Chapter two

Convergence (Behavior) in Exact Arithmetic

We start with preliminaries.

1. Convergence (better behavior)

2. Hermitian case

3. Non-Hermitian, but normal case

4. Nonnormal case

2.1 Convergence (better behavior)

Convergence of iterative methods:

$x_0, x_1, \ldots, x_n \longrightarrow x$, size of the error $\|x - x_n\|$.

Nonlinear problems:

$$\lim_{n\to\infty} \frac{\|x - x_n\|}{\|x - x_{n-1}\|^p} = \text{const} < 1\,.$$

Linear stationary methods of the first order

$$x_n = x_{n-1} + M^{-1}(b - Ax_{n-1}), \quad Mx_n = Nx_{n-1} + b, \quad A = M - N\,.$$

Richardson, Jacobi, Gauss-Seidel, SOR, SSOR

[Young - 51], [Varga - 62], [Young - 71], [Hageman, Young - 81], [Axelsson - 94]

Description of convergence? Linearization at infinity! After some transition phase the iterates converge with an almost linear rate predicted by the asymptotic convergence factor.

The description of convergence is linear.

$$x - x_n \;=\; (I - M^{-1}A)(x - x_{n-1}); \qquad \|x - x_n\| \;\leq\; \|(I - M^{-1}A)^n\| \|x - x_0\| \,.$$

Asymptotic convergence:

$$\lim_{n\to\infty} \left(\frac{\|x - x_n\|}{\|x - x_0\|} \right)^{\frac{1}{n}} \;=\; \lim_{n\to\infty} \|(I - M^{-1}A)^n\|^{\frac{1}{n}} \;=\; \rho \,,$$

where ρ is equal to the spectral radius of $(I - M^{-1}A)$.
Asymptotically,

$$\|x - x_n\| \;\approx\; \rho^n \, \|x - x_0\| \,.$$

In our course $\|\cdot\| \equiv \|\cdot\|_2$. In general it may happen

$$\rho < 1 < \|I - M^{-1}A\|.$$

Then the norm of the error may grow for some number of steps, before it eventually start to decrease. Related problems on how many steps we need to find whether $\rho < 1$ from the behavior of $\|(I - M^{-1}A)^n\|$ were studied by Pták, which led to the theory of the critical exponent.

ρ alone is not sufficient to describe the transient behavior, unless $M^{-1}A$ is normal. Then $\rho^n = \|(I - M^{-1}A)^n\|$, and the power of the spectral radius gives the tight upper bound for the norm of the error from the first step.

Chebyshev semiiterative method

Suppose that $x_1, \ldots, x_n$ have been generated via the linear stationary method of the first order. Define

$$y_n = \sum_{i=0}^{n} \nu_i^{(n)} x_i \,, \quad \sum_{i=0}^{n} \nu_i^{(n)} = 1 \,.$$

Then a simple calculation gives

$$\|x - y_n\| = \|\varphi_n(I - M^{-1}A)\,(x - x_0)\| \le \|\varphi_n(I - M^{-1}A)\|\,\|x - x_0\| \,.$$

Assume that $I - M^{-1}A$ is Hermitian with the eigenvalues between α and β. Then minimizing the bound for the norm of the matrix polynomial using the information on the interval containing the spectrum leads to the Chebyshev method.

[Golub, Varga - 1961], [Varga - 1962]

Please note that $\varphi(1) = 1$.

For the Richardson iteration $M = I$. Rearranging $p_n(A) \equiv \varphi_n(I - A)$, we get the condition $p_n(0) = 1$. The polynomial φ_n, and so p_n, is determined from the shifted and normalized Chebyshev polynomial.

The method needs a-priori information about the interval containing the spectrum (with the Richardson iteration the matrix must be positive definite). The Chebyshev method does not use any information about distribution of the eigenvalues within the given interval – it aims at minimizing the norm of the matrix polynomial.

Krylov subspace methods:

$$\mathcal{K}_n \equiv \mathcal{K}_n(A, r_0) \equiv \mathsf{span}\,\{r_0, \cdots, A^{n-1} r_0\}\,.$$

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

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

Generalization of the asymptotic convergence factor idea?

Let $\mathcal{S}$ be a compact set in the complex plane not containing zero and not separating it from the point at infinity. The asymptotic estimated convergence factor associated with $\mathcal{S}$ is defined by

$$E_n(\mathcal{S}) \equiv \min_{p \in \Pi_n} \max_{z \in \mathcal{S}} |\, p(z) \,|, \quad \rho(\mathcal{S}) = \lim_{n \to \infty} (E_n(\mathcal{S}))^{\frac{1}{n}} < 1 \,.$$

How to link with convergence of Krylov subspace methods and how the important set $\mathcal{S}$ should be chosen?

In order to be relevant, it assumes a large number of steps.

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: $A = S\,J\,S^{-1}\,.$

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

Chapter 2: Convergence (behavior) in exact arithmetic

2.2 Hermitian case

1. Basic relationships

2. Characterization of convergence

3. Ritz values

4. Matrix polynomial and worst case bound

5. Minimal polynomial idea can be misleading

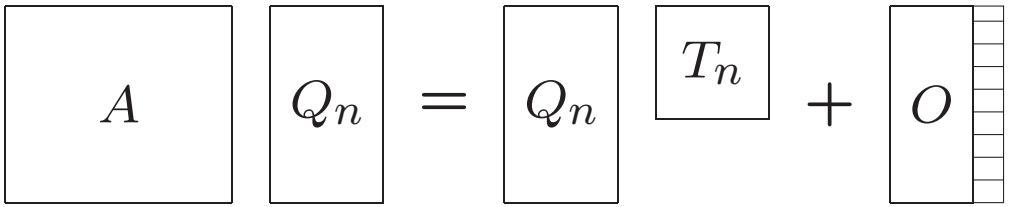
6. Measuring convergence

2.2.1 Basic relationships

Lanczos basis of $K_n(A, r_0)$

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

Three–term recurrence for generating orthonormal basis of Krylov subspaces



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

Jacobi matrix

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

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

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

Relation to orthogonal polynomials

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

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

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

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

$$\downarrow$$

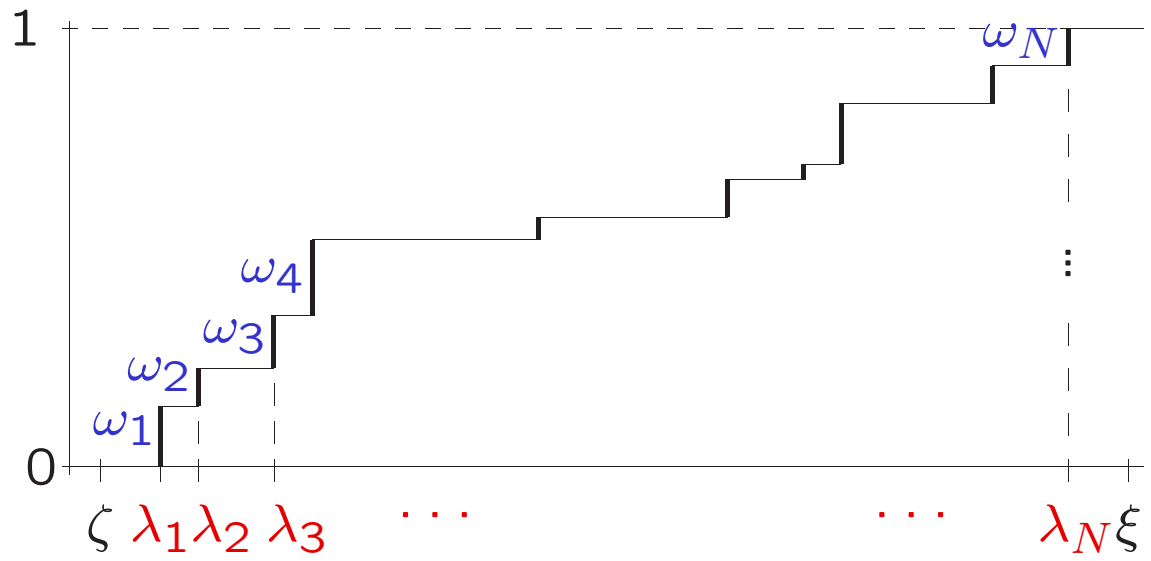
$$\sum_{i=1}^{N} (u_i, q_1)^2 \, \psi_n^2(\lambda_i) \;=\; \min_{\psi \in \mathcal{M}_n} \sum_{i=1}^{n} (u_i, q_1)^2 \, \psi^2(\lambda_i) \,.$$

Riemann-Stieltjes integral

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

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

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



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

T_n is determined by the Lanczos process for the matrix T_n and the starting vector e_1, the monic polynomials $\{1, \psi_1, \ldots, \psi_n\}$ are orthogonal with respect to

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

The n-th Riemann-Stieltjes integral,

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

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

Lanczos process:

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

Relationship between the original

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

and the n-th R-S integral $\int_{\zeta}^{\xi} f(\lambda)\, d\omega^{(n)}(\lambda) = \sum_{i=1}^{n} \omega_i^{(n)} f\left(\theta_i^{(n)}\right)$

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, Tich y - 02], [S, Liesen - 05]

Conjugate gradient method (CG)

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

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

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 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 individual $\omega^{(n)}(\lambda)$.

This relationship is all but trivial !

The essence of the CG method

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

Gauss quadrature !

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

2.2.2 Characterization of convergence

Conjugate gradient method, A Hermitian positive definite

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

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 $[0, \ldots, 0, \beta_{n+1}]$.

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

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

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

[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 – 01], [Liesen, Tichý - 04]

The value of the bound is known analytically [Greenbaum - 79], though in a rather complicated form

$$\min_{p\in\Pi_n} \max_i |p(\lambda_i)| \;=\; \left(\sum_{j=1}^{n+1} \prod_{k=1,k\neq j}^{n+1} \frac{|\mu_k|}{|\mu_k-\mu_j|} \right)^{-1},$$

where $\{\mu_1, \ldots \mu_{n+1}\}$ is some subset of the distinct eigenvalues of A.

2.2.3 Ritz values

Roots of the normalized Lanczos polynomial (which is equivalent to the CG polynomial)

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

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]

2.2.4 Matrix pol. and worst case 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,

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

2.2.5 Minimal polynomial idea can mislead

Linear bounds based on the Chebyshev method, see, e.g., [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" (which, in addition, totally ignore delays due to rounding errors . . .).

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").

The minimal polynomial idea is often linked with several tight clusters of eigenvalues. Representing each cluster with a single point, it is believed that the polynomial having roots at these points (a single root within each cluster) gives a good approximation to the minimal polynomial. Consequently, it is believed that a good approximate solution should be obtained in m steps, where m is the number of clusters.

The idea is applied to general Krylov subspace methods. However, without considering the distribution of clusters together with their diameter, the general statements, though found in good references, are incorrect even for the Hermitian positive definite case and in exact arithmetic.

The trouble has nothing to do with non-normality!

Quiz:

A Hermitian, positive definite, exact precision.

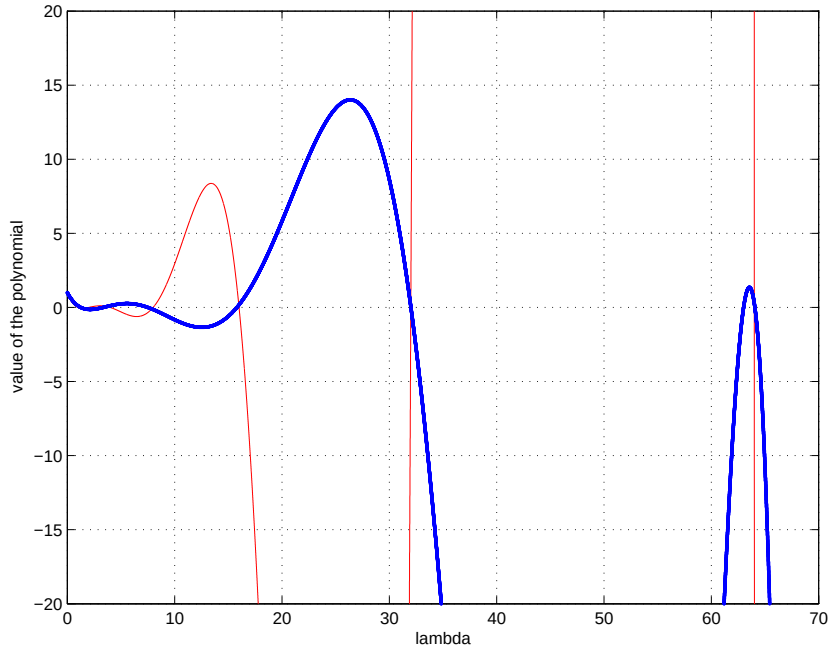
Do the following characterizations of eigenvalue distribution guarantee fast convergence of CG, i.e., a reasonably accurate (with the relative error, say, 10^{-4}) approximate solution in t + few (fixed number) steps?

1. Cluster arround one & t large eigenvalues.

2. Cluster arround one & t tight clusters of large eigenvalues, with diameters of the clusters bounded by, say, 10^{-12}.

Consider, e.g., a polynomial

$p(\lambda) = (1/2^{21})(\lambda - 1)(\lambda - 2) \cdots (\lambda - 64)$:



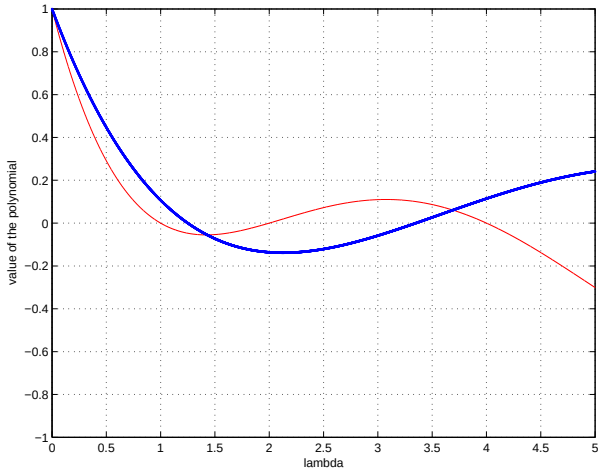
The minimization property determining the CG polynomial enforces redistribution of roots.

In the presence of well separated clusters of large eigenvalues the CG polynomial places multiple roots in single clusters.

When single eigenvalues are replaced by tight clusters (or, vice versa, when tight clusters are represented by single points), the behavior of CG in exact arithmetic can change dramatically, in dependence on the interplay between the distribution and the diameters of the clusters. Surprisingly, we will hear more on this in the part on numerical stability analysis of CG.

In contrast to simpler iterative methods, including the Chebyshev method, CG in each iteration uses the complete information about the spectrum.

Example - well separated cluster of large eigenvalues affect the values of the CG polynomial at the smallest eigenvalues, see the detail:



When CG is not much better than Chebyshev? When the spectral information is determined locally (e.g. the spectrum is close to uniform).

2.2.6 Measuring convergence

The CG example

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

For $n = 1, 2, \ldots$

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

For most elliptic PDE, a natural measure of convergence in solving the discretized problem is $\|x - x_n\|_A$.

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],

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

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

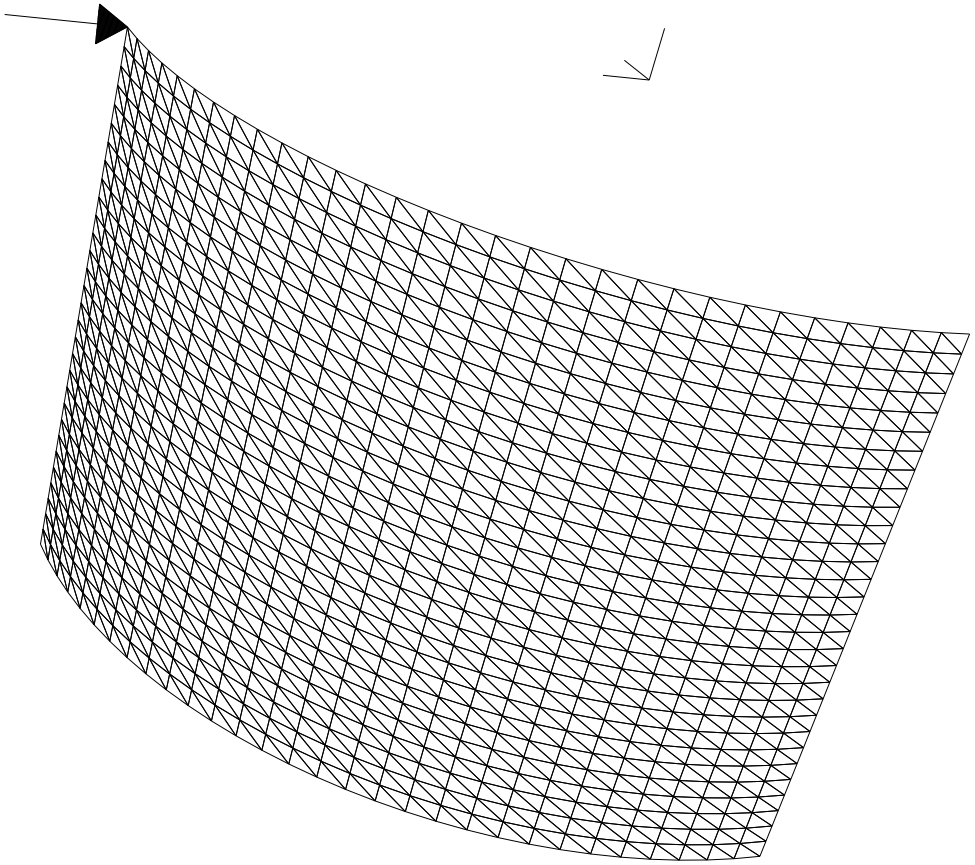
Mathematically equivalent formulas for EST^2 :

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

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

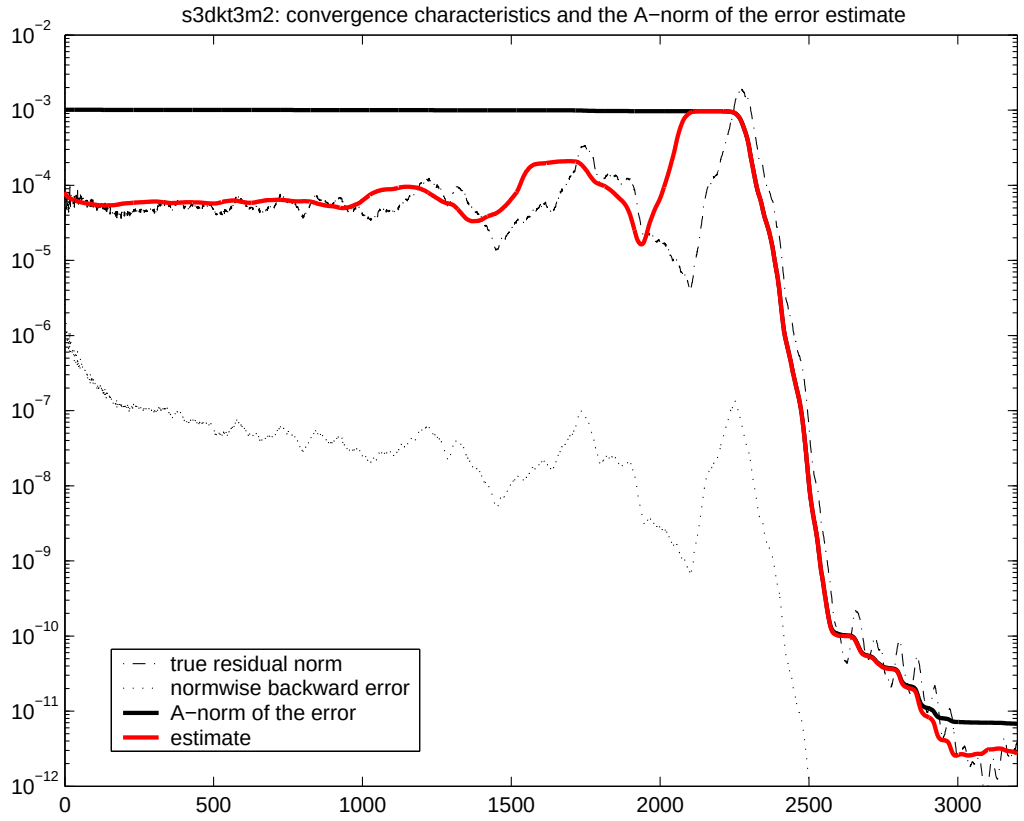
[Hestenes, Stiefel - 52], after fifty years found and extended in [S, Tichý 2002, 04] with justification for finite precision computations

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



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