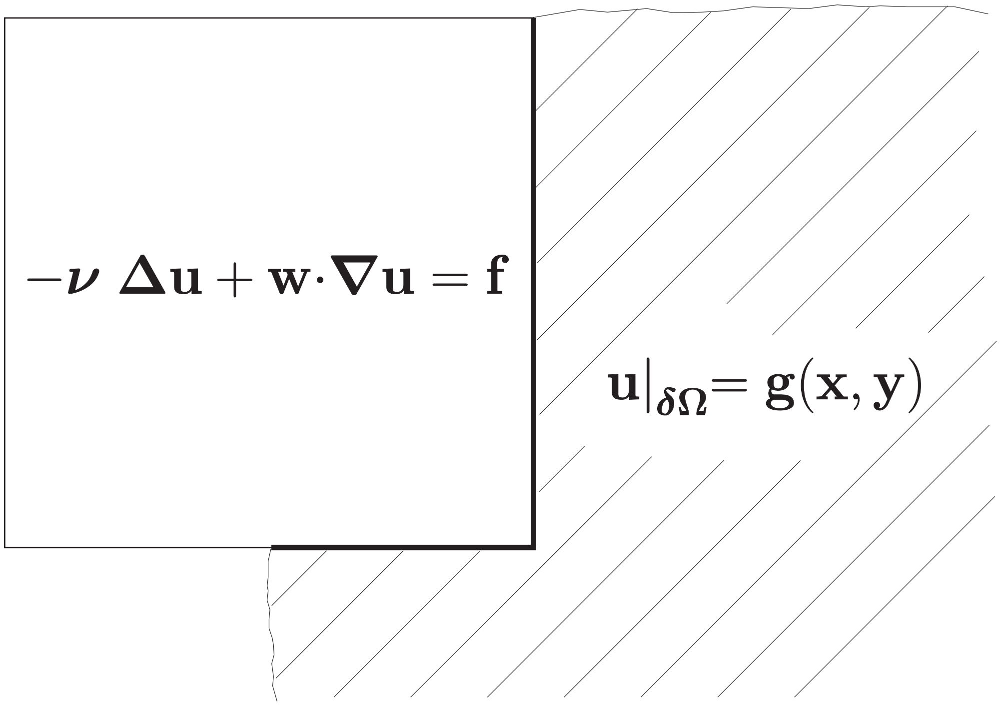
- In modern iterative methods we have to study early stage of computation (transient phase) which represents a nonlinear phenomenon in a finite dimensional space of a small dimensionality [Pták]

- Operator approach leads to asymptotics and can not describe the transient phase

- Solving $Ax = b$ for a given, problem-related b is often different from solving a system with the matrix A and some worst-case right hand side!

Example

Convection–diffusion model problem



$\uparrow \quad \mathbf{w}=[\mathbf{0},\mathbf{1}]^{\mathbf{T}}$

Convection dominated: $\quad \nu \ll \|w\|$

Discretization

- regular $h \times h$ grid, bilinear finite elements, mesh Peclet number $\quad P_h \equiv \frac{h\|w\|}{2\nu}$

- $P_h > 1$, then Galerkin discretization produces wiggles (nonphysical oscillations near the boundary layers)

- Streamline Upwind Petrov Galerkin (SUPG) equivalent to adding stabilizing diffusion in the direction of the flow (wind)

- wind parallel to the mesh; here the vertical wind

$$w = [0, 1]^T$$

Long list of authors, papers and books

Brooks, Hughes, Raithby, Roos, Stynes,
Tobiska, Morton, Axelsson, . . .

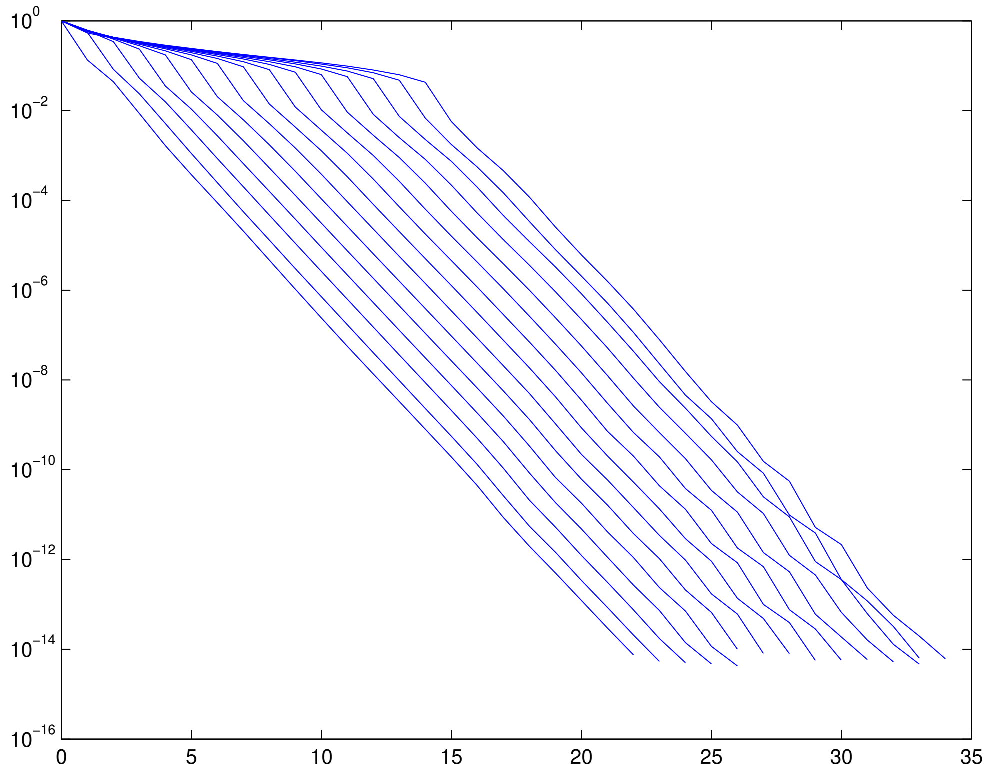
Convergence behaviour of GMRES

which is affected by stabilization

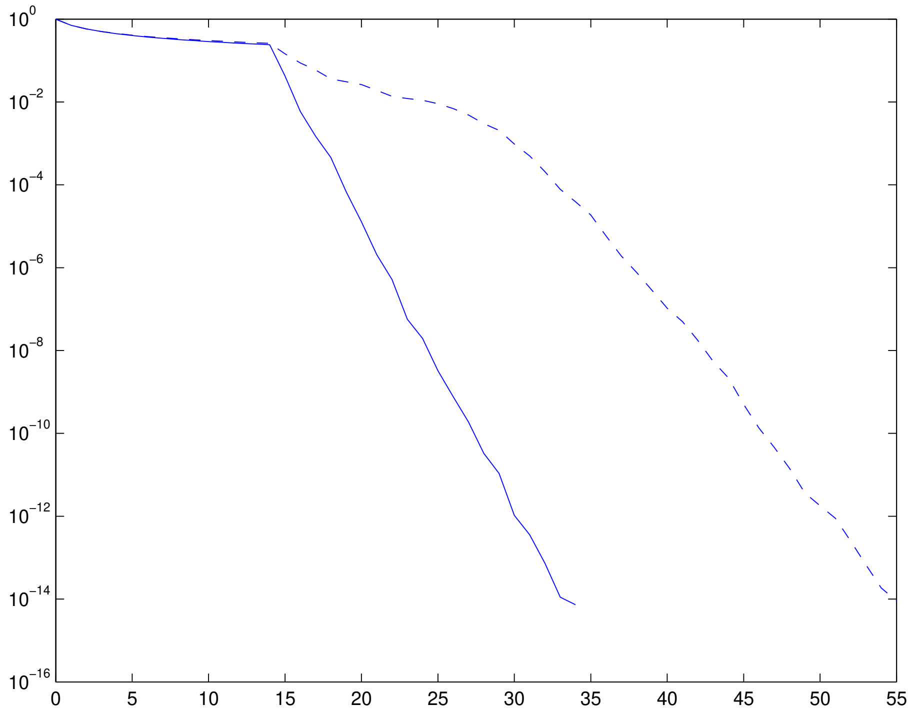
Trefethen, Eiermann, Elman, Ernst, Ramage,
Fischer, Silvester and Wathen,

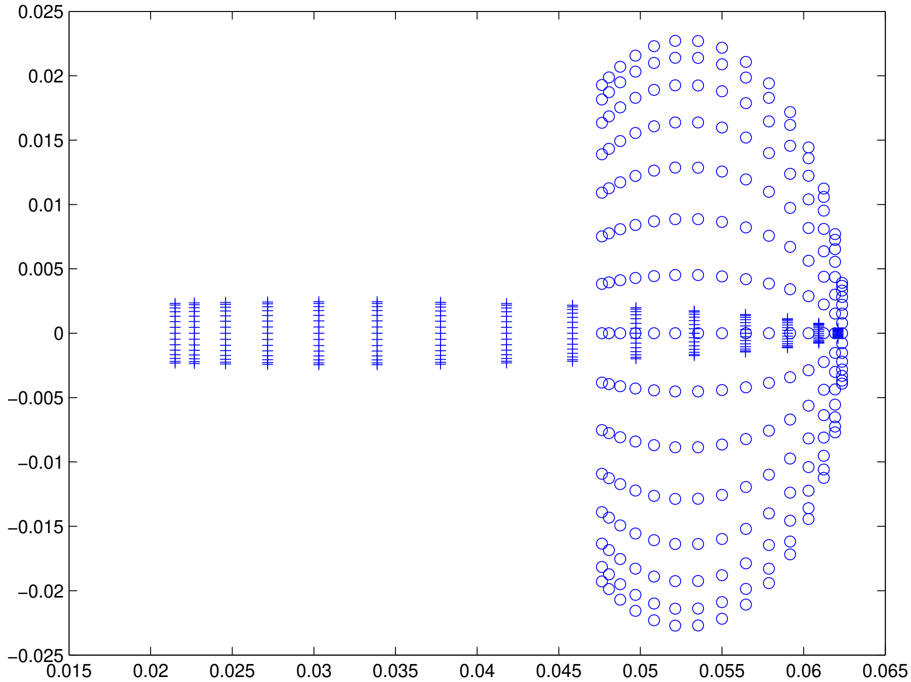
Liesen, S, SIAM J. on Matrix Anal. Appl.,
to appear, 2004
Liesen, S, SIAM J. on Sci. Comput.,
submitted, 2003

- Cost of computation depends (for the same operator) on boundary conditions



- Spectral analysis of the operator is not of much help


10^0
10^-2
10^-4
10^-6
10^-8
10^-10
10^-12
10^-14
10^-16
0 5 10 15 20 25 30 35 40 45 50 55


0.025
0.02
0.015
0.01
0.005
0
−0.005
−0.01
−0.015
−0.02
−0.025
0.015 0.02 0.025 0.03 0.035 0.04 0.045 0.05 0.055 0.06 0.065

2. Backward stability

How good is an approximate solution $\widehat{x}$ of the linear algebraic problem $A\,x = b$?

Perturbation theory:

$$(A + \Delta A)\,\widehat{x} = b + \Delta b\,.$$

Backward error:

Given $\widehat{x}$, construct ΔA , Δb such that both $\|\Delta A\|/\|A\|$ and $\|\Delta b\|/\|b\|$ are minimal;

$$\widehat{x} \longrightarrow \frac{\|\Delta A\|}{\|A\|} = \frac{\|\Delta b\|}{\|b\|} = \frac{\|b - A\widehat{x}\|}{\|b\| + \|A\|\|\widehat{x}\|}\,.$$

Measuring convergence:

- Backward error $\|r_n\| \,/\, (\, \|b\| + \|A\| \, \|x_n\| \,)$

We ask and answer the question

"How close is the problem $\hat{A}\, x_n = \hat{b}$
which is solved by x_n accurately
to the original problem $A\, x = b$?"

Perhaps this is what we need –
the matrix A and the right hand side b
are inaccurate anyway.

Is the computed convergence curve
close to the exact one?

- The role of the relative residual norm and the initial approximation x_0 ?

Backward error restricted to the right hand side only is given by

$$\|r_n\| / \|b\|$$

For unwise choice of x_0 this may differ greatly from the relative residual norm

$$\|r_n\| / \|r_0\| \ !$$

Example

Liesen, Tichý

$$-\Delta u \;=\; 32(y - y^2 + x - x^2)$$

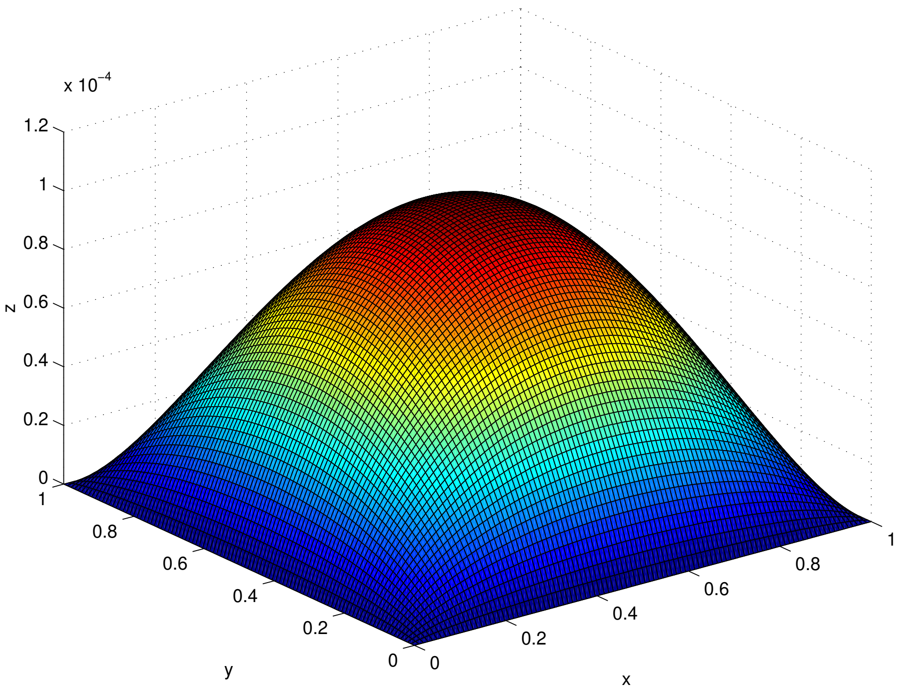
on a unit square with zero Dirichlet boundary conditions. Exact solution

$$u(x,y) \;=\; 16(xy - xy^2 - x^2y + x^2y^2)$$

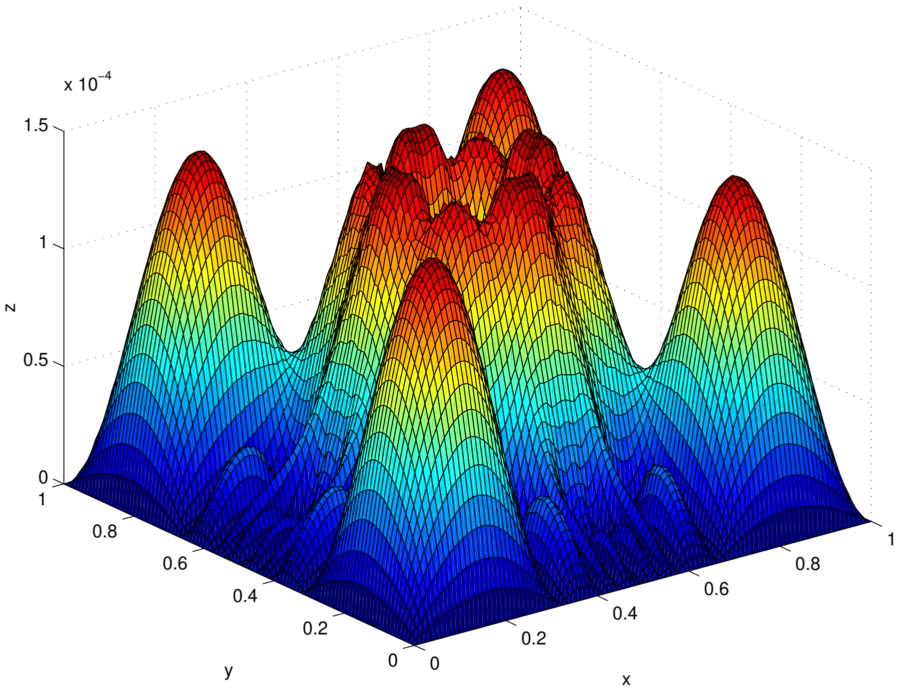
Linear FEM, discretization error in forming the linear algebraic system $\approx h^{-2}$. Stopping criterion for the CG computation

$$\frac{\|r_n\|}{\|b\| + \|A\|\,\|x_n\|} < \; h^{-3}$$

$|u-u_h|$, u_h is "exact" solution of $A_h\ u_h = b_h$



$|u-u_{it}|$, u_{it} is approximate solution of $A_h\ u_h = b_h$, tol=h^3



- Then the computational error does not contribute significantly to the total error measured by the accuracy of the computed approximate solution

- The criterion is cost–efficient. A similar reasoning based on the relative residual norm approximately doubles the number of iterations

- For a different and thorough approach see Arioli et al.

3. Intermediate quantities

How does an accuracy of the intermediate quantities affect the final accuracy of the computed result?

Surprising observation [Parlett, Wilkinson]

The number of significant digits in the intermediate quantities generated in a computation may be quite irrelevant to the accuracy of the final output

Vital correlations between (inaccurately) computed quantities lead to highly accurate final results. Understanding gained via rounding error analysis can guarrantee final accuracy

Example

Lanczos method for solving
Hermitian eigenvalue problem

Principle of the Lanczos method

(consider steps 1 through n)

Find an N by n matrix Q_n having orthonormal columns such that

$$Q_n^* A Q_n = T_n \,,$$

where T_n is Hermitian tridiagonal. Eigenvalues of T_n are then considered approximations of the (dominant) eigenvalues of A

In the presence of rounding errors, however, Q_n does not have orthonormal columns. The columns may even become (numerically) linearly dependent. Even worse, for the computed quantities

$$Q_n^* A Q_n \neq T_n\,,$$

and most of the entries in T_n may not have a single digit of accuracy, i.e.

$$T_n - T_n \quad \text{is} \quad \textbf{large.}$$

Does this mean disaster? No!
The magic is called backward error,
see [Wilkinson, Paige, Greenbaum]

For a given Lanczos computation there exist:

- An M by M matrix $\hat{A}$ having **all its eigenvalues close to the eigenvalues of** A, $M \geq N$, possibly $M \gg N$

- An M by n matrix $\widehat{Q}_n$ having orthonormal columns such that

$$\widehat{Q}_n^* \hat{A} \, \widehat{Q}_n \; = \; T_n$$

Results of the **finite precision** Lanczos computation for the matrix A

are equivalent to

the results of the **exact** Lanczos computation for the matrix $\hat{A}$ having nearby eigenvalues

Consequently, T_n is used for computing the eigenvalues of A to full machine precision

4. Cost and numerical stability

Towards a mathematical foundation of numerical analysis – a quest for formal mathematical model of computing with real numbers see [Smale, Cucker, Blum, Shub 1997]:

- Complexity theory of numerical analysis – study of the number of arithmetic operations required to pass from the input to the output of a numerical problem

- Upper bounds aspect – **worst or average** case analysis of basic algorithms

- Lower bounds aspect – examination of efficiency for **all algorithms** solving a given problem (the intrinsic difficulty of solving a problem)

Complications

- Ill-posed problems

- Conditioning

- Round-off errors

- Problems are by their nature solved only to a certain precision (eigenvalue problems, . . . , iterative methods in general)

Conclusion

There are practically no results linking complexity and numerical stability of computing over real numbers [Cucker 99]

Our stand

In numerical analysis, algorithms are means to solve practical problems [SCBS 1997, p.23]

In addition to that, practical problems mean particular (class of) data, for which approximate solutions should be found

Consequently, questions related to a particular problem (including data) are much more specific than worst-case or average-case bounds. We do not use complexity and restrict ourselves to the cost of particular computations

There are many results linking cost of computation to numerical stability!

4.1 Lanczos and CG basics

Consider A Hermitian (positive definite)

$$A = U \,\mathrm{diag}\,(\lambda_i)\, U^*, \quad U^*U = UU^* = I$$

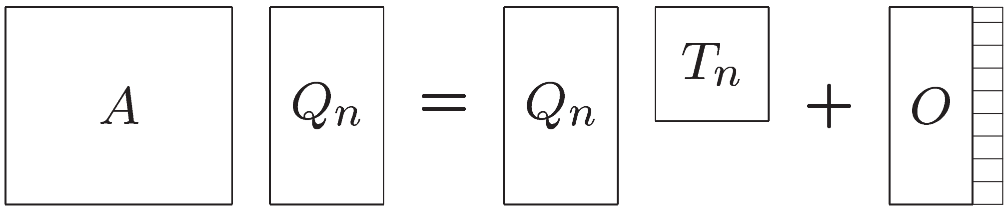
$$U = [u_1, \ldots, u_N]$$

Lanczos process

$$AQ_n \;=\; Q_n T_n + \beta_{n+1} q_{n+1}\, e_n^T$$

$$Q_n = [q_1, \ldots, q_n]$$

Three–term recurrence for generating orthonormal basis of Krylov subspaces



$$T_n = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix}$$

Jacobi matrix

$$T_n \;=\; S_n \,\mathsf{diag}\,(\theta_j^{(n)})\; S_n^*$$

$$S_n = [s_1^{(n)}, \ldots, s_n^{(n)}],$$
$$S_n^* S_n = S_n S_n^* = I$$

Relation to orthogonal polynomials

$$q_{n+1} = \psi_n(A)\, q_1/(\beta_2\beta_3 \ldots \beta_{n+1})$$

$$\psi_n \; : \; \| \psi_n(A)\, q_1 \| = \min_{\psi \in \mathcal{M}_n} \| \psi(A)\, q_1 \|$$

$$\downarrow$$

$$\sum_{i=1}^{N} (u_i, q_1)^2 \psi_n^2(\lambda_i) =$$

$$= \min_{\psi \in \mathcal{M}_n} \sum_{i=1}^{n} (u_i, q_1)^2 \psi^2(\lambda_i)$$

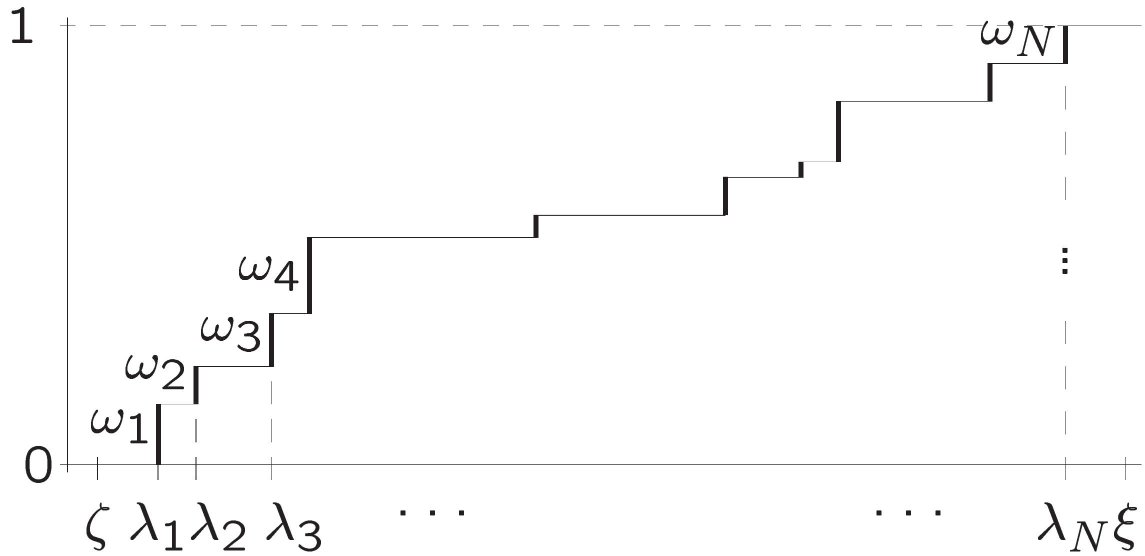
$\{1, \psi_1, \ldots, \psi_n\}$ monic orthogonal polynomials wrt

$$(\varphi, \psi) = \sum_{i=1}^{N} \omega_i\, \varphi(\lambda_i)\psi(\lambda_i)\,, \quad \omega_i = (u_i, q_1)^2$$

Riemann-Stieltjes integral

$$\sum_{1}^{N} \omega_i f(\lambda_i) = \int_{\zeta}^{\xi} f(\lambda) d\omega(\lambda)$$

$$\begin{aligned}
\omega(\lambda) &= 0 && \zeta \le \lambda < \lambda_1 \\
\omega(\lambda) &= \sum_{j=1}^{l} \omega_j && \lambda_l \le \lambda < \lambda_{l+1} \\
\omega(\lambda) &= \sum_{j=1}^{N} \omega_j && \lambda_N \le \lambda \le \xi
\end{aligned}$$



Repeating the argument for the matrix T_n with the initial vector e_1:

T_n is determined by the Lanczos process for the matrix T_n and the starting vector e_1,

$\{1, \psi_1, \ldots, \psi_n\}$ are orthogonal wrt

$$(\varphi, \psi) = \sum_{i=1}^{n} \omega_i^{(n)} \, \varphi\left(\theta_i^{(n)}\right) \, \psi\left(\theta_i^{(n)}\right) \, ,$$

$$\omega_i^{(n)} = \left(s_i^{(n)}, e_1\right)^2$$

The n-th Riemann-Stieltjes integral

$$\sum_{i=1}^{n} \omega_i^{(n)} f\left(\theta_i^{(n)}\right) = \int_{\zeta}^{\xi} f(\lambda)\, d\omega^{(n)}(\lambda)$$

$$\begin{array}{lcll}
\omega^{(n)}(\lambda) & = & 0 & \zeta \le \lambda < \theta_1^{(n)} \\
\omega^{(n)}(\lambda) & = & \displaystyle\sum_{j=1}^{l} \omega_j^{(n)} & \theta_l^{(n)} \le \lambda < \theta_{l+1}^{(n)} \\
\omega^{(n)}(\lambda) & = & \displaystyle\sum_{j=1}^{n} \omega_j^{(n)} & \theta_n^{(n)} \le \lambda < \xi
\end{array}$$

Lanczos process:

- sequence of orthonormal vectors
  $$\{q_1, \ldots, q_n\}$$

- sequence of Jacobi matrices
  $$\{T_1, \ldots, T_n\}$$

- sequence of monic orthogonal polynomials
  $$\{1, \ldots, \psi_n\}$$

- sequence of R-S integrals with the distribution functions
  $$\{\omega^{(1)}, \ldots, \omega^{(n)}\}$$

- sequence of continued fractions
  $$\{C_1, \ldots, C_n\}$$

Relationship between the original

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda)$$

and the n-th R-S integral

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega^{(n)}(\lambda) \;=\; \sum_{i=1}^{n} \omega_i^{(n)}\, f\left(\theta_i^{(n)}\right)$$

$\Longrightarrow$ **Gauss Quadrature !**

Lanczos algorithm
$\equiv$ matrix formulation of the Gauss quadrature

Conjugate gradient method (CG)

$$\|x - x_n\|_A = \min_{u \in x_0 + K_n(A, r_0)} \|x - u\|_A$$

where $\quad K_n(A, r_0) = \text{span}\{r_0, \ldots, A^{n-1} r_0\}$

- $\min\limits_{z \in K_n(A, r_0)} \|(x - x_0) - z\|_A$,
- $x - x_n = (x - x_0) - z_n \perp_A K_n(A, r_0)$,
- $r_n = b - Ax_n = A(x - x_n) \perp K_n(A, r_0)$,
- $r_n \perp \text{span}\{q_1, \ldots, q_n\}$.

The CG approximation is determined by

$$x_n = x_0 + Q_n\, y_n, \qquad T_n\, y_n = \|r_0\|\, e_1\,.$$

The essence of CG is nothing but
Gauss quadrature !

Consequence

Everything is determined by $\omega(\lambda)$. The way the eigenvalues are linked to convergence is given by the way $\omega(\lambda)$ determines the individual $\omega^{(n)}(\lambda)$.

This relationship is all but trivial !

The essence of the CG method

$$Ax = b,\ x_0 \qquad \longrightarrow \qquad \int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda)$$

$$\uparrow \qquad\qquad\qquad\qquad \uparrow$$

$$T_n y_n = \|r_0\|\, e_1 \qquad \longleftrightarrow \qquad \sum_{i=1}^{n} \omega_i^{(n)} f\left(\theta_i^{(n)}\right)$$

$$x_n = x_0 + Q_n y_n$$

Gauss quadrature !

$$\boxed{\omega^{(n)} \longrightarrow \omega(\lambda)}$$

4.2 Lanczos accuracy

Loss of orthogonality

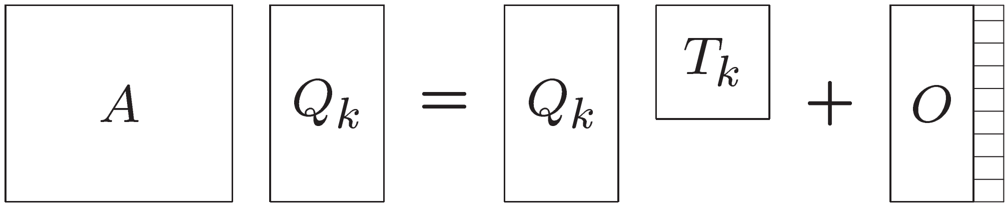
Lanczos vectors are computed using three term recurrence. Consequently, orthogonality (even linear independence) may be lost quickly!

However, [Paige 1971, 76, 80]:

The loss of orthogonality follows regular structure!

Exact arithmetic

$$AQ_n = Q_n T_n + \beta_{n+1} q_{n+1} e_n^T$$



Finite precision computation

$$AQ_n = Q_n T_n + \beta_{n+1} q_{n+1} e_n^T + F_n$$

$$\| F_n \| \leq k^{1/2} \| A \| \varepsilon_1$$

$$Q_n^T Q_n \neq I$$

T_n computed by FP $L(A, q_1)$ may be far from the theoretical counterpart eigendecompositions:

$A = U \text{ diag}(\lambda_i)\, U^*$

$$U = [u_1, \ldots, u_N], \qquad U^*U = I$$

$T_n = S_n \text{ diag}\,(\theta_j^{(n)})\; S_n^*$

$$S_n = [s_1^{(n)}, \ldots, s_n^{(n)}]$$

$$s_j^{(n)} = \begin{pmatrix} s_{1j}^{(n)} \\ \vdots \\ s_{nj}^{(n)} \end{pmatrix}$$

$s_{1j}^{(n)}$ top element - weight

$s_{nj}^{(n)}$ bottom element - approx. bound

$\theta_j^{(n)}$ *Ritz value*

$z_j^{(n)} = Q_n s_j^{(n)}$ *Ritz vector*

Accuracy of the computed Ritz values in the Lanczos method

Exact arithmetic:

$$\min_l |\lambda_l - \theta_j^{(n)}| \;\le\; \|Az_j^{(n)} - \theta_j^{(n)} z_j^{(n)}\| \;\le\; \delta_{nj}$$

Finite precision arithmetic:

$$\begin{aligned}\min_l |\lambda_l - \theta_j^{(n)}| \;&\le\; \|Az_j^{(n)} - \theta_j^{(n)} z_j^{(n)}\| \,/\, \|z_j^{(n)}\| \\ &\le\; (\,\delta_{nj} + n^{1/2}\|A\|\varepsilon_1\,) \,/\, \|z_j^{(n)}\|\end{aligned}$$

Due to the loss of orthogonality
it can happen $\|z_j^{(n)}\| \rightarrow 0$!

The quantity δ_{nj} is easy to compute with negligible additional rounding errors. Does it tell anything about FP convergence of $\theta_j^{(n)}$?

$$|\lambda_i - \theta_j^{(n)}| \;\leq\; \max\Big\{2.5\left[\delta_{nj} + n^{1/2} \parallel A \parallel \varepsilon_1\right],\\ (n+1)^3 \parallel A \parallel \varepsilon_2\Big\}$$

$$\boxed{\delta_{nj} = \beta_{n+1}|s_{nj}^{(n)}|}$$

Fascinating result!

Result of FP computation

verified at no cost!

Bounds for $|s_{nj}^{(n)}|$, δ_{nj}?

[Parlett 80], [Greenbaum, S 90],
[Wülling 03], [Zemke 03], [Meurant 03]

Loss of orthogonality among Lanczos vectors?

Paige:

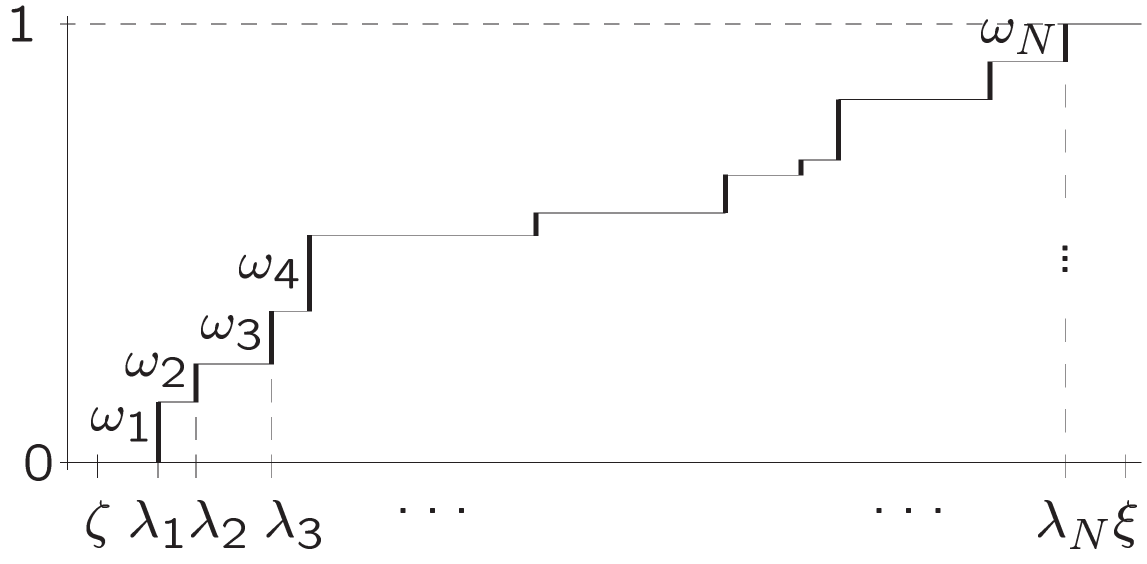
$$\left|(z_j^{(n)}, q_{n+1})\right| = \frac{\left|\varepsilon_{jj}^{(n)}\right|}{\delta_{nj}}, \qquad \left|\varepsilon_{jj}^{(n)}\right| \le n \parallel A \parallel \varepsilon_2$$

orthogonality among $z_j^{(n)}$, q_{n+1} is lost
$\Rightarrow \theta_j^{(n)}$ have converged to some λ_i

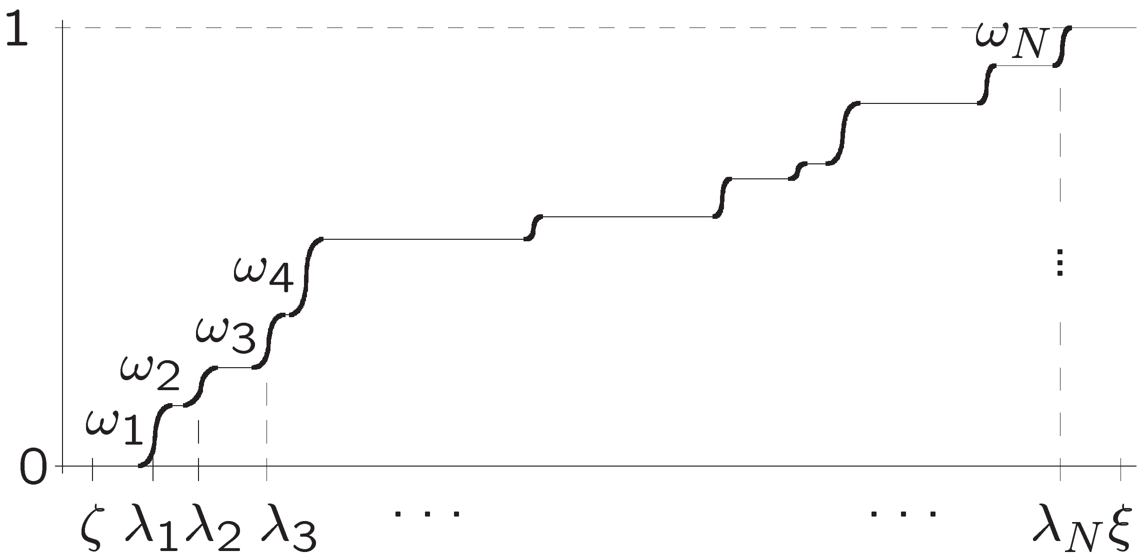
Bai: non-Hermitian case
No implication about convergence!

Multiple copies and delay of convergence?

Finite precision Lanczos (CG) computation (corresponding to the stepwise RS distribution function)



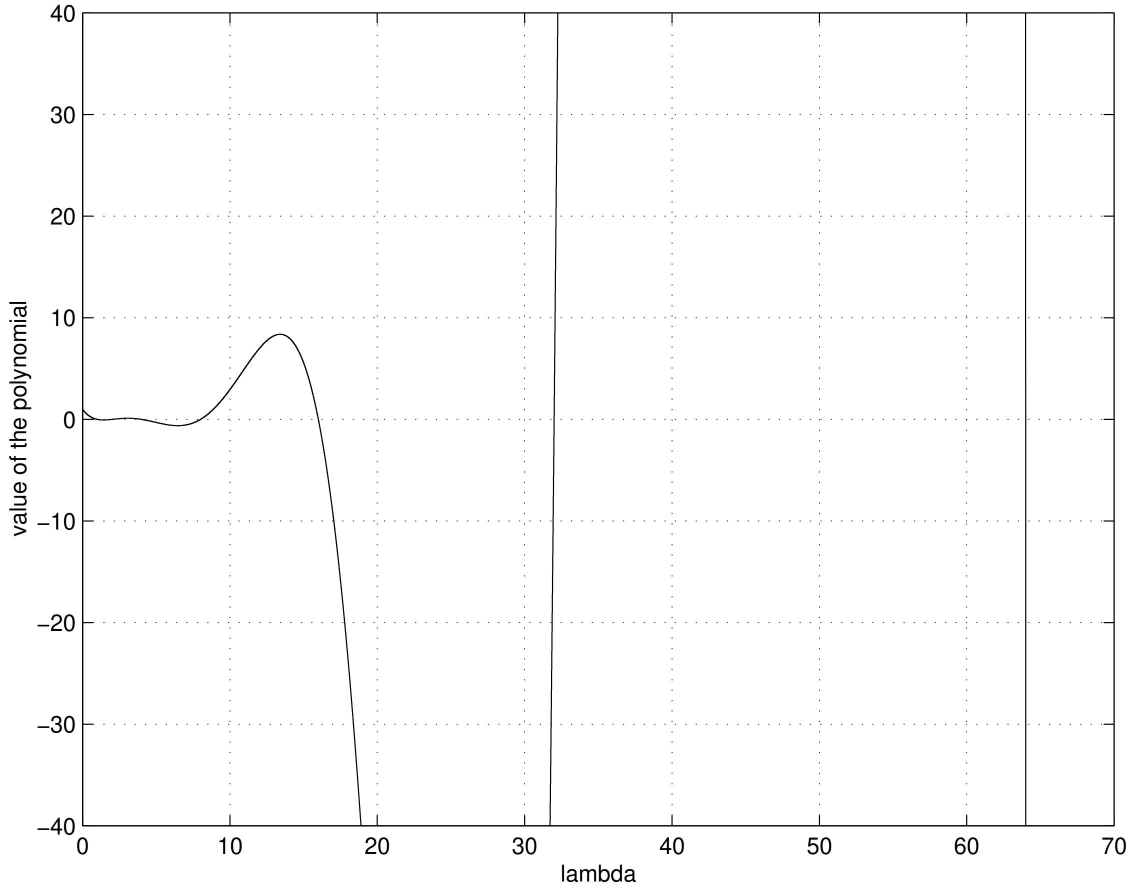
can be viewed as exact Lanczos (CG) corresponding to the blurred distribution function



Consequence

Exact precision theory applied to the blurred problem explains computing multiple copies of single dominant eigenvalues and delay of convergence in the CG method

Example



another root is required
by the minimization property

4.3 CG accuracy

Hestenes and Stiefel (1952)

Given x_0, $r_0 = b - Ax_0$, $p_0 = r_0$

For $n = 1, 2, \ldots$

$$
\begin{aligned}
& \gamma_{n-1} = (r_{n-1}, r_{n-1})/(p_{n-1}, Ap_{n-1}) \\
x_n &= x_{n-1} + \gamma_{n-1}\, p_{n-1} \\
r_n &= r_{n-1} - \gamma_{n-1}\, Ap_{n-1} \\
& \delta_n = (r_n, r_n)/(r_{n-1}, r_{n-1}) \\
p_n &= r_n + \delta_n\, p_{n-1}.
\end{aligned}
$$

A natural measure of convergence is

$$\|x - x_n\|_A$$

Idea of estimating $\|x - x_n\|_A$ at the price of d extra steps comes from [Golub, S 1994]. It was developed into a practical algorithm in [Golub, Meurant 1997]

$$\|x - x_n\|_A^2 = \text{EST}^2 + \|x - x_{n+d}\|_A^2$$

When $\|x - x_n\|_A^2 \gg \|x - x_{n+d}\|_A^2$, EST gives a tight (lower) estimate for $\|x - x_n\|_A$, with the inaccuracy bounded by $\|x - x_{n+d}\|_A$

Mathematically equivalent estimates:

[Golub, S 1994], [Golub, Meurant 1997]

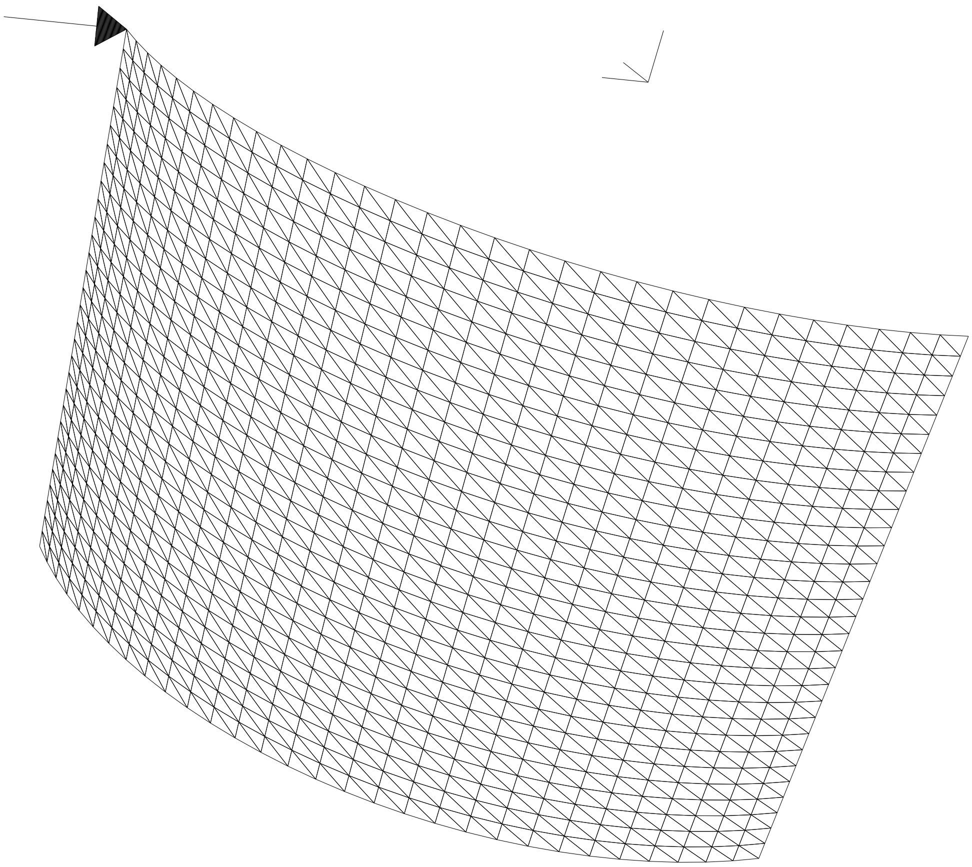
$$\eta_{n,d}^2 \;=\; \|r_0\|^2 \,[\, C_{n+d} - C_n \,]$$

[Warnick 2000]

$$\mu_{n,d}^2 \;=\; r_0^T (x_{n+d} - x_n)$$

[Hestenes, Stiefel 1952], after fifty years found and extended in [S, Tichý 2002, 03] with justification for finite precision computations

$$\nu_{n,d}^2 \;=\; \sum_{l=n}^{n+d-1} \gamma_l \, \|r_l\|^2$$

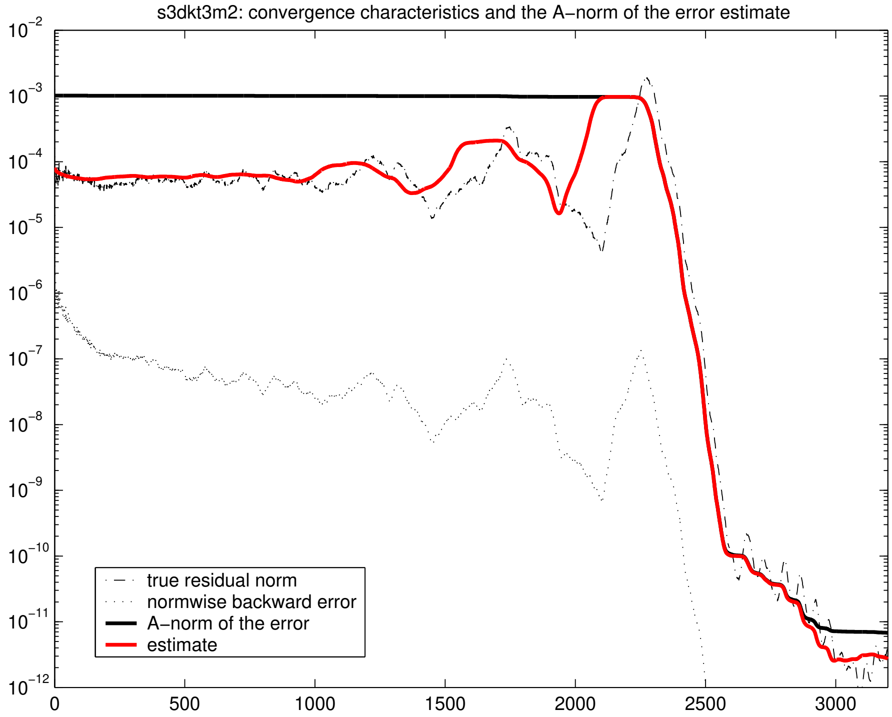


R. Kouhia, collection Cylshell,

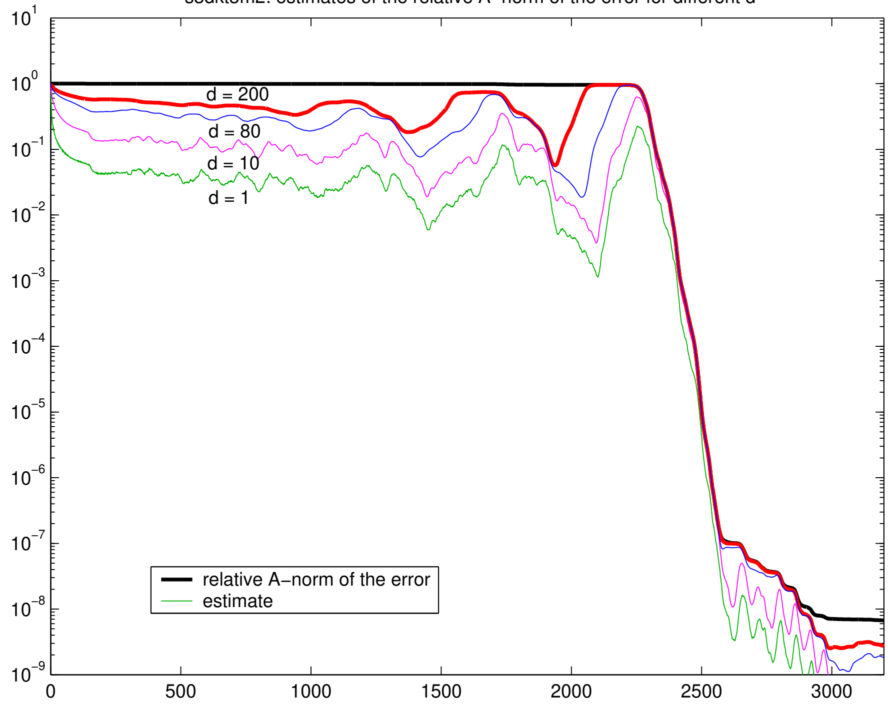
$$N = 90449, \kappa(A) = 3.62e + 11$$

Incomplete Choleski preconditioned CG

s3dkt3m2: convergence characteristics and the A−norm of the error estimate



s3dkt3m2: estimates of the relative A−norm of the error for different d



4.4 GMRES accuracy

$$\begin{aligned}\|r_n\| &= \min_{u \in x_0 + K_n(A, r_0)} \|b - Au\| \\ &= \min_{z \in AK_n(A, r_0)} \|r_0 - z\| \\ &\Leftrightarrow \; r_n \perp AK_n(A, r_0)\end{aligned}$$

MINRES [Paige, Saunders 1975]
and its **simplification**
GMRES [Saad, Schultz 1986]

(mathematically equivalent to GCR
analyzed in [Elman 1982])

MINRES $\boxed{\text{IS NOT}}$ a symmetric variant of GMRES!

Results of [Liesen, Rozložník, S 2002]:

Computing basis of a properly chosen subspace
is fundamental for numerical stability
of Krylov subspace methods

Even the best orthogonalization technique
in computing the basis (Householder or
Givens transformations)
can not compensate for instabilities
due to a bad choice of the subspace.

It can even make things worse!

Saad and Schultz described
the right implementation!

Numerical stability of the GMRES implementation based on the orthonormal basis of $K_n(A, r_0)$

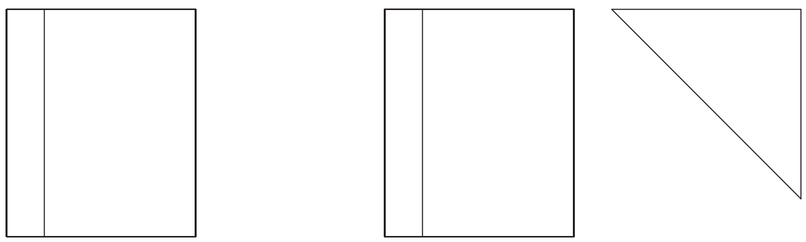
[Björck 1967], [Björck, Paige 1992]
[Karlson 1991]
[Arioli, Fassino 1996]

[Drkošová, Greenbaum, Rozložník, S 1995]
[Greenbaum, Rozložník, S 1996]
[Rozložník 1997]
[Paige, S 2002/1, 02/2, 02/3, 02/4, 02/5]
[Rozložník, Paige, S . . . in progress]

Householder - reflections based implementation of GMRES is backward stable

Arnoldi process $\equiv$ recursive columnwise
QR decomposition

$$[r_0, AV_n] \quad = \quad V_{n+1} \quad R_{n+1}$$



Implementation using Householder transformations

[Walker 1988, 89]

MGS GMRES

$$
\begin{aligned}
v_{n+1} &= (I - v_n v_n^* - \ \ldots \ - v_1 v_1^*)\, A v_n \\
&= (I - v_n v_n^*)\ \ldots\ (I - v_1 v_1^*)\, A v_n
\end{aligned}
$$

A common belief:

MGS is a good compromise between
propagation of errors (loss of orthogonality)
and
algorithm efficiency (computational cost)

Comparison with classical Gram-Schmidt,
Householder reflections, Givens rotations

However:

Despite the loss of orthogonality, some algorithms with MGS provide results as good as algorithms using the most stable orthogonalization processes (with the loss of orthogonality among the basis vectors kept
close to the machine precision level)

Theoretical justification?

- Linear least squares: [Björck, Paige 1992]
- Our case: MGS GMRES

In MGS GMRES, loss of orthogonality
is **controlled by convergence**.

Statement:

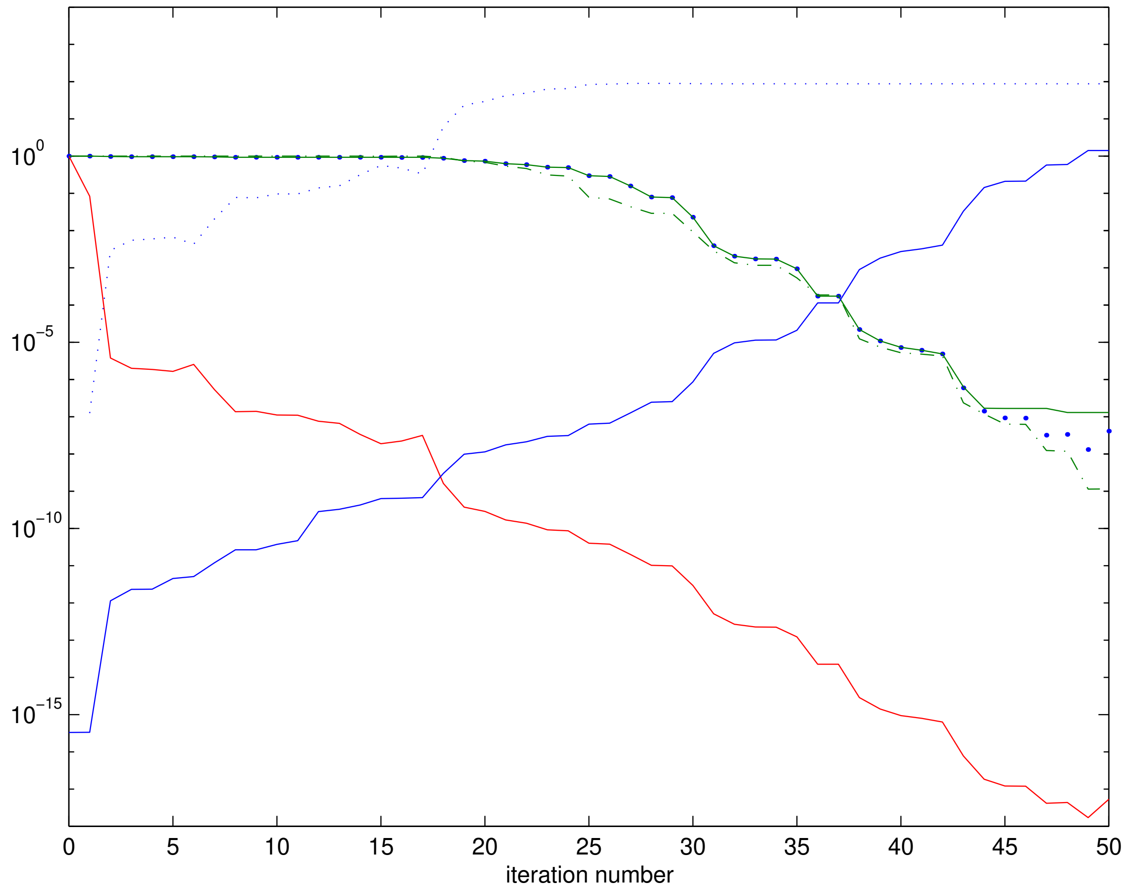
Loss of orthogonality $\|I - V_{n+1}^{*} V_{n+1}\|_F$
in the modified Gram–Schmidt
Arnoldi process

is inversely proportional

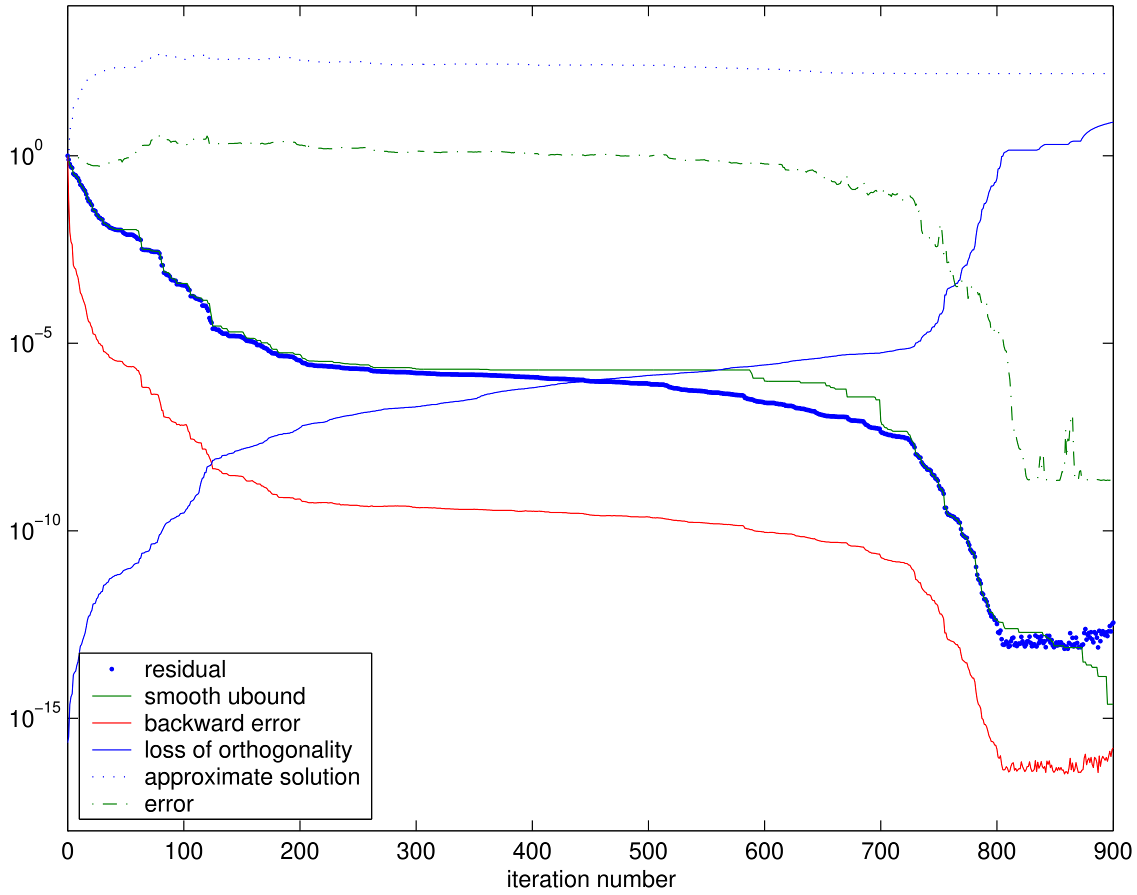
to the value
of the **GMRES backward error**

$$\frac{\|b - Ax_n\|}{\|b\| + \|A\|\,\|x_n\|}\ !$$

FS1836, $b =$ ones, $x_0 = 0$

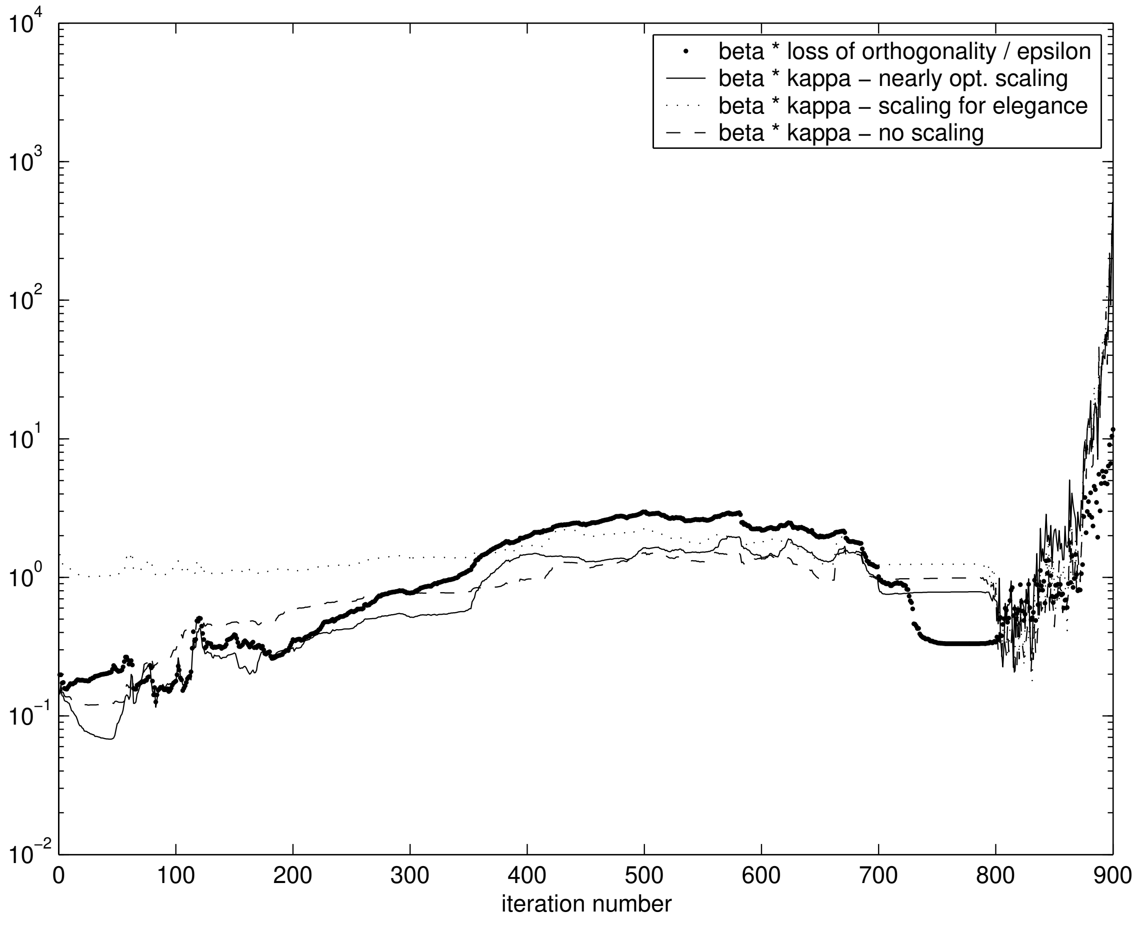


Sherman 2, b MM, $x_0 = 0$



Proof? A long story,
links with **Scaled Total Least Squares**

Sherman 2, b MM, $x_0 = 0$



5. Summary

- Nonlinear problems
  in numerical linear algebra

- Backward stability

- Intermediate quantities
  and accuracy of final results

- Theoretical results link computational cost
  and numerical stability

6. Closure

- Analysis of iterative methods:

  Convergence bounds, rounding error estimates are tools, not the goals.

- CG, Lanczos – Gauss Q of R-S integral; Scaled TLS fundamentals - GMRES:

  Unexpected, revealing links, analytic and computational (finite precision arithmetic) parts deeply connected!

- Lanczos, MGS GMRES:

  Loss of orthogonality means convergence!