

# Modern Methods of Numerical Linear Algebra – Principles, Trends and Open Problems

Zdeněk Strakoš  
Institute of Computer Science, Academy of Sciences,  
and  
Faculty of Mathematics and Physics, Charles University,  
Prague, Czech Republic

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

May 19-23, 2008

### **Basic idea of the course:**

An attempt for an advanced “essay” course. It should not be a replacement for any basic course. The goal is to promote **understanding of principles and connections**, not to derive formulas (they can be found elsewhere).

### **Prerequisites:**

Standard math, a bit of numerical math, good linear algebra.

### **Basic assumption:**

This is a **dialogue**. Please ask questions, react whenever something is unclear. Do not throw flowers (or other object), but, please, feel free to be very active.

## NLA as a simple (linear ???) tool ?

A possible stand might ask for “NUMERICAL RECIPES” i.e. clear descriptions what to do in specific situations.

Such approach works in textbook situations. Life is more complicated. No numerical recipes approach works without **achieving fundamental understanding**.

Part of our course have been used elsewhere before. The content and speed our exposition depend on all of us.

# SYLLABUS

## **I. General setting and introduction**

- I./1. Computational mathematics and numerical linear algebra
- I./2. Eigenvalues
- I./3. Linear models
- I./4. General methodology

## **II. Model reduction and SVD, bidiagonalization**

- II./1. SVD and what does it reveal
- II./2. TSVD solution of the deblurring problem
- II./3. Computation of SVD – bidiagonalization
- II./4. Application to solving  $Ax \approx b$

### **III. Principles of Krylov subspace methods**

III./1. Power method and finding the dominance

III./2. The essence of Krylov subspace methods

III./3. A symmetric positive definite example

### **IV. Gauß quadrature**

IV./1. Interpolatory quadrature on  $n$  points

IV./2. Gauß-Christoffel quadrature

IV./3. Relationship with the SPD example in III./3.

## **V. Lanczos algorithm and the conjugate gradient method**

V./1. Lanczos, CG and Gauß quadrature

V./2. Characterization of convergence

V./3. Measuring convergence

## **VI. GMRES behaviour and eigenvalues**

VI./1. A counterintuitive theory

VI./2. Convection – diffusion model problem

## **VII. Numerical behaviour of Krylov subspace methods**

VII./1. Delay of convergence

VII./2. Maximal attainable accuracy

## **VIII. Roundoff error analysis a mathematical discipline?**

VIII./1. Tedious bounds

VIII./2. Mathematical rigor and beauty

## Motto:

- Folklore (Oxford User's Guide to Mathematics, E. Zeidler (ed.), OUP, 2004, 1.14.7 The calculus of residues and the calculation of integrals, p. 549)

Mathematics is the art of avoiding calculations.

What about computational mathematics (numerical linear algebra)?

- E. Study (1862-1930, Leipzig, Marburg, Göttingen, Greifswald, Bonn, successor of Lipschitz, Einleitung in die Theorie der Invarianten lineare Transformationen auf Grund der Vektorenrechnung, Friedr. Vieweg & Sohn Akt.-Ges., Braunschweig, 1923)

Mathematics is neither the art of calculation  
nor the art of avoiding calculations.  
Mathematics, however, entails the art  
of avoiding superfluous calculations  
and conducting the necessary ones skilfully.  
In this respect one could have learned  
from the older authors.

- A. Einstein (Oxford User's Guide to Mathematics, E. Zeidler (ed.), OUP, 2004, p. 3)

Everything should be made as simple as possible,  
but not simpler.

Simple does not mean easy.

Simple does not contradict rigor.

The problem is often in interpretation.

(See, e.g., CG convergence behavior and its estimates via the Chebyshev polynomials).

# Lecture I

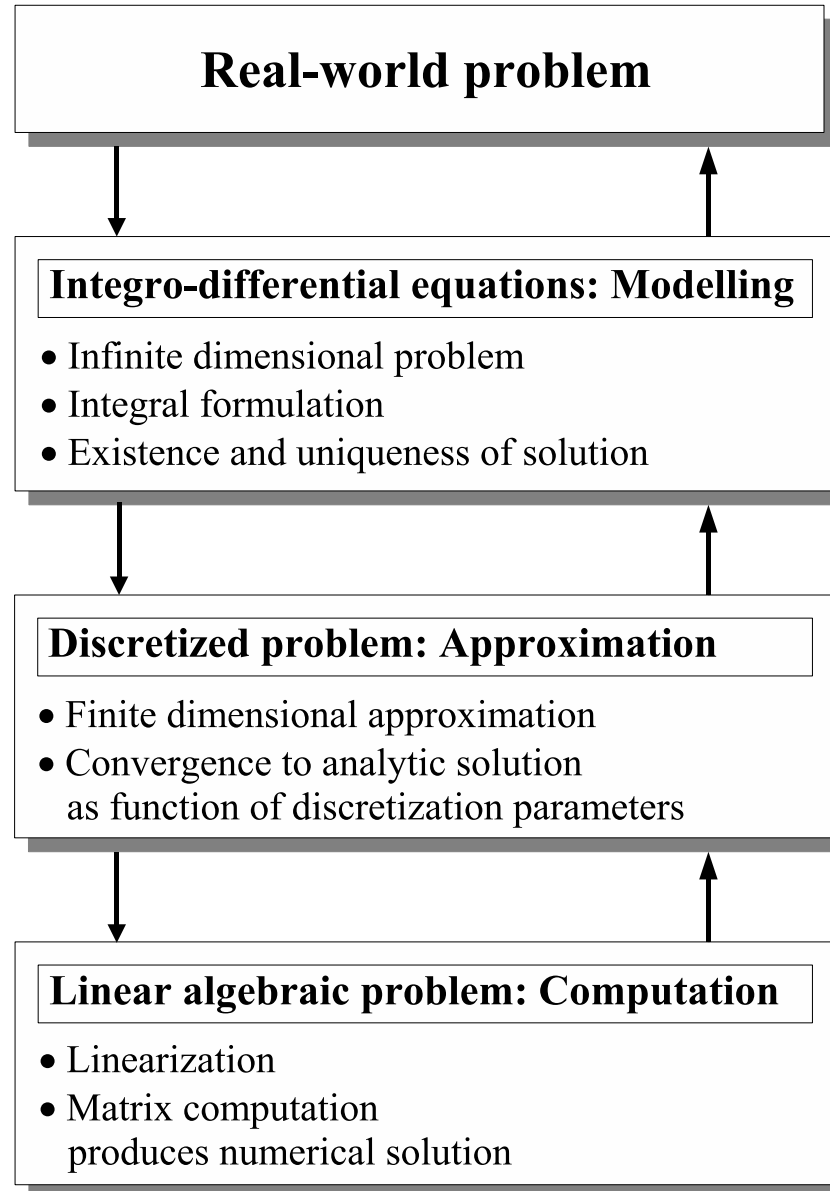
## GENERAL SETTING AND INTRODUCTION

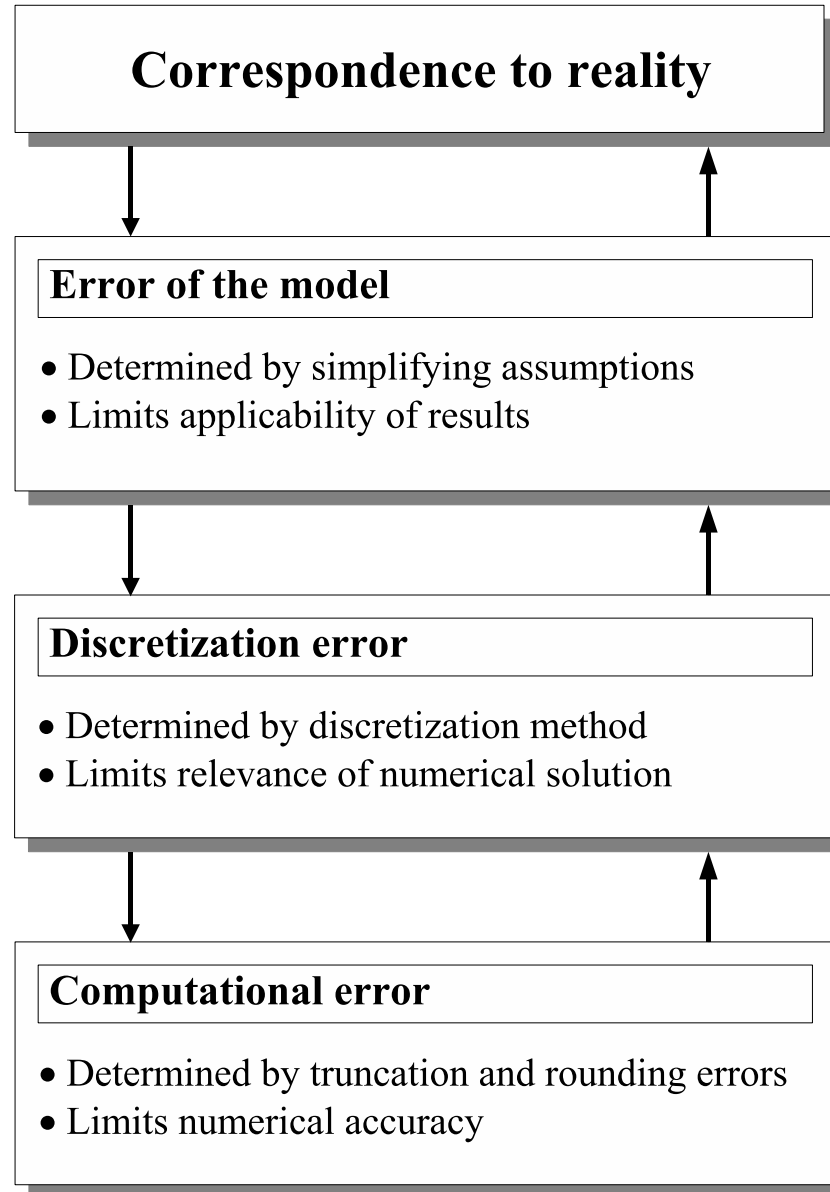
## I/1 Computational mathematics and NLA

Scientific computing, computational sciences, numerical analysis (INA at UCLA, founded in 1947 by the National Bureau of Standards), numerical linear algebra a part of these.

[Baxter, Iserles-03], [Trefethen-92]:

Given a real-world problem, the goal is to produce **reliably, robustly** and **affordably** a solution which is **within a user specified tolerance**. A quest for mathematical structure, rigor and beauty.





- **mathematical aspects:** pattern, rigor and beauty are not lost with discretization

With discretization, the mathematical problem is, if anything, more intricate. Fundamental analytical tools such as limit often need a special setting (a sequence of problems), or they can not be applied at all.

**Example:** Analysis of the “transient convergence” in Krylov subspace (Ksp) methods

- **computational aspects:**

- *errors and inaccuracy* are inherent parts of computations, see the **goals above**.

- \* can we eliminate all errors by computing exactly?

- No – in principle, see the **eigenvalue** computations, or solving **ill-posed problems** below.

- We do not wish to compute “exact solution” to the wrong problem!

- \* can we eliminate rounding errors by computing intermediate results to an arbitrary accuracy?

In principle yes, but we fail to reach the **goals** above.

There is no need for it if we understand the **mathematics** behind computation: **Numerical stability**.

↓

We can stay with finite precision arithmetic and a fixed **round off** unit. The axioms of arithmetic are, however, lost.

↓

**Mathematical** concepts such as perturbation theory, conditioning, backward error analysis and errors-in-variables, adaptivity etc.

– *computer architectures and sw tools*

principles of parallel comp. architectures (important both for control of operations and for processing the data):

\* parallel processing –  
    pipelining  
    array processing  
    multiprocessing

\* memory organization –  
    interleaving  
    distributed memory  
    hierarchical memory

\* information exchange – interconnection networks

- **practical aspects**

- *structure of the problem* (discretized PDE, control, image processing, signal processing);

- *user-specified tolerance* and accuracy of the data.

- Mathematical concepts of **complexity** (computational cost).

- Rounding error analysis – numerical stability (stability get a new content).

I. **Babuška**, SINUM 9, No 1, 1972, pp. 53–77,

“Numerical stability of problems of linear algebra” **p. 66**

“STABLE PROCESS does not significantly increase the uncertainty of the results caused by the **UNCERTAINTY OF THE INPUT DATA**”

“The notion GOOD (or BAD) is relative to the uncertainty of the results caused by uncertainty of the input data alone”

## I/2 Eigenvalues

**Definition.**  $A \in C^{N,N}$ ,  $\lambda \in C$ . If  $(A - \lambda I)$  is singular, then  $\lambda$  is called an eigenvalue of  $A$ .

$$\begin{aligned} A - \lambda I \text{ singular} &\Leftrightarrow \exists v \in C^N, Av = \lambda v, v \neq 0 \\ &\Leftrightarrow \det(A - \lambda I) = 0 \end{aligned}$$

### Computation of all eigenvalues?

A general math principle: translate the problem by a proper mapping into the form which reveals the **solution**

REMEMBER THE FOLKLORE

A proper mapping?

– **unitary** similarity transformation  $ZAZ^{-1}$ ,  $Z^{-1} = Z^*$

if  $A$  contains errors, then for an unitarity invariant norm

$$\begin{aligned} A &= A_{ex} + E, & UAU^* &= UA_{ex}U^* + UEU^* \\ \|UA_{ex}U^*\| &= \|A_{ex}\|, \\ \|UEU^*\| &= \|E\| \end{aligned}$$

The least constrained matrix form (with the smallest amount of prescribed constraints) which reveals the spectrum?

A nice example that world makes a sense:

## Schur theorem

Let  $A \in C^{N,N}$ . Then there exist a unitary matrix  $U \in C^{N,N}$  and an upper triangular matrix  $R \in C^{N,N}$  such that

$$U^*AU = R.$$

$U$  can be chosen in such a way that  $R$  contains on its diagonal the eigenvalues of  $A$  in an arbitrary prescribed order.

Proof: By induction – **nonconstructive**.

**Definition.**  $A \in C^{N,N}$  is normal if  $A^*A = AA^*$

**Theorem.**

Let  $A \in C^{N,N}$  **normal**. Then its Schur decomposition gives a diagonal matrix  $U^*AU = \text{diag}(\lambda_i)$ .

$U^*AU = D \Leftrightarrow AU = UD \Leftrightarrow U = (u_1, \dots, u_N)$  is composed of the normalized eigenvectors of  $A \Leftrightarrow$  the eigenvectors of  $A$  form an orthonormal basis of  $C^N$

$$A = UDU^* \Leftrightarrow \boxed{A = \sum_{i=1}^N u_i \lambda_i u_i^*} \quad \text{dyadic form}$$

$$\|u_i \lambda_i u_i^*\| = |\lambda_i|.$$

**Definition.**  $A$  **Hermitian** iff  $A^* = A$ .

$A$  Hermitian  $\Leftrightarrow A$  normal and  $\lambda_i \in R, i = 1, \dots, N$ .

**Definition.**  $A = C^{N,N}$  diagonalizable if there exist  $X \in C^{N,N}$  such that  $X^{-1}AX = \text{diag}(\lambda_i)$  .

**Theorem.** The set of all (diagonalizable) matrices with **distinct eigenvalues** is a dense and open subset of  $C^{N,N}$ .

**Proof:** Schur theorem, continuity of **eigenvalues**.

**Theorem** (Jordan canonical form).

For any  $A \in C^{N,N}$  there exist  $X \in C^{N,N}$  such that

$$X^{-1}AX = \text{diag} (J_{n_1}, \dots, J_{n_l}),$$

$$J_{n_k} = C^{n_k, n_k}, \quad J_{n_k} = \text{bidiag} (\lambda_k, 1) .$$

Jordan canonical form reveals the structure of **invariant subspaces** (functional analytic view, Marek, Žitný, ... ).

## Observation.

There is no (computer, numerical) algorithm which would give Schur decomposition of a general matrix  $A$  in a finite number of steps. It follows from the Abel-Galois theorem (1822, 1830) which states that polynomial equations are for  $N \geq 5$  generally unsolvable in radicals

(there is no formula using  $+$ ,  $-$ ,  $\times$ ,  $\sqrt[m]{\phantom{x}}$  giving the roots as functions of the coefficients).

## I/3 Linear models:

- the concept of solution

$Ax \approx b$ , in general  $A \in C^{N,M}$ ,  $N \gtrless M$

**Numerical PDE:**  $A \in R^{N,N}$ ,  $A$  nonsingular

More general setting: statistics, control, image proc.

|              |     |          |
|--------------|-----|----------|
| an input $x$ | ... | unknown  |
| system $A$   | ... | given    |
| output $b$   | ... | observed |

$b$  might contain errors; noise (control, image processing), measurement errors (Gauß: geodetic measurements).

$A$  might not be known accurately:

- inaccuracy in the model description (physical constants ...),
- uncertainty in choosing the model.

## Linear least squares:

closest approximation to  $b$  in  $\mathcal{R}(A)$ ,

$$Ax = b + g, \quad \|g\| \text{ min .}$$

## Total least squares (Errors-in-variables):

Find the smallest change of both  $A$  and  $b$  such that the modified problem is compatible

$$(A + E)x = b + g, \quad \|[g, E]\|_F \text{ min .}$$

- **numerical difficulty**

Consider  $A$  square, nonsingular,  $Ax = b \rightarrow x = A^{-1}b$ .

Let  $b$  is corrupted by small errors (noise). Can  $A^{-1}b$  change the proportion of the useful information to the noise?

↓

**conditioning**

$$\kappa(A) = \max_{\|z\|=1} \|Az\| / \min_{\|z\|=1} \|Az\|$$

$$= \max \text{mag}(A) / \min \text{mag}(A)$$

$$\kappa(A^{-1}) = (1 / \min \text{mag}(A)) / (1 / \max \text{mag}(A)) = \kappa(A)$$

It describes how the geometry of the space is deformed by the mapping defined by the operator  $A$ .

If the noise is magnified as  $(1/\min \text{mag}(A))$  while the useful information is mainly in the directions magnified only as  $\sim (1/\max \text{mag}(A))$ , and  $\kappa(A)$  is **large**  $\Rightarrow$  **TROUBLE**

Ill-conditioning is sometimes identified with conceptual mistakes in setting the model. In **numerical PDE** it can be related to wrong model or improper discretization. Not always, however. See [Babuška 72], some adaptive mesh refinements in the presence of singularity.

Ill conditioning **can not always be avoided!**

Example: ill posed inverse problems,

# Ill-posed problem

## Example

Numerical deblurring (Computer Tomography, Image Reconstruction and Restoration)

$$A x + \eta = b$$

$b$  ... observed data (output signal)

$A$  ... describes the system,  
(in IRR the Point Spread Function)

$\eta$  ... noise (unknown)

$x$  ... original image (input data)  
which is to be reconstructed

This problem is correctly very ill-conditioned, it can not be solved naively by computation of

$$x \approx A^{-1} b.$$

- $b - \eta$  unknown ,
- singular values of  $A$  clustered near zero; no numerical rank and no simple model reduction.

Such problem must be *regularized* – useful information must be found without **an uncontrolled amplification of the noise** present in the observed data.

## Example (J. Nagy, Emory University)

Original image (the unknown  $x$ )

$x$  = true image

**Jonathan Swift**  
**Vision is the  
art of seeing  
what is  
invisible to  
others.** 

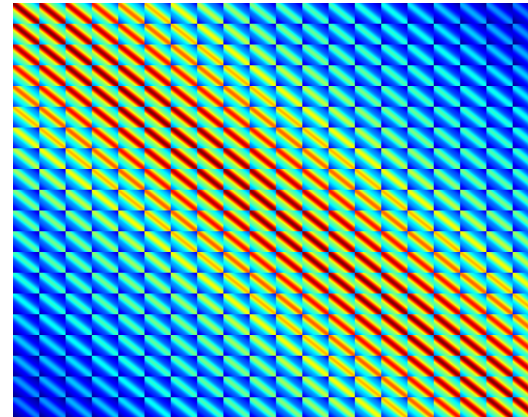
Observed image  
(the right hand side  $b$ )

$b$  = blurred, noisy image



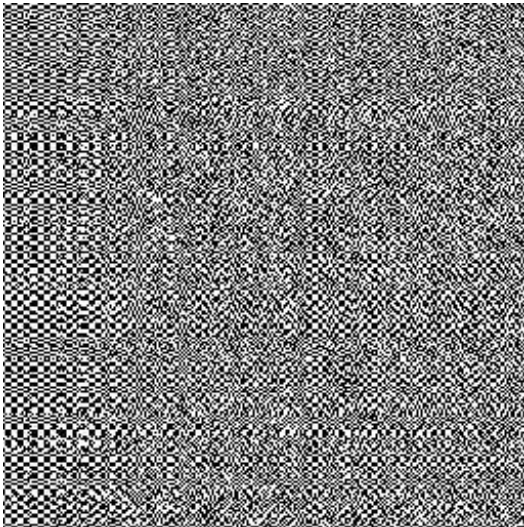
Matrix  $A$  describing  
the Point Spread Function

$A$  = matrix



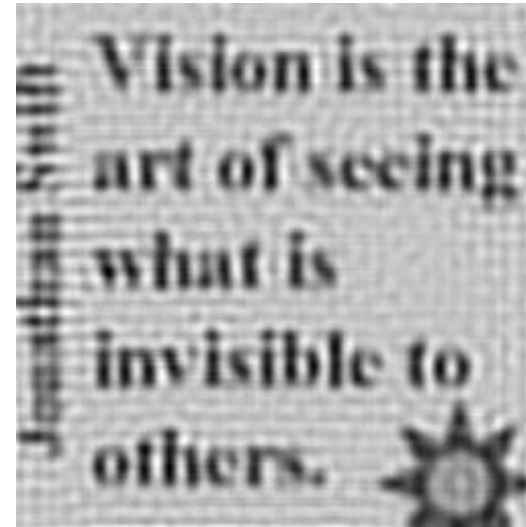
The naive exact solution  
of  $Ax = b$

x = inverse solution



Regularized solution  
via TSVD or CGLS

659 iterations



- **size and sparsity**

- Large scale computations means

Large scale **mathematical approach**, not using large scale computing environment with standard methods and algorithms (sometimes SW) which are in use for decades!

↑

This extensive way is used in sciences (physics, comp. chemistry, medical imaging) **too often!**

- Sparsity and structure adds another dimension, partial goals
  - ... matching the computer architecture,
  - ... keeping the cost under control – computational efficiency,
  - ... numerical stability,

often **in conflict!** See SPD/indefinite direct solvers.

A nice historical perspective:

D. Heller, “A Survey of Parallel Algorithms in Numerical Linear Algebra”, SIAM Review, 20, No 4, 1978, pp. 740-773.

## I/4 General methodology

- **correct definition and rigorous analysis of the problem**

Even with  $Ax \approx b$  it is **not always simple!** For example

$$\min \| [g, E] \|_F \quad \text{subject to} \quad (A + E)x = b + g$$

IS NOT in general a correctly defined problem!

- design, choice and application of **methods and algorithms**,  
computing an approximate solution

- **error analysis.** Though often ignored, it is crucial. A computation without an error analysis is like jumping out of the plane without checking whether the backpack contains a parachute. In numerical solution of PDE, algebraic errors are very rarely considered.

A key for finding a good approximate solutions in large scale **NLA**

## MODEL REDUCTION

# Lecture II

## MODEL REDUCTION AND SVD, BIDIAGONALIZATION

## II/1 SVD and what does it reveal

A symmetric, then  $A = V \text{diag} (\lambda_j) V^T$ ,  $VV^T = V^T V = I$

$$\begin{array}{ccc} & A & \\ v_1 & \xrightarrow{\lambda_1} & v_1 \\ v_2 & \xrightarrow{\lambda_2} & v_2 \\ & \vdots & \\ v_N & \xrightarrow{\lambda_N} & v_N \end{array}$$

One dimensional invariant subspaces, mutually orthogonal.

$A \in R^{N,M}$ , consider with no loss of generality  $N \geq M$ . Then

$$A = U\Sigma V^T = U_r \Sigma_r V_r^T,$$

$$UU^T = U^T U = I, \quad V^T V = V V^T = I_M, \quad \Sigma = \text{diag}(\sigma_1, \dots, \sigma_r, 0),$$

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0,$$

$$U = [U_r, \dots], \quad V = [V_r, \dots], \quad \Sigma_r = \text{diag}(\sigma_1, \dots, \sigma_r).$$

$$\begin{array}{ccccc}
& & A & & A^T \\
& & \mathcal{R}(A) & & \mathcal{R}(A^T) \\
v_1 & \xrightarrow{\sigma_1} & u_1 & \xrightarrow{\sigma_1} & v_1 \\
v_2 & \xrightarrow{\sigma_2} & u_2 & \xrightarrow{\sigma_2} & v_2 \\
\vdots & \vdots & \dots & \vdots & \vdots \\
v_r & \xrightarrow{\sigma_r} & u_r & \xrightarrow{\sigma_r} & v_r \\
\mathcal{N}(A) & \left. \begin{array}{c} v_{r+1} \\ \vdots \\ v_M \end{array} \right\} & \rightarrow 0, & \mathcal{N}(A^T) & \left. \begin{array}{c} u_{r+1} \\ \vdots \\ u_N \end{array} \right\} \rightarrow 0,
\end{array}$$

$$\boxed{A = \sum_1^r u_i \sigma_i v_i^T}, \quad \|u_i \sigma_i v_i^T\| = \sigma_i.$$

Enormously powerful, **theoretically** and **computationally**.

**Theorem.** The closest rank- $k$  approximation to  $A$ , measured by  $\|A - A_k\|$ , is given by

$$A_k = \sum_1^k u_i \sigma_i v_i^T, \quad \|A - A_k\| = \sigma_{k+1}.$$

**Proof:** the minimality needs some exercise, Watkins, Fundamentals of Matrix Comp., Thm. 7.3.6.

**Theorem.** The set of matrices of full rank is an open dense subset of  $R^{N,M}$ .

$$\|A\| = \sigma_1, \quad \kappa(A) = \sigma_1/\sigma_M \quad (\text{consider } \sigma_M > 0).$$

Distance to a nearest singular matrix (rank-deficient matrix):

$$\|A - A_{M-1}\| = \sigma_m, \quad \frac{\|A - A_{M-1}\|}{\|A\|} = \frac{\sigma_m}{\sigma_1} = 1/\kappa(A).$$

## Numerical rank and reduction of the model:

$$\underbrace{\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k}_{A_k = \sum_{i=1}^k u_i \sigma_i v_i^T} \gg \sigma_{k+1} \geq \cdots \geq \sigma_r,$$

we have a good rank- $k$  approx. to  $A$

If  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \gg \underbrace{\sigma_{k+1} \sim O(\varepsilon) \|A\|}_{\text{proportional to machine precision } \times \sigma_1},$

then  $k$  is called the **numerical rank of  $A$** .

## II/2 TSVD solution of the deblurring problem

$Ax \approx b$ ,  $b$  corrupted by noise,  $A \in R^{N,N}$ ,

$Ax + \eta = b$ ,  $\eta$  is **unknown**

denote  $b = \hat{b} + \eta$ ,  $\hat{b} = Ax$  the noise free response.

Try using SVD for solving  $A\tilde{x} = b$

$$U\Sigma V^T \tilde{x} = b \quad \Rightarrow \quad \tilde{x} = \sum_1^N v_i \frac{u_i^T b}{\sigma_i} \quad \text{is getting large.}$$

$$\tilde{x} = \underbrace{\sum_1^N v_i \frac{u_i^T \hat{b}}{\sigma_i}}_{x, \text{ but } \hat{b} \text{ unknown.}} + \sum_1^N v_i \frac{u_i^T \eta}{\sigma_i}$$

The discrete Piccard condition guarantees  $\frac{u_i^T \hat{b}}{\sigma_i} \rightarrow 0$ .

On average, the size of the components of the exact observation vector  $\hat{b}$  in the individual left singular vector subspaces of  $A$  decay faster than the corresponding singular values.

The problem is in  $\frac{u_i^T \eta}{\sigma_i}$  **for small**  $\sigma_i$ .

Truncated SVD:

$$\tilde{x} = \underbrace{\sum_1^k v_i \frac{u_i^T \hat{b}}{\sigma_i} + \sum_1^k v_i \frac{u_i^T \eta}{\sigma_i}}_{\approx x} + \underbrace{\sum_{k+1}^N v_i \frac{u_i^T \hat{b}}{\sigma_i} + \sum_{k+1}^N v_i \frac{u_i^T \eta}{\sigma_i}}_{\text{Neglect}}$$

Here  $k$  chosen so that the effect of noise is still small.

Piccard condition: the neglected information about the true solution  $x$ ,

$$\sum_{k+1}^N v_i \frac{u_i^T \hat{b}}{\sigma_i} \approx 0 ?$$

## II/3 Computation of SVD – bidiagonalization

Observation: SVD decomposition **can not** be in general computed by a **finite** algorithm. Computation must in general be iterative (singular values of  $A$  are the nonzero eigenvalues of  $A^T A$ ,  $AA^T$  is the argument complete?).

**Main point:** SVD computation is numerically reliable. For some structured matrices, **all** singular values can be determined (for a given machine precision) to the full number of **digits**.

Two steps:

- A unitary reduction of  $A$  to **lower (upper) bidiagonal** matrix,  $\tilde{U}^T A \tilde{V} = B$ ,  $B = \text{bidiag}(\beta_j, \alpha_j)$
- SVD of  $B$  via the implicit QR algorithm  
**out of the scope here.**

**Bidiagonalization:**

- Householder reflections
- (Lanczos)-Golub-Kahan iterative algorithm.

## Lanczos, Golub, Kahan bidiagonalization

[Golub, Kahan - 65]

Given  $\tilde{u}_1$ ,  $\|\tilde{u}_1\| = 1$ ,  $\beta_1 \tilde{v}_0 \equiv 0$  compute

$$\alpha_i \tilde{v}_i = A^T \tilde{u}_i - \beta_i \tilde{v}_{i-1}, \quad \|\tilde{v}_i\| = 1$$

$$\beta_{i+1} \tilde{u}_{i+1} = A \tilde{v}_i - \alpha_i \tilde{u}_i, \quad \|\tilde{u}_{i+1}\| = 1$$

stop whenever  $\alpha_i = 0$  or  $\beta_{i+1} = 0$ .

Suppose  $\alpha_i, \beta_{i+1}$  nonzero for  $i = 1, \dots, k$ ,

$$\tilde{U}_k = [\tilde{u}_1, \dots, \tilde{u}_k], \quad \tilde{V}_k = [\tilde{v}_1, \dots, \tilde{v}_k], \quad L_k = \ell - \text{bidiag}(\beta_i, \alpha_i) \in R^{k,k}.$$

Then

$$\begin{aligned} A^T \tilde{U}_k &= \tilde{V}_k L_k^T, & A \tilde{V}_k &= \tilde{U}_k L_k + \beta_{k+1} u_{k+1} e_k^T, \\ \tilde{U}_k^T A \tilde{V}_k &= \tilde{U}_k^T \tilde{U}_k L_k + \beta_{k+1} \tilde{U}_k^T u_{k+1} e_k^T = L_k \tilde{V}_k^T \tilde{V}_k. \end{aligned}$$

Let  $\alpha_{k+1} \neq 0$ , i.e.  $v_{k+1}$  can also be computed. Then, appending one more column to  $\tilde{U}_k, \tilde{V}_k$  we get  $\tilde{U}_{k+1}, \tilde{V}_{k+1}$ , and appending a row  $[0, \beta_{k+1}]$  to  $L_k$ , giving  $L_{k+1}$ ,

$$A^T \tilde{U}_{k+1} = \tilde{V} L_{k+1}^T + \alpha_{k+1} \tilde{v}_{k+1} e_{k+1}^T, \quad A \tilde{V}_k = \tilde{U}_{k+1} L_{k+1}$$

$$\tilde{V}_k^T A^T \tilde{U}_{k+1} = \tilde{V}_k^T \tilde{V}_k L_{k+1}^T + \alpha_{k+1} \tilde{V}_k^T \tilde{v}_{k+1} e_{k+1} = L_{k+1}^T \tilde{U}_{k+1}^T \tilde{U}_{k+1}.$$

**Now magic:**

By induction:  $U_k^T U_k = V_k^T V_k = I$ , then

$$\tilde{U}_k^T \tilde{u}_{k+1} = 0 \quad \tilde{V}_k^T \tilde{v}_{k+1} = 0$$

i.e. the algorithm generates **orthonormal vectors**

$$\{\tilde{u}_1, \dots, \tilde{u}_{k+1}\}, \quad \{\tilde{v}_1, \dots, \tilde{v}_{k+1}\}.$$

## II/4 Application to solving $Ax \approx b$

Assume

$$[b, A] = \left[ \begin{array}{c|c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array} \right].$$

Then

$$\begin{aligned} Ax \approx b &\Leftrightarrow A_{11} x_1 \approx b_1 \\ &A_{22} x_2 \approx 0 \end{aligned}$$

with the meaningful solution  $x_2 = 0$ ,  $x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$ .

## Core problem within $\tilde{A}\tilde{x} \approx \tilde{b}$

Our suggestion is to find an orthogonal transformation

$$Ax \equiv (P^T \tilde{A} Q) (Q^T \tilde{x}) \approx P^T \tilde{b} \equiv b ,$$

$$P^T [\tilde{b} , \tilde{A} Q] = \left[ \begin{array}{c|c|c} b_1 & A_{11} & 0 \\ \hline 0 & 0 & A_{22} \end{array} \right] , \quad P^{-1} = P^T , \quad Q^{-1} = Q^T$$

so that  $A_{11}$  has minimal dimensions. Then solve  $A_{11}x_1 \approx b_1$ , and take the original problem solution to be

$$\tilde{x} = Q \begin{bmatrix} x_1 \\ 0 \end{bmatrix} .$$

Such an orthogonal transformation is given (assuming  $\tilde{A}^T \tilde{b} \neq 0$ ) by reducing  $[\tilde{b}, \tilde{A}]$  to an upper bidiagonal matrix. In fact,  $A_{22}$  need not be bidiagonalized,  $[b_1, A_{11}] = P_1^T [\tilde{b}, \tilde{A} Q_1]$  has nonzero bidiagonal elements and is either

$$[b_1 | A_{11}] = \left[ \begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \cdot & \cdot & \\ & & & \beta_p & \alpha_p \end{array} \right], \quad \beta_i \alpha_i \neq 0, \quad i = 1, \dots, p$$

if  $\beta_{p+1} = 0$  or  $p = N$ , or

$$[b_1 \mid A_{11}] = \left[ \begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \cdot & \cdot & \\ & & & \beta_p & \alpha_p \\ & & & & \beta_{p+1} \end{array} \right], \quad \beta_i \alpha_i \neq 0, \beta_{p+1} \neq 0$$

if  $\alpha_{p+1} = 0$  or  $p = M < N$ .

$[b_1, A_{11}]$  has full row rank and  $A_{11}$  has full column rank.

The first case happens if and only if the original problem is compatible, the second case if and only if the original problem is incompatible.

## Theorem

- (a)  $A_{11}$  has no zero or multiple singular values, so any zero singular values or repeats that  $\tilde{A}$  has must appear in  $A_{22}$ ;
- (b)  $A_{11}$  has minimal dimensions, and  $A_{22}$  maximal dimensions, over all orthogonal transformations of the form given above;
- (c) All components of  $b_1$  in the left singular vector subspaces of  $A_{11}$  are nonzero.

The core problem approach consists of three steps:

1. Orthogonal transformation  $[b, A] = P^T [\tilde{b}, \tilde{A}Q]$ , where the upper bidiagonal block  $[b_1, A_{11}]$  is as above, and  $A_{22}$  is not bidiagonalized. All irrelevant and multiple information is filtered out to  $A_{22}$ .
2. Solving the minimally dimensioned  $A_{11}x_1 \approx b_1$ .
3. Setting  $\tilde{x} = Qx \equiv Q \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$ .

The core problem approach does not need to complete the SVD of all of  $[\tilde{b}, \tilde{A}]$ . When the bidiagonalization stops, we use only the **necessary** (and sufficient) information for computing the solution.

For orthogonally invariant approximation problems, the formulations for the original data  $[\tilde{b}, \tilde{A}]$  and the orthogonally transformed data  $[b, A]$  are equivalent. Consequently, the core problem approach always gives meaningful solutions by setting  $x_2 = 0$  (neglecting the  $A_{22}$  block).

## Approximating TSVD

Stopping the bidiagonalization when TSVD can be well approximated from the *computed part* of  $A_{11}$ : it can approximate soon the large singular values, more slowly the smallest singular values

→

regularization effect of **Lanczos – TSVD**

[Fierro, Golub, Hansen, O’Leary – 97]

[Hansen, Kilmer, Jensen – 07]

[**Paige, Strakoš – 05**],

[**Hnnětynková, Strakoš – 07**], ...

## Filtration factors

$$x^r = \sum_1^N v_i \frac{u_i^T b}{\sigma_i} \phi_i$$

Do we solve linear algebraic systems arising from PDE **in this** way?

Yes, we do! **A key is**

**APPROXIMATING DOMINANT INFORMATION FAST**

# Lecture III

## PRINCIPLES OF KRYLOV

## SUBSPACE METHODS

User wants to know, which method should be applied under given circumstances. No simple template answer. Modern iterative methods **are complicated**.

Consider  $A \in R^{N,N}$ ,  $b \in R^N$ ;

We restrict our exposition to  $A$  **nonsingular**.

### III/1 Power method for finding an eigenvector corresponding to the dominant eigenvalue

$$v_1; (Av_i/\|Av_i\|) = v_{i+1}, \quad i = 1, 2, \dots$$

If  $|\mu_1| > |\mu_j|, j \neq 1, \{\mu_1, \dots, \mu_N\}$  eigenvalues of  $A$  with  $\{w_1, \dots, w_N\}$  the corresponding eigenvectors or Jordan canonical vectors, and if  $v_1$  has nonzero component in the direction of  $w_1$ ,

then  $v_i$  converges to  $w_1$  with the *asymptotic rate of convergence*

$$\max_{j \neq 1} |\mu_j|/|\mu_1|.$$

Local procedure ( $i \rightarrow i + 1$ ), it aims at a single dominant information.

Krylov subspace methods represent **global procedures** which at each iteration step **remember the history** and **anticipate the future**.

Though we use them as **iterative** methods, they **are finite** (when we consider  $Ax = b$ ). Unlike the Power method and unlike the stationary iterative methods, they work with the **global** information, see **later**.

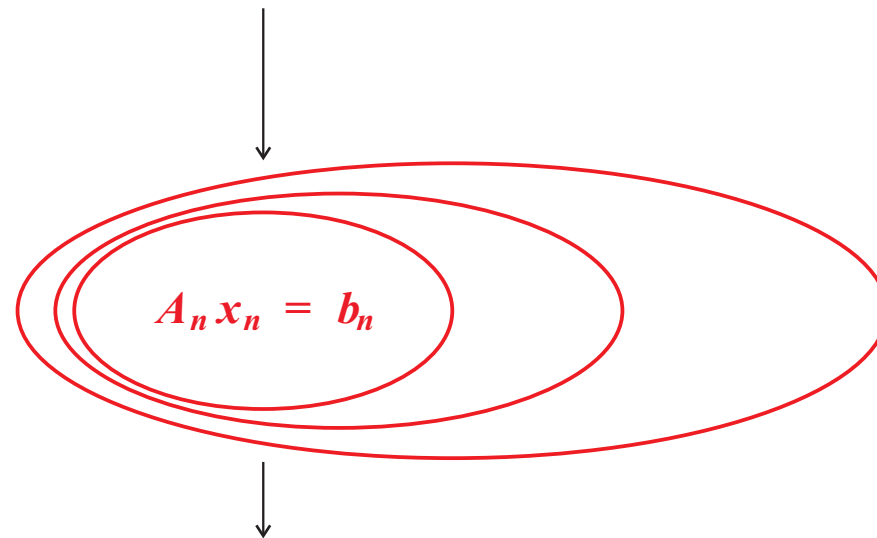
## III/2 The essence of Krylov subspace methods

### The essence of KSM

Projections of the  $N$ -dimensional problem onto nested Krylov subspaces of increasing dimension.

Step  $n$ : **Model reduction** from dimension  $N$   
to dimension  $n$ ,  $n \ll N$ .

$$Ax = b$$



$x_n$  approximates the solution  $x$   
using the subspace of small dimension.

## Projection processes

$$x_n \in x_0 + \mathcal{S}_n, \quad r_0 \equiv b - Ax_0$$

where the constraints needed to determine  $x_n$  are given by

$$r_n \equiv b - Ax_n \in r_0 + A\mathcal{S}_n, \quad r_n \perp \mathcal{C}_n.$$

Here  $\mathcal{S}_n$  is called the **search space**,  $\mathcal{C}_n$  is called the **constraint space**.

$r_0$  decomposed to  $r_n$  + the part in  $A\mathcal{S}_n$ . It should be called **orthogonal** projection if  $\mathcal{C}_n = A\mathcal{S}_n$ , **oblique** otherwise.

## Krylov subspace methods:

$$\mathcal{S}_n \equiv \mathcal{K}_n \equiv \mathcal{K}_n(A, r_0) \equiv \text{span} \{r_0, \dots, A^{n-1}r_0\}.$$

Algebraic and polynomial formulation:

$$\begin{aligned}x_n &\in x_0 + \mathcal{K}_n(A, r_0), \\x - x_n &= p_n(A)(x - x_0),\end{aligned}$$

$$\begin{aligned}r_n &\equiv b - Ax_n = p_n(A)r_0 \\&\in r_0 + A\mathcal{K}_n(A, r_0), \quad p_n(0) = 1.\end{aligned}$$

More general setting possible.

Krylov subspaces tend to contain the **dominant information** of  $A$  with respect to  $r_0$ . Unlike in the power method for computing the dominant eigenspace, here all the information accumulated along the way is used [Parlett - 80, Example 12.1.1].

Discretization means approximation of a continuous problem by a finite dimensional one. Computation using Krylov subspace methods means nothing but further model reduction. Well-tuned combination has a chance for being efficient.

The idea of Krylov subspaces is in a fundamental way linked with the **problem of moments**.

## Riemann-Stieltjes integral:

$\omega(\lambda)$  a nondecreasing function on  $[a, b]$ ,

$$\int_a^b f(\lambda) d\omega(\lambda)$$

is defined as the Riemann integral with the size of the elementary intervals  $[t_i, t_{i+1}]$  given by

$$\omega(t_{i+1}) - \omega(t_i).$$

In **Stieltjes'** formulation, a sequence of numbers  $\xi_k$ ,  $k = 0, 1, \dots$ , is given and a non-decreasing distribution function  $\omega(\lambda)$ ,  $\lambda \geq 0$ , is sought such that the Riemann-Stieltjes integrals satisfy

$$\int_0^{\infty} \lambda^k d\omega(\lambda) = \xi_k, \quad k = 0, 1, \dots .$$

Here  $\int_0^{\infty} \lambda^k d\omega(\lambda)$  represents the  $k$ -th moment of the distribution function  $\omega(\lambda)$ .

[Shohat, Tamarkin - 43], [Akhiezer - 65], [Karlin, Shapley - 53]

## Vector moment problem of Vorobyev:

Find a linear operator  $A_n$  on  $\mathcal{K}_n$  such that

$$\begin{aligned}A_n r_0 &= Ar_0, \\A_n (Ar_0) &= A^2 r_0, \\&\vdots \\A_n (A^{n-2} r_0) &= A^{n-1} r_0, \\A_n (A^{n-1} r_0) &= Q_n (A^n r_0),\end{aligned}$$

where  $Q_n$  projects onto  $\mathcal{K}_n$  orthogonally to  $\mathcal{C}_n$ .

[Vorobyev - 65], [Brezinski - 97], [S - 08]

Please notice that here  $A^n r_0$  is decomposed into the part  $Q_n(A^n r_0) \in \mathcal{K}_n$  and a part orthogonal to  $\mathcal{C}_n$ .

Therefore  $Q_n$  is the **orthogonal** projector if  $\mathcal{C}_n = \mathcal{K}_n$ , **oblique** otherwise.

Some important comments:

## Direct and iterative

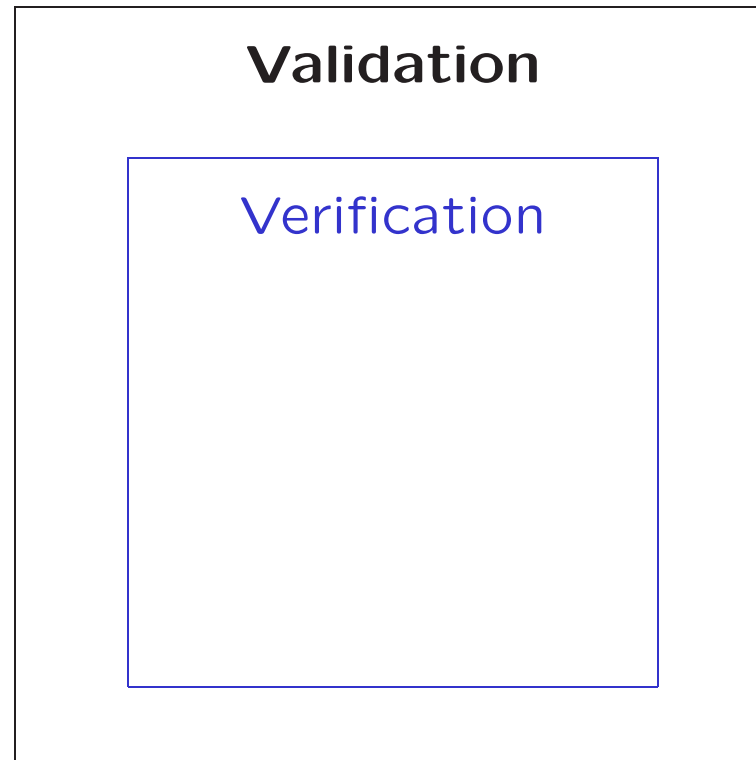
- Some practitioners: when **enough computer resources** are available, then direct methods should be preferred to iterative. They compute **accurately**. Do they?
- If our goal is to improve methodology for solving given problem, then the questions about having or not having enough computer resources do not make sense.
- The statement should be read: “We wish to focus on a particular step and not to be disturbed by possible problems of the other steps.”

- “and” means combination for eliminating the disadvantages and strengthening the advantages.
- Principal advantage of the iterative part is the possibility of stopping the computation at the **desired accuracy level**.
- It requires a meaningful stopping criterion. The errors of the **model, discretization** error and the **computational** error should be of the same order.
- Due to difficulties with the previous point this (potential) principal advantage is often presented as a disadvantage (**a need for a stopping criteria ...**).

The point was presented by the founding fathers, it is well understood in the context of multilevel methods. But it is not accepted in practical computational mathematics in general. See the following quote from a negative referee report:

“... the author give a misguided argument. The main advantage of iterative methods over direct methods does not primarily lie in the fact that the iteration can be stopped early [whatever this means], but that their memory (mostly) and computational requirements are moderate.”

It is sad to see that - how clearly things were explained, e.g. in the introduction to the [Hestenes, Stiefel - 52]!



[Babuška - 03], [Oden et al - 03]

... a linear problem?

Methods and phenomena in NLA need not be linear!

How fast we get an acceptable approximate solution?

- In modern iterative methods we have to study transient phase (early stage of computation) which represents a nonlinear phenomenon in a **finite dimensional space** of small dimensionality. [Pták, Trefethen, Baxter and Iserles, ... ]
- Operator approach can not describe the transient phase! Linearization by asymptotic tools is (even with adaptation) of limited use.

## Another look

An information-based argument:  $A^{-1}$  uses global information from  $A$ . Consequently, good iterative solution requires that:

- Either the global information is taken care for by a good preconditioner. An extremely good preconditioner, transforming the system matrix almost to identity, reduces the number of iterations to  $\mathcal{O}(1)$
- Or the global information is gathered together by many (close to  $N$ ) matrix-vector multiplications.

What is wrong?

Solving  $Ax = b$  for some particular **meaningful**  $b$  can be different from solving the system with the matrix  $A$  and some worst-case right hand side! The data have typically some meaning and are correlated.

For an acceptable accuracy we may not need the full global communication.

Operator approach (in analytical considerations working with an approximation of  $A^{-1}$ ) leads to asymptotics. Solving  $Ax = b$  for the particular meaningful  $\{A, b\}$  is different from computing  $A^{-1}$ .

[Beckermann, Kuijlaars – 02], [Kuijlaars – 06]

## Principal question:

When does  $K_n(A, r_0)$  contain enough information about the original  $N$ -dimensional linear algebraic problem in order to provide a good approximate solution  $x_n$ ?



**Convergence (better Behavior).**

## Orthogonality in generating Krylov subspaces

- Refining the multidimensional information (unlike in the power method), and computing projections from  $b, Ab, \dots, A^{n-1}b$  ?



Mutual orthogonalization, or orthogonalization against some auxiliary vectors.

Goal: getting in an affordable way a good basis ("good" does not necessarily mean "the best possible")

- Practical computations means **limited accuracy** and must be justified by numerical stability analysis.

## Consistency of the whole solution process

Elliptic PDE can serve as a nice example. The weak formulation leads to a SPD bilinear form, with the **energy** as the quantity in charge. The Galerkin FEM discretized problem is again SPD, and, consequently, an algebraic iterative method consistent with the whole solution process should minimize the **energy norm of the error** of the finite-dimensional approximate solution at each iteration step. The world makes a sense - the **conjugate gradient** method represents such consistency.

For some pioneering work on "cascadic CG" see [Deuffhard – 93 (94)]. Recently, M. Arioli and his co-workers, see [Arioli et al. – 04]. Algebraic level – [Hestenes and Stiefel - 52], [S, Tichý - 02], see below.

### III/3 A symmetric positive definite example

$$Ax = b, \quad r_0 = b - Ax_0$$

$q_1 = r_0/\|r_0\|$ ,  $\{q_1, q_2, \dots, q_n\}$  an orthonormal basis of  $K_n(A, r_0)$

Using Gram-Schmidt orthogonalization:

$$\begin{aligned}AQ_n &= Q_n H_{n,n} + \beta_{n+1} q_{n+1} e_1^T \\ Q_n^T A Q_n &= H_{n,n}\end{aligned}$$

If  $A$  is symmetric, then  $H_{n,n}$  must be symmetric tridiagonal.

For  $A$  symmetric, the **Lanczos** orthonormal basis is computed via the 3-term recurrence

$$\beta_{n+1}q_{n+1} = Aq_n - \alpha_n q_n - \beta_n q_{n-1}$$

Where

$$\beta_1 q_0 = 0, \quad \alpha_n = (Aq_n - \beta_n q_{n-1}, q_n), \quad \|q_{n+1}\| = 1.$$

**Matrix form** (we will see it in detail below):

$$AQ_n = Q_n T_n + \beta_{n+1} q_{n+1} e_n^T, \quad T_n = \text{tridiag}(\beta_j, \alpha_j, \beta_{j+1}) \in R^{n,n}.$$

In terms of polynomials

(here  $A$  SPD)

$$q_{n+1} = \psi_n(A)q_1 / (\beta_2 \dots \beta_{n+1}), \quad \psi_n \text{ monic}$$

$$(\psi_i(A)q_1, \psi_j(A)q_1) = \delta_{ij}$$

With the spectral decomposition  $A = U \text{diag} (\lambda_l) U^T$

$$\psi_i(A)q_1 = \sum_{l=1}^N (q_1, u_l) \psi_i(\lambda_l) u_l$$

$$\delta_{ij} = (q_{i+1}, q_{j+1}) = \sum_{l=1}^N (q_1, u_l)^2 \psi_i(\lambda_l) \psi_j(\lambda_l)$$

$\{1, \psi_1, \dots, \psi_n\}$  monic orthogonal polynomials with respect to

$$(\varphi, \psi) = \sum_{l=1}^N \omega_l \varphi(\lambda_l) \psi(\lambda_l), \quad \omega_l = (q_1, u_l)^2$$

**Polynomials orthogonal** with respect to the R-S integral with the distribution function  $\omega(\lambda)$  with the finite points of increase  $\{\lambda_1, \dots, \lambda_N\}$  and weights  $\{\omega_1, \dots, \omega_N\}$ .

What does it offer?     **See next.**

# Lecture IV

# GAUß QUADRATURE

## IV/1 Interpolatory quadrature formula on $n$ points

Consider a nondecreasing function  $\omega(\lambda)$  on  $[\zeta, \xi]$ . Let  $\mu_1, \mu_2, \dots, \mu_n$  be  $n$  distinct points in  $[\zeta, \xi]$ . Then there exist weights  $\tilde{\omega}_1, \dots, \tilde{\omega}_n$  such that the quadratic formula

$$\int_{\zeta}^{\xi} f(\lambda) d\omega(\lambda) = \sum_{l=1}^n \tilde{\omega}_l^{(n)} f(\mu_l)$$

is exact for any polynomial of degree at most  $n - 1$ .

Indeed, take  $\mathcal{L}_f(\mu_1, \dots, \mu_n)$  the Lagrange interpolating polynomial on the points  $\mu_1, \dots, \mu_n$ :

$$\mathcal{L}_f(\mu_1, \dots, \mu_n) = \sum_{l=1}^n f(\mu_l) \frac{\prod_{m \neq l} (\lambda - \mu_m)}{\prod_{m \neq l} (\mu_l - \mu_m)} \equiv \sum_{l=1}^n f(\mu_l) \phi_l(\lambda)$$

Then

$$\tilde{\omega}_l^{(n)} = \int_{\zeta}^{\xi} \phi_l(\lambda) d\omega(\lambda)$$

and since every polynomial of degree  $\leq (n - 1)$  is identical with its Lagrange interpolating polynomial, we get the result.

## IV/2 Gauß-Christoffel Quadrature

Interpolatory quadrature formula on the  $n$  points  $\mu_1, \dots, \mu_n$ , weights

$$\omega_l^{(n)} = \int_{\zeta}^{\xi} \phi_l(\lambda) d\omega(\lambda)$$

Add a new node  $\eta_1$  (different from  $\mu_1, \dots, \mu_n$ ).

Observation: its weight proportional to

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

$$\psi_n(\lambda) = (\lambda - \mu_1)(\lambda - \mu_2) \dots (\lambda - \mu_n).$$

Add another new node  $\eta_2$  (different from  $\mu_1, \dots, \mu_n, \eta_1$ ),  
its weight proportional to

$$\int_{\zeta}^{\xi} \psi_n(\lambda)(\lambda - \eta_1) d\omega(\lambda);$$

Add  $2n$ th new node  $\eta_n$  (different from all previous nodes)  
its weight proportional to

$$\int_{\zeta}^{\xi} \psi_n(\lambda)(\lambda - \eta_1) \dots (\lambda - \eta_{n-1}) d\omega(\lambda)$$

The resulting quadrature is of degree  $2n - 1$  ( $2n$  nodes).

## The ingenious idea of Gauß :

Take  $\mu_1, \dots, \mu_n$  the roots of the monic orthogonal polynomial  $\psi_n(\lambda)$  defined by the inner product

$$(\varphi, \psi) = \int_{\zeta}^{\xi} \varphi(\lambda)\psi(\lambda) d\omega(\lambda),$$

THEN ALL WEIGHTS CORRESPONDING TO THE ADDITIONAL  $n$  NODES  $\eta_1, \dots, \eta_n$  are 0 , WE EVEN DO NOT NEED TO KNOW  $\eta_1, \dots, \eta_n!$

How does it show up in **SPD** Krylov subspace methods for solving  $Ax = b$ ?

# Lecture V

## LANCZOS ALGORITHM AND THE CONJUGATE GRADIENT METHOD

## V/1 Lanczos, CG and Gauß quadrature

**Lanczos** orthonormal basis of  $K_n(A, r_0)$  is generated by the three-term recurrence

$$AQ_n = Q_n T_n + \beta_{n+1} q_{n+1} e_n^T, \quad Q_n = [q_1, \dots, q_n].$$

The diagram illustrates the matrix equation  $AQ_n = Q_n T_n + O$ . It shows a square matrix  $A$  on the left, followed by a tall rectangular matrix  $Q_n$ , an equals sign, another tall rectangular matrix  $Q_n$ , a square matrix  $T_n$ , a plus sign, and a tall rectangular matrix  $O$ . The matrix  $O$  has a shaded right side, indicating it is a matrix with a specific structure or property.

$$T_n = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & & & \\ & \dots & \dots & & \\ & & & \dots & \\ & & & & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix} \quad \text{Jacobi matrix}$$

$$T_n = S_n \Theta_n S_n^*,$$

$$\Theta_n = \text{diag} (\theta_1^{(n)}, \dots, \theta_n^{(n)}),$$

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

## Relationship with orthogonal polynomials

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

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

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

From orthogonality

$$\psi_n : \|\psi_n(A) q_1\|^2 = \min_{\psi \in \mathcal{M}_n} \|\psi(A) q_1\|^2,$$

↓

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

**Riemann-Stieltjes integral determined by  $A, q_1$**

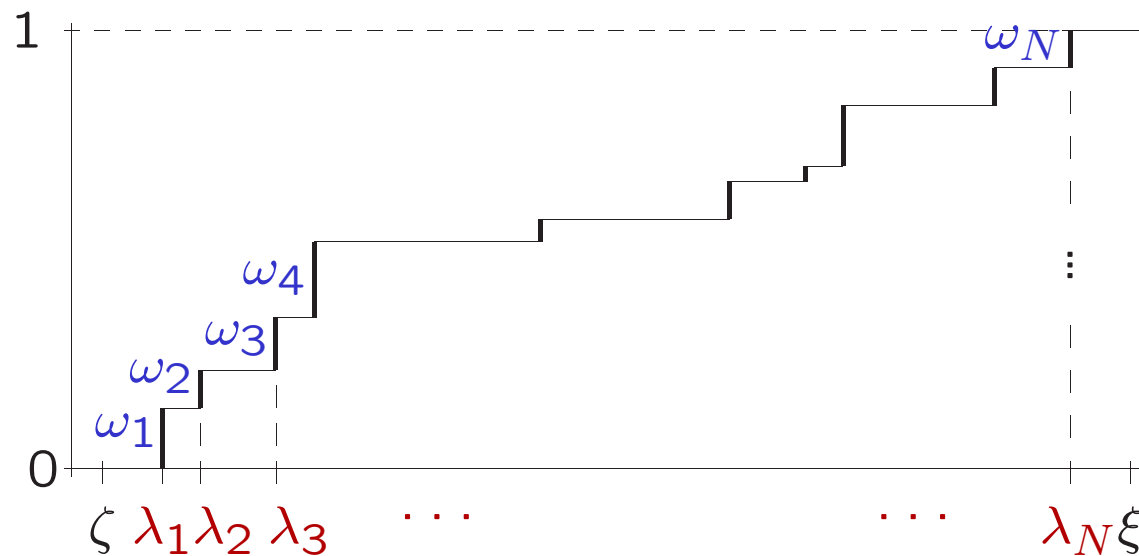
$$\sum_1^N \omega_i f(\lambda_i) = \int_{\zeta}^{\xi} f(\lambda) d\omega(\lambda),$$

$$\omega(\lambda) = 0 \quad \zeta \leq \lambda < \lambda_1,$$

$$\omega(\lambda) = \sum_{j=1}^l \omega_j \quad \lambda_l \leq \lambda < \lambda_{l+1},$$

$$\omega(\lambda) = \sum_{j=1}^N \omega_j \quad \lambda_N \leq \lambda \leq \xi.$$

Piecewise constant distribution function  $\omega(\lambda)$  with the finite number of points of increase, recall the [spectral decomposition](#) of the corresponding operator,



## Gauß quadrature interpretation?

Concentrate for a second on the matrix  $T_n$  computed at the step  $n$  of the Lanczos algorithm.

Given the matrix  $T_n$ , consider the Lanczos algorithm applied to  $T_n$  with the starting vector  $e_1$ . Then we get the Lanczos recurrence for  $n$ -dimensional vectors, but with exactly the same recurrence coefficients as before. Consequently, the  $n$ -dimensional Lanczos recurrence must in the steps 1 to  $n$  generate the same monic orthogonal polynomials as the original  $N$ -dimensional Lanczos recurrence for  $A, q_1$ .

Summarizing,

$T_n$  is determined by the Lanczos process for the matrix  $T_n$  and the starting vector  $e_1$ , the monic polynomials  $\{1, \psi_1, \dots, \psi_n\}$  are orthogonal with respect to the (new) innerproduct determined by the eigenvalues of  $T_n$  and the corresponding weights

$$(\varphi, \psi)_n = \sum_{i=1}^n \omega_i^{(n)} \varphi(\theta_i^{(n)}) \psi(\theta_i^{(n)}), \quad \omega_i^{(n)} = |(s_i^{(n)}, e_1)|^2.$$

The  $n$ -th Riemann-Stieltjes integral,

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

$$\omega^{(n)}(\lambda) = 0 \quad \zeta \leq \lambda < \theta_1^{(n)},$$

$$\omega^{(n)}(\lambda) = \sum_{j=1}^l \omega_j^{(n)} \quad \theta_l^{(n)} \leq \lambda < \theta_{l+1}^{(n)},$$

$$\omega^{(n)}(\lambda) = \sum_{j=1}^n \omega_j^{(n)} \quad \theta_n^{(n)} \leq \lambda < \xi.$$

## Lanczos algorithm:

- sequence of orthonormal vectors  $\{q_1, \dots, q_n\}$
- sequence of Jacobi matrices  $\{T_1, \dots, T_n\}$
- sequence of monic orthogonal polynomials  $\{1, \dots, \psi_n\}$
- sequence of R-S integrals with  $\{\omega^{(1)}, \dots, \omega^{(n)}\}$
- sequence of continued fractions  $\{C_1, \dots, C_n\}$

Relationship between  $\int_{\zeta}^{\xi} f(\lambda) d\omega(\lambda)$  given by  $A, q_1$

and  $\int_{\zeta}^{\xi} f(\lambda) d\omega^{(n)}(\lambda) = \sum_{i=1}^n \omega_i^{(n)} f(\theta_i^{(n)})$  given by  $T_n, e_1$

is nothing but the **Gauß quadrature !**

The Lanczos process determining the orthonormal basis of Krylov subspaces is therefore the **matrix formulation of the Gauss quadrature**. [S, Tichý - 02], [S, Liesen - 05], [Meurant, S - 06].

# Conjugate gradient method (CG)

The unique method which minimizes the discrete energy norm of the error over Krylov subspaces (see **Appendix**)

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

$$\min_{z \in K_n(A, r_0)} \|(x - x_0) - z\|_A,$$

$$x - x_n = (x - x_0) - z_n \perp_A K_n(A, r_0),$$

$$r_n = b - Ax_n = A(x - x_n) \perp K_n(A, r_0), \quad r_n \perp \text{span}\{q_1, \dots, q_n\}.$$

The CG approximation is determined by

$$0 = Q_n^T (b - Ax_n) = \|r_0\| e_1 - Q_n^T A Q_n y_n,$$

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

### Consequence:

Again, the essence of CG is nothing but the Gauß quadrature! Everything is determined by  $\omega(\lambda)$ . The way the **eigenvalues** are linked to **convergence** is given by the way  $\omega(\lambda)$  determines the sequence  $\omega^{(n)}(\lambda)$ ,  $n = 1, 2, \dots$ . This relationship is nothing but trivial !

## The essence of the CG method

$$\begin{array}{ccc} Ax = b, x_0 & \longrightarrow & \int_{\zeta}^{\xi} f(\lambda) d\omega(\lambda) \\ \uparrow & & \uparrow \\ T_n y_n = \|r_0\| e_1 & \longleftrightarrow & \sum_{i=1}^n \omega_i^{(n)} f(\theta_i^{(n)}) \\ x_n = x_0 + Q_n y_n & & \end{array}$$

**Gauß quadrature !**

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

## V/2 Characterization of convergence

In practical applications, **preconditioned** Krylov subspace methods search for the **sufficiently accurate** approximate solution of the finite dimensional problem in **a small number of steps** (much smaller than the system dimension).

“**Convergence**” must be understood differently from the classical iterative methods [Hackbush - 94]. We must study the behavior from the very beginning. No limit, no escape to infinity. We are interested in the transition period itself [Driscoll, Toh and Trefethen - 98].

In early iterations convergence behavior can strongly depend on the **initial residual (right hand side)**. Consequently, no analysis based on the operator (system matrix) only can be sufficient for achieving a complete understanding.

Very complex phenomenon. In general, no single approach is sufficient.

Role of the most frequently used **eigenvalue - eigenvector structure in relation to the particular initial residual ?**

Nick Trefethen [Trefethen-97]:

Any use of eigenvalues to derive physical predictions relies on an implicit transformation to eigenvector coordinates. If the matrix is (even moderately) far from normal, the change to eigenvector coordinates may involve an extreme distortion with a superposition of huge eigen-components that nearly cancel. The state of the system may be determined by the **pattern of cancellation**, rather than by the size of the individual eigen-components.

Without further transformation the eigenvalue - eigenvector structure can in such cases hardly be useful!

## Spectral decompositions

$A$  Hermitian:  $A = U\Lambda U^*$ ,  $UU^* = U^*U = I$ ,  $\Lambda = \bar{\Lambda}$ .

$A$  Normal:  $A = U\Lambda U^*$ ,  $UU^* = U^*U = I$ .

$A$  Diagonalizable:  $A = X\Lambda X^{-1}$ .

$A$  General:  $A = SJS^{-1}$ .

**Goal:** Show the difference in our understanding when the system matrix changes from **Hermitian** to **general nonnormal**.

## Conjugate gradient method, $A$ HPD

- $\|x - x_n\|_A = \|b - Ax_n\|_{A^{-1}}$  minimal
- $x_n = x_0 + Q_n y_n, \quad T_n y_n = \|r_0\| e_1$
- $\|r_n^{\text{CG}}\|_{A^{-1}} / \|r_0^{\text{CG}}\|_{A^{-1}} \leq \min_{p \in \Pi_n} \max_i |p(\lambda_i)|$

## Miminal residual method (MINRES), $A$ Hermitian

- $\| b - Ax_n \|$  minimal

- $x_n = x_0 + Q_n y_n,$

where  $\| \| r_0 \| e_1 - T_{n+1,n} y_n \| = \min_y \| \| r_0 \| e_1 - T_{n+1,n} y \|$

- $\| r_n^M \| / \| r_0^M \| \leq \min_{p \in \Pi_n} \| p(A) \| = \min_{p \in \Pi_n} \max_i |p(\lambda_i)|$

Here  $T_{n+1,n}$  represents the upper Hessenberg tridiagonal matrix obtained from  $T_{n,n}$  by appending a row  $[0, \dots, 0, \beta_{n+1}]$ .

Please notice that  $\| r_n^{\text{CG}} \|_{A^{-1}}$ ,  $\| r_n^{\text{CG}} \|_{A^{-2}}$  and  $\| r_n^M \|$  decrease monotonically, but  $\| r_n^{\text{CG}} \|$  does not. The CG residual can exhibit erratic behavior or increase in norm until the last step!

[Hestenes and Stiefel - 52], [Gutknecht, S - 01]

$$\| r_n^{\text{CG}} \| = \frac{\| r_n^M \|}{\sqrt{1 - \left( \| r_n^M \| / \| r_{n-1}^M \| \right)^2}}$$

[Cullum, Greenbaum - 96], (previously [Brown - 91] for FOM – GMRES). Residual as a measure of convergence for CG? For a HPD system, MINRES is strictly monotonic.

## Conclusion :

All is determined by the eigenvalues and by the components of the initial residual in the individual (invariant) eigenspaces. The last factor can play a significant role only if the individual components differ in magnitude.

[Beckerman, Kuijlaars – 02], [Liesen, Tichý - 04]

However, they are initial residuals which can pathologically affect convergence, see [Scott – 79]. The same residuals **suppress completely the loss of orthogonality in finite precision computations!**

## Ritz values

Roots of the normalized Lanczos polynomial (which is a multiple of the CG polynomial)

$$p_n^{\text{CG}}(\mu) = \psi_n(\mu)/\psi_n(0)$$

are given by the eigenvalues of  $T_n$  i.e. **Ritz values**. Roots of the MINRES polynomial are **harmonic Ritz values**.

[Paige, Parlett and van der Vorst - 95]

Convergence of Ritz values (harmonic Ritz values) explains the acceleration of convergence of CG (MINRES).

[van der Sluis, van der Vorst - 86]

## Worst case (operator) bound

For CG and MINRES, minimizing the matrix polynomial (independent on the initial residual) gives the worst case bound. The worst case initial residual may differ for different  $n$ . MINRES example – for each  $n$ ,

$$\begin{aligned} \frac{\|r_n\|}{\|r_0\|} &= \min_{p \in \Pi_n} \|p(A) q_1\| \leq \max_{\|q\|=1} \min_{p \in \Pi_n} \|p(A) q\| \\ &= \min_{p \in \Pi_n} \max_{\|q\|=1} \|p(A) q\| \\ &= \min_{p \in \Pi_n} \|p(A)\|. \end{aligned}$$

Linear bounds based on the Chebyshev method, see, e.g. [Hageman, Young - 80], [Fischer - 96], [Saad, van der Vorst - 00],

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq 2 \left[ \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right]^n$$

can not be identified, except for some special cases, with the true behavior of the CG method. Various misleading conclusions about “complexity of CG”.

Unless  $\kappa(A)$  is close to one, the **distribution of eigenvalues** between the maximal and minimal ones (not only  $\kappa(A)$ ) is important;

(here we can see a trouble with the term "**preconditioning**").

Does CG significantly outperform Chebyshev?

Unless the spectrum is very special, the **global character** of CG takes advantage of the eigenvalue distribution between the minimal and the maximal eigenvalue !

## V/3 Measuring convergence

### The CG example (see Appendix)

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

For  $n = 1, 2, \dots$

$$\gamma_{n-1} = (r_{n-1}, r_{n-1}) / (p_{n-1}, Ap_{n-1})$$

$$x_n = x_{n-1} + \gamma_{n-1} p_{n-1}$$

$$r_n = r_{n-1} - \gamma_{n-1} Ap_{n-1}$$

$$\delta_n = (r_n, r_n) / (r_{n-1}, r_{n-1})$$

$$p_n = r_n + \delta_n p_{n-1}.$$

For most elliptic self-adjoint PDEs, it is natural to measure convergence in solving the discretized problem by  $\|x - x_n\|_A$ .

The idea of estimating  $\|x - x_n\|_A$  at the price of  $d$  extra steps comes from [Golub, S - 94]. It was developed into a practical algorithm in [Golub, Meurant - 97], all based on [Gauß quadrature](#) relationship,

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

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 determined by  $\|x - x_{n+d}\|_A$ .

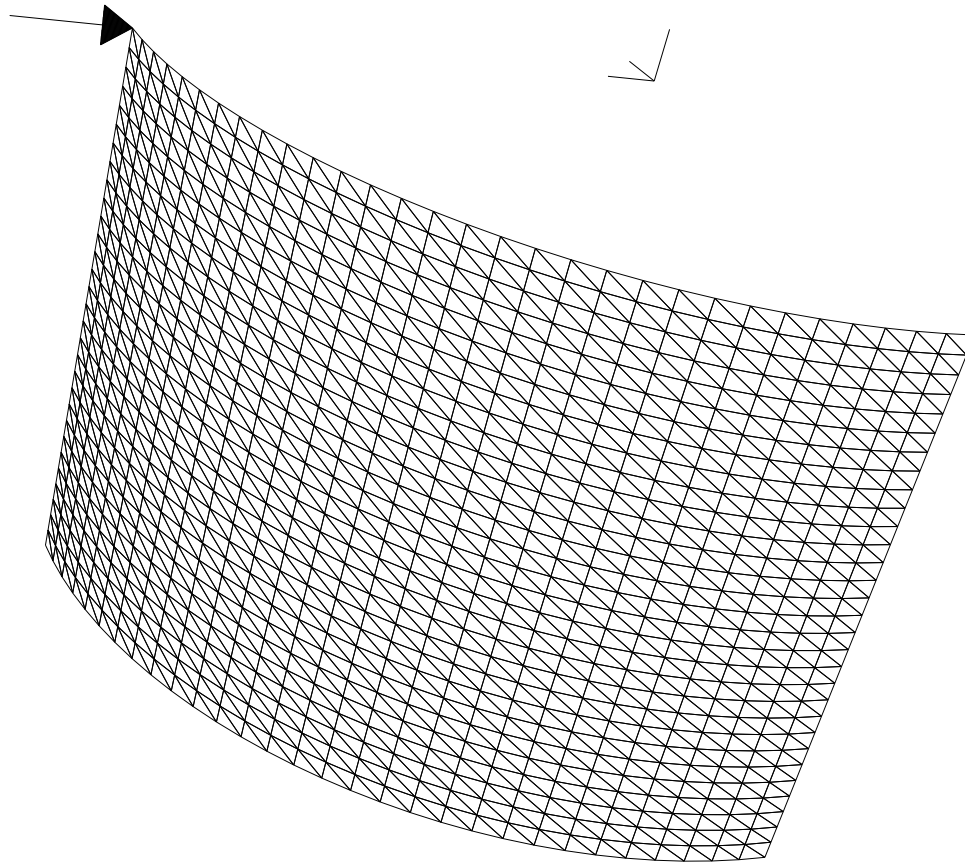
**Mathematically equivalent formulas for  $EST^2$  :**

[Golub, S - 94], [Golub, Meurant - 97]  $\|r_0\|^2 [C_{n+d} - C_n]$

[Warnick - 00]  $r_0^T (x_{n+d} - x_n)$

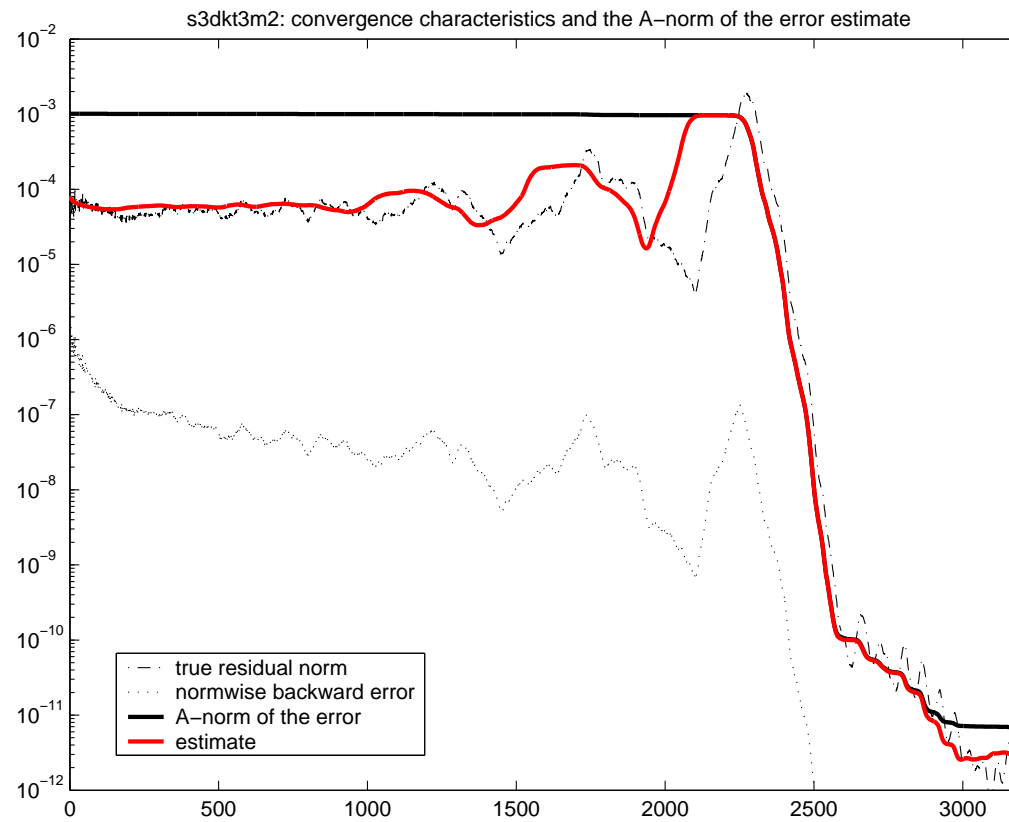
[Hestenes, Stiefel - 52], after fifty years found and extended in  
[S, Tichý 2002, 04] [with justification for finite precision computations](#)

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

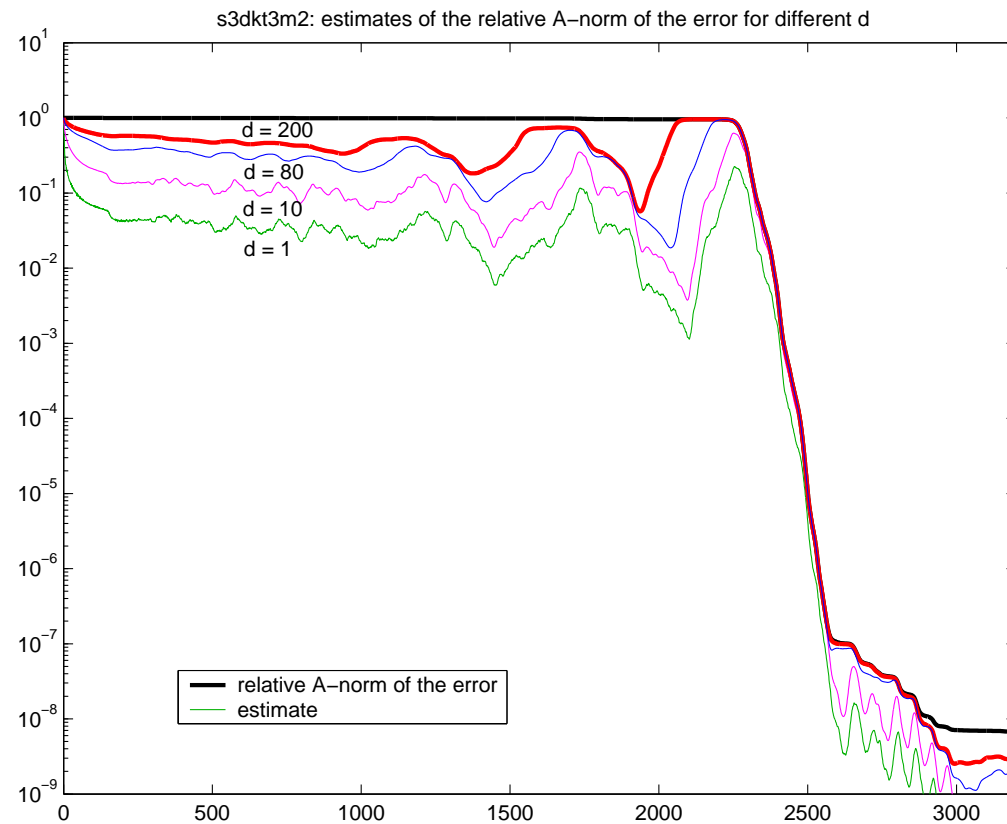


R. Kouhia, collection Cylshell,  $N = 90449$ ,  $\kappa(A) = 3.62e + 11$

# Incomplete Choleski preconditioned CG, convergence characteristics and estimate for the A-norm of the error



Estimates for the relative A-norm of the error with different values of the parameter  $d$



Will the spectral information still determine the convergence behaviour when we move to the non-Hermitian case?

**See next.**

# Lecture VI

## GMRES BEHAVIOUR AND EIGENVALUES

## Spectral information and convergence

Normal matrices have a full set of eigenvectors forming a basis of  $\mathbb{C}^N$  which can be chosen **orthonormal**. Therefore the change to (orthonormal) eigenvector coordinates does not involve any **distortion of geometry**.

Substantial difference from the Hermitian case which causes enormous technical difficulties in proofs and in deriving bounds - **the eigenvalues are not real**. However, principal difficulties come with nonnormality.

We restrict ourselves to the GMRES method.

Given  $A \in \mathbb{C}^{N \times N}$ ,  $b \in \mathbb{C}^N$ ,  $A$  nonsingular, we wish to solve  $Ax = b$ .

Consider  $x_0 \in \mathbb{C}^N$ ,  $r_0 = b - Ax_0$ ,

construct the sequence of Krylov subspaces

$$K_j(A, r_0) = \text{span} \{r_0, Ar_0, \dots, A^{j-1}r_0\}, \quad j = 1, 2, \dots$$

and look for  $x_j \in x_0 + K_j(A, r_0)$ .

## Minimal residual methods

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

(Hermitian) MINRES [Paige, Saunders - 75]  
and its general **simplification** GMRES [Saad, Schultz - 86];  
mathematically equivalent to GCR analyzed in [Elman - 1982]  
and to many other (mostly numerically inferior) methods.

MINRES **IS NOT** a symmetric variant of GMRES!

## Implementation of GMRES [Saad, Schultz - 86]

- Arnoldi basis  $\left\{ v_1 \equiv \frac{r_0}{\|r_0\|}, v_2, \dots, v_n \right\}$ ,  $AV_n = V_{n+1}H_{n+1,n}$ .

- $x_n = x_0 + V_n y_n$ ,

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

Other implementations (GCR, simpler GMRES, ORTHODIR) suffer from possible numerical difficulties.

**Bound by Elman step by step for  $A$  normal:**

$$\begin{aligned}
 \|r_n\| &= \|p_n(A)r_0\| = \min_{p \in \Pi_n} \|p(A)r_0\| = \min_{p \in \Pi_n} \|Y [p(\Lambda) Y^* r_0]\| \\
 &= \min_{p \in \Pi_n} \|p(\Lambda) Y^* r_0\| = \min_{p \in \Pi_n} \left\{ \sum_i |(y_i^* r_0) p(\lambda_i)|^2 \right\}^{\frac{1}{2}} \\
 &\leq \|r_0\| \min_{p \in \Pi_n} \max_i |p(\lambda_i)|.
 \end{aligned}$$

$p_n(\lambda_i)$  represents a multiplicative correction to the values of the individual components of  $r_0$  in the orthonormal basis  $\{y_1, \dots, y_N\}$  in order to minimize the sum of squares.

**Bound by Elman step by step for  $A$  diagonalizable:**

$$\begin{aligned}
 \|r_n\| &= \|p_n(A)r_0\| = \min_{p \in \Pi_n} \|p(A)r_0\| = \min_{p \in \Pi_n} \|Y [p(\Lambda) Y^{-1}r_0]\| \\
 &\leq \|Y\| \min_{p \in \Pi_n} \|p(\Lambda) Y^{-1}r_0\| \\
 &= \|Y\| \min_{p \in \Pi_n} \left\{ \sum_i |[Y^{-1}r_0]_i p(\lambda_i)|^2 \right\}^{\frac{1}{2}} \\
 &\leq \|Y\| \|Y^{-1}r_0\| \min_{p \in \Pi_n} \max_i |p(\lambda_i)| \\
 &\leq \|r_0\| \kappa(Y) \min_{p \in \Pi_n} \max_i |p(\lambda_i)|.
 \end{aligned}$$

For any  $v$ ,  $\|v\| = 1$

$$\min_{p \in \Pi_n} \|p(A)v\| \leq \min_{p \in \Pi_n} \|p(A)\| \equiv \min_{p \in \Pi_n} \max_{\|v\|=1} \|p(A)v\|,$$

therefore

$$\max_{\|v\|=1} \min_{p \in \Pi_n} \|p(A)v\| \leq \min_{p \in \Pi_n} \max_{\|w\|=1} \|p(A)w\|.$$

In the normal case  $\leq$  can be replaced by  $=$   
 and the polynomial approximation problem is attainable  
 [Joubert -93], [Gurvits, Greenbaum - 93], [Trefethen - 93]

$$\frac{\|r_n\|}{\|r_0\|} = \min_{p \in \Pi_n} \|p(A)v_1\| \leq \max_{\|v\|=1} \min_{p \in \Pi_n} \|p(A)v\| = \min_{p \in \Pi_n} \|p(A)\|.$$

For a general  $Y$ , some of the components  $Y^{-1}r_0$  can become very large. In such case  $Y [p(\Lambda) Y^{-1}r_0]$  represents a significant cancelation. The minimization problem

$$\|r_n\| = \min_{p \in \Pi_n} \| Y [p(\Lambda) Y^{-1}r_0] \|$$

reflects that, while the term in the bound

$$\|Y\| \min_{p \in \Pi_n} \| p(\Lambda) Y^{-1}r_0 \|$$

does not (cf. [Trefethen-97]).

## Convergence to the exact solution:

GMRES converges to the exact solution at the step  $m$   
( $x_m \equiv x, r_m \equiv 0$ )

if and only if

the vectors  $r_0, Ar_0, \dots, A^{m-1}r_0$  are linearly independent and the  
vectors  $r_0, Ar_0, \dots, A^m r_0$  are linearly dependent

(Krylov sequence  $r_0, Ar_0, \dots, A^N r_0$  has length  $m$ ).

## Spectrum of $A$ and convergence of GMRES

In practical computations the rate of convergence is linked to the **distribution of eigenvalues of the matrix  $A$** .

There are, however, examples showing that **any (nonincreasing) convergence curve is possible for GMRES with matrix  $A$  having any given (nonzero) eigenvalues**.

[Greenbaum, S - 94], [Greenbaum, Pták, S - 96]

Assume convergence exactly in  $N$  steps (generalization to  $m < N$  possible). For simplicity of notation  $r_0 = 0$  ( $x_0 = 0$ ).

### Question I:

Given **convergence curve**, describe the set of all  $\{A, b\}$  such that GMRES  $(A, b)$  generates the prescribed curve.

### Question II:

Given **convergence curve**, given  $N$  **nonzero eigenvalues** (not necessarily distinct), describe the set of all  $\{A, b\}$  such that GMRES  $(A, b)$  generates the curve while the spectrum of  $A$  is prescribed.

### Question III:

Given  $A$ , denote by  $\hat{m}$  the degree of the minimal polynomial of  $A$ . Describe those  $b$  for which GMRES  $(A, b)$  converges in  $\hat{m}$  steps.

## Convergence curve

$$\|r_0\| \geq \|r_1\| \geq \cdots \geq \|r_{N-1}\| > \|r_N\| = 0,$$

$$h \equiv (\eta_1, \dots, \eta_N)^T, \quad \eta_j \equiv ((\|r_{j-1}\|)^2 - \|r_j\|^2)^{1/2}.$$

$$d \equiv (\nu_1, \dots, \nu_N), \quad \nu_1 = \frac{1}{\eta_N}, \nu_2 = -\frac{\eta_1}{\eta_N}, \dots, \nu_N = -\frac{\eta_{N-1}}{\eta_N}.$$

Meaning? Let  $W = (w_1, \dots, w_j)$  be the orthonormal basis of  $AK_j(A, r_0)$ . Then

$$r_n = r_0 - \sum_{j=1}^n w_j \eta_j, \quad r_0 = \sum_{j=1}^n w_j \eta_j + r_n, \quad \|r_0\|^2 = \sum_{j=1}^n \eta_j^2 + \|r_n\|^2$$

Convergence curve companion matrix

$$\hat{H} = \begin{pmatrix} 0 & & & 1/\eta_N \\ 1 & \cdots & & -\eta_1/\eta_N \\ & \cdots & 0 & \vdots \\ & & 1 & -\eta_{N-1}/\eta_N \end{pmatrix} = \begin{pmatrix} 0 & & & \boxed{\phantom{d}} \\ 1 & \cdots & & \\ & \cdots & 0 & \\ & & 1 & \boxed{d} \end{pmatrix}$$

$$\hat{H}^{-1} = \begin{pmatrix} \eta_1 & 1 & \cdots & \\ \vdots & 0 & \cdots & \\ \vdots & & \cdots & 1 \\ \eta_N & & & 0 \end{pmatrix} = \begin{pmatrix} \boxed{\phantom{h}} & 1 & \cdots & \\ & 0 & \cdots & \\ & & \cdots & 1 \\ & & & 0 \end{pmatrix}$$

## Eigenvalues:

$$\{\lambda_1, \lambda_2, \dots, \lambda_N\}, \quad \lambda_j \neq 0, \quad j = 1, \dots, n.$$

$$q_N(z) \equiv z^N - \sum_{j=0}^{N-1} \alpha_j z^j = (z - \lambda_1)(z - \lambda_2) \dots (z - \lambda_N),$$

$$p_N(z) \equiv 1 - \sum_{j=1}^N \xi_j z^j = -\frac{1}{\alpha_0} q_N(z), \quad \xi_N = \frac{1}{\alpha_0}, \quad \xi_j = -\frac{\alpha_j}{\alpha_0},$$

$$s \equiv (\xi_1, \dots, \xi_N)^T, \quad a = (\alpha_0, \dots, \alpha_{N-1})^T$$

Spectral comparison matrix:  $q_N(z) \equiv \det(C - zI)$

$$C = \begin{pmatrix} 0 & & & \alpha_0 \\ 1 & \cdots & & \alpha_1 \\ & \cdots & 0 & \vdots \\ & & 1 & \alpha_{N-1} \end{pmatrix} = \begin{pmatrix} 0 & & & \boxed{\phantom{a}} \\ 1 & \cdots & & \\ & \cdots & 0 & \\ & & 1 & \boxed{a} \end{pmatrix}$$

$$C^{-1} = \begin{pmatrix} -\alpha_1/\alpha_0 & 1 & & \\ -\alpha_2/\alpha_0 & 0 & \cdots & \\ \vdots & & \cdots & \\ \vdots & & & 1 \\ 1/\alpha_0 & & & 0 \end{pmatrix} = \begin{pmatrix} \boxed{\phantom{s}} & 1 & & \\ \boxed{s} & 0 & \cdots & \\ & & \cdots & 1 \\ & & & 0 \end{pmatrix}$$

## Theorem 1 (Question I)

The following assertions are equivalent:

1° Residual vectors norms of GMRES( $A, b$ ) form a prescribed nonincreasing sequence  $\|r_0\| \geq \|r_1\| \geq \dots \geq \|r_{N-1}\| > \|r_N\| = 0$ .

2° Matrix  $A$  is of the form  $A = W\hat{R}\hat{H}W^*$  and  $b$  satisfies  $W^*b = h$ , where  $W$  is a unitary matrix,  $\hat{R}$  is a nonsingular upper triangular matrix and

$$\hat{H} = \begin{pmatrix} 0 & & & & \\ 1 & \dots & & & \\ & \dots & & & \\ & & 0 & & \\ & & 1 & & \boxed{d} \end{pmatrix}.$$

**Proof.** Consider the QR decomposition

$$B \equiv (Ab, A^2b, \dots, A^N b) = \tilde{W} \tilde{R}$$

Then the columns  $\tilde{W}_j = (\tilde{w}_1, \dots, \tilde{w}_j)$  represent an orthonormal basis of

$$AK_j = A \operatorname{span}\{b, \dots, A^{j-1}b\} = \operatorname{span}\{Ab, \dots, A^j b\}.$$

therefore

$$\eta_j = |\tilde{\eta}_j| = \left( \|r_{j-1}\|^2 - \|r_j\|^2 \right)^{1/2}.$$

Rescaling

$$b = \tilde{W}(\Gamma h) = (\tilde{W}\Gamma)h = Wh$$

where

$$\Gamma = \text{diag}(\gamma_i), \quad |\gamma_i| = 1,$$

we can write

$$B = WR, \quad R = \Gamma^* \tilde{R}.$$

1° is equivalent to

$$A(b, W_{N-1}) = AW \begin{pmatrix} \boxed{h} & & & \\ & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & & 0 \end{pmatrix} = AW \hat{H}^{-1}.$$

Since for some nonsingular upper triangular  $\hat{R}$

$$A(b, W_{N-1}) = (Ab, AW_{N-1}) = W\hat{R},$$

the identity  $AW\hat{H} = W\hat{R}$  finishes the proof.

## Theorem 2 (Question II)

*The following two assertions are equivalent:*

1° *The spectrum of  $A$  is  $\{\lambda_1, \dots, \lambda_N\}$  and GMRES( $A, b$ ) yields residuals with the prescribed nonincreasing sequence*

$$\|r_0\| \geq \|r_1\| \geq \dots \geq \|r_{N-1}\| > \|r_N\| = 0.$$

2° *Matrix  $A$  is of the form  $A = WRCR^{-1}W^*$  and  $b = Wh$  where  $C$  is the companion matrix corresponding to the polynomial  $q_N(z)$ ,  $W$  is unitary and  $R$  a nonsingular upper triangular matrix such that  $Rs = h$ .*

**Corollary:** Any nonincreasing convergence curve can be generated by GMRES for a matrix having any prescribed eigenvalues.

**Proof.** Assume 1°.  $A$  is annihilated by  $q_N(z)$ , therefore

$$B = (A^{-1}B)C = (b, \dots, A^{N-1}b)C,$$

$$AB = BC \quad \text{and} \quad b = BC^{-1}e_1 = Bs.$$

Similarly to Theorem 1,  $b = Wh$ ,  $B = WR$ , which gives  $Rs = h$ .  
Moreover,

$$AWR = AB = BC = WRC$$

proves 2°.

Assume 2°.

Then  $\text{sp}(A) = \{\lambda_1, \dots, \lambda_N\}$ , and, by induction,  $\{w_1, \dots, w_k\}$  represents the unitary basis of  $AK_k$  which proves 1°.

**Remark:**  $W$  represents a change of the basis.

Denote

$\mathcal{S}_1 = \mathcal{S}_1(f)$  the set of all pairs  $\{A, b\}$  determined by Theorem 1,

$\mathcal{S}_2 = \mathcal{S}_2(f, \{\lambda_1, \dots, \lambda_N\})$  the set of all pairs  $\{A, b\}$  determined by  
Theorem 2.

Clearly  $\mathcal{S}_2 \subset \mathcal{S}_1$ .

Parametrization?

## Proposition

The set  $\mathcal{S}_2$  is parametrized by  $W$  and by the nonsingular upper triangular matrix  $R$  satisfying the relation

$$Rs = h.$$

The set  $\mathcal{S}_1$  is parametrized by  $W$  and an arbitrary nonsingular upper triangular matrix  $\hat{R}$ . If, in addition, the spectrum of the matrix  $A$  is prescribed, then this additional condition is equivalent to

$$RCR^{-1} = \hat{R}\hat{H},$$

or, equivalently,  $\hat{R}$  is given by (for the proof see [Arioli, Pták, S-98])

$$\hat{R} = R \begin{pmatrix} 1 & 0 \\ 0 & \boxed{R_{1,N-1}^{-1}} \end{pmatrix}.$$

Since  $\xi_N \neq 0$ , any nonsingular upper triangular matrix  $R$  satisfying

$$Rs = h$$

has its last column uniquely determined by the entries in the left principal submatrix  $R_{1,N-1}$  representing free parameters.

Denoting

$$Y \equiv RC^{-1} = R \begin{pmatrix} \boxed{s} & 1 & & \\ & 0 & \cdots & \\ & & \cdots & 1 \\ & & & 0 \end{pmatrix} = \begin{pmatrix} \boxed{h} & \boxed{R_{1,N-1}} \\ & 0 \end{pmatrix},$$

Then

$$\begin{aligned} A &= WRCR^{-1}W^* = W(RC^{-1})C(CR^{-1})W^* = \\ &= W(RC^{-1})C(RC^{-1})^{-1}W^* = WYCY^{-1}W^* \end{aligned}$$

Assertions 1° and 2° of Theorem 2 are equivalent to

## Theorem 2 (continuation)

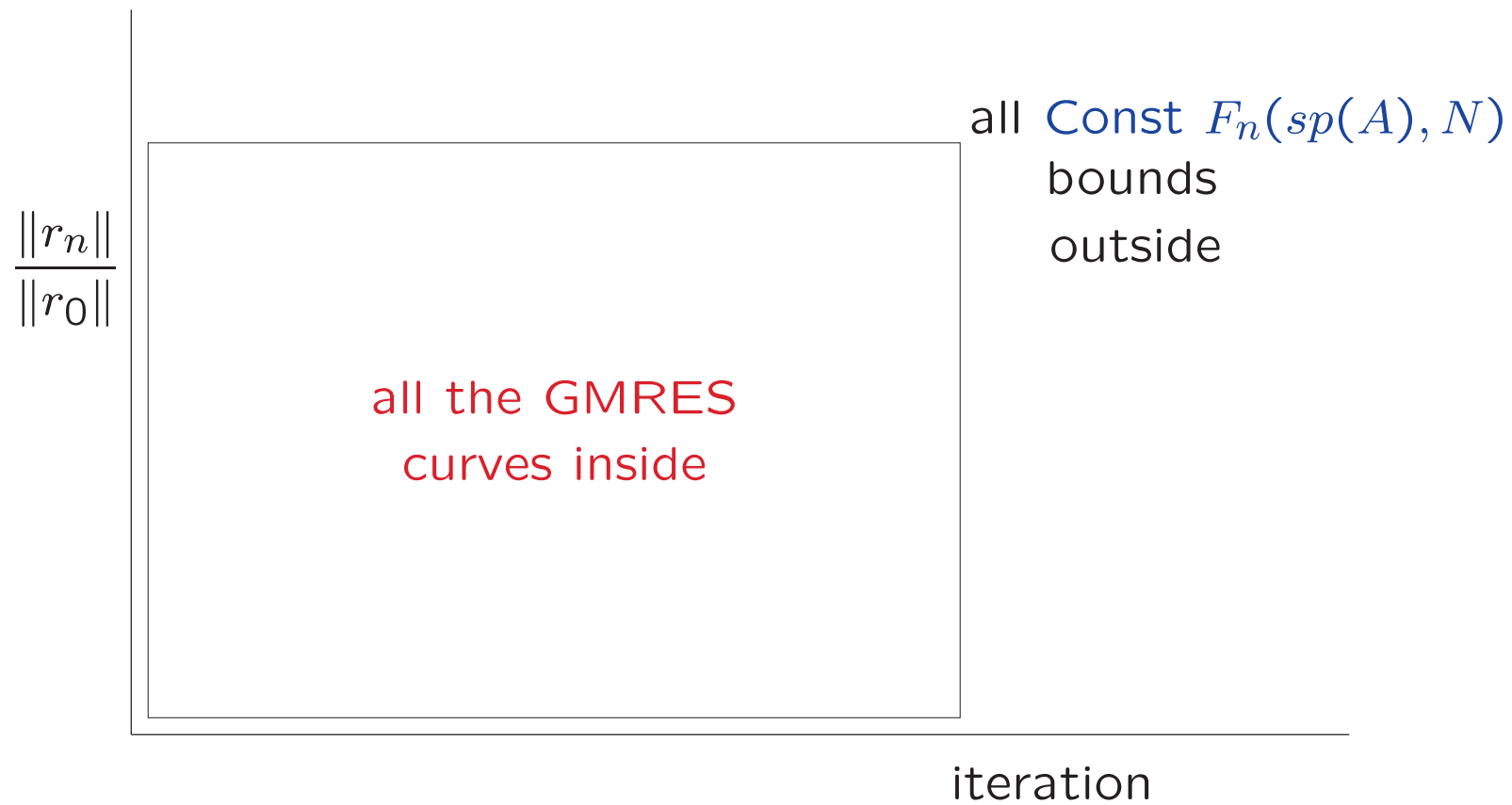
3° Matrix  $A$  is of the form  $A = WYCY^{-1}W^*$  and  $b = Wh$  where  $C$  is the companion matrix corresponding to the polynomial  $q(\lambda)$ ,  $W$  is unitary and  $R_{1,N-1}$  part of  $Y$  is **any**  $(N-1)$  by  $(N-1)$  nonsingular upper triangular matrix.

The problem of “constants” in the bounds of the type

$$\| r_n \| \leq \omega(A, r_0) F_n(sp(A), N).$$

If conclusion is based only on  $F_n(sp(A), N)$  and the dependence of  $\omega(A, r_0)$  on the data is not included, then the bound must hold **for any data**. Consequently, the bound is for any finite dimensional problem irrelevant, otherwise we get a contradiction with the Theorem.

The bound  $\text{Const } F_n(sp(A), N)$  does not intersect the rectangle  $(1, 0) - (1, N) - (0, N) - (0, 0)$ .



## Relationship to minimal polynomial

### Theorem 3

*Let  $m$  denotes the degree of the minimal polynomial  $q_A(\lambda)$  of the matrix  $A$ . Then, for any right hand side  $b$ ,  $GMRES(A, b)$  converges to the exact solution  $x$  on or before the step  $m$ . Moreover, there exist a right hand side  $\tilde{b}$ , for which  $GMRES(A, \tilde{b})$  converges to  $x$  exactly in  $m$  steps.*

Characterization of right hand sides, for which Krylov sequences have the maximal length?

Minimal polynomial  $q_A(\lambda) = (\lambda - \lambda_1)^{n_1} \dots (\lambda - \lambda_{\tilde{k}})^{n_{\tilde{k}}}$ .

Denote the nullspaces of  $(\lambda_j I - A)^{n_j}$  by  $E(\lambda_j)$ .

Then any  $b$  can be decomposed as

$$b = t_1 + t_2 + \dots + t_{n_{\tilde{k}}}, \quad t_j \in E(\lambda_j).$$

The vector  $b$  yields the Krylov sequence of the length  $m$  if and only if

$$(\lambda_j I - A)^{n_j - 1} t_j \neq 0$$

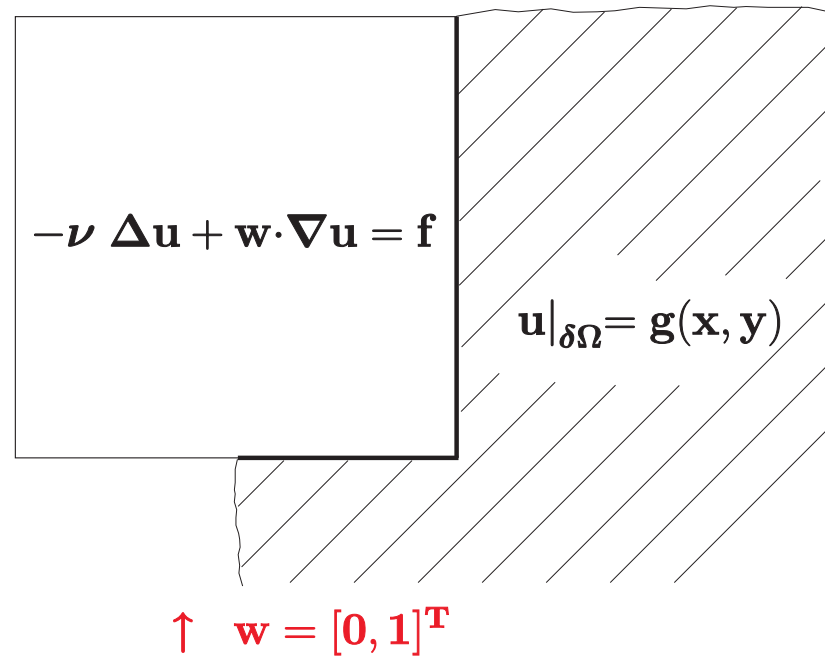
for each  $j$ ,  $j = 1, \dots, \tilde{k}$ . Equivalently, the vector  $b$  have for each  $j$  nonzero component in the direction of at least one last Jordan principal vector conformed to any of the Jordan blocks largest in size corresponding to  $\lambda_j$ .

## Pathological initial residuals?

The presented cautious view seems to be in conflict with the common wisdom – convergence is commonly related to eigenvalue distribution even for general matrices **without examining eigenvectors**. The proved facts should not be ignored (even a common knowledge can be wrong), but they should be understood and interpreted correctly! There are good reasons for linking convergence to eigenvalues in many cases, but the reasons **must be given and examined** (contrary to common practice).

The role of “**pathological initial residuals**”; just academic examples ? Not true. Convection-diffusion examples were described by Trefethen long ago, see also [Ernst - 00].

## VI/2 Convection-diffusion model problem

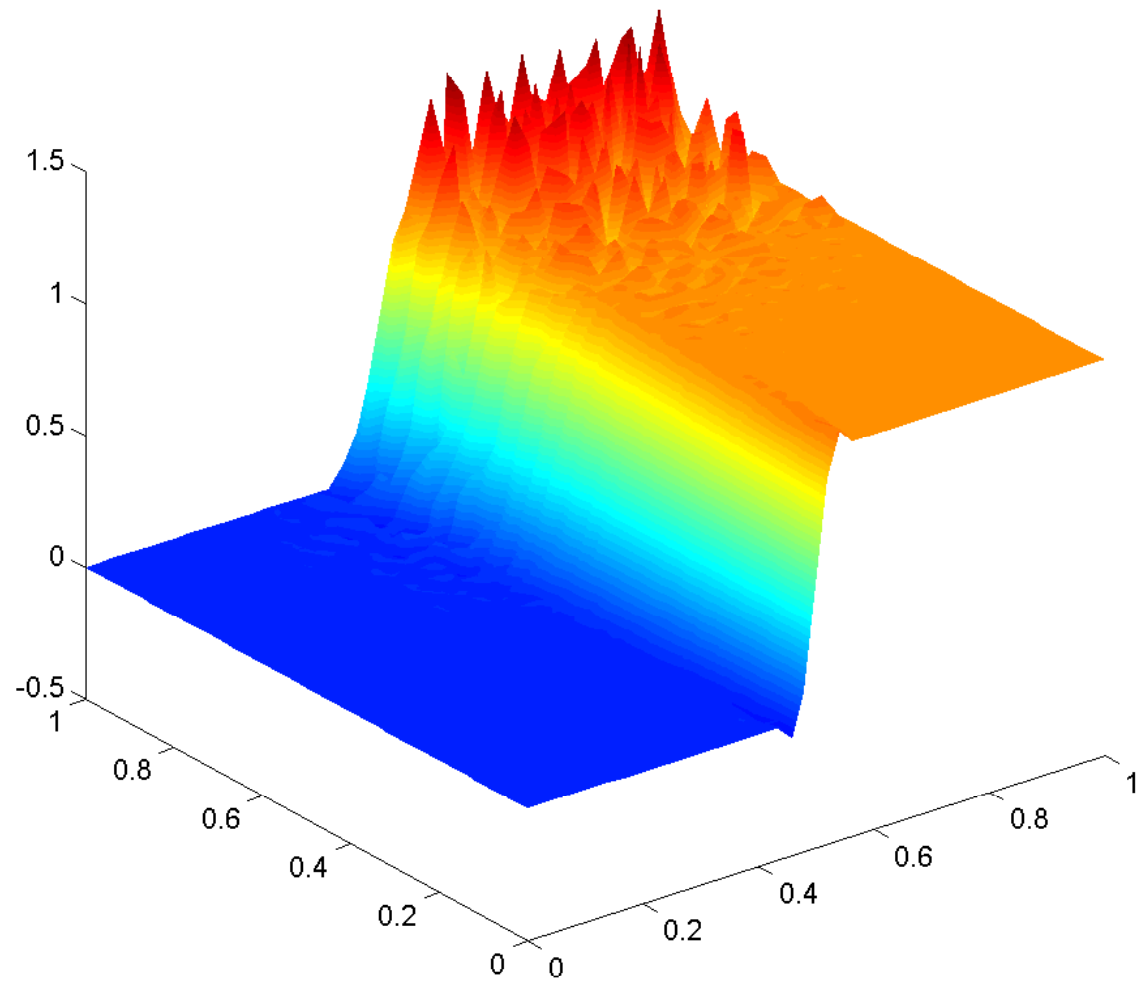


Convection dominated:  $\nu \ll \|\mathbf{w}\|$

## Discretization

- regular  $h \times h$  grid,  $h = 1/(N + 1)$ , bilinear finite elements, mesh Peclet number  $P_h \equiv (h\|w\|)/(2\nu)$ ;
- $P_h > 1$ , then Galerkin discretization produces wiggles (non-physical 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

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



The coefficient matrix of the linear algebraic system is

$$A = \nu A_d + A_c + \hat{\delta} A_s,$$

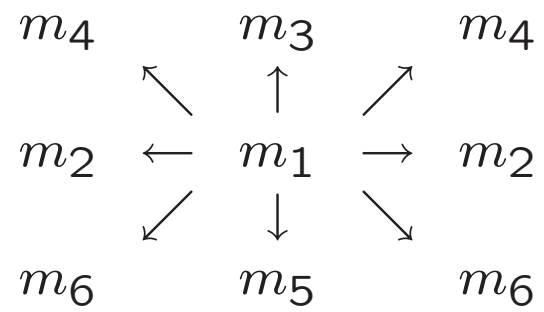
$$A_d = (\nabla \phi_j, \nabla \phi_i),$$

$$A_c = (w \cdot \nabla \phi_j, \phi_i),$$

$$A_s = (w \cdot \nabla \phi_j, w \cdot \nabla \phi_i), \quad \hat{\delta} = \delta_* h / \|w\|.$$

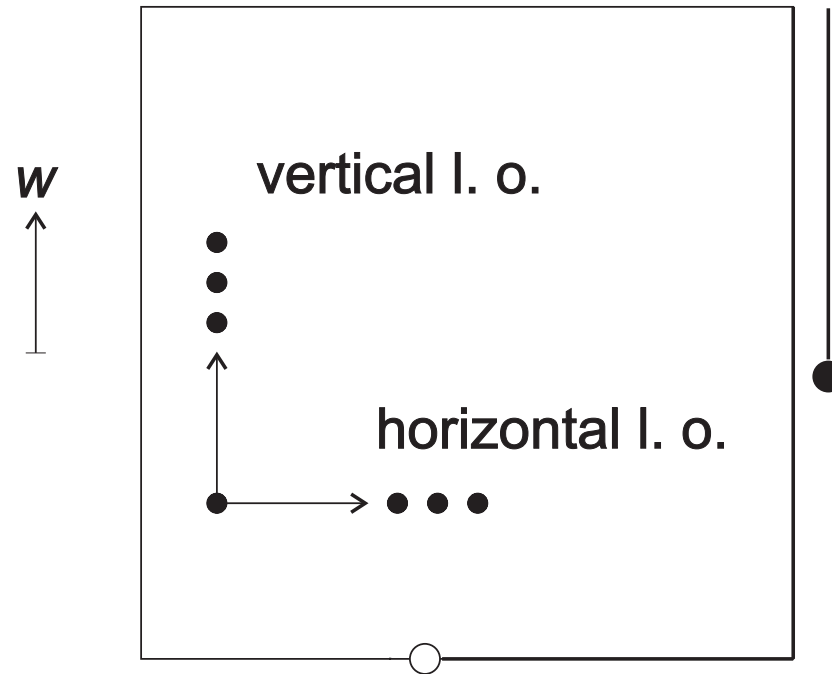
$$A = \left( (\nu I + \hat{\delta} w w^T) \nabla \phi_j, \nabla \phi_i \right) + (w \cdot \nabla \phi_j, \phi_i).$$

The constituent matrix stencil for  $A$



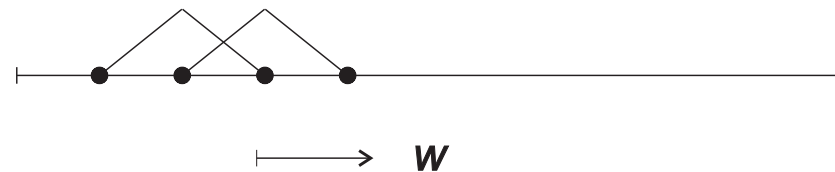
has numerical values

$$\begin{array}{ccccc}
 -\frac{\nu}{3} + \frac{h}{12}(1 - 2\delta) & & -\frac{\nu}{3} + \frac{h}{3}(1 - 2\delta) & & -\frac{\nu}{3} + \frac{h}{12}(1 - 2\delta) \\
 & \swarrow & \uparrow & \nearrow & \\
 -\frac{\nu}{3} + \frac{\delta h}{3} & \leftarrow & \frac{8}{3}\nu + \frac{4}{3}\delta h & \rightarrow & -\frac{\nu}{3} + \frac{\delta h}{3} \\
 & \swarrow & \downarrow & \searrow & \\
 -\frac{\nu}{3} - \frac{h}{12}(1 + 2\delta) & & -\frac{\nu}{3} - \frac{h}{3}(1 + 2\delta) & & -\frac{\nu}{3} - \frac{h}{12}(1 + 2\delta)
 \end{array}$$



With our choice of  $w$ , the differential equation is [separable](#), and the eigendecomposition of the discretized operator is known analytically.

Consider the mass ( $M$ ), stiffness ( $K$ ) and gradient ( $C$ ) matrices of the corresponding 1D convection-diffusion model problem discretized using linear elements with the mesh size  $h$ ,



$$M = \frac{h}{6} \text{tridiag} (1, 4, 1), \quad K = \frac{1}{h} \text{tridiag} (-1, 2, -1),$$

$$C = \frac{1}{2} \text{tridiag} (-1, 0, 1).$$

Let ' $\otimes$ ' denote the Kronecker product of matrices.

Then the 2D SUPG discretized  $N^2 \times N^2$  operator is for the horizontal line ordering of unknowns

$$A_H = \nu M \otimes K + ((\nu + \delta h)K + \mathbf{C}) \otimes M,$$

for the vertical line ordering of unknowns

$$A_V = \nu K \otimes M + M \otimes ((\nu + \delta h)K + \mathbf{C}).$$

$A_H$  and  $A_V$  are orthogonally similar,

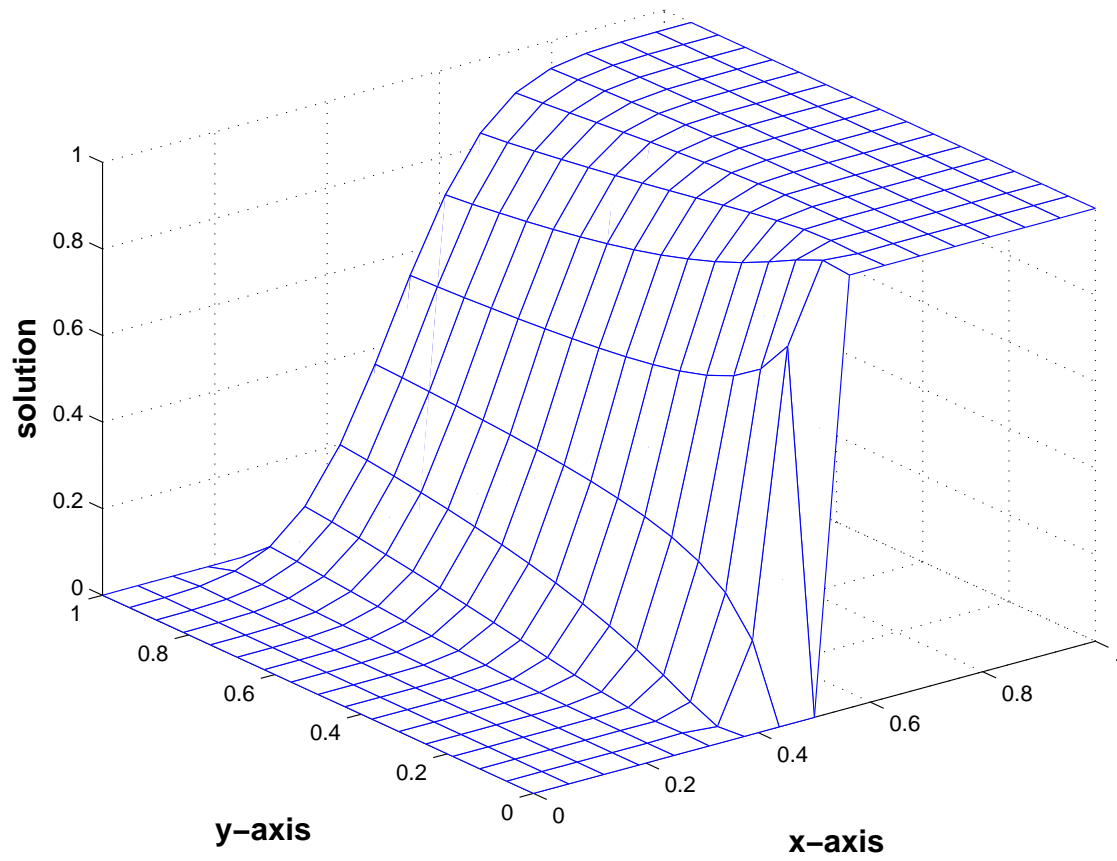
$$A_V = P A_H P, \quad P = [I \otimes e_1, \dots, I \otimes e_n], \quad P = P^T, \quad P^2 = I.$$

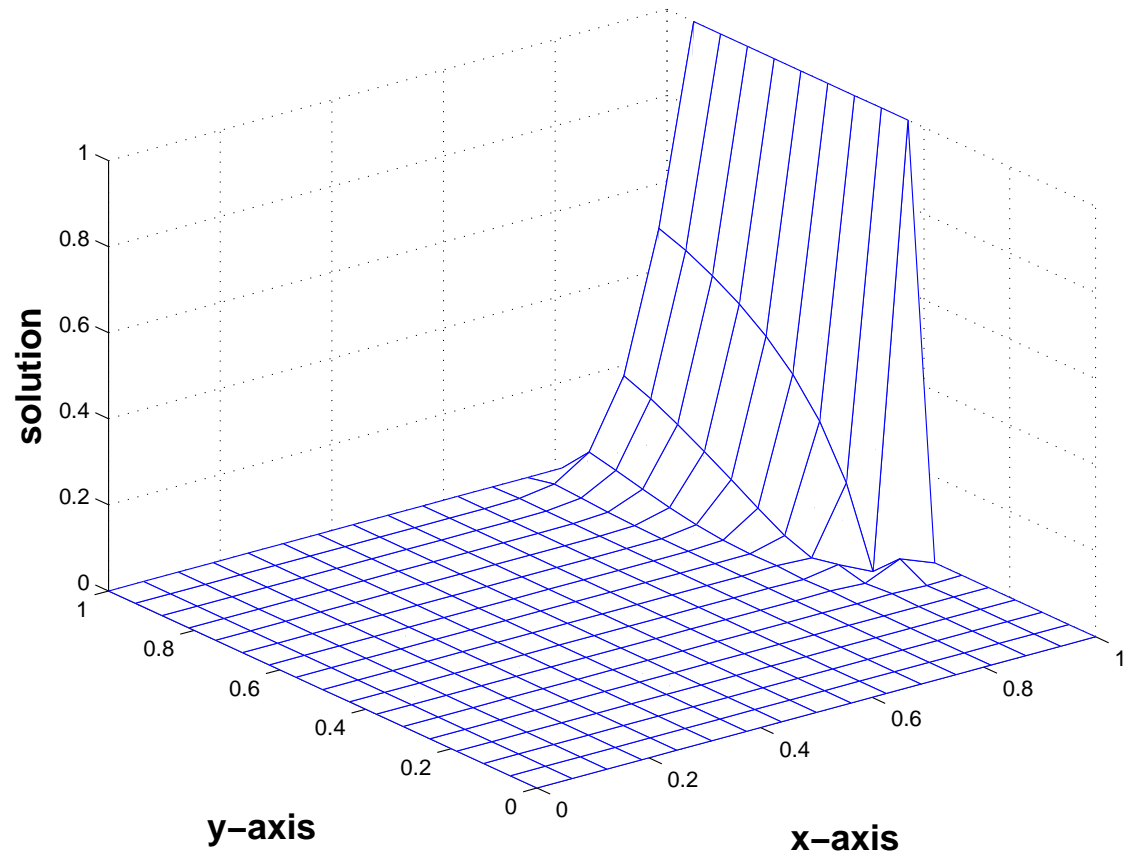
$\approx$  optimal stabilization parameter  $\delta_* \equiv \frac{1}{2} \left( 1 - \frac{1}{P_h} \right)$  affects

- smoothing of the discretized solution,
- behavior of the linear algebraic solver (convergence behavior of GMRES).

Examples of boundary conditions:

- Raithby (discontinuous inflow),
- partial right side of the domain.





## Long list of authors, papers and books

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

GMRES convergence studied using the field of values and the eigendecomposition of the system matrix in particular by

[Eiermann - 89], [Ernst - 00], [Eiermann, Ernst - 02], [Fisher, Ramage, Silvester and Wathen - 99], [Elman, Ramage - 01, 02].

Different approach suggested in [Liesen, S - 04], [Liesen, S - 05].

Eigendecomposition of  $A_H$  ( $A_V$ ) does not lead to useful bounds due to the **ill-conditioned eigenvectors and cancelation**. Instead: The matrices  $K$  and  $M$  are symmetric tridiagonal Toeplitz. The matrix of eigenvectors

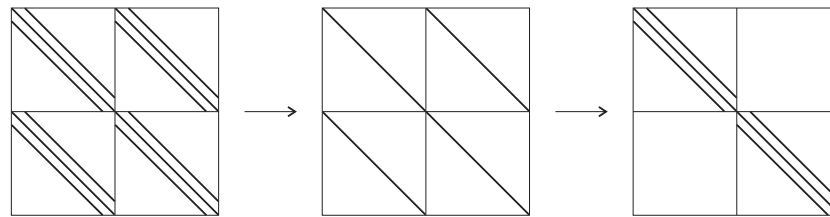
$$U = [u_1, \dots, u_N], \quad U = U^T, \quad U^2 = I$$

$$u_j = (2h)^{1/2} [\sin(jh\pi), \dots, \sin(Njh\pi)]^T, \quad j = 1, \dots, N,$$

represents the Fourier basis. Consider the Fourier transformation of unknowns in the direction perpendicular to the wind. Subsequent reordering of the transformed unknowns by vertical lines gives

$$P(I \otimes U) A_H (I \otimes U) P P(I \otimes U) x_H = P(I \otimes U) b_H.$$

The transformation above corresponds to the simultaneous diagonalization of the symmetric tridiagonal Toeplitz blocks in the block tridiagonal matrix  $A_H$ , with the subsequent permutation of the rows and columns.



The approach using  $A_V x_V = b_V$  is even more straight,  $A_H$  was used here for historical reasons. Resulting system:

$$\mathbf{T} \mathbf{y} = \mathbf{f}$$

$$T = \text{diag}(T_j), \quad T_j = \text{tridiag}(\gamma_j, \lambda_j, \mu_j), \quad j = 1, \dots, N.$$

Thus, the original discretized system transforms to  $N$  non-symmetric tridiagonal Toeplitz systems

$$\mathbf{T}_j \mathbf{y}_j = \mathbf{f}_j, \quad j = 1, \dots, N$$

representing  $N$  discretized one-dimensional convection-diffusion problems (in the vertical direction of the original mesh, but accounting for the diffusion in the horizontal direction).

## GMRES convergence behavior:

$$\|r_n\|^2 = \min_{p \in \Pi_n} \|p(A)b\|^2 = \min_{p \in \Pi_n} \|p(T)f\|^2 = \min_{p \in \Pi_n} \sum_{j=1}^N \|p(T_j)f_j\|^2.$$

GMRES for non-symmetric tridiagonal Toeplitz systems? **Interesting case:** the superdiagonal ( $\mu_j$ ) substantially smaller in magnitude than the two others,  $|\gamma_j| \approx \lambda_j \gg \mu_j$ . Relating the problem for  $T_j, f_j$  to convergence of GMRES for **scaled Jordan blocks**, we proved (and quantified)

### Theorem

Let  $l$  be the index of the first significant nonzero entry in  $f_j$ . Let  $|\gamma_j| \approx \lambda_j \gg \mu_j$ . Then GMRES for  $T_j y_j = f_j$  must converge slowly for at least  $N - l$  steps.

## Slow initial convergence:

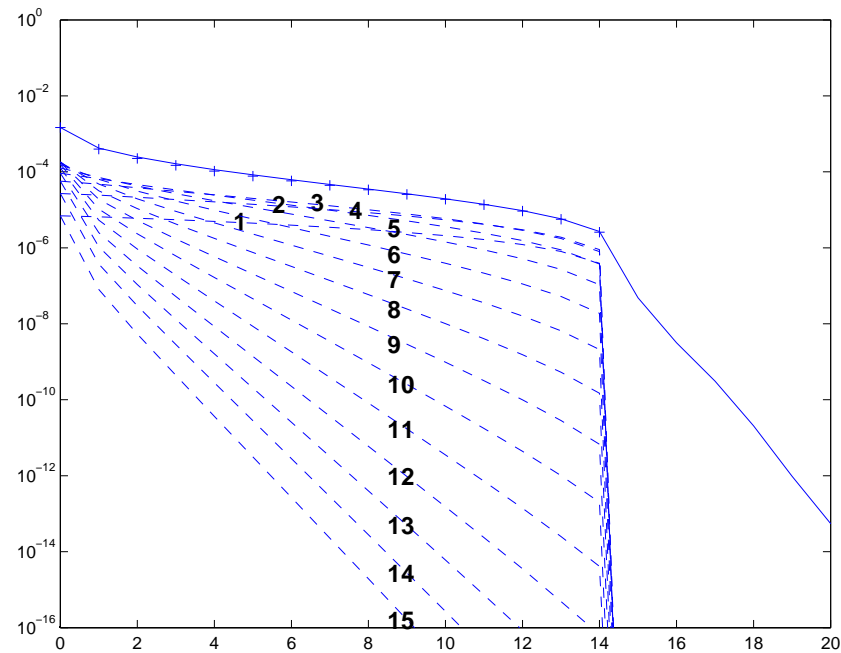
$$\|r_n\|^2 = \min_{p \in \Pi_n} \sum_{j=1}^N \|p(T_j) f_j\|^2 \geq \sum_{j=1}^N \min_{p \in \Pi_n} \|p(T_j) f_j\|^2.$$

If the theorem applies at least for one  $j$ , then the convergence of GMRES for  $Ax = b$  must be slow for at least  $N - l$  steps.

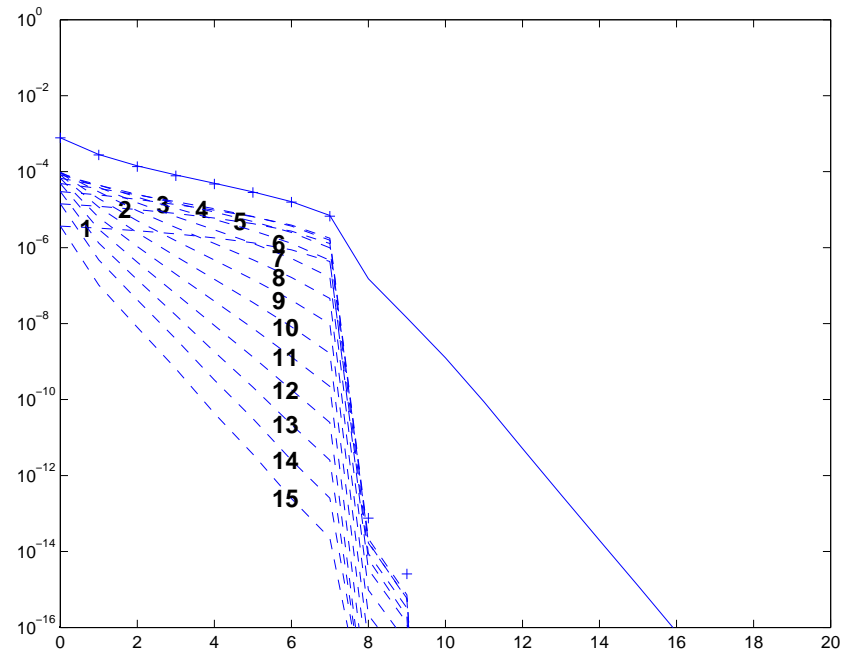
## Acceleration of convergence:

The technique developed in [Greenbaum, Duintjer Tebbens and S - 05?] leads to tight upper and lower bounds which capture the sharp convergence acceleration.

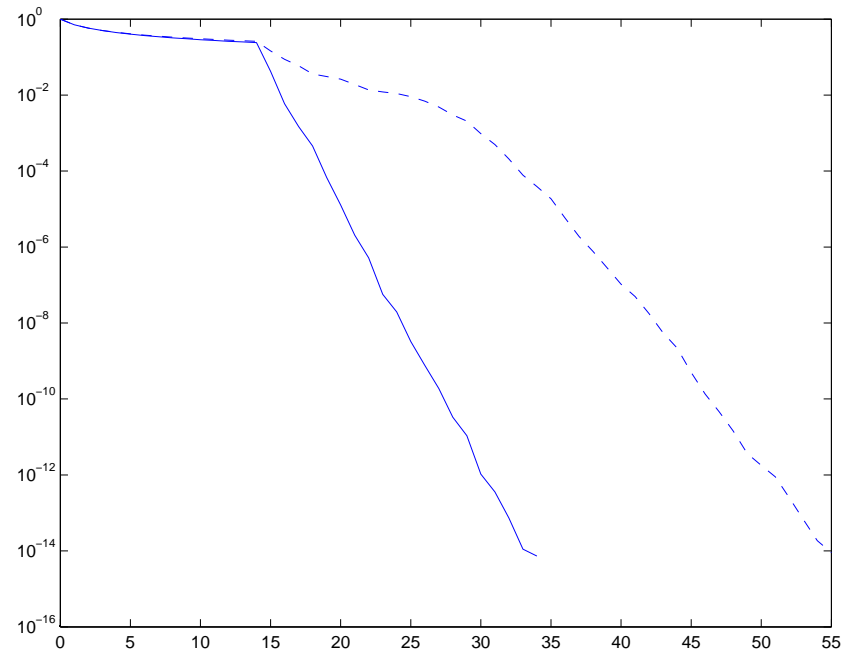
Nonzero boundary conditions on the full right side of the domain, GMRES convergence for the whole system (solid line) and for the individual tridiagonal Toeplitz blocks.



Nonzero boundary conditions on the part of the right side of the domain, GMRES convergence for the whole system (solid line) and for the individual tridiagonal Toeplitz blocks.



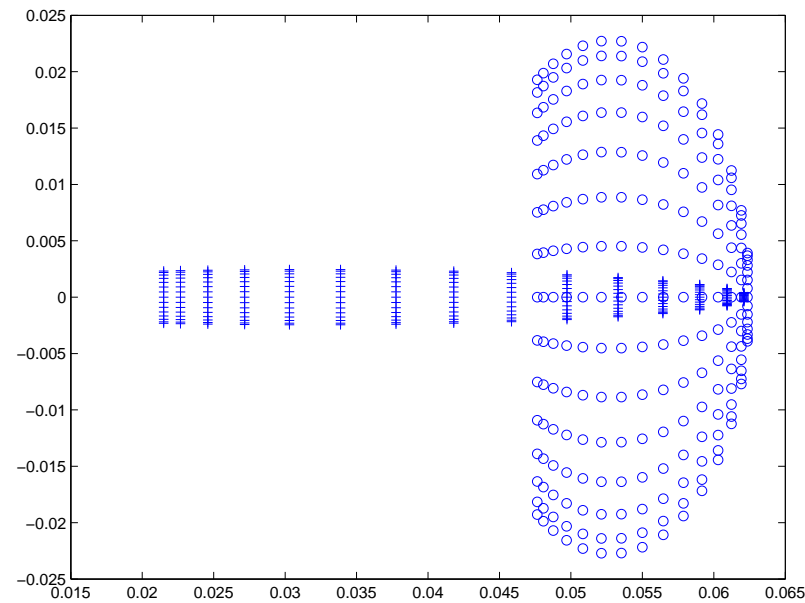
Discontinuous inflow boundary conditions (Raithby), two different values of the diffusion coefficient  $\nu = 0.01$  and  $\nu = 0.0001$  correspond to the solid and to the dashed line, respectively.



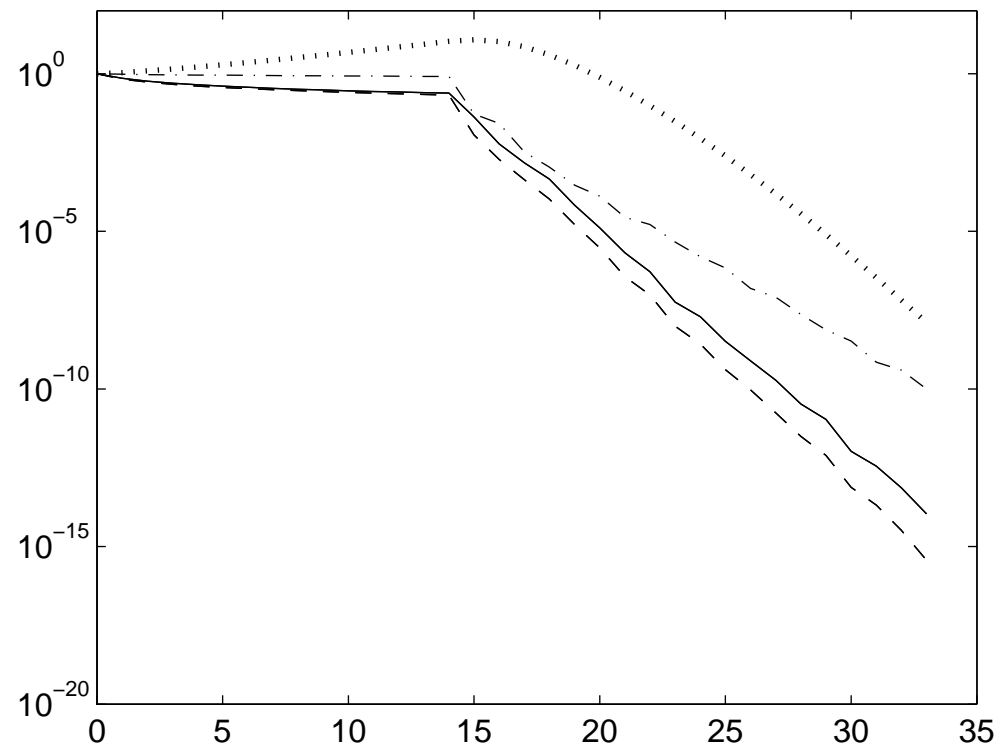
$$\sigma_{jk} = \lambda_j + (\gamma_j \mu_j)^{1/2} \omega_k, \quad \omega_k = 2 \cos(kh\pi), \quad k = 1, \dots, N.$$

Which spectrum corresponds to which convergence curve?

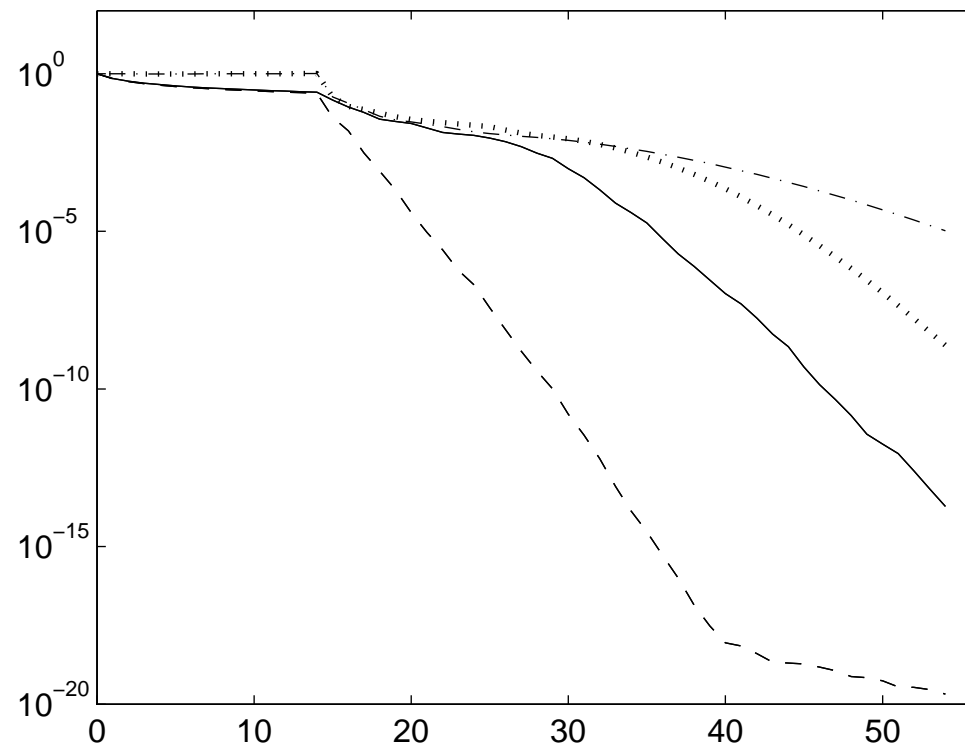
$$\lambda_j > 0, \quad \gamma_j \mu_j < 0.$$



GMRES Convergence curve, upper and lower bounds for  $\nu = 0.01$ .



GMRES Convergence curve, upper and lower bounds  
for  $\nu = 0.0001$ .



## Concluding remarks

- initial phase is important, it depends on the right hand side!
- technique: **orthonormal** transformation to **Jordan-like-structure**  
(the problem is **diagonalizable!**)
- generalizations? Many ways . . . ?
- analytical study of preconditioning?

What can be saved in finite precision **arithmetic**? Does the analysis of finite precision behaviour belong to mathematics?

See in Lectures VII and VIII.

# Lecture VII

## NUMERICAL BEHAVIOUR

### I.

## General considerations

1. Can't we compute exactly?
2. Intermediate quantities and desired accuracy
3. Computational cost and numerical stability

## 3.1 Can't we compute exactly?

No, some problems cannot be solved exactly **in principle**. For example, eigenvalues cannot in general be computed exactly because of the Abel theorem. Consequently, the Schur decomposition cannot in general be computed exactly - in a finite number of steps. There is an unavoidable **truncation error**.

Limited accuracy of performing elementary computer operations (**storing data, +, -, \*, /**) leads to **rounding errors**. This we call **precision**, and speak about **finite precision arithmetic**. We can emulate arbitrary precision arithmetic, but we cannot use it widely in solving practical problems.

Will that issue be resolved by the progress in technology? Hardly. Accuracy of a computed result is determined by the way the elementary rounding errors on the machine precision level are **amplified** in the computational process. Machine precision will always be limited, and it influences the resulting accuracy linearly, while the growth can be exponential.

However, the amplification of elementary rounding errors **is not random, it can be analyzed and understood!**

Rounding errors **are not always bad**;  
see, e.g., breaking the symmetry in shifted QR algorithm which can theoretically suffer from infinite oscillations (relation to dynamical systems, [Batterson, Smilie -90]), or generating nonzero components in invariant subspaces in the Lanczos method.

If not under control, elementary rounding errors can grow and cause a large computational (numerical) error, which can invalidate the whole solution process. Interestingly, this fact was well understood by the founding fathers Von Neumann, Goldstine, Turing, Wilkinson, Forsythe . . . However, it has largely been ignored in most numerical PDE literature.

[Nash, Golub - 90, quote by Parlett], [Babuška - 03], [Oden et al - 03], [Wohlmuth, Hoppe - 99], [Stein(ed) - 03]

Possible consequences of not including computational error in the error analysis of the whole solution process?

- Either the computation of the approximate solution of the algebraic problem consumes unnecessary time and resources due to aiming at **unnecessary high accuracy**,
- Or the computational error which is not under control can impinge the other stages of the solution process and **spoil the numerical solution**.

Efficient and computable a-posteriori error bounds which include algebraic errors [Jiránek, S, Vohralík - 08].

A philosophical difficulty of rounding error analysis - it can not be done **mechanically without a deep knowledge of the analyzed method and algorithm**.

## 3.2 Intermediate quantities and desired accuracy

Do we need in general highly accurate intermediate quantities in order to guarantee a required (high or not) accuracy of the computed final result? No, we do not.

**Surprising observation** Parlett, Wilkinson, see [Parlett - 90]

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 (recall, amplification of rounding errors is not random) can lead to highly accurate final results. Understanding gained via rounding error analysis can **guarantee** final accuracy close to the machine precision level.

Such understanding is based on deep mathematical knowledge about the analyzed method and algorithm.

### **Example:**

The Lanczos method for solving Hermitian eigenvalue problems.

## Principle of the Lanczos method

Ideally, find in steps 1 through  $n$  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  $A$ .

Computationally, in the presence of rounding errors,  $Q_n$  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  $T_n$  may even not represent a matrix of the operator  $A$  projected on the Krylov subspace generated by the computed Lanczos vectors. Most of the entries in  $T_n$  may even not have a single digit of accuracy, i.e.

$$T_n - \tilde{T}_n \text{ can be } \mathbf{large}.$$

Does this mean a total disaster? No! The magic is called **backward error**, and we know it from the work of Wilkinson, Paige and Greenbaum.

For steps 1 to  $n$  of a given Lanczos FP 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  $\hat{Q}_n$  having orthonormal columns such that

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

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, as we will see,  $T_n$  is used for computing the eigenvalues of  $A$  to close to full machine precision!

The bad part of the story is that this remarkable success is not without possible side effects. The eigenvalues of  $A$  are not approximated in the same order and with the same speed as it would be ideally (in exact arithmetic). This is caused by the fact that single eigenvalues can in finite precision Lanczos computation be approximated by multiple computed copies.

In order to prevent the side effects, we must pay the price - here not by computing the intermediate quantities using higher precision, but by applying some correction procedure such as partial reorthogonalization; for an overview see [Parlett - 92].

## 3.3 Computational cost and numerical stability

### **Towards a mathematical foundation of numerical analysis**

– quest for a formal mathematical model of computing with **real numbers**, see [Blum, Cucker, Shub, Smale - 99], [Smale - 97]:

- 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 accuracy (eigenvalue problems, iterative methods in general ...).

**Conclusion:** There are practically no results linking complexity and numerical stability of computing over real numbers.

[Cucker 99]

However, we should keep in mind that in numerical analysis, algorithms are tools for **solving practical problems** - see also [CBSS - 99, p.23], [Iserles - 00], [Baxter, Iserles - 03].

We should consider, that **a practical problem** means some **particular (class of) data**, for which we seek the approximate solution(s). The specific properties of the data (inner correlations etc.) are used in order to get the approximate solution efficiently. Consequently, questions related to a particular problem (including data) are much more specific than worst-case or average-case bounds.

We do not focus on **complexity** and restrict ourselves to the cost of particular computations. There are many results linking **the cost of a particular computation to numerical stability!**

## Short recurrences

1. Loss of orthogonality leads to delay

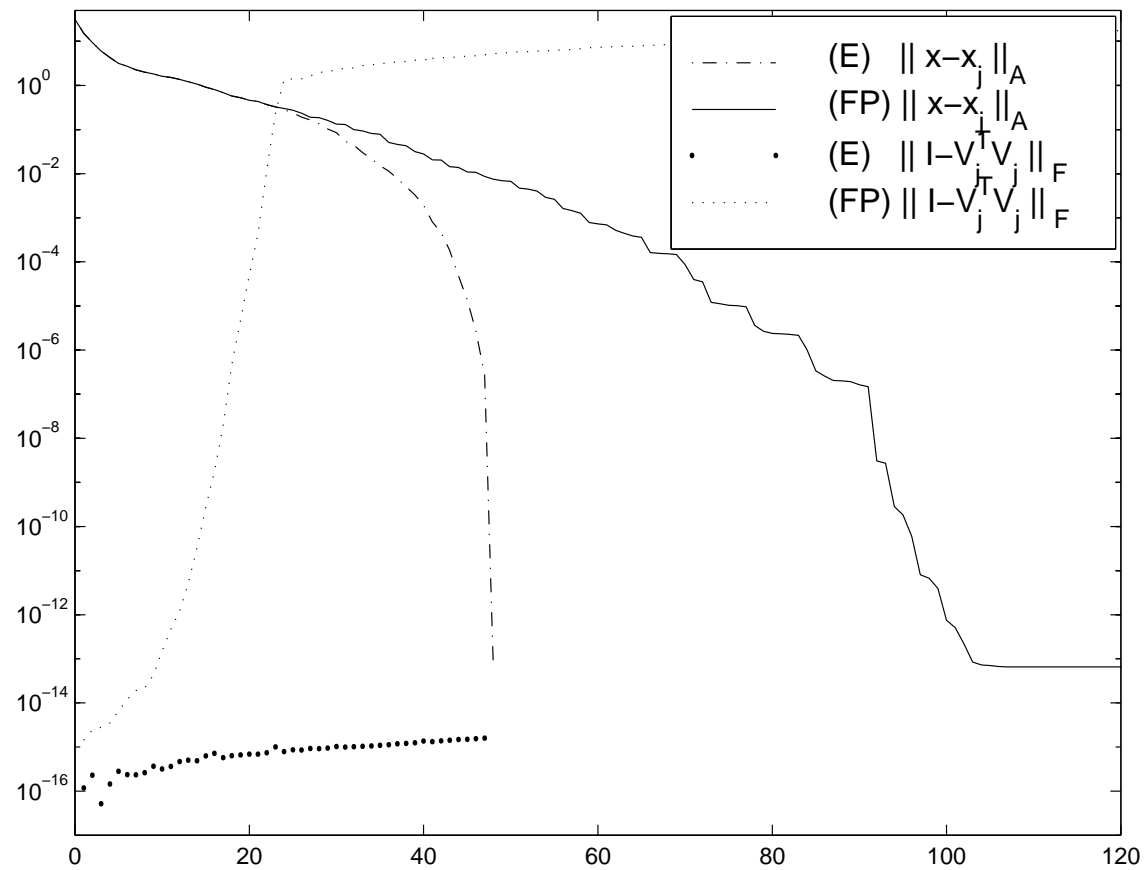
- 1.1 Hermitian systems

- 1.2 Non-Hermitian systems

2. Maximal attainable accuracy

3. Measuring convergence in FP computations

**Example:** In finite precision conjugate gradients orthogonality is lost, convergence is delayed and final accuracy is limited



## 4.1 Loss of orthogonality leads to delay

In the Hermitian case, the underlying basis (though it may not be normalized) is the **Lanczos basis**. Therefore we must study rounding error effects in computing Lanczos vectors.

Lanczos vectors are computed using a three-term recurrence, or, possibly, using two coupled two-term recurrences. Consequently, orthogonality (even linear independence) may be lost quickly. For a long time it was concluded that loss of orthogonality meant also loss of all elegant mathematical structure of orthogonal polynomials (and Gauss quadrature) which could not be extended to computational behavior of the Lanczos process.

However, [Paige - 71, 76, 80]: **Loss of orthogonality follows a regular structure**, which can be revealed!

In finite precision computation

$$AQ_n = Q_n T_n + \beta_{n+1} q_{n+1} e_n^T + F_n, \quad \|F_n\| \leq n^{1/2} \|A\| \varepsilon_1.$$

$Q_n^T Q_n \neq I$ ,  $T_n$  computed by FP  $L(A, q_1)$  may be far from the theoretical counterpart.

$$T_n = S_n \operatorname{diag}(\theta_j^{(n)}) S_n^*, \quad S_n = [s_1^{(n)}, \dots, s_n^{(n)}], \quad s_j^{(n)} = \begin{pmatrix} s_{1j}^{(n)} \\ \vdots \\ s_{nj}^{(n)} \end{pmatrix}$$

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

$s_{nj}^{(n)}$  bottom element - approx. bound,  $\delta_{nj} = \beta_{n+1} |s_{nj}^{(n)}|$ ,

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

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

## Accuracy of the Ritz values computed in the Lanczos method

Exact arithmetic:  $\min_l |\lambda_l - \theta_j^{(n)}| \leq \|Az_j^{(n)} - \theta_j^{(n)}z_j^{(n)}\| \leq \delta_{nj}$ .

Finite precision arithmetic:

$$\min_l |\lambda_l - \theta_j^{(n)}| \leq \frac{\|Az_j^{(n)} - \theta_j^{(n)}z_j^{(n)}\|}{\|z_j^{(n)}\|} \leq \frac{(\delta_{nj} + n^{1/2}\|A\|\varepsilon_1)}{\|z_j^{(n)}\|}$$

Due to the loss of orthogonality it can happen  $\|z_j^{(n)}\| \rightarrow 0!$   
The quantity  $\delta_{nj}$  is easy to compute with negligible additional rounding errors. Does it tell anything about convergence of  $\theta_j^{(n)}$  in finite precision computations?

$$|\lambda_i - \theta_j^{(n)}| \leq \max \{2.5(\delta_{nj} + n^{1/2} \|A\| \varepsilon_1), (n+1)^3 \|A\| \varepsilon_2\},$$

$$\|z_j^{(n)} - (z_j^{(n)}, u_i) u_i\| \leq \frac{(\delta_{nj} + n^{1/2} \|A\| \varepsilon_1)}{\min_{l \neq i} |\lambda_l - \theta_j^{(n)}|}$$

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

Fascinating result! Result of FP computation **verified at no cost!** Please notice that without the theory developed by Paige, the ideal relations would imply nothing about the result of FP computations!

Bounds for  $|s_{nj}^{(n)}|$ ,  $\delta_{nj}$ ? [Parlett - 80], [Greenbaum, S - 90]

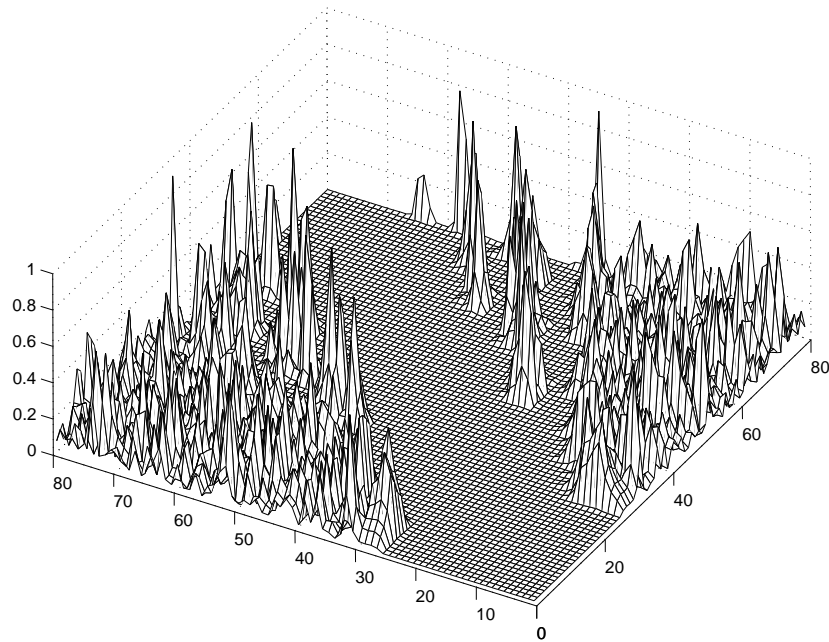
Loss of orthogonality among the Lanczos vectors? **Paige:**

$$|(z_j^{(n)}, q_{n+1})| = \frac{|\varepsilon_{jj}^{(n)}|}{\delta_{nj}}, \quad |\varepsilon_{jj}^{(n)}| \leq n \|A\| \varepsilon_2.$$

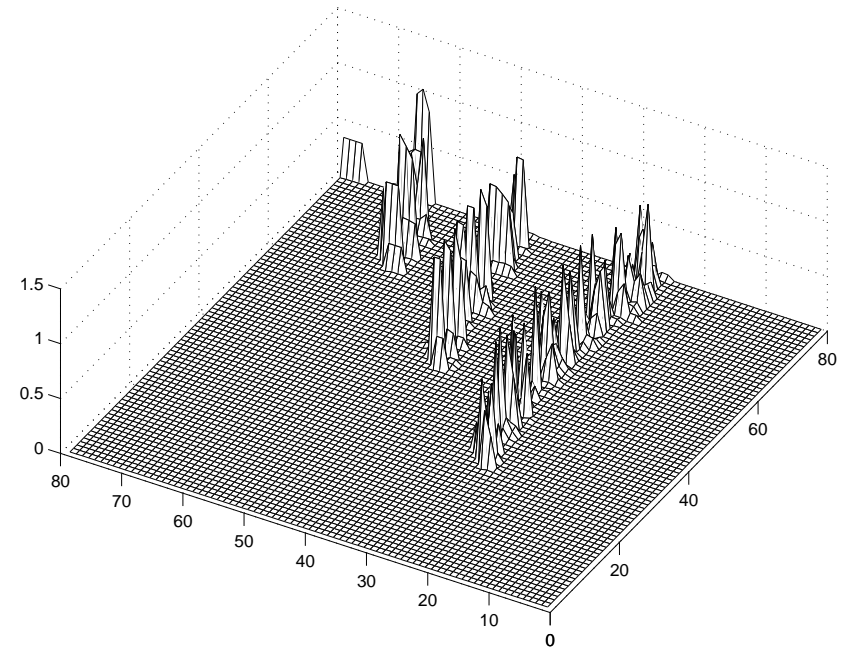
As long as there is no converged Ritz value, orthogonality must be well preserved. If the orthogonality among  $z_j^{(n)}$ ,  $q_{n+1}$  is lost, then  $\theta_j^{(n)}$  have converged to some  $\lambda_i$ .

New related results

[Wülling - 04, 05?], [Zemke - 04], [Meurant - 05?]



Orthogonality among the Lanczos vectors.



Orthogonality of the  $(n + 1)$ st Lanczos vector (right) against the Ritz vectors from the  $n$ th step (left).

Other work on the loss of orthogonality:

- In [Grcar - 81, never published] Forward analysis - when the forward error of the computed Lanczos vectors is not exceeding  $\sqrt{\varepsilon}$ , the computed Krylov subspace is correct to the  $\varepsilon$  level (the error is largely within the exact subspace). This was called **projection property**. In order to maintain the projection property, Grcar suggested **periodic reorthogonalization**. It makes sense only until the forward error is below  $\sqrt{\varepsilon}$ . Not a formal mathematical theory.
- **Berkeley**, Under the influence of Parlett, Kahan, see [Parlett - 80, 92, 94], [Parlett, Reid - 81], [Greenbaum, S - 92], [S, Greenbaum - 92]

- In [Parlett, Scott - 79]: Maintaining the strong linear independence of the Lanczos vectors - **semiorthogonality**. Orthogonalize only against converged Ritz vectors (when  $\delta_{nj} \approx \|A\|\sqrt{\varepsilon}$ ).
- In [Scott -79]: Ideally, for any matrix  $A$  there is always a starting vector  $q_1$  such that the Lanczos method **does not converge to any eigenvalue until the last step**. Construction - Ritz values at step  $n - 1$  prescribed as the midpoints of the intervals given by the eigenvalues.

Works also **computationally** (from experiments). Consequence: Rounding error amplification can strongly depend on the **initial vector!**

- In [Simon - 84, 84]: Monitoring semiorthogonality via simple scalar recurrence, **partial reorthogonalization**. Semiorthogonality ensures, that the computed matrix  $T_n$  represents, up to the terms  $\approx \|A\|_\varepsilon$ , the orthogonal projection of  $A$  onto the computed Krylov subspace.
- In [Parlett - 92]: Full reorthogonalization makes sense only until the semiorthogonality is maintained.
- Tight clusters of eigenvalues - [Dhillon - 97], [Parlett - 96], [Ye - 95], [Dhillon, Parlett - 04, 04]

## Delay of convergence: Backward error - like analysis of the symmetric Lanczos and CG

[Greenbaum - 89]

Finite precision behavior is explained using exact precision results for a larger problem.

It uses the relationship between Lanczos method, Jacobi matrices and Orthogonal polynomials.

**First**, recall that any  $n$  distinct points  $\{\theta_j^{(n)}\}_{j=1}^n$  with weights  $\{\omega_j^{(n)}\}_{j=1}^n$ ,  $\omega_j^{(n)} > 0$ ,  $\sum_{j=1}^n \omega_j^{(n)} = 1$ , define the unique set of monic polynomials

$$1, \psi_1, \dots, \psi_n$$

orthogonal with respect to the innerproduct

$$(\varphi, \psi)_n = \sum_{j=1}^n \omega_j^{(n)} \varphi(\theta_j^{(n)}) \psi(\theta_j^{(n)}).$$

Recall the R-S integrals!

If  $\omega_j^{(n)} = (s_{1j}^{(n)})^2$ , then  $\psi_l$  are the characteristic polynomials of  $T_l$  (Lanczos polynomials), satisfying the minimization property

$$\|\psi_l\|_n = \min \{ \|\psi\|_n, \psi \text{ monic of degree } \leq l \}, \quad l = 1, \dots, n.$$

Please notice the interpretation of top elements of  $T_n$ 's eigenvectors.

Selection of related work: [Karlin, Shapley - 53], [Fischer, Freund - 93], [Freund, Hochbruck - 93], [Golub, S - 94], [Golub, Meurant - 94, 97], [Gautschi - 03], ...

**Second**, please notice that **exact or FP**  $L(A, q_1)$  generates in steps 1 to  $n$  a sequence  $T_1 - T_n$  which is exactly the same as in **exact**  $L(B, p_1)$ ,

$$B = V \text{diag} (\theta_j^{(n)}) V^*, \quad V^*V = I, \quad p_1 = V \left( s_{11}^{(n)}, s_{12}^{(n)}, \dots, s_{1n}^{(n)} \right)^T,$$

e.g., for  $V \equiv S_n$ ,  $p_1 = e_1$ ,  $B \equiv T_n$ ,  $T_n$  is generated by the exact  $L(T_n, e_1)$ .

**FP Lanczos in steps 1 to  $n$   $\rightarrow$  Exact Lanczos**

**[Greenbaum - 89] much stronger:**

Let  $J$  steps of FP  $L(A, q_1)$  produce  $T_J$ . Then  $T_J$  is generated in  $J$  steps of **exact** Lanczos algorithm applied to some  $\bar{A}_J, \bar{q}_J^1$ .  $\bar{A}_J$  is of dimension  $N + l(J)$ ; all its eigenvalues lie within tiny intervals about the eigenvalues of  $A$ .

Similarly for the norm of the residuals in the CG method.

[S - 91]: For any eigenvalue of  $A$  there must be at least one eigenvalue of  $\bar{A}_J$  close to it.

Exact distribution of  $\bar{A}_J$ 's eigenvalues depends on the actual rounding errors.

[Greenbaum, S - 92]:

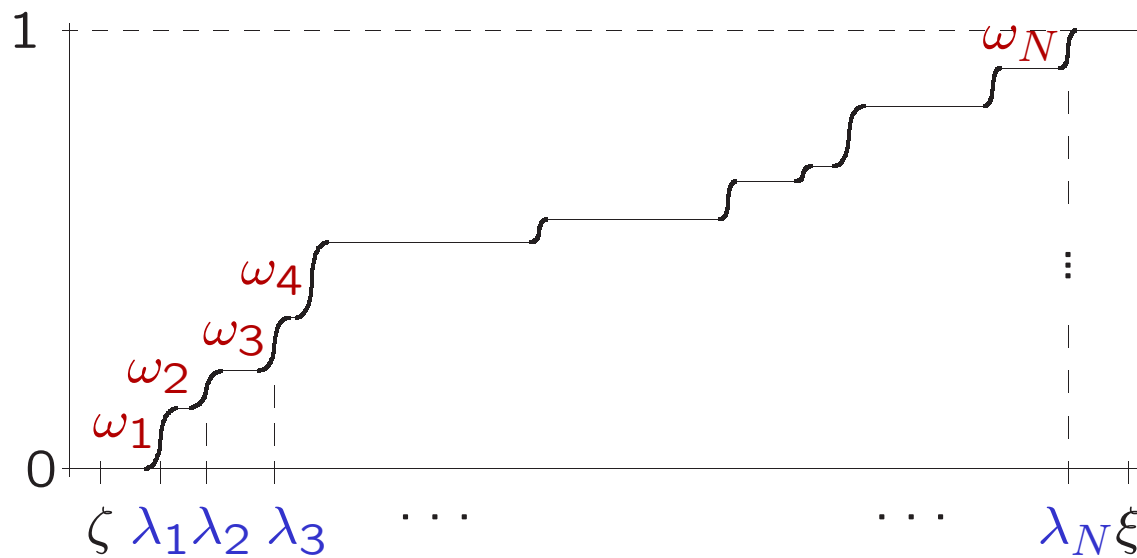
FP  $L(A, q^1)$  and FP  $CG(A, q^1)$  behave **very similarly** as the exact algorithms applied to any  $\hat{A}, \hat{q}^1$  from a certain class  $\hat{A}$  is of dimension  $Nl$ ,

where  $Nl$  eigenvalues are spread throughout tiny intervals about the eigenvalues of  $A$  while each tight cluster has the total weight of the original eigenvalue.

This model is valid for any **reasonable** number of steps, practically no dependence on  $l$  (if sufficiently large), small dependence on the size of intervals.

In terms of the R-S integrals (theory still not finished):

Finite precision Lanczos (CG) is (with some inaccuracy) the matrix formulation of the exact Gauss quadrature of the R-S integral for some blurred distribution function  $\hat{\omega}(\lambda)$ , which represents the spectral decomposition of some infinitely dimensional problem.

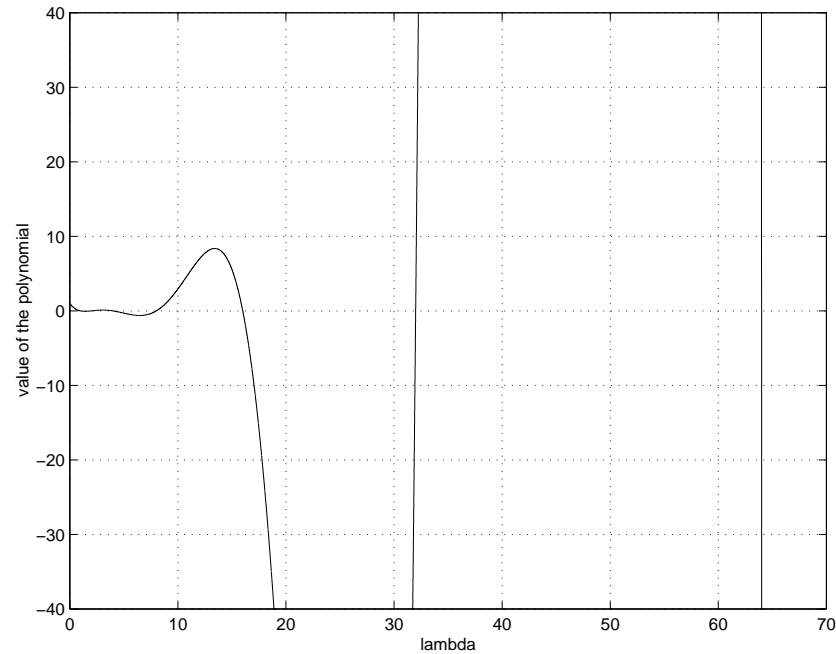


## Consequences:

- a small Hermitian (HPD) perturbation causes only a small change of the Lanczos (CG) behavior.
- Finite precision  $L$  (CG) for  $A, q_1$  corresponds to the exact precision  $L$  (CG) for  $\hat{A}, \hat{q}_1$ .

By applying exact precision theory (convergence bounds) to  $\hat{A}, \hat{q}_1$  we obtain a quantitative description of FP  $L$  (CG) for  $A, q_1$ . [Greenbaum, S - 92], [Notay - 93]

- Approximation to the minimal polynomial is for the case with **individual well-separated eigenvalues** very different from the approximation in the case of **tight clusters of eigenvalues**.



Several close roots are placed in well separated tight clusters due to the minimization property. But it means that, at the given step, we do not have enough Ritz values to approximate some eigenvalues in the other parts of the spectrum; the CG convergence can be for these two cases **very different**.

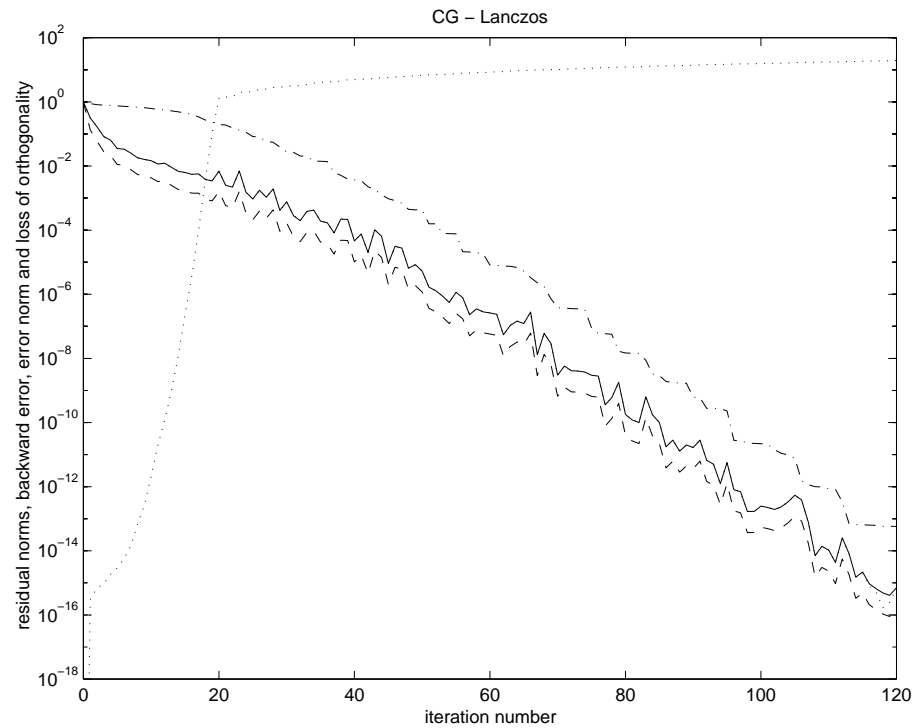
Linear independence of the computed generating vectors is lost with multiple Ritz values. **Delay of convergence:** each loss of linear independence costs one iteration!

iteration  $\dots n$ ,  
dimension of computed  $K_n \dots n - i$ ,  
**delay  $\dots i$ .**

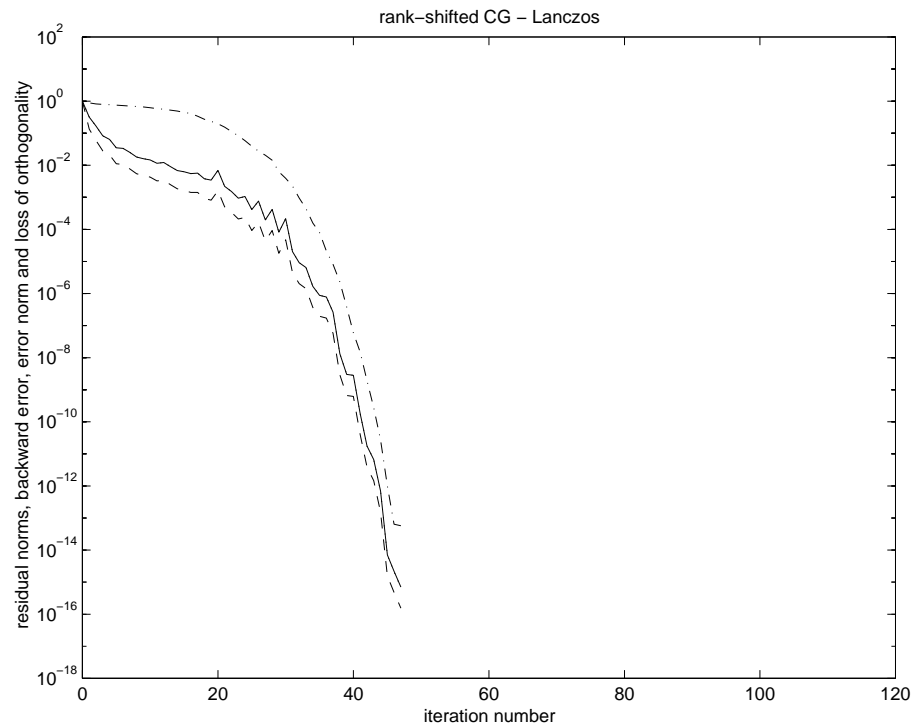
It is extremely difficult to estimate for the given step  $n$

$$\left| \|r_n^{FP}\| - \|r_n\| \right| \leq \|r_n^{FP} - r_n\|$$

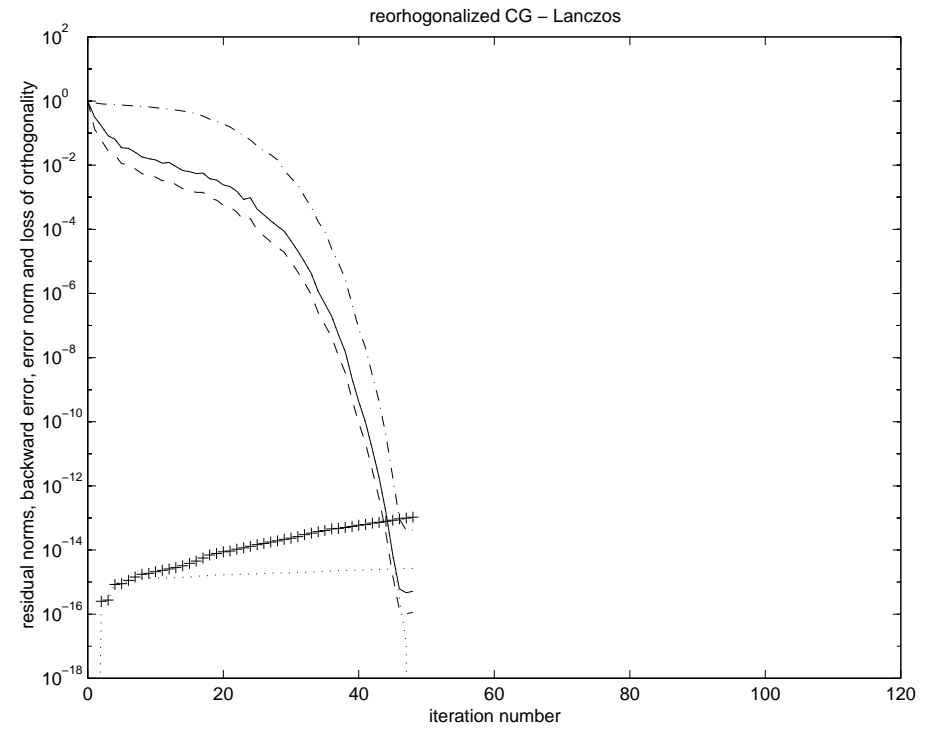
even for HH GMRES (it assumes solving the question about the stability of Krylov subspaces). We have to rotate our view.



For example, for  $n = 95$   $\dim(K_n)$  is not 95, but only 43. When we shift back each point on each convergence curve  $i = i(n)$  steps, we obtain



shifted CG convergence curves, which can be compared with



the double reorthogonalized CG [Paige - 80].

## 4.1.2 Non-Hermitian systems

Small perturbation of  $A$  can cause a large perturbation of the spectrum - the role of eigenvalues?!

- Though some results **formally** correspond to those of Paige, their interpretation must be different. [Bai - 94]
- Similarly the re-biorthogonalization (maintaining semiduality). [Day - 99]
- Backward error - like result?

## 4.2 Maximal attainable accuracy

recursively computed residual  $\times$  true residual  $b - Ax_n$

recursive residuals  $\rightarrow 0$ . Final accuracy?

**Two term recurrences:**

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

$$r_{n+1} = r_n - \omega_n A p_n, \quad p_{n+1} = r_{n+1} + \psi_n p_n$$

[Sleijpen et al. - 94], [Greenbaum - 97]:

$$e_n = (b - Ax_n) - r_n, \quad e_n = e_{n-1} + l_{n-1},$$

where  $l_{n-1}$  counts for the local errors in computing  $x_n, r_n$  from  $x_{n-1}, r_{n-1}$ .

Consequently,

$$e_{n+1} = e_0 + \sum_{j=0}^n l_j,$$

global error is given as the sum of local errors.

**Bound:**

$$\frac{\|e_k\|}{\|A\|\|x\|} \leq \text{const } k \theta_k \varepsilon + \mathcal{O}(\varepsilon^2), \quad \theta_k = \max_{j \leq k} \|x^j\| / \|x^0\|.$$

**Consequence:** Oscillations of the size of the approximate solution may damage the final accuracy (BiCG - like methods!).

Extension to LS: [Björck, Elfving, S - 97].

Backward stability (based on the assumption  $\|r_k\| \rightarrow 0$ ).

## Three-term recurrences - a different story:

Coupled two-term recurrences replaced by

$$\begin{aligned}x_{n+1} &= -(r_n + \alpha_n x_n + \beta_{n-1} x_{n-1})/\gamma_n, \\r_{n+1} &= -(Ar_n + \alpha_n r_n + \beta_{n-1} r_{n-1})/\gamma_n,\end{aligned}$$

where  $\gamma_n = -(\alpha_n + \beta_{n-1})$ .

Examples: Hestenes & Stiefel CG × Rutishauser CG; BiCG  
× BIORes; (QMR variants).

**Observation:** three-term implementations “less stable” than the coupled two-term ones (for nonsingular systems) - the final accuracy can be much worse.

**Explanation:** [Gutknecht, S - 97]

$e_n = (b - Ax_n) - r_n$ ,  $l_{n-1}$  local error analogous to two-term case (not equal!).

$$e_{n+1} = - \left( e_n \frac{\alpha_n}{\gamma_n} + e_{n-1} \frac{\beta_{n-1}}{\gamma_n} + l_n \right)$$

Local errors are potentially amplified by the recurrence.

Global error in terms of local errors - multiplicative factors may become large!

$$\begin{aligned} e_{n+1} &= e_0 - \sum_{j=0}^n l_j \\ &= l_0 \left( \frac{\beta_0}{\gamma_1} + \dots + \frac{\beta_0 \dots \beta_{n-1}}{\gamma_1 \dots \gamma_n} \right) \\ &= l_1 \left( \frac{\beta_1}{\gamma_2} + \dots + \frac{\beta_1 \dots \beta_{n-1}}{\gamma_2 \dots \gamma_n} \right) \\ &\quad \vdots \\ &= l_{n-1} \frac{\beta_{n-1}}{\gamma_n} . \end{aligned}$$

## Example: three term CG (HPD case)

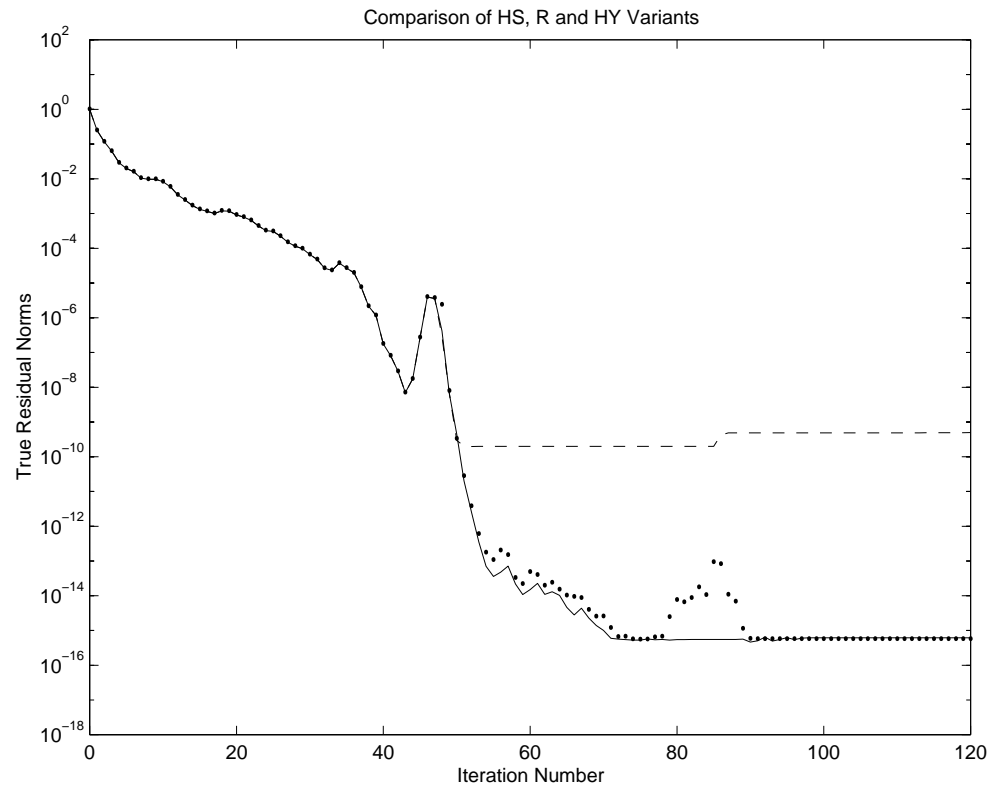
Amplification factors:

$$(1-\vartheta) \frac{1}{\kappa(A)} \frac{\|r^k\|^2}{\|r^{i-1}\|^2} \leq \prod_{j=i}^k \frac{\beta_{j-1}}{\gamma_j} \leq (1+\vartheta) \kappa(A) \frac{\|r^k\|^2}{\|r^{i-1}\|^2}, \quad \vartheta \ll 1.$$

Note: holds for the [computed values](#);

here [Greenbaum - 89], [Greenbaum, S - 92] results used.

**Consequence:** Oscillations of the size of the [recursive residuals](#) may extensively damage the final accuracy.



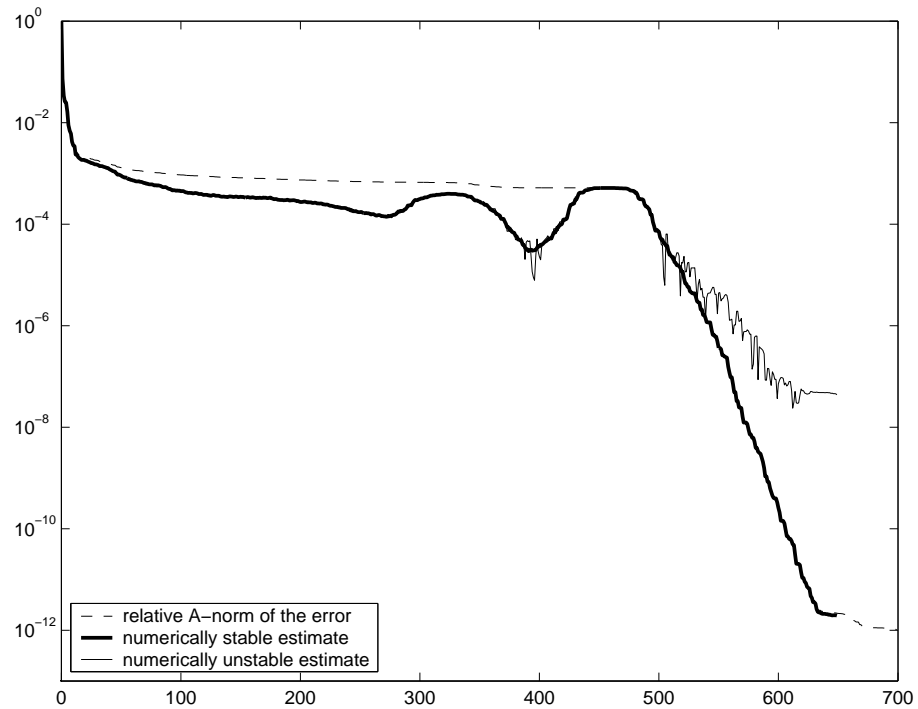
The CG relative residual can indeed very strongly oscillate!

## 4.3 Measuring convergence in FP CG

We present a CG example, matrix s3rmt3m3 from the Cyllshell collection by Reijo Kouhia, incomplete Choleski preconditioner. Ideally (in exact arithmetic)

$$\text{EST}^2 = \sum_{l=n}^{n+d-1} \gamma_l \|r_l\|^2 = r_0^T (x_{n+d} - x_n).$$

Computationally, though the second estimate is evaluated accurately, it gives misleading information.



For the numerically unstable estimate, the identity is in finite precision computations **not valid**. Rounding error analysis is fundamental, it should not be ignored!

## Among the issues not covered:

- Breakdowns and their influence [Van Den Eshof - 03];
- Closeness to singularity and incompatible systems;
- Can short recurrences produce a well-conditioned basis? (interpretation of look-ahead techniques in FP computations);
- Inaccurate Krylov subspace methods [Sleijpen et al. - 02], [Simoncini, Szyld - 02].

# Lecture VIII

## NUMERICAL BEHAVIOUR

### II.

1. Orthogonality and numerical stability
2. Householder GMRES
3. Modified Gram-Schmidt GMRES
4. What should follow

## 5.1 Orthogonality and numerical stability

[Liesen, Rozložník, S - 02]: The choice of the subspace (basis) which is used in computations can have a fundamental impact.

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

## The straightforward approach

Based on the orthonormal basis of  $AK_n(A, r_0)$ . Define  $w_1 \equiv Ar_0/\|Ar_0\|$ ,  $v_1 = r_0/\|r_0\|$ . Then the recursive columnwise  $QR$ -factorization yields

$$[Av_1, AW_{n-1}] = A[v_1, W_{n-1}] \equiv W_n R_n,$$

$$W_n \equiv [w_1, \dots, w_n], \quad W_n^T W_n = I_n,$$

$$\text{span} \{w_1, \dots, w_n\} = AK_n(A, r_0),$$

$$\kappa([v_1, W_{n-1}]) \leq \kappa(R_n) \leq \kappa(A) \kappa([v_1, W_{n-1}]).$$

Using the orthonormal basis  $w_1, \dots, w_n$  of  $AK_n(A, r_0)$ :

$$x_n = x_0 + [v_1, W_{n-1}] t_n \in x_0 + \mathcal{K}_n(A, r_0);$$

$$\Rightarrow r_n = r_0 - A [v_1, W_{n-1}] t_n = r_0 - W_n R_n t_n;$$

$$\Rightarrow t_n = (W_n R_n)^+ r_0 = R_n^{-1} W_n^T r_0.$$

How does this affect the numerical stability?

## Theorem

$$\frac{\|r_n\|}{\|r_0\|} = \sigma_{n+1}([v_1, W_n]) \sigma_1([v_1, W_n]) = \frac{2\kappa([v_1, W_n])}{\kappa([v_1, W_n])^2 + 1},$$

$$\frac{\|r_0\|}{\|r_n\|} \leq \kappa([v_1, W_n]) \leq 2 \frac{\|r_0\|}{\|r_n\|},$$

$$\frac{\|r_0\|}{\|r_n\|} \leq \kappa(R_n) \leq 2\kappa(A) \frac{\|r_0\|}{\|r_n\|}.$$

Consequently,  $\kappa(R_n)$  must inevitably increase as  $\|r_n\|$  decreases, even for small  $\kappa(A)$  and with the most stable way of computing  $w_1, \dots, w_n$ .

$\Rightarrow$  Computation of  $t_n = R_n^{-1} W_n^T r_0$  is inherently unstable!

**Surprise:** Numerical behaviour gets worse when orthogonality of  $w_1, \dots, w_n$  is maintained better; Householder implementation performs worse than Modified-Gram-Schmidt implementation.

The straightforward approach is used in “Simpler GMRES” [Walker, Lu Zhou - 94], and it is related to other implementations, e.g. Orthodir.

## Classical GMRES implementation

Based on the orthonormal basis of  $\mathcal{K}_n(A, r_0)$ . Let  $v_1 \equiv r_0/\|r_0\|$ .  
Then the Arnoldi process yields

$$AV_n = V_{n+1} H_{n+1,n},$$

$$V_n \equiv [v_1, \dots, v_n], \quad V_n^T V_n = I_n,$$

$$\text{span} \{v_1, \dots, v_n\} = \mathcal{K}_n(A, r_0),$$

$$\kappa(H_{n+1,n}) \leq \kappa(A).$$

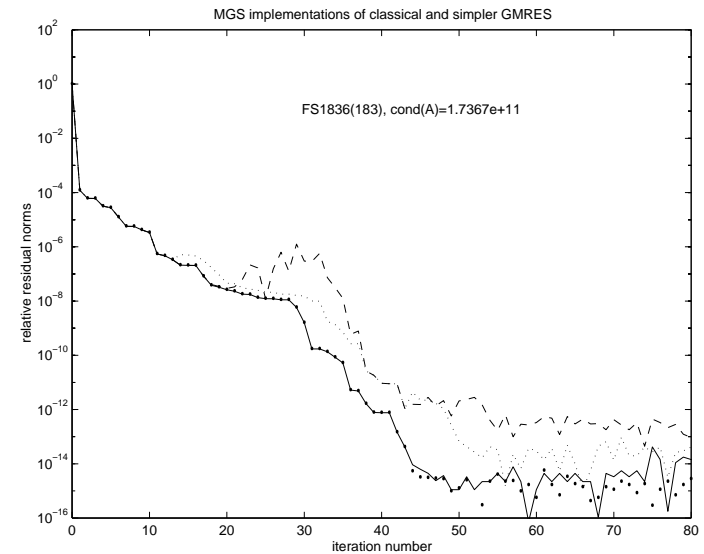
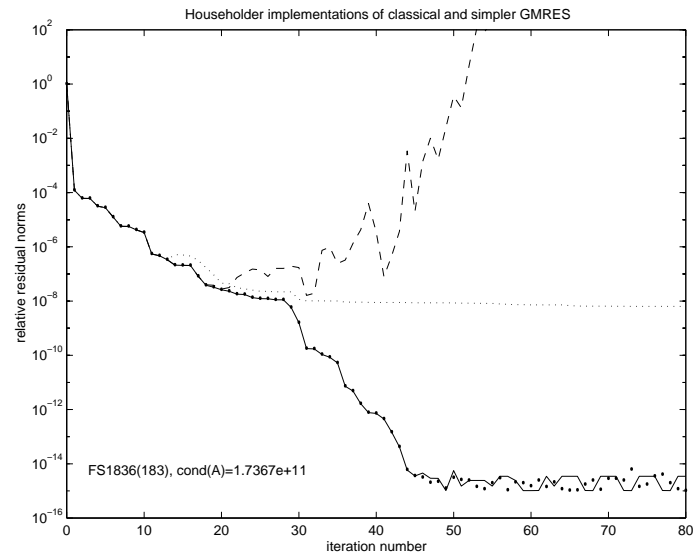
Using the orthonormal basis  $v_1, \dots, v_n$  of  $\mathcal{K}_n(A, r_0)$ :

$$x_n = x_0 + V_n z_n \in \mathcal{K}_n(A, r_0),$$

$$\Rightarrow r_n = r_0 - AV_n z_n = r_0 - V_{n+1} H_{n+1,n} z_n,$$

$$\Rightarrow z_n = (V_{n+1} H_{n+1,n})^+ r_0 = (H_{n+1,n})^+ V_{n+1}^T r_0.$$

How does this affect the numerical stability?



Householder implementations (residual and backward error). Excessive ill-conditioning of computed  $R_n$  leads in the straightforward implementation to divergence.

MGS implementations (residual and backward error). Due to rounding errors the identity is violated, and the computed  $R_n$  is not so badly ill-conditioned as it **ideally** should be!

## Moral

The choice of the right subspace (basis) is fundamental for numerical stability of the Krylov subspace methods.

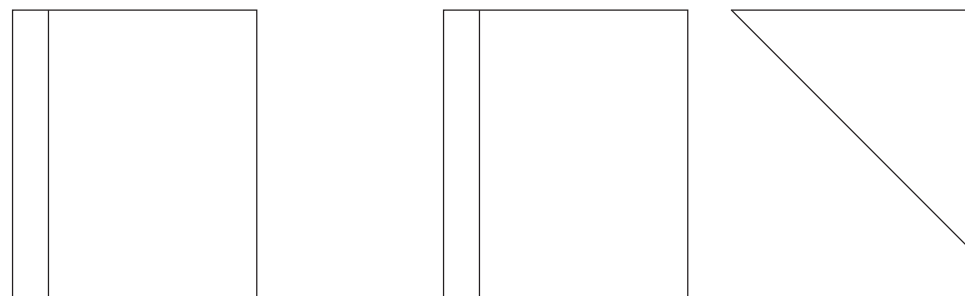
Even the best orthogonalization technique in computing the basis (here Householder reflections) can not compensate for instabilities artificially created due to a bad choice of the subspace. Paradoxically - preserving orthogonality of the computed basis can even **make things worse!**

In the rest of the text we concentrate on classical GMRES.

Numerical stability of the GMRES implementation based on the (ideally) orthonormal basis of  $K_n(A, r_0)$  – many related publications, e.g.

[Björck - 67], [Björck, Paige - 92], [Karlson - 91], [Arioli, Fassino - 96], [Drkošová, Greenbaum, Rozložník, S - 95], [Greenbaum, Rozložník, S - 96], [Rozložník 1997], [Paige, S - 02, 02, 02], [Giraud, Langou, Rozložník, van der Eshof - 05], [Rozložník, Paige, S - in progress]

Arnoldi process  $\equiv$  recursive columnwise QR decomposition

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


## 5.2 Householder GMRES

Implementation using Householder reflections was suggested in [Walker 1988, 89].

A tedious but straightforward proof in [DGRS -95]:

Householder - reflections based implementation of GMRES computes **a backward stable approximate solution.**

## 5.3 Modified Gram-Schmidt GMRES

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

**A common general belief:** MGS orthogonalization is a good compromise between propagation of errors (**loss of orthogonality**) and algorithm efficiency (**computational cost**). Price - the computation **is recursive!**

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 - 92];
- Our case: MGS GMRES.

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

### Statement:

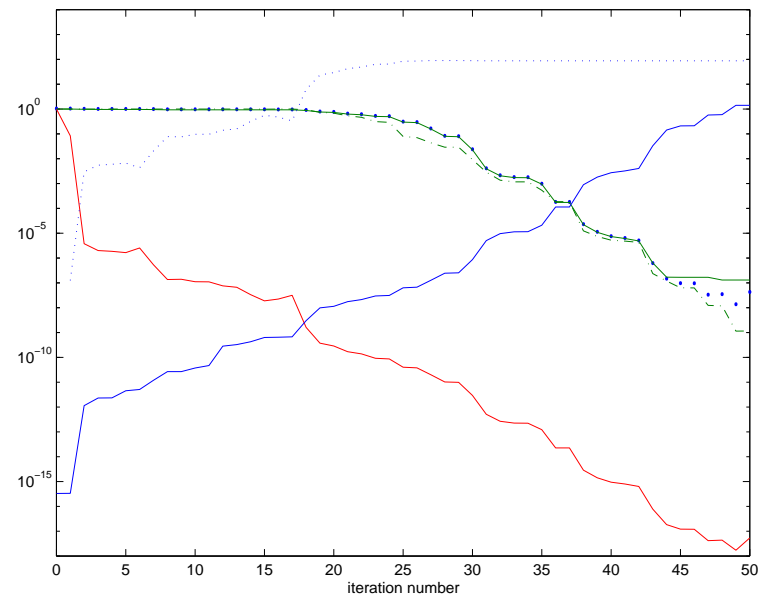
**Loss of orthogonality**  $\|I - V_{n+1}^* V_{n+1}\|_F$  in the modified Gram–Schmidt Arnoldi process

is inversely proportional

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

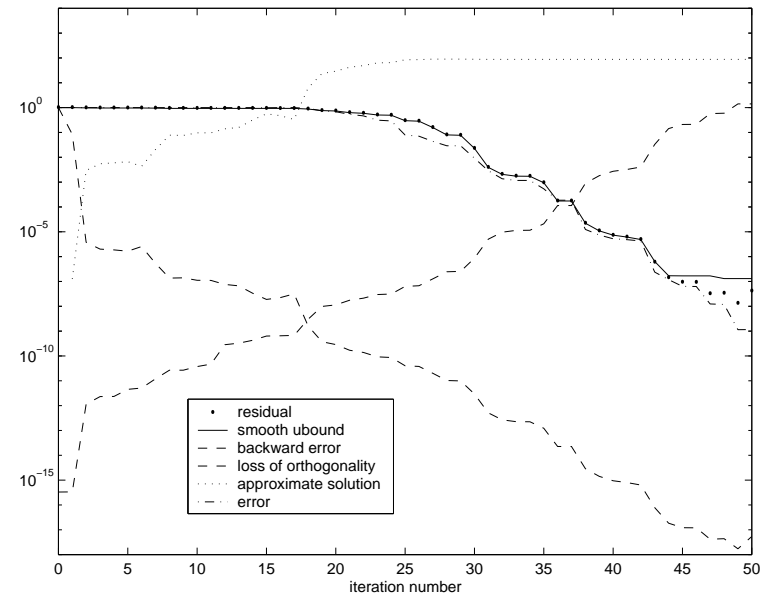
$$\frac{\|b - Ax_n\|}{\|b\| + \|A\|\|x_n\|}.$$

FS1836,  $b = \text{ones}$ ,  $x_0 = 0$

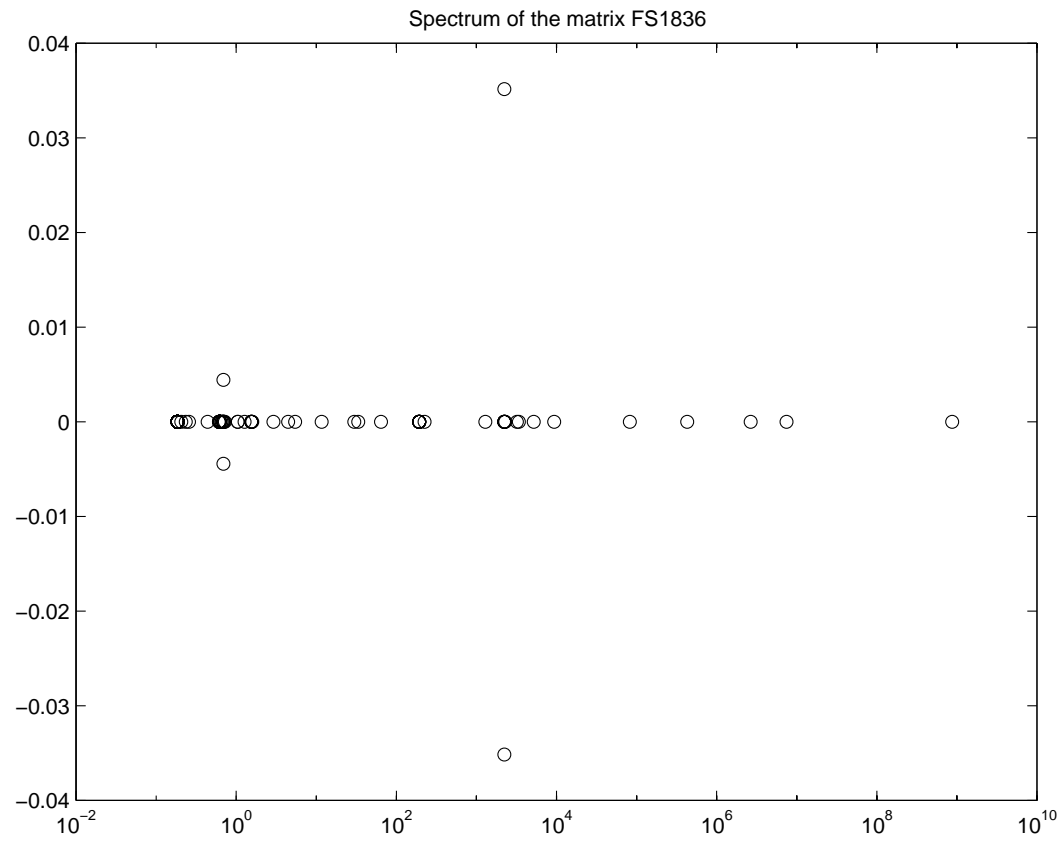


See the difference between the relative residual and backward error (it is interesting to see the same for  $x_0 \neq 0$ ).

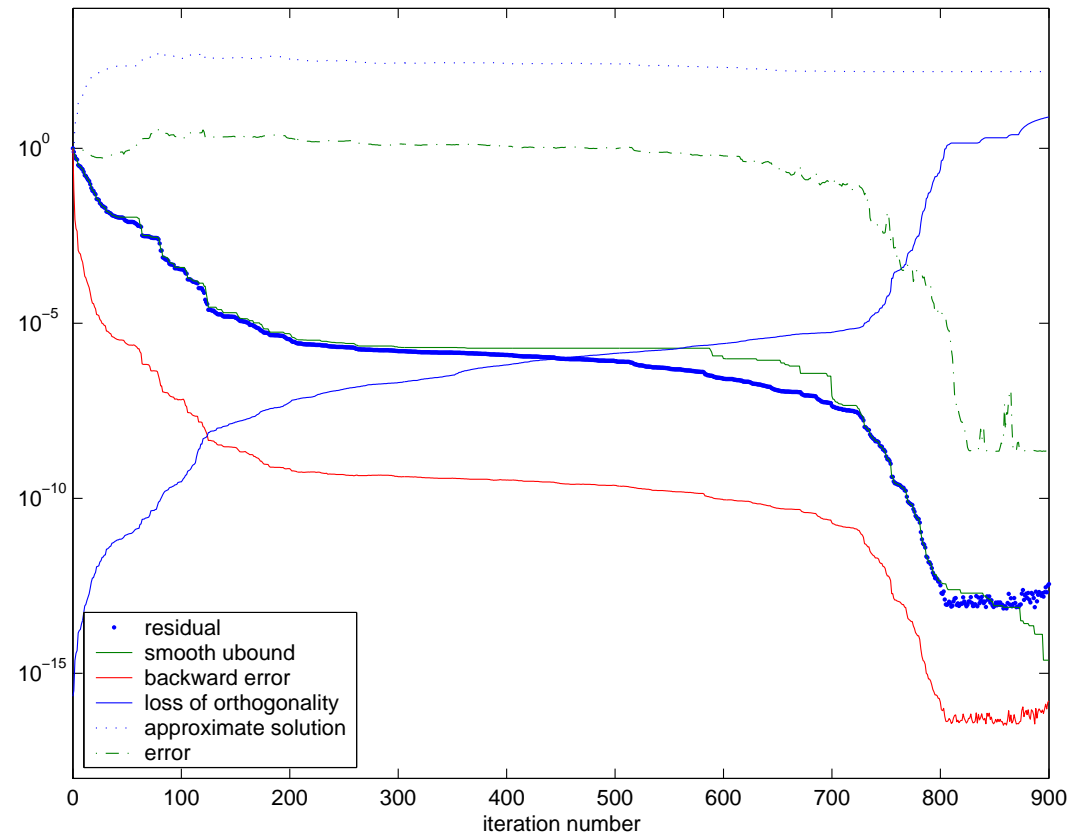
FS1836,  $b = \text{ones}$ ,  $x_0 = \text{randn}$



link the spectrum with convergence ?



## Sherman 2, $b$ MM, $x_0 = 0$

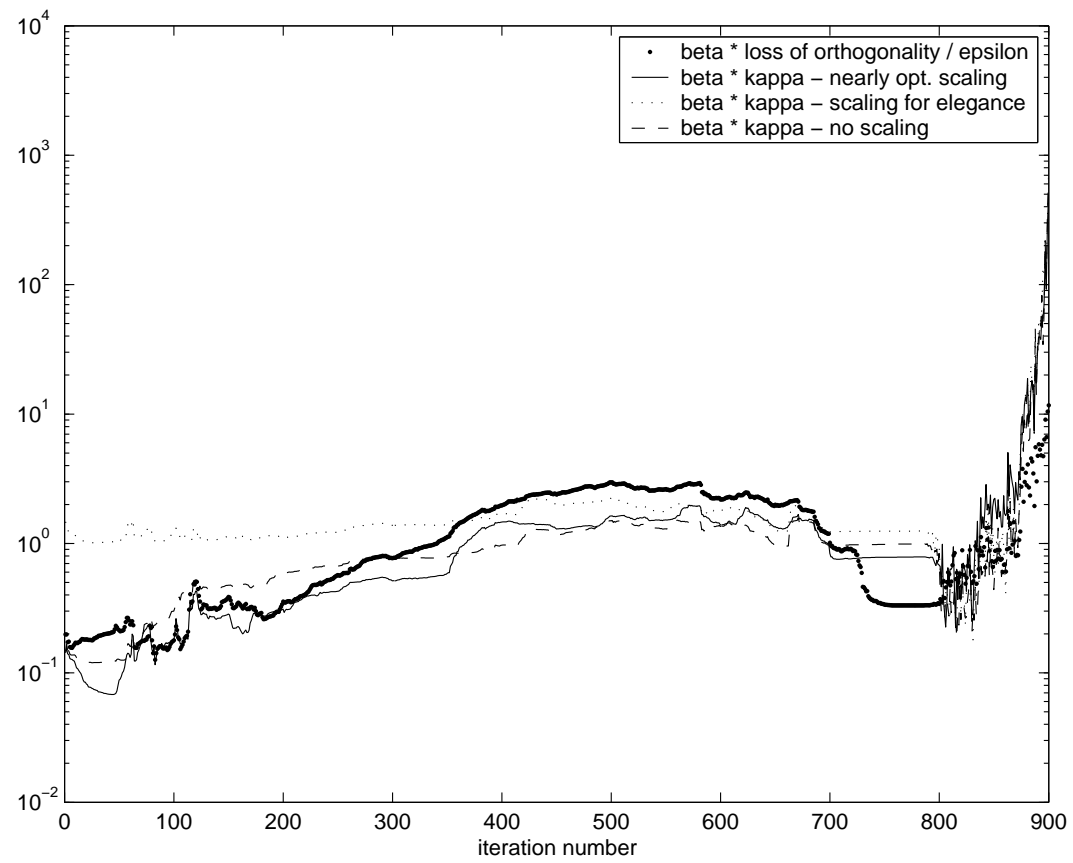


**Proof?** A long and yet unfinished story, links with **Scaled Total Least Squares** and other strange problems.

[Paige, Rozložník, S - 07]

**MGS GMRES is mbackward stable !**

## Sherman 2, $b$ MM, $x_0 = 0$



## 5.4 What should follow

Loss of orthogonality in MGS Arnoldi

$$\|I - V_{k+1}^T V_{k+1}\|_F \approx \kappa([r_0 \gamma, AV_k D_k]) \mathcal{O}(\varepsilon).$$

Loss of orthogonality in CGS Arnoldi (could be based on recent result [GLRvdE - 05])

$$\|I - V_{k+1}^T V_{k+1}\|_F \approx \kappa([r_0 \gamma, AV_k D_k])^2 \mathcal{O}(\varepsilon).$$

Numerical stability of **restarted CGS GMRES**.

# Closure

## Repeatedly appearing motives:

- Nonlinear problems in numerical linear algebra;
- Backward stability;
- Intermediate quantities and accuracy of final results;
- Theoretical results link computational cost and numerical stability;

- 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.

**Possible work (in progress):**

- RS integral model of FP CG and Lanczos;
- Instability of Krylov subspaces;
- Core problem in FP arithmetic, its use for regularization;
- Regularization property of Ksp methods;
- Using GMRES parametrization in [Arioli, Pták, S - 98] for mapping the sets of problems;
- Restarted CGS GMRES ...

The goal is **understanding**, and putting it into service of practical computations.

**From practical point of view, we need**

- Analytical theory of acceleration (preconditioning),
- Reliable and efficient stopping criteria,
- Theoretical justification of numerical stability.

## Selected references

The presented papers can be downloaded from <http://www.cs.cas.cz/strakos>, or obtained upon request from the author. They represent a complementary material for the course, and most of them contain an extensive list of references, which can be used for further reading.

Z.S. and J. Liesen, *On numerical stability in large scale numerical computations*, *Z. Angew. Math. Mech.* 85, pp. 307–325, 2005.  
*Relevant to parts II, VII, VIII.*

*P. Jiránek, Z.S. and M. Vohralík*, A posteriori error estimates including algebraic error: computable upper bounds and stopping

criteria for iterative solvers, submitted to SIAM J. Sci. Comput., 2008. Relevant to part II.

C. C. Paige and Z. S., *Core problems in linear algebraic systems, SIAM J. Matrix Anal. Appl.* 27, pp. 861–875, 2007. Relevant to part II.

C. C. Paige and Z. S., Scaled total least squares fundamentals, *Numer. Math.* 91, pp. 117–146, 2002. Relevant to part II.

I. Hnětynková, and Z. S., *Lanczos tridiagonalization and core problems, Linear Algebra Appl.* 421, pp. 243–251, 2007. Relevant to parts II, V.

*Z. S. and Petr Tichý, On efficient numerical approximation of the scattering amplitude of  $c^*A^{-1}b$  via matching moments, submitted to SIAM J. Sci. Comput., 2009. Relevant to parts III, IV, V.*

*Z. S., Model reduction using the Vorobyev moment problem, Numerical Alg., published electronically in September 2008. Relevant to parts III, IV, V.*

*J. Liesen and Z. S., On optimal short recurrences for generating orthogonal Krylov subspace bases, SIAM Review, 50, pp. 485-503, 2008. Relevant to part III.*

*D. P. O'Leary, Z. S. and Petr Tichý, On sensitivity of Gauss-Christoffel quadrature, Num. Math. 107, pp. 147-174, 2007. Relevant to parts IV, V.*

*G. Meurant and Z. S., The Lanczos and conjugate gradient algorithms in finite precision arithmetic, Acta Numerica 15, pp. 471–542, 2006. Relevant to parts IV, V, VII, VIII.*

*Z. S. and P. Tichý, On Error Estimation in the Conjugate Gradient Method and Why It Works In Finite Precision Computations, Electronic Trans. Numer. Anal. (ETNA) 13, pp. 56-80, published online, 2002. Relevant to parts IV, V, VII.*

*Z. S. and P. Tichý, Error Estimation in Preconditioned Conjugate Gradients, BIT Numerical Mathematics, 45, pp. 789-817, 2005. Relevant to parts V, VII.*

*M. Arioli, V. Ptak and Z. S., Krylov Sequences of Maximal Length and Convergence of GMRES, BIT 38, pp. 636–643. 1998. Relevant to part VI.*

*A. Greenbaum, V. Pták, and Z. S., Any Convergence Curve is Possible for GMRES, SIAM Matrix Anal. Appl. 17, pp. 465-470, 1996. Relevant to part VI.*

*J. Liesen and Z. S., GMRES Convergence Analysis for a Convection-Diffusion Model Problem, SIAM J. on Sci. Comput., 26, pp. 1989-2009, 2005. Relevant to part VI.*

*J. Liesen and Z. S., Convergence of GMRES for Tridiagonal Toeplitz Matrices, SIAM J. Matrix Anal. Appl., 26, pp. 233-251, 2004. Relevant to part VI.*

*M. H. Gutknecht and Z. S., Accuracy of Two Three-Term and Three Two-Term Recurrences for Krylov Space Solvers, SIAM*

*J. Matrix Anal. Appl.* 22, pp. 213–229, 2001. Relevant to parts VII, VIII.

*J. Liesen, M. Rozložník and Z. S.*, Least Squares Residuals and Minimal Residual Methods, *SIAM J. Sci. Comput.* 23, pp. 1503–1525, 2002. Relevant to parts VII, VIII.

C.C. Paige, M. Rozložník and Z. S., *Modified Gram-Schmidt (MGS), Least Squares, and Backward Stability of MGS-GMRES*, *SIAM J. Matrix Anal. Appl.* 28, pp. 264–284, 2006. Relevant to parts VII, VIII.