What does it mean to solve Ax=b iteratively? On what theoretical and practical criteria do you compare different solvers? What are the best solvers based on these criteria?

Zdeněk Strakoš
Charles University, Prague
Faculty of Mathematics and Physics
Jindřich Nečas Center for Mathematical Modelling

CMC Seminar, October 13, 2021

I. Questions beyond matrix computations. Numerical PDE (PCG):

$\mathcal{A}x = b, \quad \mathcal{A} : V \to V^{\#}, \quad x \in V, \quad b \in V^{\#}, \quad$ approximations $x_0, x_1, \ldots.$ to x

I. Questions beyond matrix computations. Numerical PDE (PCG):

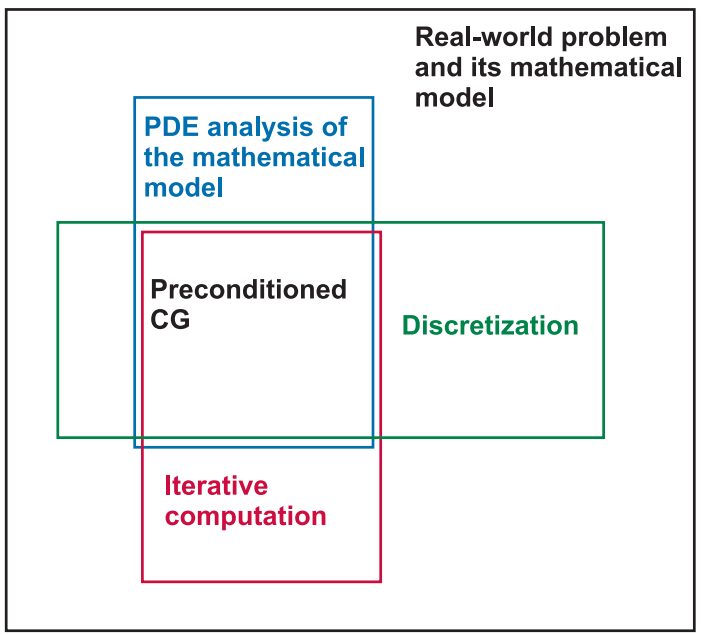
$\mathcal{A}x = b, \quad \mathcal{A} : V \to V^{\#}, \quad x \in V, \quad b \in V^{\#}, \quad$ approximations $x_0, x_1, \ldots.$ to x

Common approach: $\mathcal{A}$, $b \to \mathbf{A}_h, \mathbf{b}_h \to$ preconditioner $\to \mathbf{x}_n \approx \mathbf{x}_h \to x_n \approx x$

I. Questions beyond matrix computations. Numerical PDE (PCG):

$\mathcal{A}x = b, \quad \mathcal{A} : V \to V^{\#}, \quad x \in V, \quad b \in V^{\#}, \quad$ approximations $x_0, x_1, \ldots.$ to x

Common approach: $\mathcal{A},\ b \ \to \mathbf{A}_h, \mathbf{b}_h \to$ preconditioner $\ \to \mathbf{x}_n \approx \mathbf{x}_h \ \to \ x_n \approx x$



$\{\mathcal{A}, b, \tau\} \to \{\mathcal{A}_h, b_h, \tau\} \to \{\mathbf{A}_h, \mathbf{b}_h, \mathbf{M}_h\} \to \mathbf{x}_n \approx \mathbf{x}_h \ \to \ x_n \approx x$

II. Criteria: Adaptivity and Cost(acceptable accuracy). PCG:

A nested nonlinear hierarchy of problems based on projections onto Krylov subspaces. The essence can be formulated through the link with the Stieltjes problem of moments and Gauss quadrature, recall Hestenes and Stiefel (1952):

II. Criteria: Adaptivity and Cost(acceptable accuracy). PCG:

A nested nonlinear hierarchy of problems based on projections onto Krylov subspaces. The essence can be formulated through the link with the Stieltjes problem of moments and Gauss quadrature, recall Hestenes and Stiefel (1952):

Spectral decomposition determines the distribution function and moments

$$m_\ell \;=\; w_1^*(\tau\mathcal{A})^\ell\, w_1 \;=\; w_1^*\left(\int\, \lambda^\ell\, d\mathcal{E}(\lambda)\right) w_1 = \int\, \lambda^\ell\, d\omega(\lambda)\,, \quad \ell = 0, 1, 2, \ldots$$

II. Criteria: Adaptivity and Cost(acceptable accuracy). PCG:

A nested nonlinear hierarchy of problems based on projections onto Krylov subspaces. The essence can be formulated through the link with the Stieltjes problem of moments and Gauss quadrature, recall Hestenes and Stiefel (1952):

Spectral decomposition determines the distribution function and moments

$$m_\ell \;=\; w_1^*(\tau\mathcal{A})^\ell\, w_1 \;=\; w_1^*\left(\int\, \lambda^\ell\, d\mathcal{E}(\lambda)\right) w_1 = \int\, \lambda^\ell\, d\omega(\lambda)\,, \quad \ell = 0, 1, 2, \ldots$$

CG at step n implicitly solves system of $2n$ nonlinear equations for $2n$ unknowns (it matches the first $2n$ moments)

$$\sum_{i=1}^n \omega_i^{(n)} \{\theta_i^{(n)}\}^\ell \;=\; m_\ell\,, \qquad \ell \;= 0, 1, 2, \ldots, 2n-1,$$

II. Criteria: Adaptivity and Cost(acceptable accuracy). PCG:

A nested nonlinear hierarchy of problems based on projections onto Krylov subspaces. The essence can be formulated through the link with the Stieltjes problem of moments and Gauss quadrature, recall Hestenes and Stiefel (1952):

Spectral decomposition determines the distribution function and moments

$$m_\ell \;=\; w_1^*(\tau\mathcal{A})^\ell\, w_1 \;=\; w_1^*\left(\int\, \lambda^\ell\, d\mathcal{E}(\lambda)\right) w_1 = \int\, \lambda^\ell\, d\omega(\lambda)\,, \quad \ell = 0, 1, 2, \ldots$$

CG at step n implicitly solves system of $2n$ nonlinear equations for $2n$ unknowns (it matches the first $2n$ moments)

$$\textstyle\sum_{i=1}^{n}\, \omega_i^{(n)} \{\theta_i^{(n)}\}^\ell \;=\; m_\ell\,, \qquad \ell \;= 0, 1, 2, \ldots, 2n-1,$$

and provides the minimal energy norm of the error over the shifted n-th Krylov subspace $x_0 + \mathcal{K}_n(\tau\mathcal{A}, \tau r_0) = x_0 + \mathrm{span}\{\tau r_0, \tau\mathcal{A}\,\tau r_0, \ldots, (\tau\mathcal{A})^{n-1}\tau r_0\}$

$$\frac{\|x - x_0\|_{\mathcal{A}}^2}{\|\tau r_0\|^2} \;=\; \sum_{i=1}^{n}\, \omega_i^{(n)}\, \frac{1}{\theta_i^{(n)}} \;+\; \frac{\|x - x_n\|_{\mathcal{A}}^2}{\|\tau r_0\|^2}\,.$$

II. Criteria: Adaptivity and Cost(acceptable accuracy). PCG:

A nested nonlinear hierarchy of problems based on projections onto Krylov subspaces. The essence can be formulated through the link with the Stieltjes problem of moments and Gauss quadrature, recall Hestenes and Stiefel (1952):

Spectral decomposition determines the distribution function and moments

$$m_\ell \;=\; w_1^*(\tau\mathcal{A})^\ell\, w_1 \;=\; w_1^*\left(\int\, \lambda^\ell\, d\mathcal{E}(\lambda)\right) w_1 = \int\, \lambda^\ell\, d\omega(\lambda)\,, \quad \ell = 0, 1, 2, \ldots$$

CG at step n implicitly solves system of $2n$ nonlinear equations for $2n$ unknowns (it matches the first $2n$ moments)

$$\textstyle\sum_{i=1}^{n}\, \omega_i^{(n)} \{\theta_i^{(n)}\}^\ell \;=\; m_\ell\,, \qquad \ell \;= 0, 1, 2, \ldots, 2n-1,$$

and provides the minimal energy norm of the error over the shifted n-th Krylov subspace $x_0 + \mathcal{K}_n(\tau\mathcal{A}, \tau r_0) = x_0 + \mathrm{span}\{\tau r_0, \tau\mathcal{A}\,\tau r_0, \ldots, (\tau\mathcal{A})^{n-1}\tau r_0\}$

$$\frac{\|x - x_0\|_{\mathcal{A}}^2}{\|\tau r_0\|^2} \;=\; \sum_{i=1}^{n}\, \omega_i^{(n)}\, \frac{1}{\theta_i^{(n)}} \;+\; \frac{\|x - x_n\|_{\mathcal{A}}^2}{\|\tau r_0\|^2}\,.$$

III. *"a way that does justice to the inner nature of the problem."*

Lanczos to Einstein, March 1947. Nine days later Einstein writes in his reply:
" importance of **adapted** *approximation methods ... a fruitful mathematical aspect, and not just a utilitarian one".*

III. *"a way that does justice to the inner nature of the problem."*

Lanczos to Einstein, March 1947. Nine days later Einstein writes in his reply: *" importance of* **adapted** *approximation methods ... a fruitful mathematical aspect, and not just a utilitarian one".*

A quest for the (near to) best choice involves very complex issues, e.g.:

- Interplay of infinite and finite dimensional. Approximation of possibly continuous spectra of operators by matrix eigenvalues. Operator preconditioning, discretization and algebraic preconditioning.
- Analysis based on spectral information ... t tight clusters of eigenvalues do not necessarily mean reaching good approximation to the solution x in t steps. Smaller condition number does not necessarily mean faster decrease of error and lower computational cost. Analysis in the non-normal case is intriguing.
- Long vs. short recurrences dilemma.
- Rounding errors do matter.

Matching moments model reduction

The nth convergent

$$\mathcal{F}_n(\lambda) \equiv \cfrac{1}{\lambda-\gamma_1-\cfrac{\delta_2^2}{\lambda-\gamma_2-\cfrac{\delta_3^2}{\lambda-\gamma_3-\ldots\cfrac{\ddots}{\lambda-\gamma_{n-1}-\cfrac{\delta_n^2}{\lambda-\gamma_n}}}}} = \frac{\mathcal{R}_n(\lambda)}{\mathcal{P}_n(\lambda)} \, .$$

Stieltjes (1894): "*we shall determine in which cases this convergent tends to a limit for* $n \to \infty$ *and we shall investigate more closely the nature of this limit regarded as a function of* λ*.*"

Here we use notation different from Stieltjes (1894), in particular $\lambda \equiv -z$.

Remarkable history

- Euclid (300BC), Hippassus from Metapontum (before 400BC), ... ,
- Bhascara II (around 1150), Brouncker and Wallis (1655-56):
  Three term recurrences (for numbers)
- Euler (1737, 1748), ... , Brezinski (1991), Khrushchev (2008)
- Gauss (1814), Jacobi (1826), Christoffel (1858, 1857), ... ,
  Chebyshev (1855, 1859), Markov (1884), Stieltjes (1884, 1893-94):
  Orthogonal polynomials, quadrature, analytic theory of continued fractions, problem of moments, minimal partial realization, Riemann-Stieltjes integral
  Gautschi (1981, 2004), Brezinski (1991), Van Assche (1993), Kjeldsen (1993)
- Hilbert (1906, 1912), ... , Von Neumann (1927, 1932), Wintner (1929):
  resolution of unity, integral representation of operator functions, mathematical foundation of quantum mechanics

Matrix computation and control theory context

- Krylov (1931), Lanczos (1950, 1952, 1952c), Hestenes and Stiefel (1952), Rutishauser (1953), Henrici (1958), Stiefel (1958), Rutishauser (1959), ... , Vorobyev (1954, 1958, 1965), Golub and Welsch (1968), ... , Laurie (1991 - 2001), ...

- Gordon (1968), Schlesinger and Schwartz (1966), Steen (1973), Reinhard (1979), ... , Horáček (1983 - ...), Simon (2007 - ...)

- Paige (1971, 1972, 1976, 1980), Reid (1971), Greenbaum (1989), ...

- Magnus (1962a,b), Gragg (1974), Kalman (1979), Gragg, Lindquist (1983), Gallivan, Grimme, Van Dooren (1994), ...

Moment problem considered by Stieltjes (1894)

Consider an infinite sequence of real numbers $m_0, m_1, m_2, \ldots$.

Find the necessary and sufficient conditions for the existence of the Riemann-Stieltjes integral with the (positive nondecreasing) distribution function $\omega(\lambda)$ such that

$$\int_0^\infty \lambda^\ell \, d\omega(\lambda) \;=\; m_\ell \,, \qquad \ell = 0, 1, 2, \ldots$$

and determine $\omega(\lambda)$.

Moment problem considered by Stieltjes (1894)

Consider an infinite sequence of real numbers $m_0, m_1, m_2, \ldots$.

Find the necessary and sufficient conditions for the existence of the Riemann-Stieltjes integral with the (positive nondecreasing) distribution function $\omega(\lambda)$ such that

$$\int_0^\infty \lambda^\ell \, d\omega(\lambda) \;=\; m_\ell \,, \qquad \ell = 0, 1, 2, \ldots$$

and determine $\omega(\lambda)$.

Related moment problem can also be formulated while approximating bounded linear (positive definite self-adjoint) operators in Hilbert spaces; see Vorobyev (1958, 1965).

Method of moments in Hilbert space

Let $\mathcal{B}$ be a bounded linear operator on Hilbert space V. Choosing an initial element z_0, we first form a sequence of elements $z_1, z_2, \ldots, z_n, \ldots$ such that

$$z_0,\ z_1 = \mathcal{B}z_0,\ z_2 = \mathcal{B}z_1 = \mathcal{B}^2 z_0,\ \ldots\ , z_n = \mathcal{B}z_{n-1} = \mathcal{B}^n z_{n-1},\ \ldots$$

At the the present time, $z_1, \ldots, z_n$ are assumed to be linearly independent. Determine a sequence of operators $\mathcal{B}_n$ defined on the sequence of nested subspaces V_n generated by $z_0, z_1, z_2, \ldots, z_{n-1}$, $n = 1, 2, \ldots$ such that

$$\begin{aligned}
z_1 &= \mathcal{B}z_0 = \mathcal{B}_n z_0, \\
z_2 &= \mathcal{B}^2 z_0 = (\mathcal{B}_n)^2 z_0, \\
&\vdots \\
z_{n-1} &= \mathcal{B}^{n-1} z_0 = (\mathcal{B}_n)^{n-1} z_0, \\
E_n z_n &= E_n \mathcal{B}^n z_0 = (\mathcal{B}_n)^n z_0.
\end{aligned}$$

Interpretation as model reduction using Krylov subspaces

Using the projection E_n onto V_n we can write for the operators constructed above (here we need the linearity of $\mathcal{B}$)

$$\mathcal{B}_n \;=\; E_n \mathcal{B} E_n \,.$$

The finite dimensional operators $\mathcal{B}_n$ can be used to obtain approximate solutions to various linear problems. The choice of the elements $z_0, \ldots, z_n, \ldots$ as above gives Krylov subspaces that are determined by:

- the operator (given by, e.g., a partial differential equation)
- and the initial element z_0 (given by, e.g., boundary conditions and outer forces).

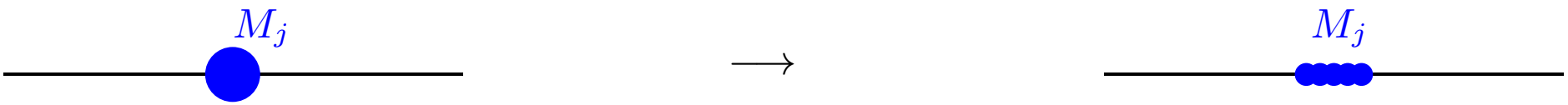
Two key ingrediences:

I. Krylov subspaces,
II. Projections that can lead to optimality.

See the method of conjugate gradients using orthogonal projections (to follow).

Outlying eigenvalues and the clustering argument

clustering of eigenvalues can make a difference



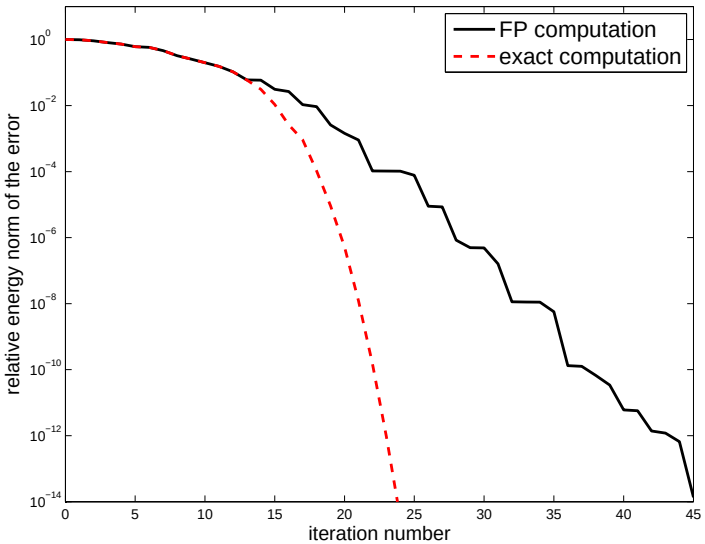
single eigenvalue
λ_j

$\longrightarrow$

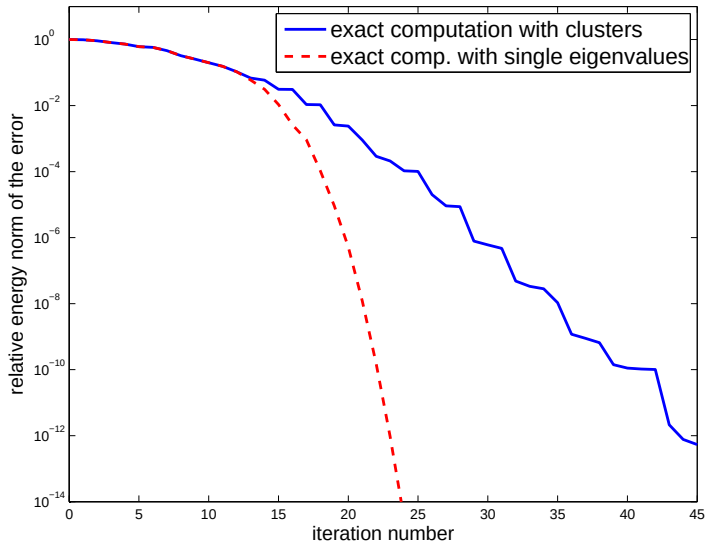
many close eigenvalues
$\widehat{\lambda}_{j_1}, \widehat{\lambda}_{j_2}, \ldots, \widehat{\lambda}_{j_\ell}$

Replacing a single eigenvalue by a tight cluster can make a substantial difference; Greenbaum (1989); Greenbaum, S (1992); Golub, S (1994). This was revealed due to the investigation of the propagation of rounding errors.

FP CG and clustering of eigenvalues in EXACT CG



Rounding errors in finite precision CG computations can cause a large delay of convergence.

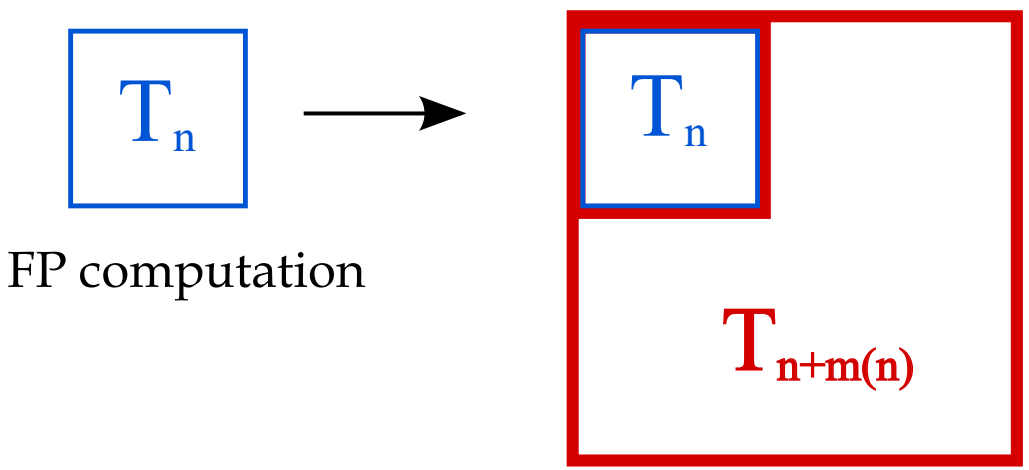


Exact CG computation for a matrix, where each eigenvalue is replaced by a tight cluster.

Understanding is based on the spectral information in the sequence of the computed (nested) Jacobi matrices $\mathbf{T}_n$, $n = 1, 2, \ldots$.
Seminal contribution of C.C. Paige (1971–80).

Beautiful idea of A. Greenbaum (1989)

- Consider the Jacobi matrix $\mathbf{T}_n$ computed in n steps of CG in FP arithmetic. This matrix can be extended to a larger Jacobi matrix $\mathbf{T}_{n+m(n)}$ having all its eigenvalues close to the eigenvalues of the matrix $\mathbf{A}$.
- Then the EXACT CG (Lanczos) applied to this extended Jacobi matrix and the initial residual e_1 gives in the first n steps $\mathbf{T}_n$.



In this way, finite precision computation is viewed and analyzed as exact computation for the problem having clusters of eigenvalues.

Why is consideration of rounding errors so fundamental?

1. CG should be used when it has a chance to accelerate its convergence due to adaptation to the information (hidden) in data, e.g., when the eigenvalues are far from being uniformly spread throughout the spectral interval. Presence of large outlying eigenvalues may seem as the most favourable case.

2. Hovewer, apart from the trivial situation mentioned next, in such cases CG convergence behaviour is typically substantially affected by rounding errors due to the loss of orthogonality among the direction vectors (and residuals).

3. If CG behaviour is not affected by rounding errors, then either we are lucky because convergence is so fast that rounding errors have not enough iterations to amplify (trivial cases), or CG convergence is hopelessly linear with no chance to accelerate. In the latter case linear methods would with high probability be more efficient in terms of computing time (energy consumption).

Why is consideration of rounding errors so fundamental?

1. CG should be used when it has a chance to accelerate its convergence due to adaptation to the information (hidden) in data, e.g., when the eigenvalues are far from being uniformly spread throughout the spectral interval. Presence of large outlying eigenvalues may seem as the most favourable case.

2. Hovewer, apart from the trivial situation mentioned next, in such cases CG convergence behaviour is typically substantially affected by rounding errors due to the loss of orthogonality among the direction vectors (and residuals).

3. If CG behaviour is not affected by rounding errors, then either we are lucky because convergence is so fast that rounding errors have not enough iterations to amplify (trivial cases), or CG convergence is hopelessly linear with no chance to accelerate. In the latter case linear methods would with high probability be more efficient in terms of computing time (energy consumption).

4. There seems to be no escape from the implications of the presented facts.

In which sense is the CG polynomial $\varphi_n^{\mathrm{CG}}(\lambda)$ optimal?

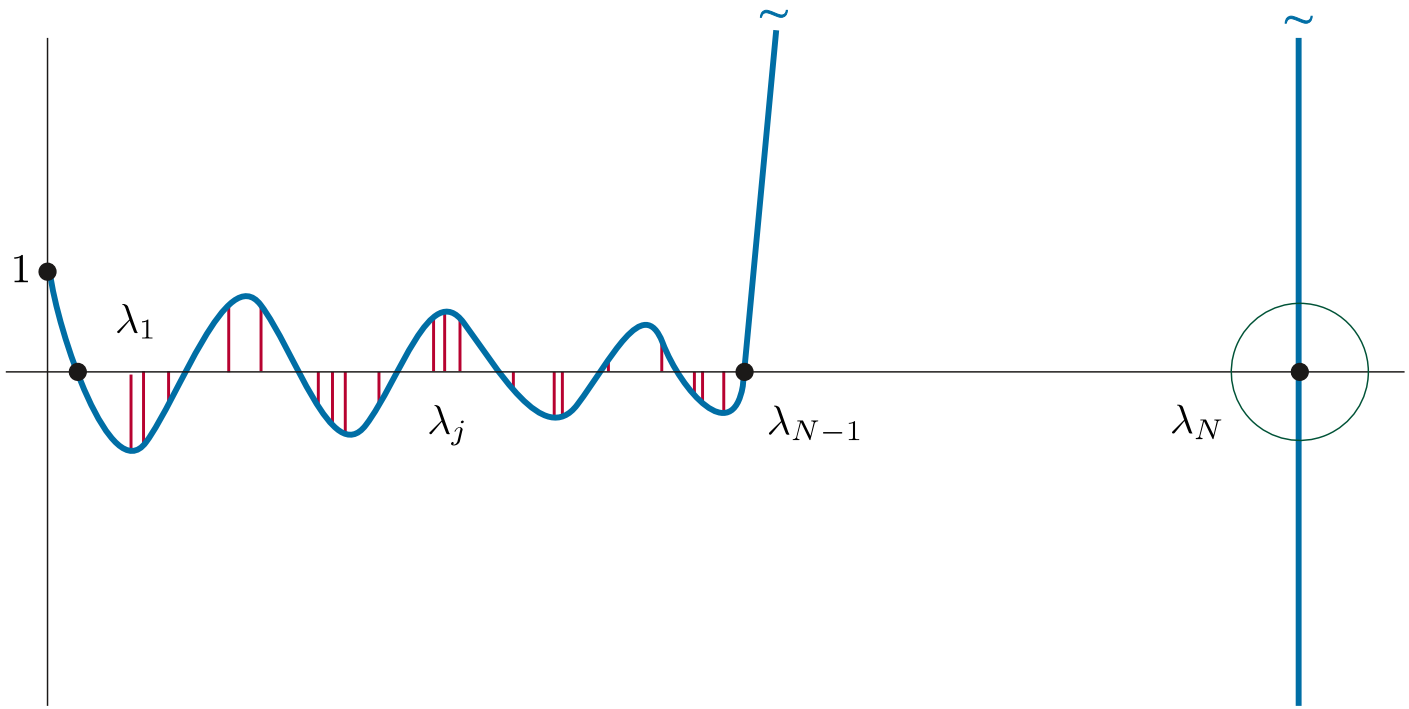
$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_n\|_{\mathbf{A}}^2 &= \min_{\varphi \in \Pi_n} \|\varphi(\mathbf{A})(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{A}}^2 \\ &= \sum_{j=1}^{N} \lambda_j \, \zeta_j^2 \, \varphi_n^{\mathrm{CG}}(\lambda_j)^2 \,, \quad j = 1, 2, \ldots \end{aligned}$$

Here

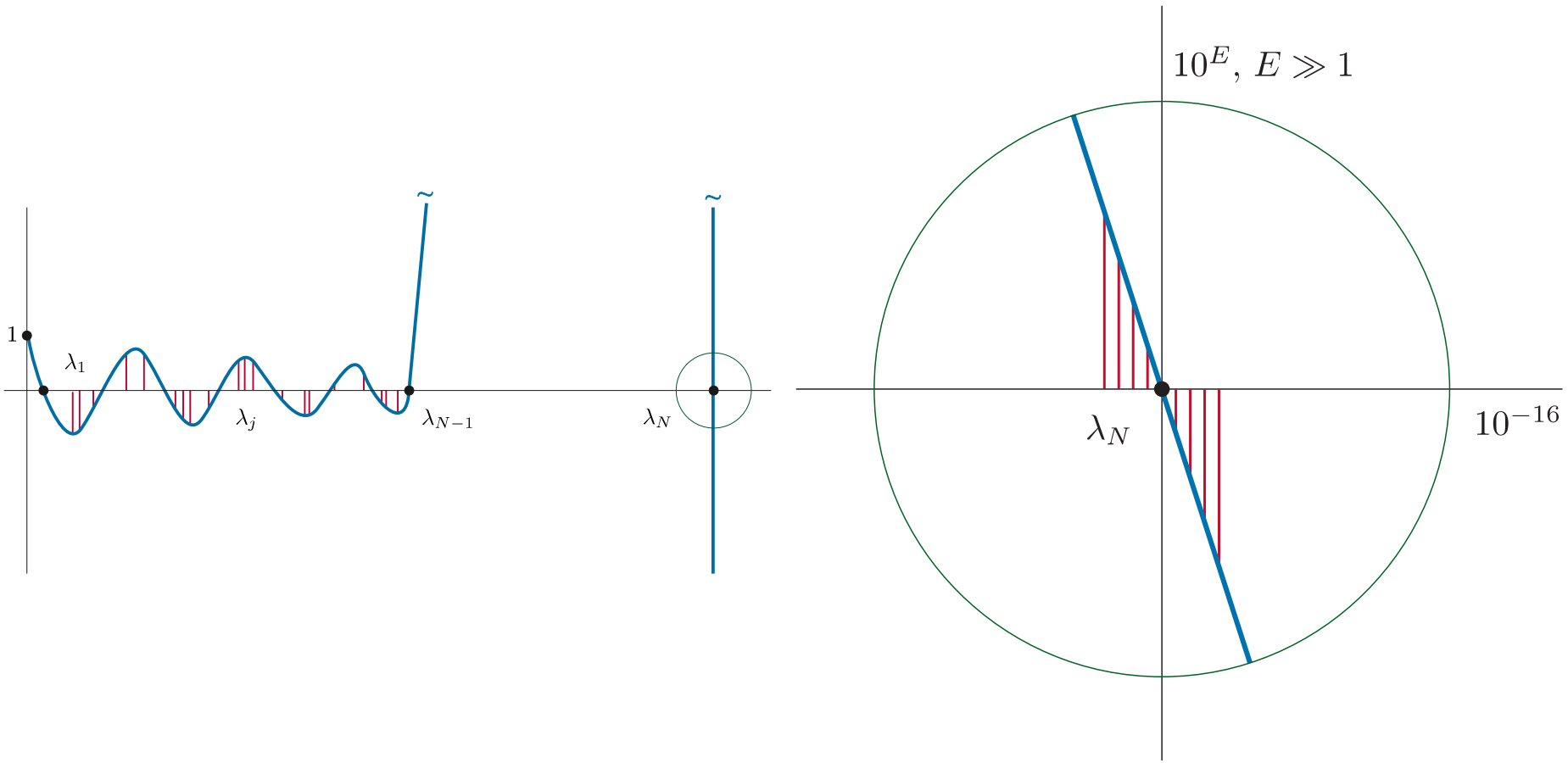
$$\varphi_n^{\mathrm{CG}}(\lambda) = \frac{(\lambda - \theta_1^{(n)}) \cdots (\lambda - \theta_n^{(n)})}{(-1)^n \, \theta_1^{(n)} \cdots \theta_n^{(n)}}$$

is determined by the eigenvalues of the orthogonally restricted operator, i.e., by the eigenvalues $\theta_1^{(n)}, \ldots, \theta_n^{(n)}$ of $\mathbf{T}_n$ (Ritz values).

Illustration of $\varphi_n^{\mathrm{CG}}(\lambda)$, $\mathbf{Ax} = \mathbf{b}$, a single large outlier λ_N



Replacing λ_N by a tight cluster significantly affects $\varphi_n^{\mathrm{CG}}(\lambda)$



Since E in the illustration of the slope is enormous, many roots close to λ_N are needed.

Chebyshev polynomial bound for CG ... remarkable history

r

- Markov (1890)
- Flanders and Shortley (1950)
- Lanczos (1952–53), Kincaid (1947), Young (1954, ...)
- Stiefel (1958), Rutishauser (1959)
- Meinardus (1963), Kaniel (1966)
- Daniel (1967a, 1967b)
- Luenberger (1969)

Derivations are repeated in recent textbooks and monographs and the resulting bound is identified with the convergence of CG without noticing severe limitations.

Statement published in 1952 is right to the point.

C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. of Research of the National Burreau of Standards, 49 (1952), pp. 33–53:

"The principle by which this process [the conjugate gradient method] *gives good attenuation, is quite different from the previous one.* ['Purification' based on Tschebyshev polynomials.] *Here we take heed of the specific nature of the matrix A and operate in a selective way. The polynomials of this process ... have the peculiarity that they attenuate due to the nearness of their zeros to those λ-values* [eigenvalues] *which are present in A. These polynomials take advantage of the fact that the spectrum to be attenuated is a line spectrum and not a continuous spectrum. They work efficiently in the neighbourhood of the λ_i of the matrix but not for the intermediate values."*

Composite bounds considering large outliers

The condition-number-based bound should be used with a great care in connection with the behaviour of CG unless $\kappa(A) = \lambda_N/\lambda_1$ is really small or unless the (very special) distribution of eigenvalues makes the bound tight.

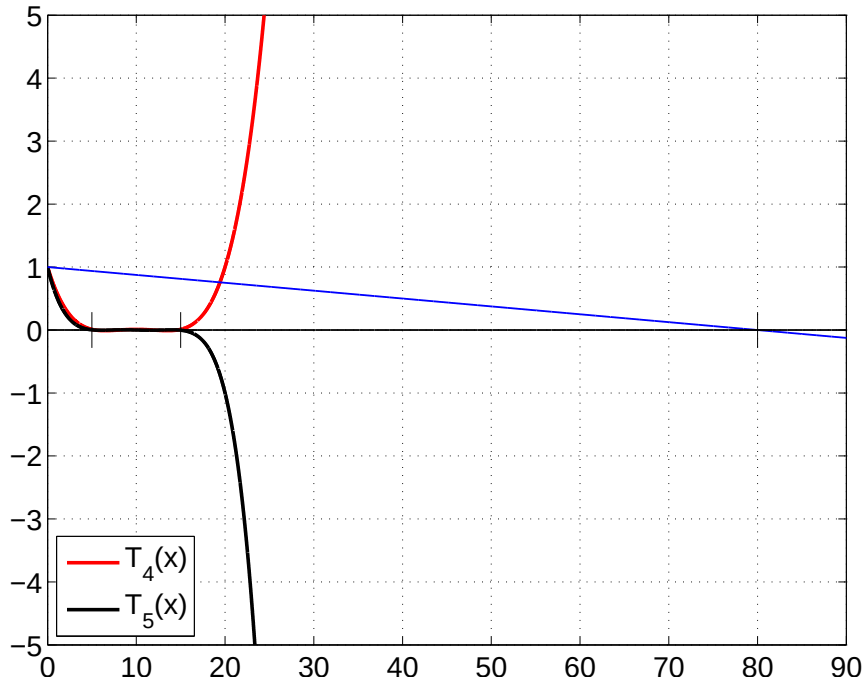
In particular, one should be very careful while using it as a part of a composite bound in the presence of large outlying eigenvalues

$$\min_{\substack{p(0)=1 \\ \deg(p)\le n-s}} \max_{1\le j\le N} |\, q_s(\lambda_j)\, p(\lambda_j)\,| \quad \le \quad \max_{1\le j\le N} |q_s(\lambda_j)| \left| \frac{T_{n-s}(\lambda_j)|}{T_{n-s}(0)} \right|$$

$$< \quad \max_{1\le j\le N-s} \left| \frac{T_{n-s}(\lambda_j)}{T_{n-s}(0)} \right| .$$

This Chebyshev method bound for the spectral interval $[\lambda_1, \lambda_{N-s}]$ is then valid after s initial steps.

Polynomial $q_s(\lambda)$ has the desired root, but look at $T_{4-5}(\lambda)$



A single large outlying eigenvaue:

The shifted and scaled Chebyshev polynomials $T_4(\lambda), T_5(\lambda)$, and the polynomial $q_1(\lambda)\,, q_1(0) = 1$ having the root at the large outlying eigenvalue.

Misconception when applied to practical computations

Consider the desired accuracy ϵ, $\kappa_s(A) \equiv \lambda_{N-s}/\lambda_1$. Then, assuming exact arithmetic, n CG steps, where

$$n = \mathbf{s} + \left\lceil \frac{\ln(2/\epsilon)}{2} \sqrt{\kappa_s(A)} \right\rceil ,$$

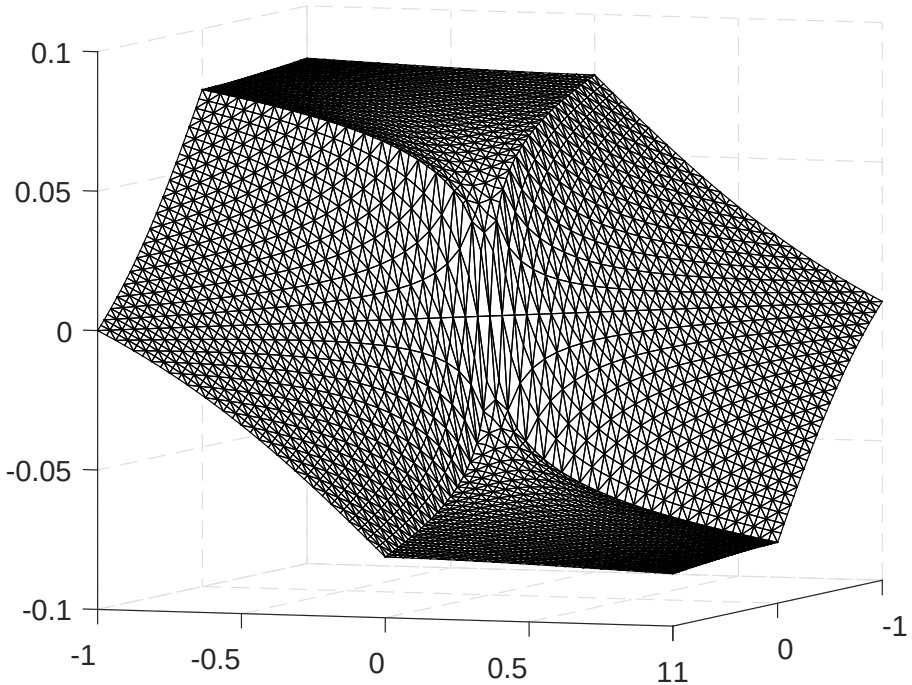
will produce the approximate solution x_n satisfying

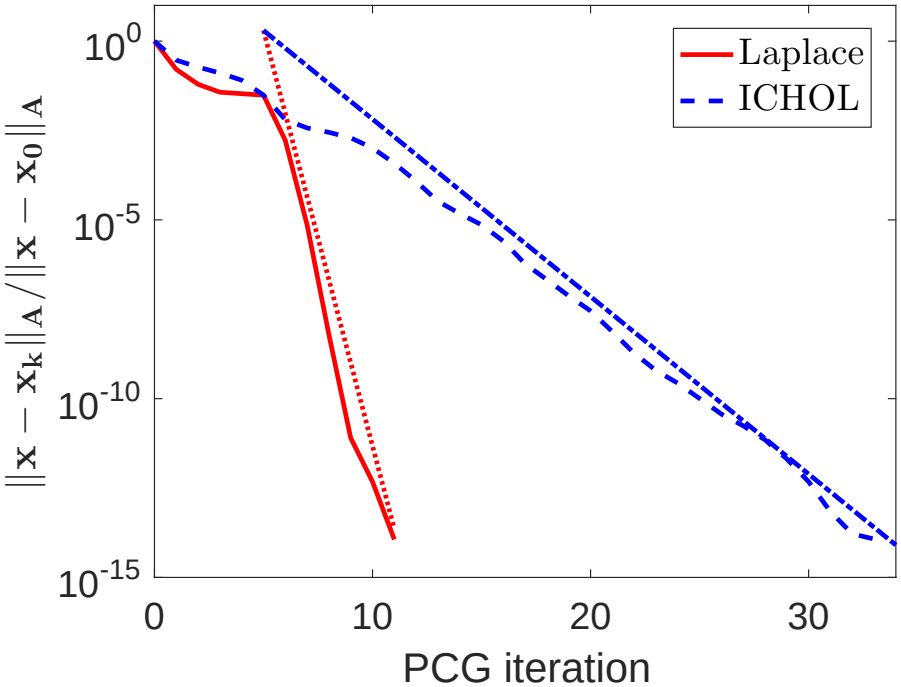
$$\|x - x_n\|_A \leq \epsilon \, \|x - x_0\|_A .$$

This statement has been used to explain superlinear convergence of CG at the presence of large outliers in the spectrum. Due to rounding errors, this concept can not be applied in a meaningful way to practical computations. Recall the mathematical model of FP computations that is based on replacing large outlying eigenvalues by large clusters.

Condition number does not tell the story

Class of elliptic PDEs, frequently used example



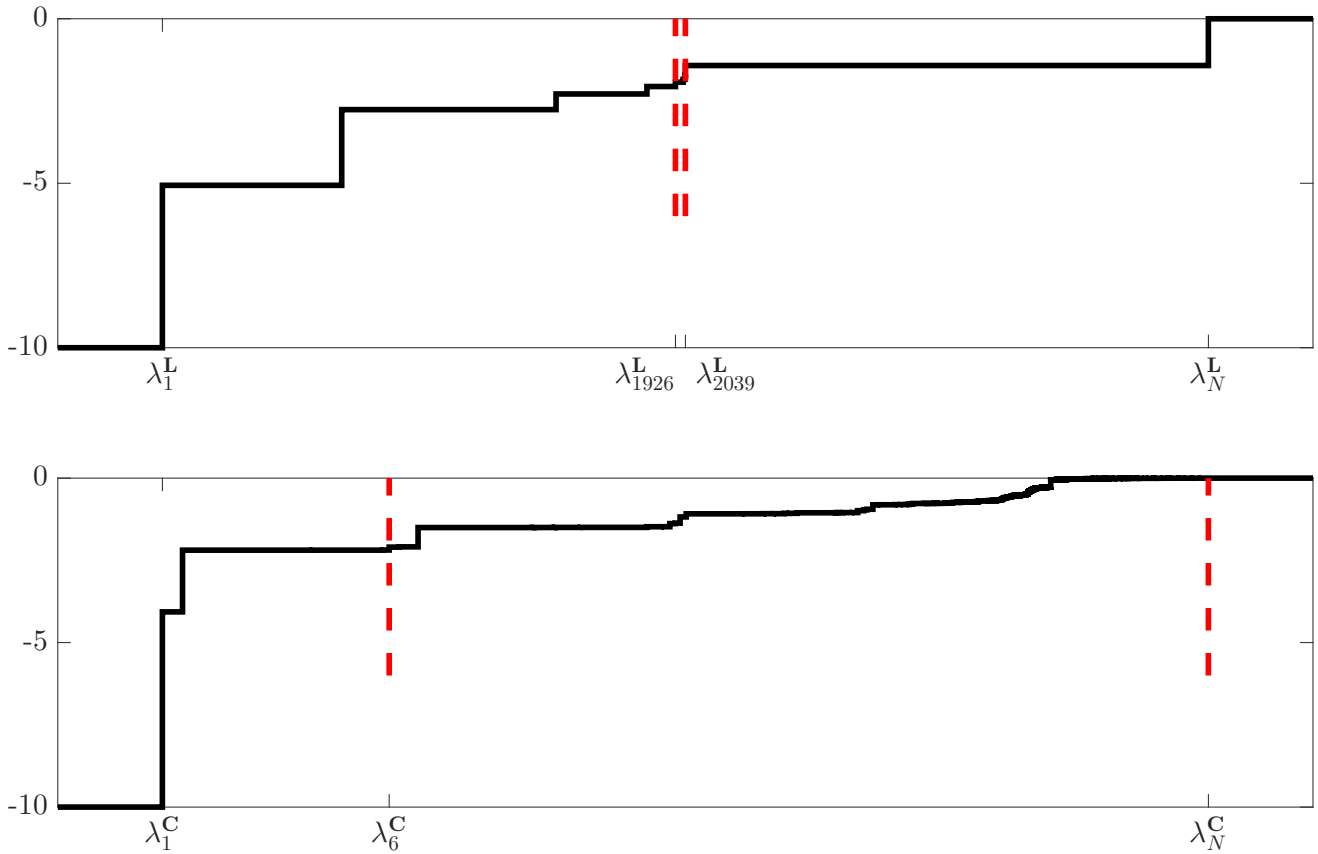


$$-\nabla\cdot(\,k(x)\,\nabla u\,)\;=\;0\,,$$

Morin, Nocheto, Siebert, SIREV (2002),
linear FE, standard uniform triangulation, $N = 3969$ DOF.

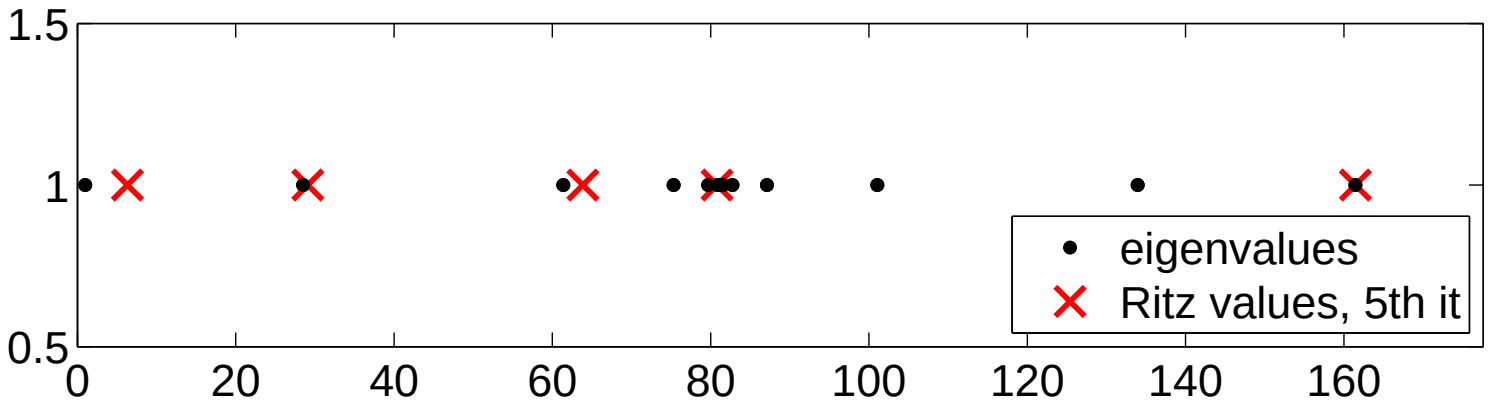
ICHOL PCG (drop-off tolerance 1e-02), $\kappa \approx 16$;
Laplace operator PCG, $\kappa \approx 160$.

1 Parts of the spectra and convergence behavior



(The horizontal scales are different.)

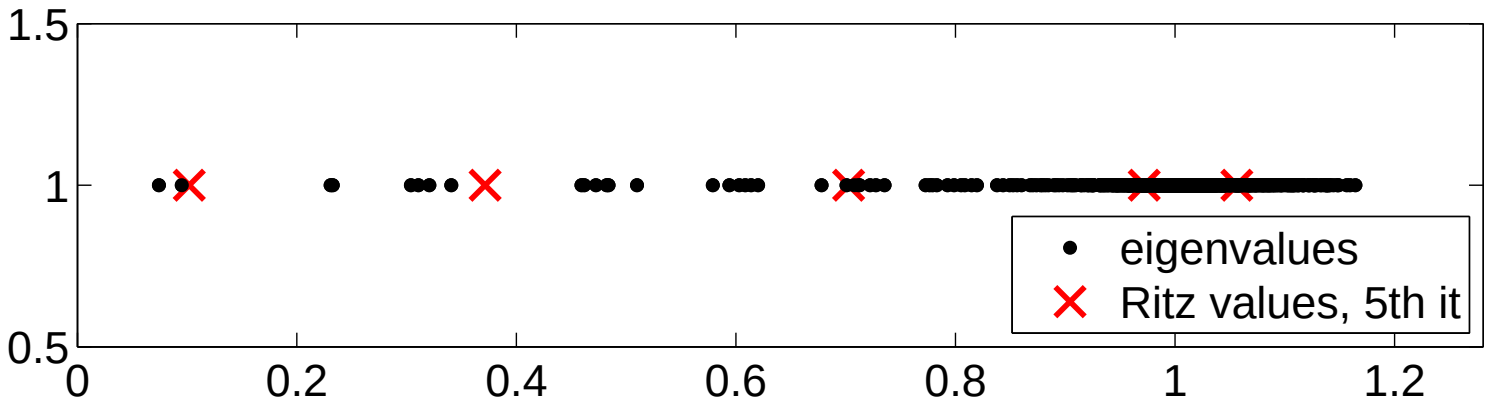
1 Ritz values at the 5th CG iteration - LAPL



Index	$1 - 1922$	1923	1924	1925	1926
Eigenvalues	1	28.508	61.384	75.324	$\lambda^{\mathbf{L}}_{1926} = 79.699$
Total weight	9×10^{-6}	$\approx 10^{-3}$	$\approx 10^{-3}$	$\approx 10^{-3}$	$\approx 10^{-3}$

Index	$1927 - 1930$	$1931 - 2039$	$2040 - 2047$	$2048 - 3969$
Eigenvalues	$80.875 - 81.222$	$\lambda^{\mathbf{L}}_{2039} = 81.224$	$81.226 - 133.94$	161.45
Total weight	$\approx 10^{-3}$	1.8×10^{-2}	8×10^{-10}	0.96

1 Ritz values at the 5th CG iteration - ICHOL



Index	1	2	3	4
Eigenvalues	0.074	0.095	0.231	0.233
Total weight	8×10^{-5}	6.4×10^{-3}	8×10^{-7}	10^{-8}

Index	5	6	7 – 3969
Eigenvalues	0.304	$\lambda_6^C = 0.311$	$0.321 - \lambda_{3969}^C = 1.1643$
Total weight	6×10^{-5}	1.5×10^{-3}	0.992

Thank you very much for your kind patience!