Chapter three

Numerical behavior

– 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.

Work in this direction will have to be done. Possible candidates for trouble? Mesh refinements close to singularity.

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.

$$T_n - T_n \quad \text{can be} \quad \textbf{large.}$$

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!

*

Chapter four

Numerical behavior

– Short recurrences

1. Loss of orthogonality leads to delay

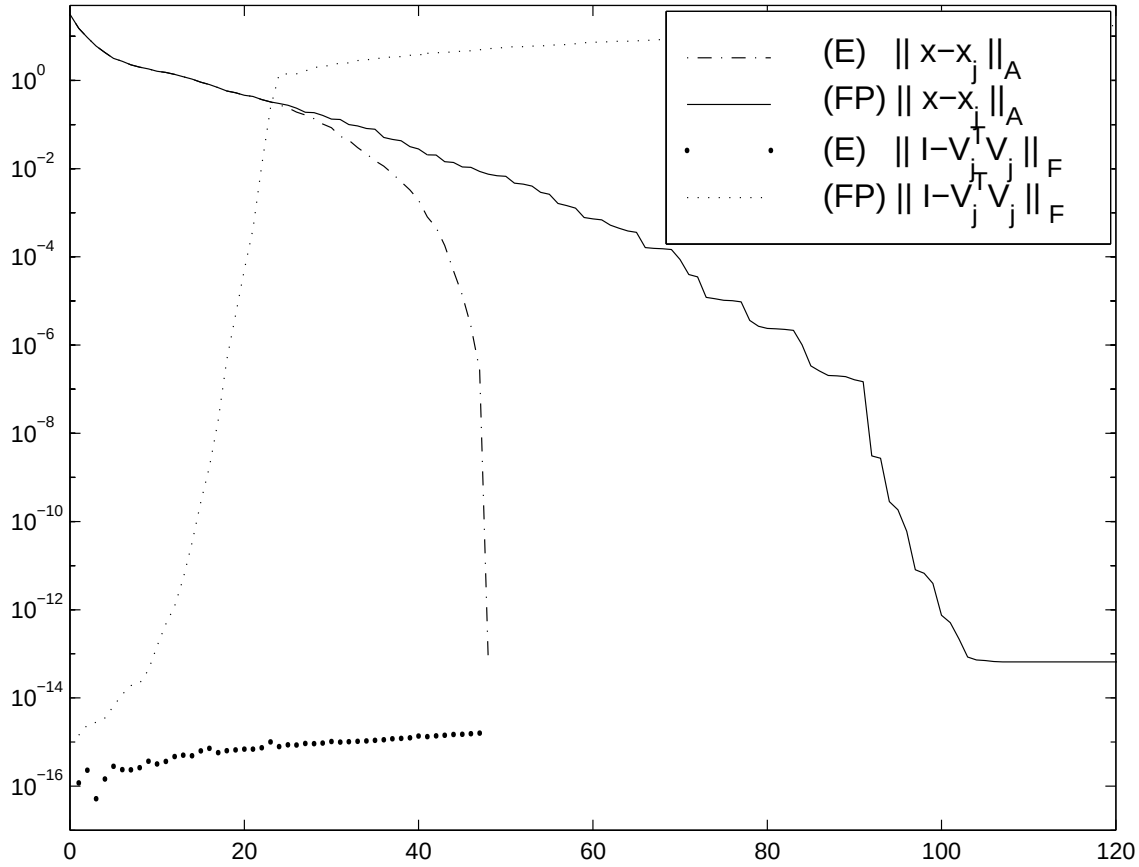
   1.1 Hermitian systems

   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

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

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

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

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

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

$\theta_j^{(n)}$ *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 $\|z_j^{(n)}\| \to 0$!
The quantity δ_{nj} is easy to compute with negligible additional rounding errors. Does it tell anything about convergence of $\theta_j^{(n)}$ in finite precision computations?

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

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

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

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!

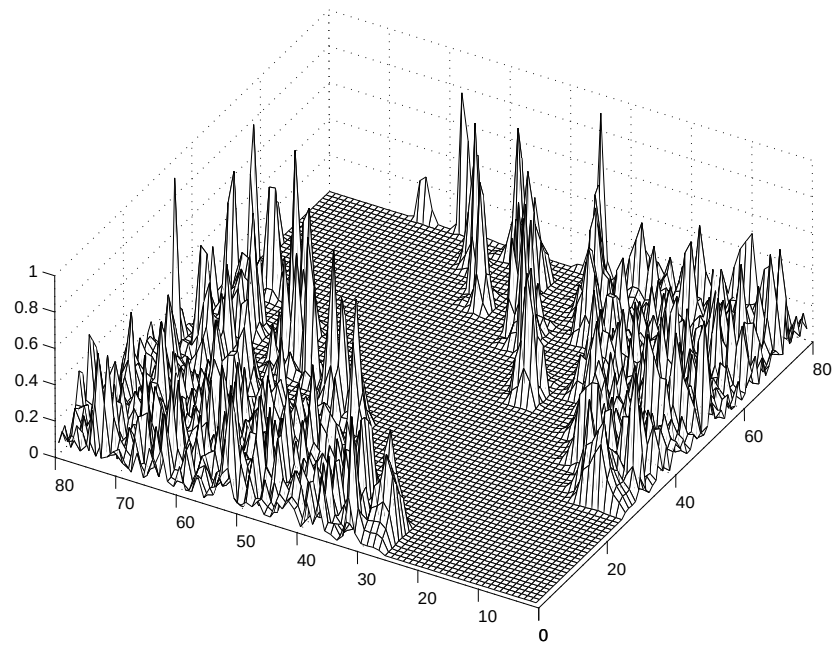
Bounds for $|s_{nj}^{(n)}|$, δ_{nj}? [Parlett - 80], [Greenbaum, S - 90]

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

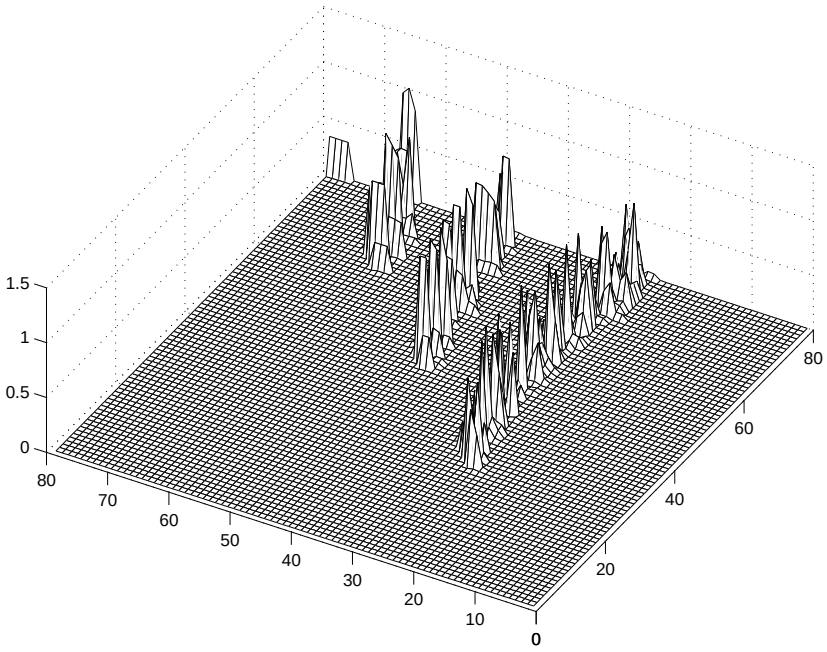
$$|(z_j^{(n)}, q_{n+1})| = \frac{|\varepsilon_{jj}^{(n)}|}{\delta_{nj}}, \quad |\varepsilon_{jj}^{(n)}| \leq n \parallel A \parallel \varepsilon_2 .$$

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 λ_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 nth 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 ε 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

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

$$1, \psi_1, \ldots, \psi_n$$

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 ψ_l *are* the characteristic polynomials of T_l (Lanczos polynomials), satisfying the minimization property

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

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 $L(B, p_1)$,

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

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 $\overline{A}_J$, $\overline{q}_J^1$. $\overline{A}_J$ 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 $\overline{A}_J$ close to it.

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

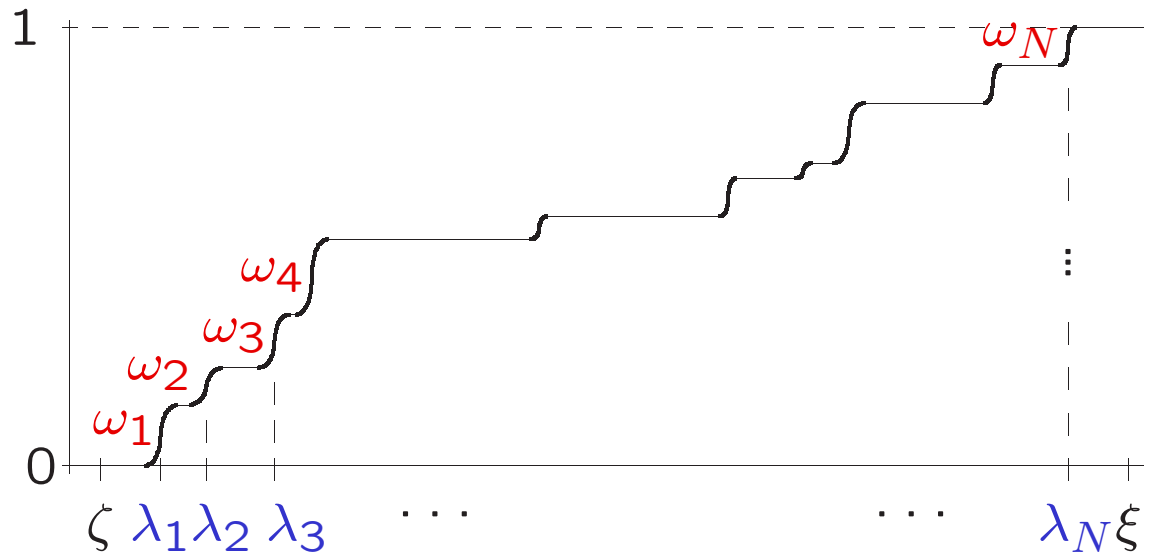
[Greenbaum, S - 92]:

FP $L(A, q^1)$ and FP $CG(A, q^1)$ behave very similarly as the exact algorithms applied to any $\widehat{A}$, $\widehat{q}^1$ from a certain class $\widehat{A}$ 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 $\widehat{\omega}(\lambda)$, which represents the spectral decomposition of some infinitely dimensional problem.



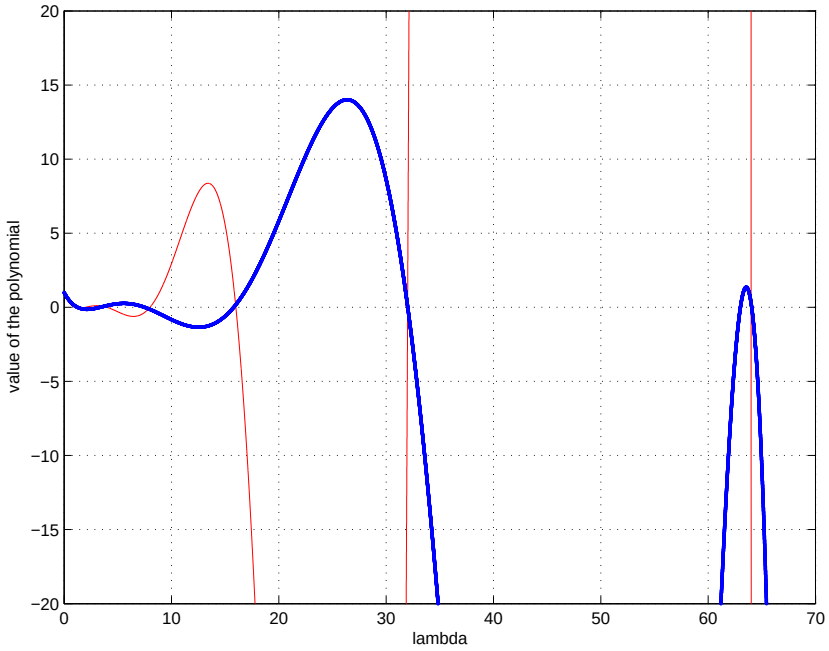
Consequences:

- 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 $\widehat{A}$, $\widehat{q}_1$.

  By applying exact precision theory (convergence bounds) to $\widehat{A}$, $\widehat{q}_1$ 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.

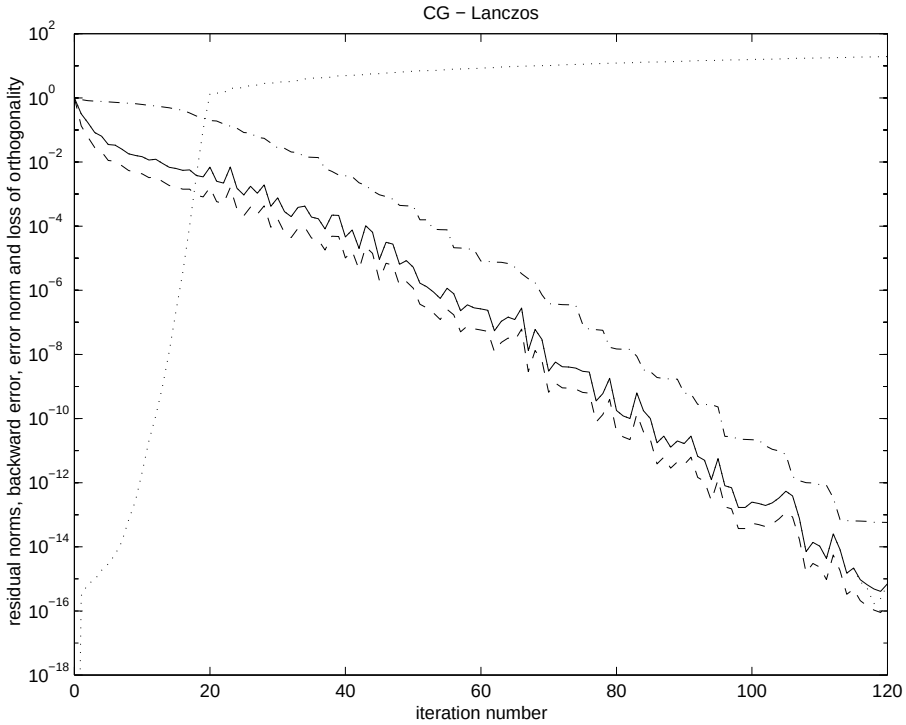
Linear independence of the computed generating vectors is lost with multiple Ritz values. Delay of convergence: each loss of linear independence costs one iteration!

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

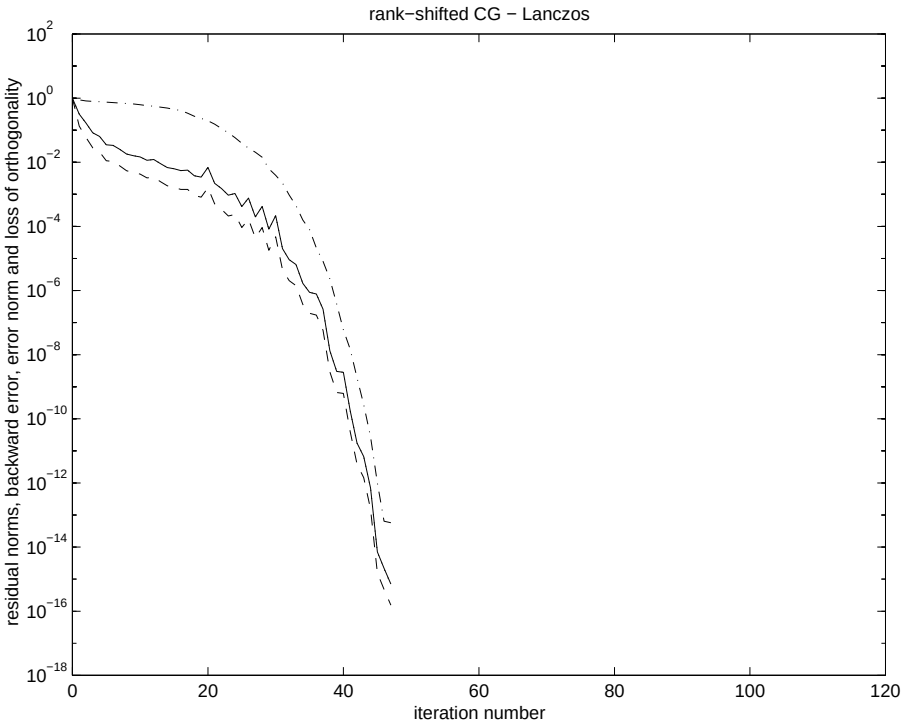
It is extremely difficult to estimate for the given step n

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

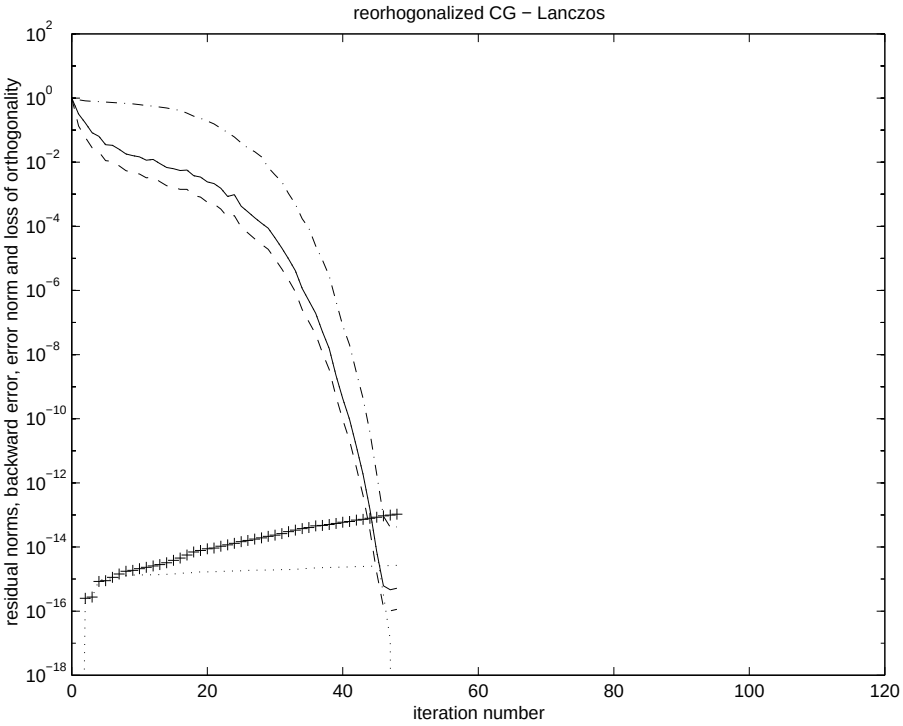
even for HH GMRES (it assumes solving the question about the stability of Krylov subspaces). We have to rotate our view.



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 reorhogonalized 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 $\longrightarrow 0$. 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 $\|r_k\| \to 0$).

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 $\times$ Rutishauser CG; BiCG $\times$ 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]

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

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

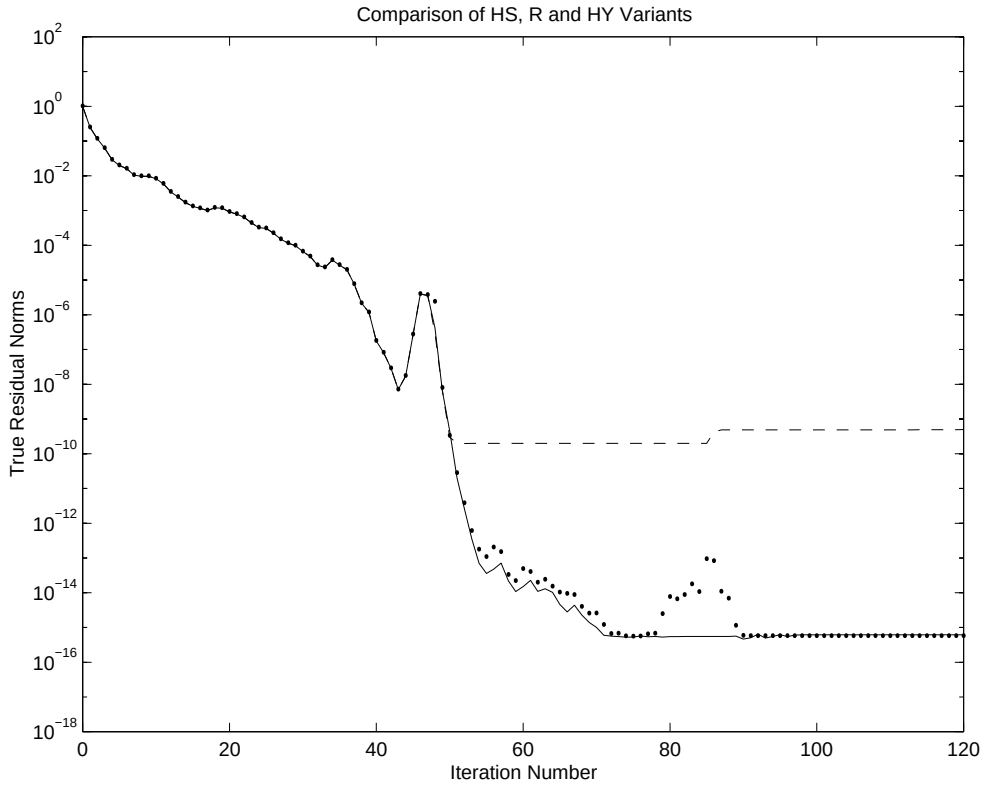
Example: three term CG (HPD case)

Amplification factors:

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

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.



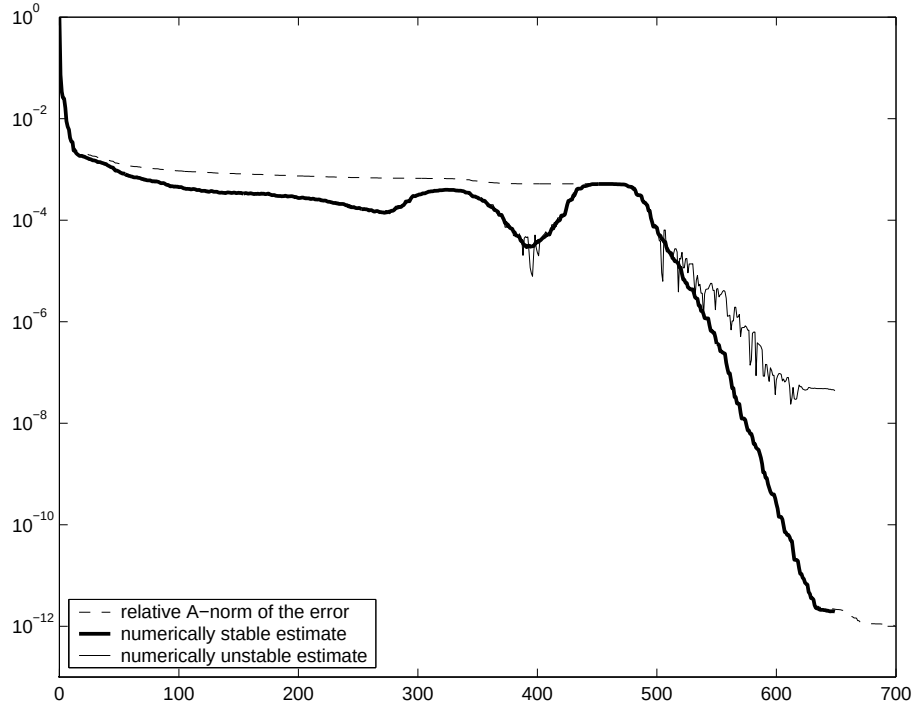
The CG relative residual can indeed very strongly oscillate!

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)

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

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].