Matrix iterative methods from the historical, analytic, application, and computational perspective

Zdeněk Strakoš
(Charles University and Academy of Sciences of the Czech Republic, Prague)

IDGK Compact Course, TU Münich, November 2016

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods

  Matrices and operators in infinite dimensional Hilbert spaces.

  - Euler
  - Gauss
  - Jacobi
  - Chebyshev, Markov
  - Stieltjes
  - Hilbert, von Neumann
  - Krylov, Gantmakher
  - Lanczos, Hestenes, Stiefel

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods

Cornelius Lanczos, *Why Mathematics*, 1966

" In a recent comment on mathematical preparation an educator wanted to characterize our backwardness by the following statement: "Is it not astonishing that a person graduating in mathematics today knows hardly more than what Euler knew already at the end of the eighteenth century?". On its face value this sounds a convincing argument. Yet it misses the point completely. Personally I would not hesitate not only to graduate with first class honors, but to give the Ph.D. (and with summa cum laude) without asking any further questions, to anybody who knew only one quarter of what Euler knew, provided that he knew it in the way in which Euler knew it. "

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods
- *historical:* Without understanding the history we are confused in the presence and we will get lost in the future. This holds also for mathematics.

Cornelius Lanczos, *Linear Differential Operators*, 1961

"To get an explicit solution of a given boundary value problem is in this age of large electronic computers no longer a basic question. The problem can be coded for the machine and the numerical answer obtained. But of what value is the numerical answer if the scientist does not understand the peculiar analytical properties and idiosyncrasies of the given operator?"

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods
- *historical:* Without understanding the history we are confused in the presence and we will get lost in the future. This holds also for mathematics.
- *analytic:* The progress in computing technology and the need for solving practical problems forces us to think algorithmically and do things fast. Analytic view is by its nature slow and it does not keep up with the pace. But analytic view is absolutely crucial. It makes a little sense to progress fast in a wrong direction.

Cornelius Lanczos, *The Inspired Guess in the History of Physics*, 1964,

"Once the great mathematician Gauss was engaged in a particularly important investigation, but seemed to make little headway. His colleagues inquired when the publication was to appear. Gauss gave them an apparently paradoxical and yet perfectly correct answer: 'I have all the results but I don't know yet how I am going to get them'."

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods
- *historical:* Without understanding the history we are confused in the presence and we will get lost in the future. This holds also for mathematics.
- *analytic:* The progress in computing technology and the need for solving practical problems forces us to think algorithmically and do things fast. Analytic view is by its nature slow and it does not keep up with the pace. But analytic view is absolutely crucial. It makes a little sense to progress fast in a wrong direction.
- *application:* Development and application of mathematics lives in an unbreakable symbiosis. I do not believe in "pure" against "applied" mathematics. This division is artificial, caused by proudness and ambitions. As a malign disease it leads mathematics to fragmentation and the fields of mathematics to dangerous isolation. Applications are like a fresh water. Any application must honor the assumptions of the theory.

Meaning of the words - application

Henri Poincaré, 1909, graduate of the Polytechnique

"The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful. If nature were not beautiful, it would not be worth knowing, and if nature were not worth knowing, life would not be worth living. ...
I mean that deeper beauty coming from the harmonious order of the parts, and that a pure intelligence can grasp.

Science has had marvelous applications, but a science that would only have applications in mind would not be science anymore, it would be only cookery."

Meaning of the words

- *Matrix iterative methods* $\longrightarrow$ Krylov subspace methods
- *historical:* Without understanding the history we are confused in the presence and we will get lost in the future. This holds also for mathematics.
- *analytic:* The progress in computing technology and the need for solving practical problems forces us to think algorithmically and do things fast. Analytic view is by its nature slow and it does not keep up with the pace. But analytic view is absolutely crucial. It makes a little sense to progress fast in a wrong direction.
- *application:* Development and application of mathematics lives in an unbreakable symbiosis. I do not believe in "pure" against "applied" mathematics. This division is artificial, caused by proudness and ambitions. As a malign disease it leads mathematics to fragmentation and the fields of mathematics to dangerous isolation. Applications are like a fresh water. Any application must honor the assumptions of the theory.
- *computational:* Computing is a very involved process. Computers should serve in solving properly mathematically formulated problems. Mathematics must respect limitations of the computing technology.

Meaning of the words - computational

John von Neumann and Herman H. Goldstine, *Numerical ...* , 1947

"When a problem in pure or in applied mathematics is 'solved' by numerical computation, errors, that is, deviations of the numerical 'solution' obtained from the true, rigorous one, are unavoidable. Such a 'solution' is therefore meaningless, unless there is an estimate of the total error in the above sense.

Such estimates have to be obtained by a combination of several different methods, because the errors that are involved are aggregates of several different kinds of contributory, primary errors. These primary errors are so different from each other in their origin and character, that the methods by which they have to be estimated must differ widely from each other. A discussion of the subject may, therefore, advantageously begin with an analysis of the main kinds of primary errors, or rather of the sources from which they spring.

This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, ..."

Cornelius Lanczos, March 9, 1947

On (what are now called) the Lanczos and CG methods:

"The reason why I am strongly drawn to such approximation mathematics problems is ... the fact that a very "economical" solution is possible only when it is very "adequate".

To obtain a solution in very few steps means nearly always that one has found a way that does justice to the inner nature of the problem."

Albert Einstein, March 18, 1947

"Your remark on the importance of adapted approximation methods makes very good sense to me, and I am convinced that this is a fruitful mathematical aspect, and not just a utilitarian one."

Albert Einstein, March 18, 1947

"Your remark on the importance of adapted approximation methods makes very good sense to me, and I am convinced that this is a fruitful mathematical aspect, and not just a utilitarian one."

Main principle behind Krylov subspace methods:

Highly nonlinear adaptation of the iterations to the problem.

Conjugate Gradient (CG) method for $Ax = b$ with A HPD (1952)

$r_0 = b - Ax_0,\ p_0 = r_0\,.$ For $n = 1, \ldots, n_{\max}$:

$$
\begin{aligned}
\alpha_{n-1} &= \frac{r_{n-1}^* r_{n-1}}{p_{n-1}^* A p_{n-1}} \\
x_n &= x_{n-1} + \alpha_{n-1} p_{n-1}\,, \quad \text{stop when the stopping criterion is satisfied} \\
r_n &= r_{n-1} - \alpha_{n-1} A p_{n-1} \\
\beta_n &= \frac{r_n^* r_n}{r_{n-1}^* r_{n-1}} \\
p_n &= r_n + \beta_n p_{n-1}
\end{aligned}
$$

Here α_{n-1} ensures the minimization of $\|x - x_n\|_A$ along the line

$$z(\alpha) = x_{n-1} + \alpha p_{n-1}\,.$$

Mathematical elegance of CG

- Provided that

$$p_i \perp_A p_j, \quad i \neq j,$$

the one-dimensional line minimizations at the individual steps 1 to n result in the n-dimensional minimization over the whole shifted Krylov subspace

$$x_0 + \mathcal{K}_n(A, r_0) \;=\; x_0 + \operatorname{span}\{p_0, p_1, \ldots, p_{n-1}\}\,.$$

Mathematical elegance of CG

- Provided that

$$p_i \perp_A p_j, \quad i \neq j,$$

the one-dimensional line minimizations at the individual steps 1 to n result in the n-dimensional minimization over the whole shifted Krylov subspace

$$x_0 + \mathcal{K}_n(A, r_0) \;=\; x_0 + \operatorname{span}\{p_0, p_1, \ldots, p_{n-1}\}\,.$$

- The orthogonality condition leads to short recurrences due to the relationship to the orthogonal polynomials that define the algebraic residuals and search vectors.

Mathematical elegance of CG

- Provided that

$$p_i \perp_A p_j, \quad i \neq j,$$

the one-dimensional line minimizations at the individual steps 1 to n result in the n-dimensional minimization over the whole shifted Krylov subspace

$$x_0 + \mathcal{K}_n(A, r_0) \;=\; x_0 + \operatorname{span}\{p_0, p_1, \ldots, p_{n-1}\}\,.$$

- The orthogonality condition leads to short recurrences due to the relationship to the orthogonal polynomials that define the algebraic residuals and search vectors.

- Inexact computation?

Antoine de Saint-Exupéry , *The Wisdom of the Sands*, 1944

"I would not bid you pore upon a heap of stones, and turn them over and over, in the vain hope of learning from them the secret of meditation. For on the level of the stones there is no question of meditation; for that, the temple must have come into being. But, once it is built, a new emotion sways my heart, and when I go away, I ponder on the relations between the stones. ...

I must begin by feeling love; and I must first observe a wholeness. After that I may proceed to study the components and their groupings. But I shall not trouble to investigate these raw materials unless they are dominated by something on which my heart is set. Thus I began by observing the triangle as a whole; then I sought to learn in it the functions of its component lines. ...

So, to begin with, I practise contemplation. After that, if I am able, I analyse and explain. ...

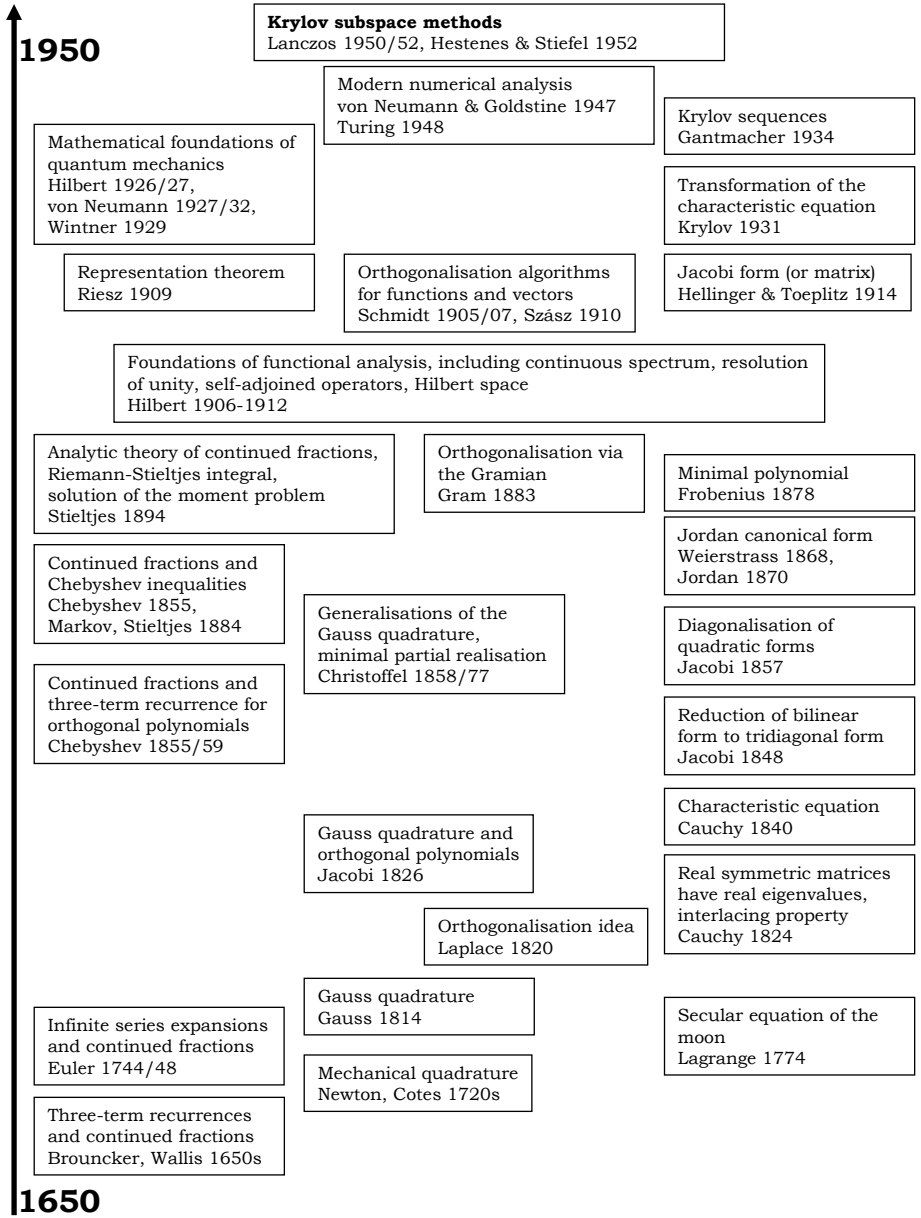
Little matter the actual things that are linked together; it is the links that I must begin by apprehending and interpreting."

Outline

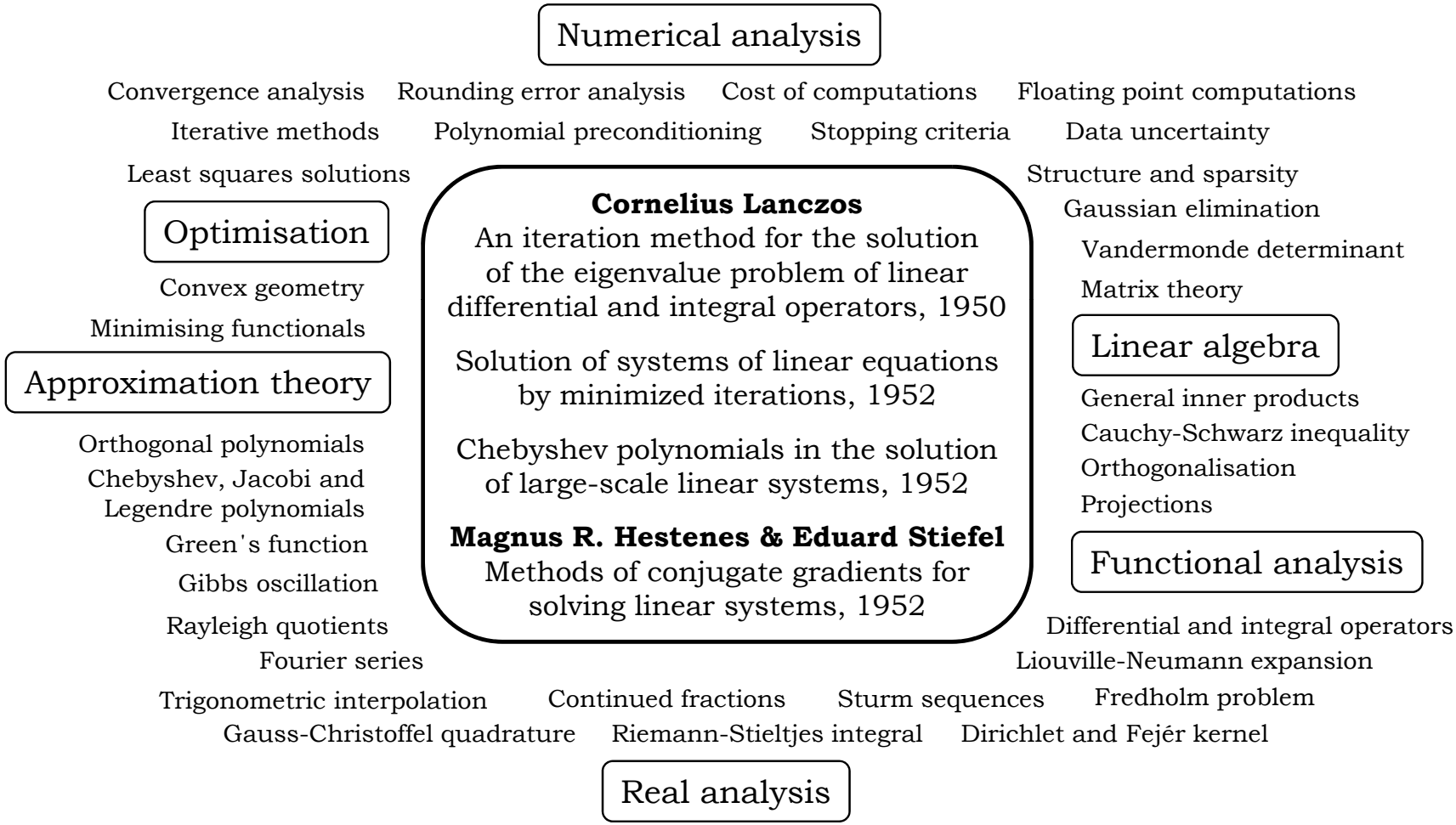
1. What are the Krylov subspace methods and what kind of mathematics is involved?
2. Linear projections onto highly nonlinear Krylov subspaces
3. Model reduction and moment matching
4. Convergence and spectral information
5. Inexact computations and numerical stability
6. Functional analysis and infinite dimensional considerations
7. Operator preconditioning, discretization and algebraic computation
8. HPC computations with Krylov subspace methods?
9. Myths about Krylov subspace methods

1. What are the Krylov subspace methods and what kind of mathematics is involved?

1 Historical development and context



2 Lanczos, Hestenes and Stiefel



1 Homework problem

Consider $2n$ real numbers $m_0, m_1, \ldots, m_{2n-1}$.
Solve the $2n$ equations

$$\sum_{j=1}^{n} \omega_j^{(n)} \{\theta_j^{(n)}\}^{\ell} = m_\ell\,, \qquad \ell = 0, 1, \ldots, 2n-1\,,$$

for the $2n$ real unknowns $\omega_j^{(n)} > 0,\ \theta_j^{(n)}$.

1 Homework problem

Consider $2n$ real numbers $m_0, m_1, \ldots, m_{2n-1}$.
Solve the $2n$ equations

$$\sum_{j=1}^{n} \omega_j^{(n)} \{\theta_j^{(n)}\}^{\ell} = m_\ell\,, \qquad \ell = 0, 1, \ldots, 2n-1\,,$$

for the $2n$ real unknowns $\omega_j^{(n)} > 0,\ \theta_j^{(n)}$.

Is this problem linear?
Does it look easy?
When does it have a solution?
How the solution can be determined?
How the solution can be computed?

Linear functional $\mathcal{L}(x)$ is positive definite on the space of polynomials $\mathcal{P}_n$ of degree at most n if its first $2n+1$ moments

$$\mathcal{L}(x^\ell) = m_\ell, \quad \ell = 0, 1, \ldots, 2n$$

are real and the Hankel matrix M_n of moments is positive definite, i.e., $\Delta_n > 0$, where

$$\Delta_n = |M_n| = \begin{vmatrix} m_0 & m_1 & \cdots & m_n \\ m_1 & m_2 & \cdots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \cdots & m_{2n} \end{vmatrix}.$$

With the positive definite $\mathcal{L}(x)$ we can restrict ourselves to real polynomials of a real variable and write, using a non-decreasing positive distribution function μ defined on the real axis having finite limits at $\pm\infty$,

$$\mathcal{L}(f) = \int f(x)\,\mathrm{d}\mu(x)\,,$$

with the inner product

$$(p,q) := \mathcal{L}(p(x)q(x)) = \int p(x)q(x)\,\mathrm{d}\mu(x)\,.$$

1 Stieltjes moment problem (1894) of order n

With the positive definite $\mathcal{L}(x)$ we can restrict ourselves to real polynomials of a real variable and write, using a non-decreasing positive distribution function μ defined on the real axis having finite limits at $\pm\infty$,

$$\mathcal{L}(f) = \int f(x)\,\mathrm{d}\mu(x)\,,$$

with the inner product

$$(p,q) := \mathcal{L}(p(x)q(x)) = \int p(x)q(x)\,\mathrm{d}\mu(x)\,.$$

Solution of the Stieltjes moment problem of order n exists and it is unique if and only if (with some $m_{2n} > 0$) we have $\Delta_n > 0$.

1 The unknown $\omega_j^{(n)}$, $\theta_j^{(n)}$?

- Cholesky decomposition of the matrix of moments $M_n = L_n L_n^T$
- The entries of the ℓth row of the the inverse L_n^{-1} give the coefficients of the ℓth orthonormal polynomial determined by the positive definite linear functional $\mathcal{L}(x)$ associated with the matrix of moments M_n .
- Roots of the ℓth orthogonal polynomial give the quadrature nodes $\theta_j^{(\ell)}$. The weights $\omega_j^{(\ell)}$ are given by the formula for the interpolatory quadrature.
- Computations are done differently

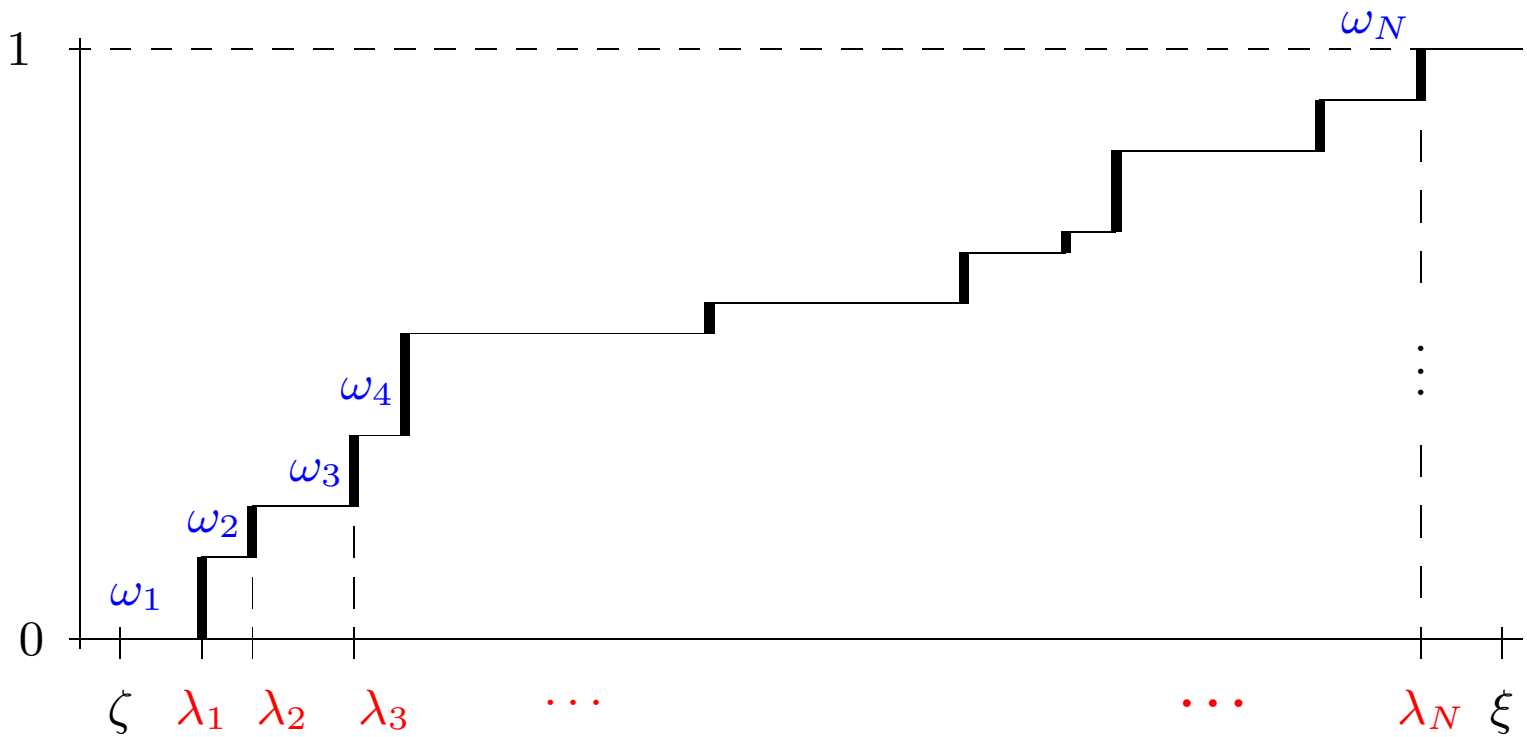
  (Gragg and Harrod, Gautschi, Laurie, ...)

  O'Leary, S, Tichý, *On Sensitivity of Gauss-Christoffel quadrature,* Numerische Mathematik, 107, 2007, pp. 147 –174

  Pranic, Pozza, S, *Gauss quadrature for quasi-definite linear functionals,* IMA J. Numer. Anal, 2016 (to appear)

Distribution function $\omega(\lambda)$ associated with $Ax = b,\ r_0 = b - Ax_0,\ A$ SPD:

λ_i, s_i are the eigenpairs of A, $\omega_i = |(s_i, w_1)|^2$



Symbolically

$$w_1^* A w_1 = w_1^* \left(\sum_{\ell=1}^{N} \lambda_\ell \, s_\ell s_\ell^* \right) w_1 \equiv w_1^* \left(\int_a^b \lambda \, dE(\lambda) \right) w_1$$
$$= \sum_{\ell=1}^{N} \lambda_\ell \, w_1^* s_\ell \, s_\ell^* w_1 = \sum_{\ell=1}^{N} \lambda_\ell \, \omega_\ell = \int_a^b \lambda \, d\omega(\lambda) \,,$$

where $dE(\lambda_\ell) \equiv s_\ell s_\ell^*$ and

$$I = \sum_{\ell=1}^{N} s_\ell s_\ell^* \equiv \int_a^b dE(\lambda) \,.$$

Hilbert (1906, 1912, 1928), Von Neumann (1927, 1932), Wintner (1929) .

2. Linear projections onto highly nonlinear Krylov subspaces

References:

J. Liesen. and Z.S., *Krylov Subspace Methods, Principles and Analysis.* Oxford University Press (2013), Chapter 2

- The Krylov sequence generated by $A \in \mathbb{C}^{N\times N}$ and $v \in \mathbb{C}^N$

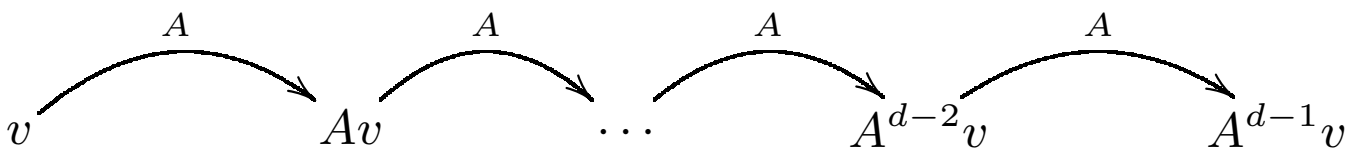
$$v, Av, A^2v, \ldots$$

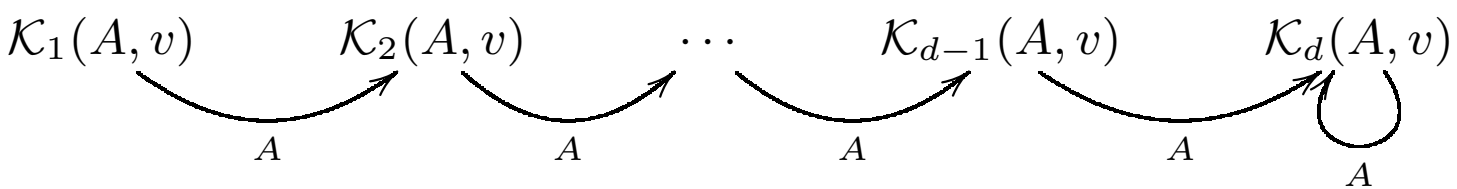
- The nth Krylov subspace generated by $A \in \mathbb{C}^{N\times N}$ and $v \in \mathbb{C}^N$

$$\mathcal{K}_n(A, v) := \operatorname{span}\{v, Av, \ldots, A^{n-1}v\}, \quad n = 1, 2, \ldots$$

- By construction,

$$\mathcal{K}_1(A, v) \subset \mathcal{K}_2(A, v) \subset \cdots \subset \mathcal{K}_d(A, v) = \mathcal{K}_{d+k}(A, v) \quad \text{for all } k \geq 1.$$





- Krylov subspace methods are based on a sequence of projections onto the nested Krylov subspaces that form the search spaces.

- Linear algebraic system $Ax = b$: x_0 (possibly zero), $r_0 = b - Ax_0$.

  $x_n \in x_0 + \mathcal{S}_n = x_0 + \mathcal{K}_n(A, r_0)$ such that $r_n = b - Ax_n \perp \mathcal{C}_n$, $n = 1, 2, \ldots$

- n-dimensional constraints space $\mathcal{C}_n$ determines the different methods.

- Eigenvalue problem $Ax = \lambda x$: v (nonzero), find (λ_n, x_n) such that

  $x_n \in \mathcal{K}_n(A, v)$ and $r_n = Ax_n - \lambda_n x_n \perp \mathcal{C}_n$.

- Examples: The Lanczos and Arnoldi methods, where $\mathcal{C}_n = \mathcal{K}_n(A, v)$.

- Krylov subspace methods are based on a sequence of projections onto the nested Krylov subspaces that form the search spaces.

- Linear algebraic system $Ax = b$: x_0 (possibly zero), $r_0 = b - Ax_0$.

  $x_n \in x_0 + \mathcal{S}_n = x_0 + \mathcal{K}_n(A, r_0)$ such that $r_n = b - Ax_n \perp \mathcal{C}_n$, $n = 1, 2, \ldots$

- n-dimensional constraints space $\mathcal{C}_n$ determines the different methods.

- Eigenvalue problem $Ax = \lambda x$: v (nonzero), find (λ_n, x_n) such that

  $$x_n \in \mathcal{K}_n(A, v) \quad \text{and} \quad r_n = Ax_n - \lambda_n x_n \perp \mathcal{C}_n.$$

- Examples: The Lanczos and Arnoldi methods, where $\mathcal{C}_n = \mathcal{K}_n(A, v)$.

- Method is well defined when x_n is uniquely determined for $n = 1, 2, \ldots, d-1$, and $x_d = x$ (in exact arithmetic).

- Conjugate gradient (CG) method: $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(A, r_0)$.
  - Well defined for HPD matrices A; short recurrences.
  - Orthogonality $r_n \perp \mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|x - x_n\|_A = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_A.$$

- GMRES method: $\mathcal{S}_n = \mathcal{K}_n(A, r_0)$, $\mathcal{C}_n = A\mathcal{K}_n(A, r_0)$.
  - Well defined for nonsingular matrices A; full recurrences.
  - Orthogonality $r_n \perp A\mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|b - Ax_n\|_2 = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|b - Az\|_2.$$

- Numerous other Krylov subspace methods. Some of them are not well defined in the above sense (e.g. BiCGStab or QMR).

- Method is well defined when x_n is uniquely determined for $n = 1, 2, \ldots, d-1$, and $x_d = x$ (in exact arithmetic).

- Conjugate gradient (CG) method: $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(A, r_0)$.
  - Well defined for HPD matrices A; short recurrences.
  - Orthogonality $r_n \perp \mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|x - x_n\|_A = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_A.$$

- GMRES method: $\mathcal{S}_n = \mathcal{K}_n(A, r_0)$, $\mathcal{C}_n = A\mathcal{K}_n(A, r_0)$.
  - Well defined for nonsingular matrices A; full recurrences.
  - Orthogonality $r_n \perp A\mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|b - Ax_n\|_2 = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|b - Az\|_2.$$

- Numerous other Krylov subspace methods. Some of them are not well defined in the above sense (e.g. BiCGStab or QMR).

- Method is well defined when x_n is uniquely determined for $n = 1, 2, \ldots, d-1$, and $x_d = x$ (in exact arithmetic).

- Conjugate gradient (CG) method: $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(A, r_0)$.
  - Well defined for HPD matrices A; short recurrences.
  - Orthogonality $r_n \perp \mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|x - x_n\|_A = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_A.$$

- GMRES method: $\mathcal{S}_n = \mathcal{K}_n(A, r_0)$, $\mathcal{C}_n = A\mathcal{K}_n(A, r_0)$.
  - Well defined for nonsingular matrices A; full recurrences.
  - Orthogonality $r_n \perp A\mathcal{K}_n(A, v)$ is equivalent to optimality:

$$\|b - Ax_n\|_2 = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|b - Az\|_2.$$

- Numerous other Krylov subspace methods. Some of them are not well defined in the above sense (e.g. BiCGStab or QMR).

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

with the formulation via the Lanczos process, $\quad w_1 = r_0/\|r_0\|\,,$

$$A\,W_n \;=\; W_n\,\mathbf{T_n} \;+\; \delta_{n+1} w_{n+1} \mathbf{e}_n^T\,, \quad \mathbf{T}_n \;=\; W_n^*(A, r_0)\,A\,W_n(A, r_0)\,,$$

and the CG approximation given by

$$\mathbf{T}_n\,\mathbf{y}_n \;=\; \|r_0\|\mathbf{e}_1\,, \quad x_n \;=\; x_0 \;+\; W_n\,\mathbf{y}_n\,.$$

$$A_n \;=\; Q_n\,A\,Q_n \;=\; W_n W_n^*\,A\,W_n W_n^* \;=\; W_n\,\mathbf{T}_n\,W_n^*\,,$$

Clearly, the projection process is very highly nonlinear in both A and r_0.

2 (Petrov-) Galerkin framework

Projection idea in Krylov subspace methods is analogous to the Galerkin framework in numerical solution of PDEs (here for convenience we take $\mathcal{C} = \mathcal{S}$).

Let $\mathcal{S}$ be an infinite dimensional Hilbert space, $a(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be a bounded and coercive bilinear form, $f : \mathcal{S} \to \mathbb{R}$ be a bounded linear functional.

- Weak formulation: Find $u \in \mathcal{S}$ with

$$a(u, v) = f(v) \quad \text{for all} \quad v \in \mathcal{S}.$$

- Discretization: Find $u_h \in \mathcal{S}_h \subset \mathcal{S}$ with

$$a(u_h, v_h) = f(v_h) \quad \text{for all} \quad v_h \in \mathcal{S}_h.$$

- Galerkin orthogonality:

$$a(u - u_h, v_h) = 0 \quad \text{for all} \quad v_h \in \mathcal{S}_h.$$

2 (Petrov-) Galerkin framework

Projection idea in Krylov subspace methods is analogous to the Galerkin framework in numerical solution of PDEs (here for convenience we take $\mathcal{C} = \mathcal{S}$).

Let $\mathcal{S}$ be an infinite dimensional Hilbert space, $a(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be a bounded and coercive bilinear form, $f : \mathcal{S} \to \mathbb{R}$ be a bounded linear functional.

- Weak formulation: Find $u \in \mathcal{S}$ with

$$a(u, v) = f(v) \quad \text{for all} \quad v \in \mathcal{S}.$$

- Discretization: Find $u_h \in \mathcal{S}_h \subset \mathcal{S}$ with

$$a(u_h, v_h) = f(v_h) \quad \text{for all} \quad v_h \in \mathcal{S}_h.$$

- Galerkin orthogonality:

$$a(u - u_h, v_h) = 0 \quad \text{for all} \quad v_h \in \mathcal{S}_h.$$

- Equivalently, there exists a bounded and coercive operator $\mathcal{A} : \mathcal{S} \to \mathcal{S}^{\#}$, with the problem formulated as the following equation in the dual space:

$$\mathcal{A}u = f.$$

- Or, using the Riesz map $\tau : \mathcal{S}^{\#} \to \mathcal{S}$ defined by the inner product in $\mathcal{S}$, as the following operator preconditioned equation in the function space

$$\tau\mathcal{A}u = \tau f.$$

- Discretization then gives

$$\tau_h\mathcal{A}_h u_h - \tau_h f_h \perp \mathcal{S}_h.$$

Krylov subspace methods (here CG for $\mathcal{A}$ self-adjoint with respect to the duality pairing) can be formulated in infinite dimensional Hilbert spaces and extended to Banach spaces.

$r_0 = f - \mathcal{A}u_0 \in \mathcal{S}^{\#}, \quad p_0 = \tau r_0 \in \mathcal{S}$. For $n = 1, 2, \ldots, n_{\max}$:

$$\alpha_{n-1} = \frac{\langle r_{n-1}, \tau r_{n-1} \rangle}{\langle \mathcal{A}p_{n-1}, p_{n-1} \rangle}$$

$u_n = u_{n-1} + \alpha_{n-1} p_{n-1}$, stop when the stopping criterion is satisfied

$$r_n = r_{n-1} - \alpha_{n-1} \mathcal{A}p_{n-1}$$

$$\beta_n = \frac{\langle r_n, \tau r_n \rangle}{\langle r_{n-1}, \tau r_{n-1} \rangle}$$

$$p_n = \tau r_n + \beta_n p_{n-1}$$

Superlinear convergence for (identity + compact) operators.

Karush (1952), Hayes (1954), Vorobyev (1958)

Here the Riesz map τ indeed serves as a preconditioner.

2 Summary and motivation

- Krylov subspace methods for solving linear algebraic problems are based on linear projections onto nested subspaces.
- Krylov subspaces and therefore the resulting methods are highly nonlinear in the data defining the problem.
- The nonlinearity allows to adapt to the problem as the iteration proceeds. This is not apparent, e.g., from the derivation of CG based on the minimization of the quadratic functional, and this fact has affected negatively the presentation of Krylov subspace methods in textbooks.
- The adaptation can be better understood via the model reduction and moment matching properties of Krylov subspace methods.

3. Model reduction and moment matching

References:

J. Liesen. and Z.S., *Krylov Subspace Methods, Principles and Analysis.* Oxford University Press (2013), Chapter 3

Jacobi matrix and the conjugate gradient method

$$T_n = \begin{pmatrix} \gamma_1 & \delta_2 & & & \\ \delta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \delta_n \\ & & & \delta_n & \gamma_n \end{pmatrix}$$

is the Jacobi matrix of the orthogonalization coefficients and the CG method is formulated by

$$T_n t_n = \|r_0\| e_1, \qquad x_n = x_0 + V_n t_n .$$

- Let the columns of $V_n = [v_1, \ldots, v_n]$ form an orthonormal basis of $\mathcal{K}_n(A, r_0)$.
- Matrix formulation of $x_n \in x_0 + \mathcal{K}_n(A, r_0)$ and $r_n \perp \mathcal{K}_n(A, r_0)$:

$$x_n = x_0 + V_n t_n$$

and $t_n \in \mathbb{C}^n$ is found by solving

$$V_n^* A V_n \, t_n = \|r_0\| e_1 .$$

- This can be viewed as a model reduction from a (large) system of order N to a (small) system of order n.
- Intuition: Projected system should capture fast a sufficient part of information contained in the original data.
- Intuition: Powering the operator tends to transfer dominant information as quickly as possible into the projected system.

- Let the columns of $V_n = [v_1, \ldots, v_n]$ form an orthonormal basis of $\mathcal{K}_n(A, r_0)$.
- Matrix formulation of $x_n \in x_0 + \mathcal{K}_n(A, r_0)$ and $r_n \perp \mathcal{K}_n(A, r_0)$:

$$x_n = x_0 + V_n t_n$$

and $t_n \in \mathbb{C}^n$ is found by solving

$$V_n^* A V_n \, t_n = \|r_0\| e_1 .$$

- This can be viewed as a model reduction from a (large) system of order N to a (small) system of order n.
- Intuition: Projected system should capture fast a sufficient part of information contained in the original data.
- Intuition: Powering the operator tends to transfer dominant information as quickly as possible into the projected system.

- Let A be HPD with spectral decomposition $A = Y\Lambda Y^*$, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_N$ (distinct eigenvalues for simplicity).
- Suppose $\omega_k = |(v_1, y_k)|^2 > 0$, $k = 1, \ldots, N$, and define the distribution function

$$\omega(\lambda) = \begin{cases} 0, & \text{if } \lambda < \lambda_1, \\ \sum_{k=1}^{\ell} \omega_k, & \text{if } \lambda_\ell \le \lambda < \lambda_{\ell+1}, \text{ for } \ell = 1, \ldots, N-1, \\ 1, & \text{if } \lambda_N \le \lambda. \end{cases}$$

- The moments of $\omega(\lambda)$ are given by

$$\int \lambda^k d\omega(\lambda) = \sum_{\ell=1}^{N} \omega_\ell \{\lambda_\ell\}^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \ldots$$

- Analogous construction applied to $T_n = V_n^* A V_n$ yields a distribution function $\omega^{(n)}(\lambda)$ with moments given by

$$\int \lambda^k d\omega^{(n)}(\lambda) = \sum_{\ell=1}^{n} \omega_\ell^{(n)} \{\lambda_\ell^{(n)}\}^k = e_1^T T_n^k e_1, \quad k = 0, 1, 2, \ldots$$

- Let A be HPD with spectral decomposition $A = Y\Lambda Y^*$, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_N$ (distinct eigenvalues for simplicity).
- Suppose $\omega_k = |(v_1, y_k)|^2 > 0$, $k = 1, \ldots, N$, and define the distribution function

$$\omega(\lambda) = \begin{cases} 0, & \text{if } \lambda < \lambda_1, \\ \sum_{k=1}^{\ell} \omega_k, & \text{if } \lambda_\ell \le \lambda < \lambda_{\ell+1}, \text{ for } \ell = 1, \ldots, N-1, \\ 1, & \text{if } \lambda_N \le \lambda. \end{cases}$$

- The moments of $\omega(\lambda)$ are given by

$$\int \lambda^k d\omega(\lambda) = \sum_{\ell=1}^{N} \omega_\ell \{\lambda_\ell\}^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \ldots$$

- Analogous construction applied to $T_n = V_n^* A V_n$ yields a distribution function $\omega^{(n)}(\lambda)$ with moments given by

$$\int \lambda^k d\omega^{(n)}(\lambda) = \sum_{\ell=1}^{n} \omega_\ell^{(n)} \{\lambda_\ell^{(n)}\}^k = e_1^T T_n^k e_1, \quad k = 0, 1, 2, \ldots$$

Let $\phi_0(\lambda) \equiv 1, \phi_1(\lambda), \ldots, \phi_n(\lambda)$ be the first $n+1$ orthonormal polynomials corresponding to the distribution function $\omega(\lambda)$.
Then, writing $\Phi_n(\lambda) = [\phi_0(\lambda), \ldots, \phi_{n-1}(\lambda)]^*$,

$$\lambda\, \Phi_n(\lambda) \;=\; T_n\, \Phi_n(\lambda) \;+\; \delta_{n+1}\, \phi_n(\lambda)\, e_n$$

represents the Stieltjes recurrence (1893-4), see Chebyshev (1855), Brouncker (1655), Wallis (1656), Toeplitz and Hellinger (1914) with the Jacobi matrix

$$T_n \equiv \begin{pmatrix} \gamma_1 & \delta_2 & & \\ \delta_2 & \gamma_2 & \ddots & \\ & \ddots & \ddots & \delta_n \\ & & \delta_n & \gamma_n \end{pmatrix}, \quad \delta_l > 0\,, \ell = 2, \ldots, n\,.$$

- $\omega^{(n)}(\lambda)$ is the distribution function determined by the n-node Gauss-Christoffel quadrature approximation of the Riemann-Stieltjes integral with $\omega(\lambda)$.

$$
\begin{array}{ccc}
A,\ v_1 = r_0/\|r_0\| & \longleftrightarrow & \omega(\lambda),\ \int f(\lambda)\, d\omega(\lambda) \\
\downarrow & & \downarrow \\
\text{Hermitian Lanczos/CG} & & \text{Gauss-Christoffel quadrature} \\
\downarrow & & \downarrow \\
T_n,\ e_1 & \longleftrightarrow & \omega^{(n)}(\lambda),\ \sum_{j=1}^{n} \omega_j^{(n)} f(\theta_j^{(n)})
\end{array}
$$

3 Continued fraction corresponding to $\omega(\lambda)$

$$\mathcal{F}_N(\lambda) \equiv \cfrac{1}{\lambda - \gamma_1 - \cfrac{\delta_2^2}{\lambda - \gamma_2 - \cfrac{\delta_3^2}{\lambda - \gamma_3 - \ldots \cfrac{\ddots}{\lambda - \gamma_{N-1} - \cfrac{\delta_N^2}{\lambda - \gamma_N}}}}}$$

The entries $\gamma_1, \ldots, \gamma_N$ and $\delta_2, \ldots, \delta_N$ represent coefficients of the Stieltjes recurrence.

$$b^*(\lambda I - A)^{-1} b \;=\; \int_L^U \frac{d\omega(\mu)}{\lambda - \mu} \;=\; \sum_{j=1}^{N} \frac{\omega_j}{\lambda - \lambda_j} \;=\; \frac{\mathcal{R}_N(\lambda)}{\mathcal{P}_N(\lambda)}\,,$$

$$\frac{\mathcal{R}_N(\lambda)}{\mathcal{P}_N(\lambda)} \;\equiv\; \mathcal{F}_N(\lambda)$$

The denominator $\mathcal{P}_n(\lambda)$ corresponding to the nth convergent $\mathcal{F}_n(\lambda)$ of $\mathcal{F}_N(\lambda)$, $n = 1, 2, \ldots$ is the nth orthogonal polynomial in the sequence determined by $\omega(\lambda)$; see Chebyshev (1855).

3 Fundamental relationship with Gauss quadrature

- The first $2n$ moments of the reduced model match those of the original model
- The n-node Gauss-Christoffel quadrature has algebraic degree $2n-1$, hence

$$v_1^* A^k v_1 = e_1^T T_n^k e_1 \quad \text{for} \quad k = 0, 1, \ldots, 2n-1.$$

- Moment matching properties can also be derived for non-Hermitian matrices using the Vorobyev method of moments
- For the infinite dimensional Hilbert spaces and self-adjoint bounded operators it was described by Vorobyev (1958, 1965).

Let $z_0, z_1, \ldots, z_n$ be $n+1$ linearly independent elements of Hilbert space V. Consider the subspace V_n generated by all possible linear combinations of $z_0, z_1, \ldots, z_{n-1}$ and construct a linear operator $\mathcal{B}_n$ defined on V_n such that

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

where $E_n z_n$ is the (orthogonal or oblique) projection of z_n onto V_n.

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

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

For the present $z_1, \ldots, z_n$ are assumed to be linearly independent. Determine a sequence of operators $\mathcal{B}_n$ defined on the sequence of nested subspaces V_n such that

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

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

Challenges:

- Convergence
- Krylov subspace methods in infinite dimensional Hilbert spaces?

4. Convergence and spectral information

References

- J. Liesen. and Z.S., *Krylov Subspace Methods, Principles and Analysis.* Oxford University Press (2013), Chapter 5, Sections 5.1 - 5.7
- T. Gergelits and Z.S., *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations,* Numer. Alg. 65, 759-782 (2014)

- The CG optimality property

$$\|x - x_n\|_A = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_A = \min_{p \in \mathcal{P}_n(0)} \|p(A)(x - x_0)\|_A$$

yields the convergence bounds

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)|$$

$$\leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n, \quad \kappa = \frac{\lambda_N}{\lambda_1}.$$

- The worst-case behavior of the method is completely determined by the distribution of the eigenvalues of A.
- The widely known κ-bound is derived using Chebyshev polynomials on the interval $[\lambda_1, \lambda_N]$. It does not depend on any other properties of A, b, x_0.
- The κ-bound is linear and it can not capture the adaptation of the CG method to the problem!

- The CG optimality property

$$\|x - x_n\|_A = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_A = \min_{p \in \mathcal{P}_n(0)} \|p(A)(x - x_0)\|_A$$

  yields the convergence bounds

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)|$$

$$\leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n, \quad \kappa = \frac{\lambda_N}{\lambda_1}.$$

- The worst-case behavior of the method is completely determined by the distribution of the eigenvalues of A.
- The widely known κ-bound is derived using Chebyshev polynomials on the interval $[\lambda_1, \lambda_N]$. It does not depend on any other properties of A, b, x_0.
- The κ-bound is linear and it can not capture the adaptation of the CG method to the problem!

Consider the desired accuracy ϵ, $\kappa_s(\mathbf{A}) \equiv \lambda_{N-s}/\lambda_1$. Then

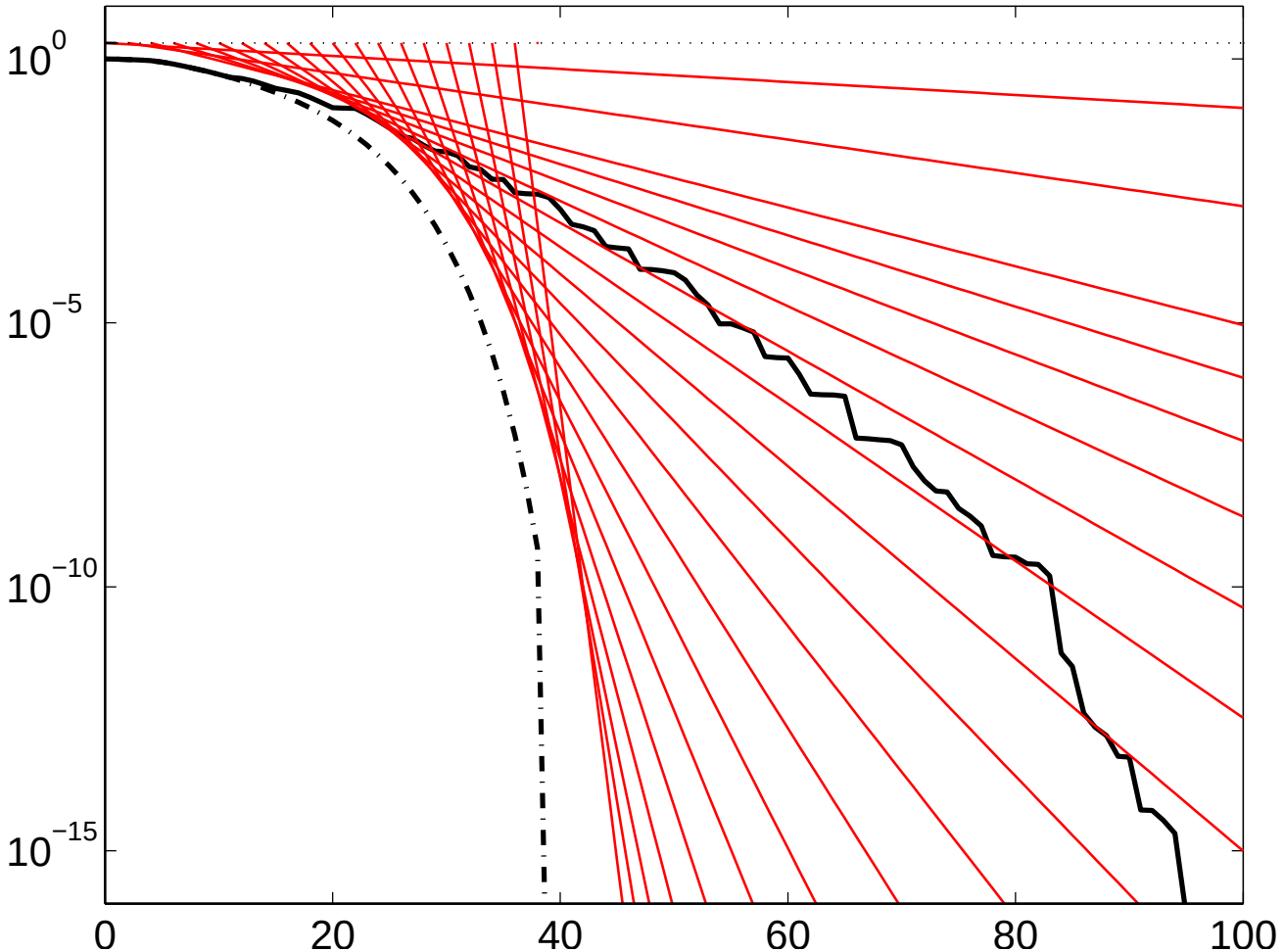
$$\mathbf{k} = \mathbf{s} + \left\lceil \frac{\ln(2/\epsilon)}{2} \sqrt{\kappa_s(\mathbf{A})} \right\rceil$$

CG steps will produce the approximate solution $\mathbf{x}_n$ satisfying

$$\|\mathbf{x} - \mathbf{x}_n\|_{\mathbf{A}} \leq \epsilon \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}} .$$

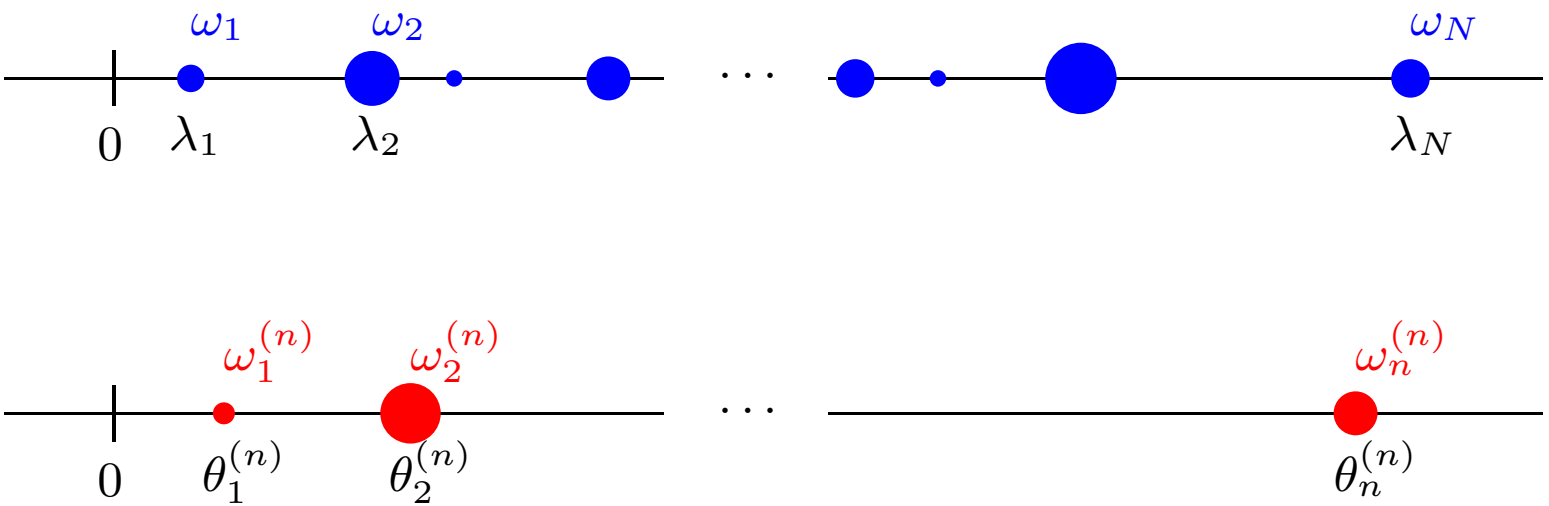
This statement qualitatively explains superlinear convergence of CG at the presence of large outliers in the spectrum, assuming exact arithmetic.

4 Adaptive Chebyshev bound?



4 Moment problem illustration

For a given n find a distribution function with n mass points in such a way that it in a best way captures the properties of the original distribution function



At any iteration step n, CG represents the matrix formulation of the n-point Gauss quadrature of the R-S integral determined by A and r_0,

$$\int f(\lambda)\, d\omega(\lambda) \;=\; \sum_{i=1}^{n} \omega_i^{(n)} f(\theta_i^{(n)}) \;+\; R_n(f)\,.$$

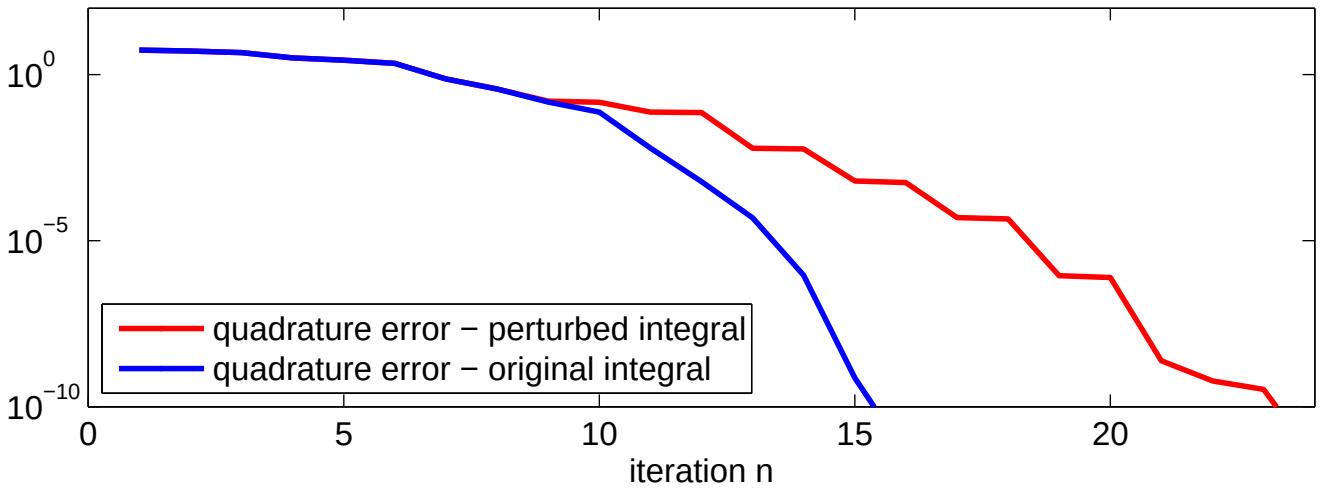
For $f(\lambda) \equiv \lambda^{-1}$ the formula takes the form

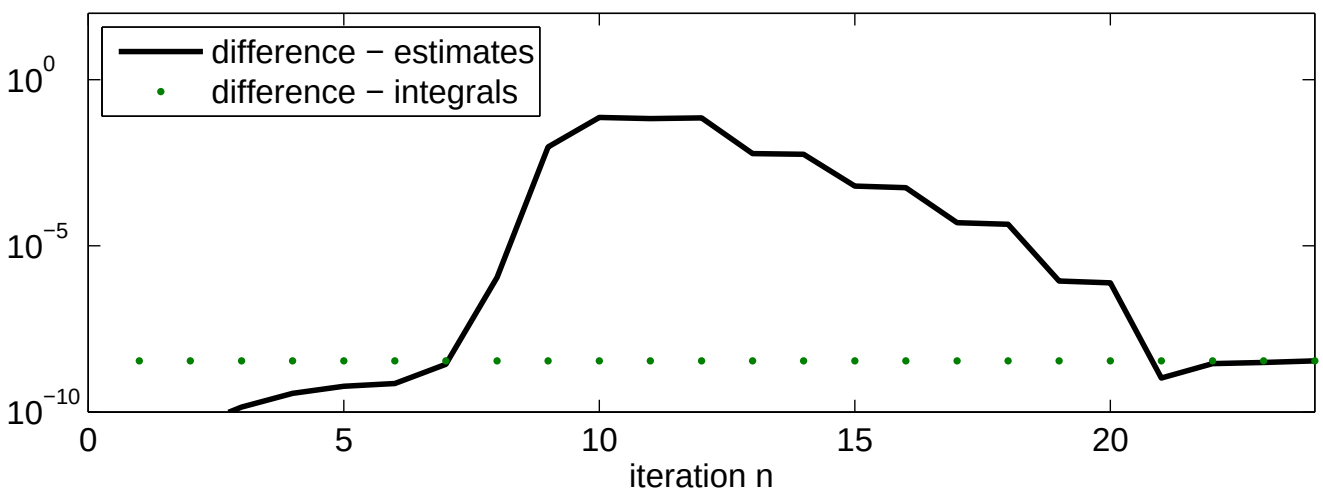
$$\frac{\|x - x_0\|_A^2}{\|r_0\|^2} \;=\; n\text{-th Gauss quadrature} \;+\; \frac{\|x - x_n\|_A^2}{\|r_0\|^2}\,.$$

This has became a base for the CG error estimation (see above); see the surveys in S and Tichý, 2002; Meurant and S, 2006; Liesen and S, 2013.

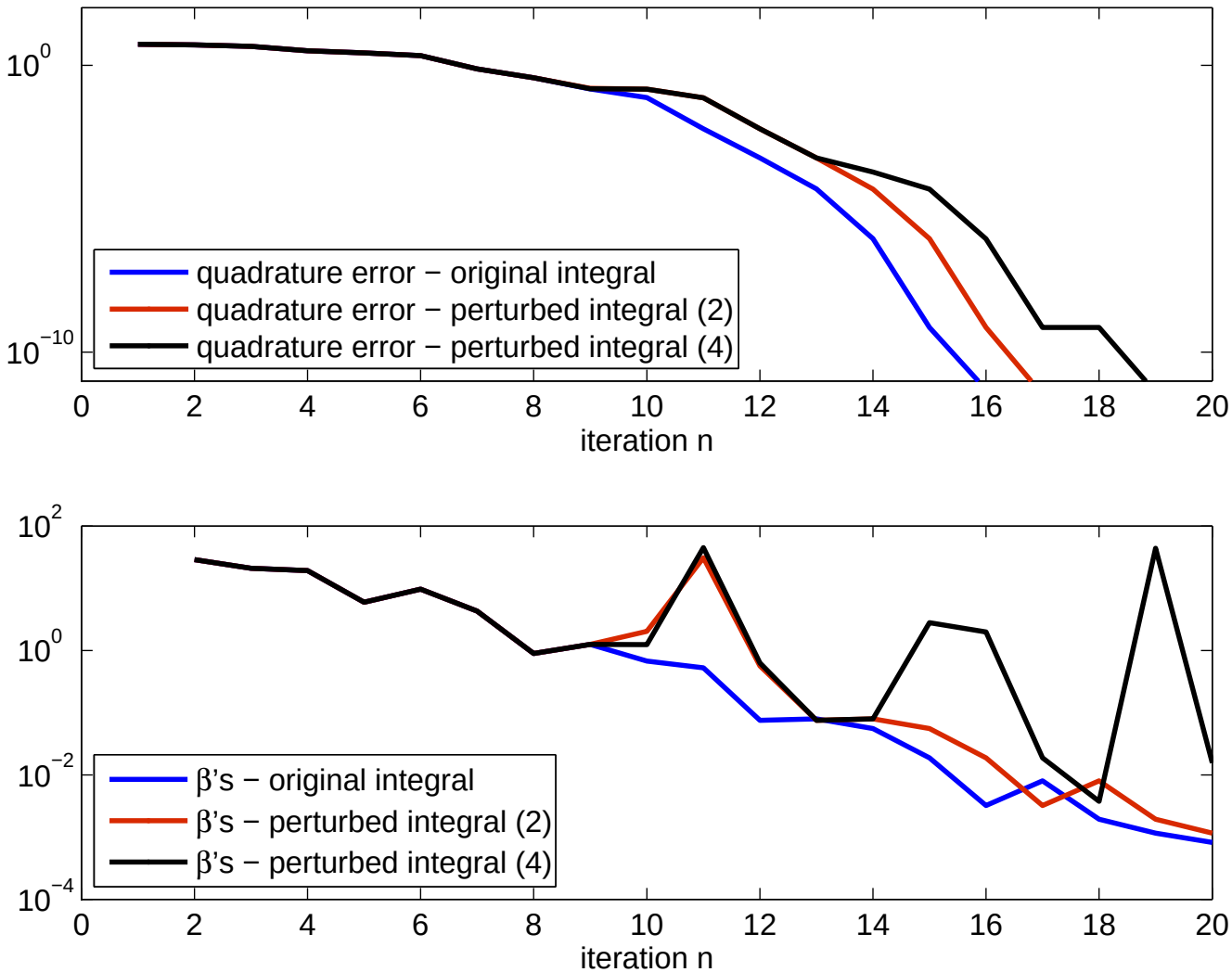
4 Gauss quadrature (CG) with tight clusters

- Replacing single eigenvalues by tight clusters can make a difference; see Greenbaum (1989); Greenbaum, S (1992); Golub, S (1994).
- The point is obvious. Orthogonal polynomials can be very sensitive to certain changes of the underlying distribution function.
- Otherwise CG behaves almost linearly and it can be described by contraction. In such case - is it worth using?


10^{0}
10^{-5}
10^{-10}
quadrature error − perturbed integral
quadrature error − original integral
0
5
10
15
20
iteration n


10^{0}
10^{-5}
10^{-10}
difference − estimates
difference − integrals
0
5
10
15
20
iteration n

4 Simplified problem - only the largest eigenvalue is replaced



4 Theorem - O'Leary, S, Tichý (2007)

Consider distribution functions $\omega(x)$ and $\tilde{\omega}(x)$. Let

$$p_n(x) = (x - x_1) \dots (x - x_n) \quad \text{and} \quad \tilde{p}_n(x) = (x - \tilde{x}_1) \dots (x - \tilde{x}_n)$$

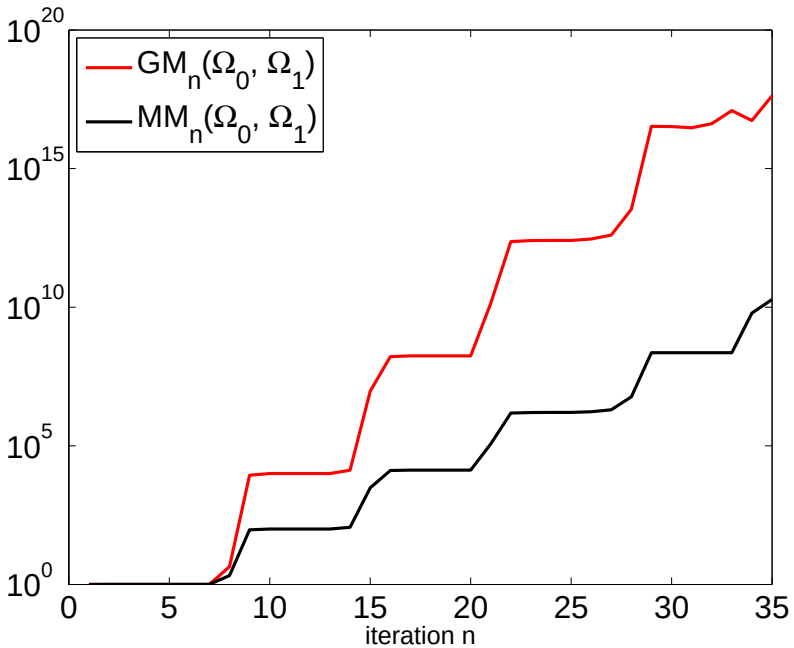
be the nth orthogonal polynomials corresponding to ω and $\tilde{\omega}$ respectively, with

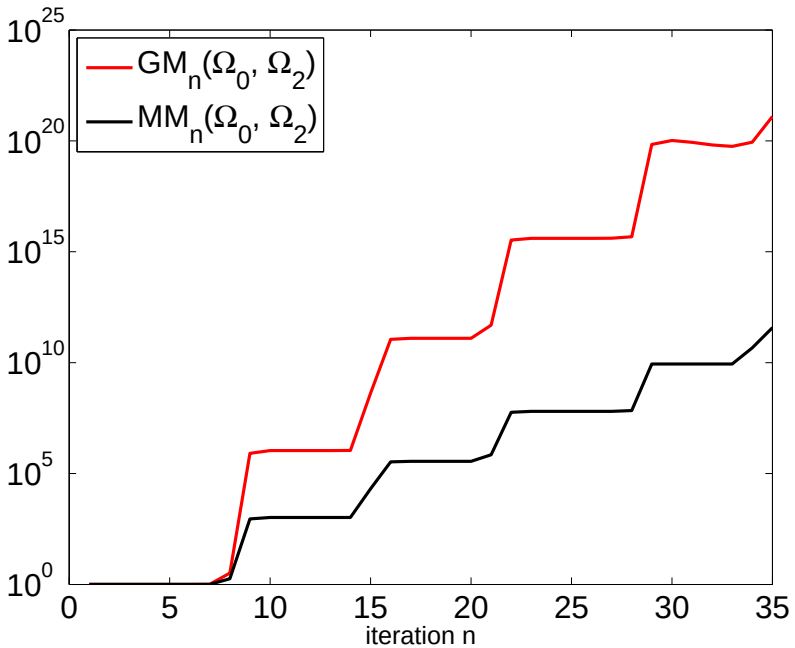
$$\hat{p}_c(x) = (x - \xi_1) \dots (x - \xi_c)$$

their least common multiple. If f'' is continuous, then the difference $\Delta^n_{\omega,\tilde{\omega}} = |I^n_\omega - I^n_{\tilde{\omega}}|$ between the approximations I^n_ω to I_ω and $I^n_{\tilde{\omega}}$ to $I_{\tilde{\omega}}$, obtained from the n-point Gauss quadrature, is bounded as

$$\begin{aligned} |\Delta^n_{\omega,\tilde{\omega}}| \quad &\leq \quad \left| \int \hat{p}_c(x) f[\xi_1, \dots, \xi_c, x] \, d\omega(x) - \int \hat{p}_c(x) f[\xi_1, \dots, \xi_c, x] \, d\tilde{\omega}(x) \right| \\ &+ \quad \left| \int f(x) \, d\omega(x) - \int f(x) \, d\tilde{\omega}(x) \right| . \end{aligned}$$

4 Modified moments do not tell the story





Condition numbers of the matrix of the modified moments (GM) and the matrix of the mixed moments (MM). Left - enlarged supports, right - shifted supports.

4 Summary of the CG/Lanczos part

1. Gauss-Christoffel quadrature for a small number of quadrature nodes can be highly sensitive to small changes in the distribution function enlarging its support.

2. In particular, the difference between the corresponding quadrature approximations (using the same number of quadrature nodes) can be many orders of magnitude larger than the difference between the integrals being approximated.

3. This sensitivity in Gauss-Christoffel quadrature can be observed for discontinuous, continuous, and even analytic distribution functions, and for analytic integrands uncorrelated with changes in the distribution functions and with no singularity close to the interval of integration.

- For diagonalizable $A = Y\Lambda Y^{-1}$ the GMRES optimality property

$$\|r_n\|_2 = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|b - Az\|_2 = \min_{p \in \mathcal{P}_n(0)} \|p(A) r_0\|_2$$

  yields the convergence bound

$$\frac{\|r_n\|_2}{\|r_0\|_2} \leq \kappa(Y) \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)|.$$

- The eigenvalue distribution and the GMRES convergence are (closely) related only when $\kappa(Y)$ is small (A is close to normal).
- In general, the eigenvalues alone do not describe GMRES convergence:
- Any non-increasing convergence curve is attainable by GMRES for a matrix having any prescribed set of eigenvalues.

- For diagonalizable $A = Y\Lambda Y^{-1}$ the GMRES optimality property

$$\|r_n\|_2 = \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|b - Az\|_2 = \min_{p \in \mathcal{P}_n(0)} \|p(A) r_0\|_2$$

  yields the convergence bound

$$\frac{\|r_n\|_2}{\|r_0\|_2} \leq \kappa(Y) \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)|.$$

- The eigenvalue distribution and the GMRES convergence are (closely) related only when $\kappa(Y)$ is small (A is close to normal).
- In general, the eigenvalues alone do not describe GMRES convergence:
- Any non-increasing convergence curve is attainable by GMRES for a matrix having any prescribed set of eigenvalues.

4 Any GMRES convergence with any spectrum

Given any spectrum and any sequence of the nonincreasing residual norms, a complete parametrization is known of the set of all GMRES associated matrices and right hand sides.

The set of problems for which the distribution of eigenvalues alone does not correspond to convergence behavior is not of measure zero and it is not pathological.

- Widespread eigenvalues alone can not be identified with poor convergence.
- Clustered eigenvalues alone can not be identified with fast convergence.

Equivalent orthogonal matrices; pseudospectrum indication.

4 Any GMRES convergence with any spectrum

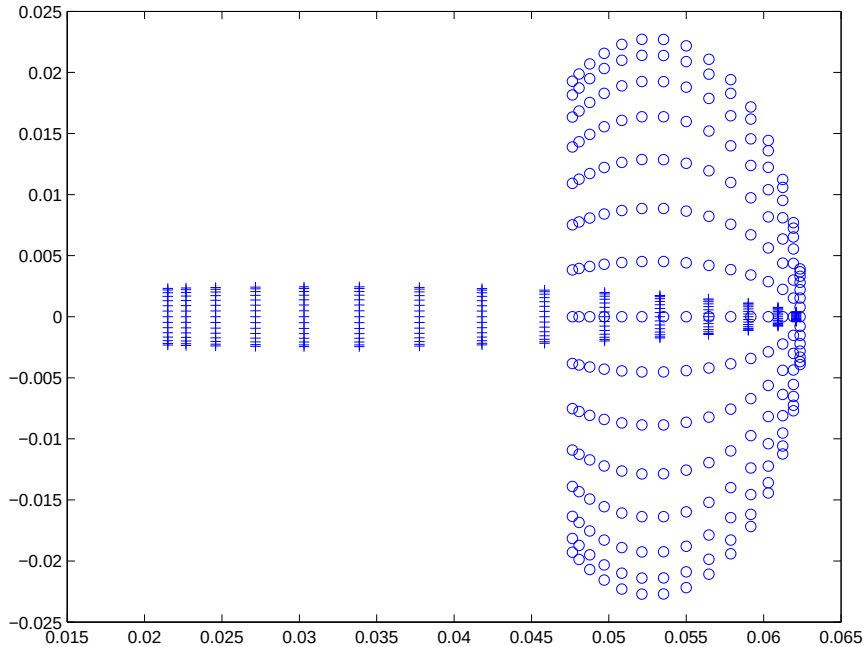
1° The spectrum of $\mathbf{A}$ is given by $\{\lambda_1, \ldots, \lambda_N\}$ and GMRES$(\mathbf{A}, \mathbf{b})$ yields residuals with the prescribed nonincreasing sequence $(x_0 = 0)$

$$\|\mathbf{r}_0\| \geq \|\mathbf{r}_1\| \geq \cdots \geq \|\mathbf{r}_{N-1}\| > \|\mathbf{r}_N\| = 0\,.$$

2° Let $\mathbf{C}$ be the spectral companion matrix, $h = (h_1, \ldots, h_N)^T$, $h_i^2 = \|\mathbf{r}_{i-1}\|^2 - \|\mathbf{r}_i\|^2$, $i = 1, \ldots, N$. Let $\mathbf{R}$ be a nonsingular upper triangular matrix such that $\mathbf{Rs} = \mathbf{h}$ with $\mathbf{s}$ being the first column of $\mathbf{C}^{-1}$, and let $\mathbf{W}$ be unitary matrix. Then

$$\mathbf{A} = \mathbf{WRCR}^{-1}\mathbf{W}^* \quad \text{and} \quad \mathbf{b} = \mathbf{Wh}\,.$$

Greenbaum, Pták, Arioli and S (1994 - 98); Liesen (1999); Eiermann and Ernst (2001); Meurant (2012); Meurant and Tebbens (2012, 2014); .....



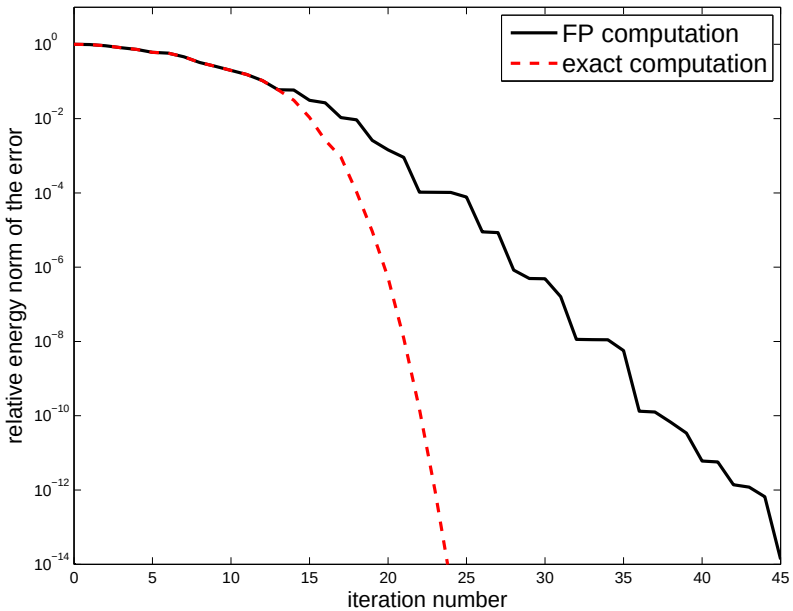
Quiz: In one case the convergence of GMRES is substantially faster than in the other; for the solution see Liesen, S (2005).

5. Inexact computation and numerical stability

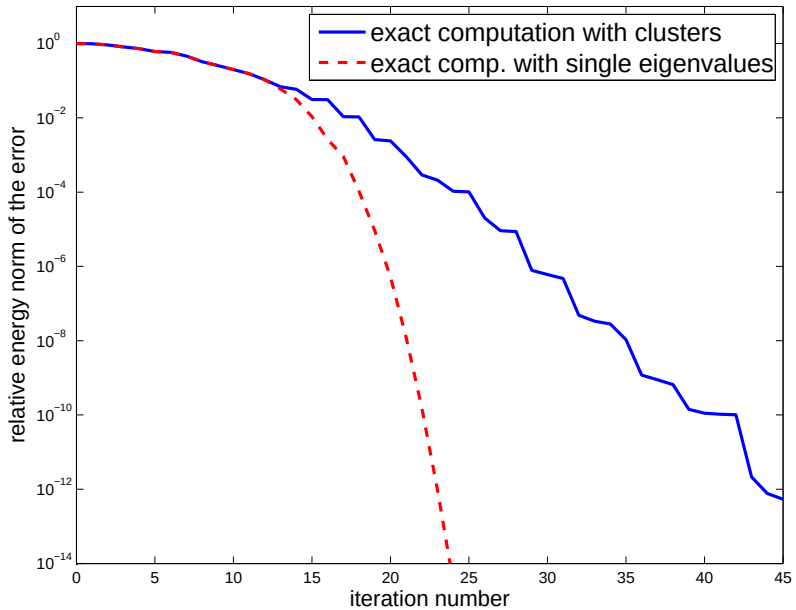
References

- J. Liesen. and Z.S., *Krylov Subspace Methods, Principles and Analysis.* Oxford University Press (2013), Chapter 5, Sections 5.8 - 5.11
- T. Gergelits and Z.S., *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations,* Numer. Alg. 65, 759-782 (2014)

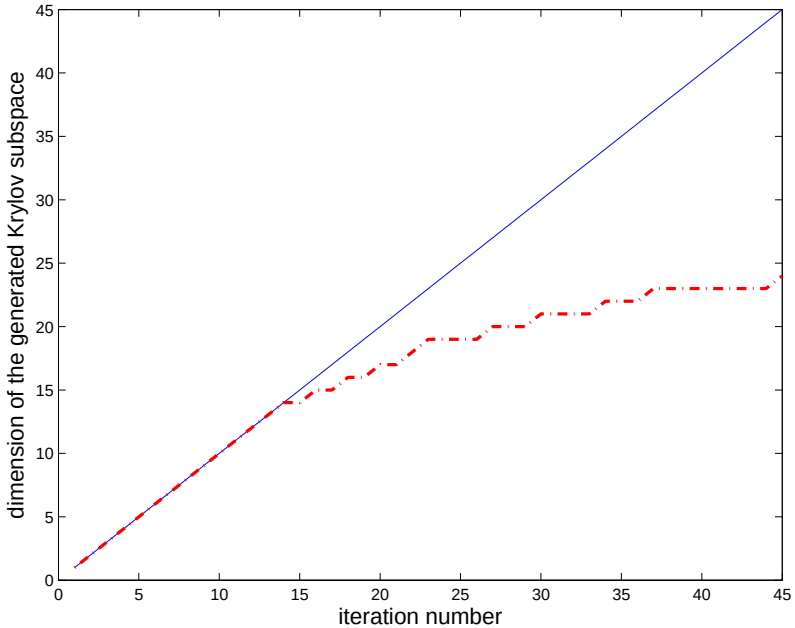
5 CG convergence behavior in finite precision arithmetic



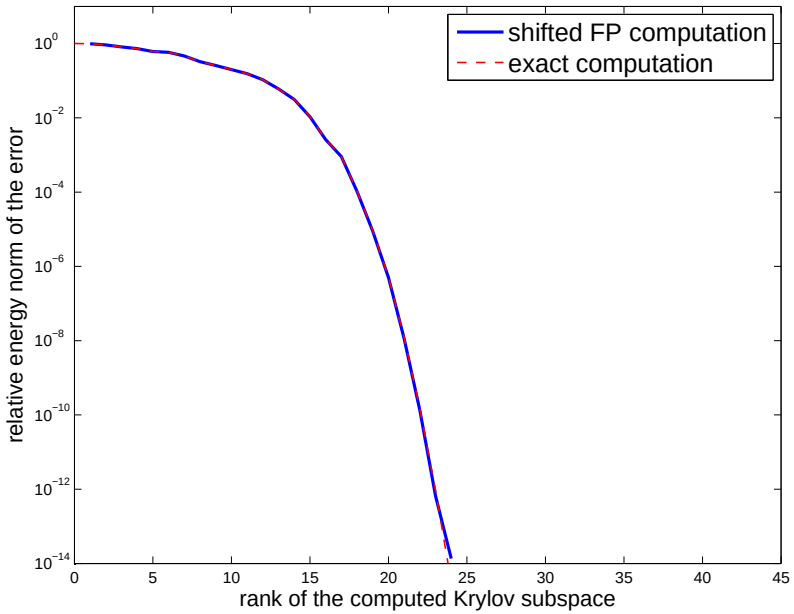
Rounding errors in finite precision CG computations cause a delay of convergence.



CG in finite precision corresponds to an exact CG computation for a matrix, where each eigenvalue is replaced by a tight cluster.



The number of steps of the delay correspond to the rank-deficiency of the computed Krylov subspaces.



Shifting the finite precision curve by the number of delayed iteration steps yields the curve for the exact computation.

- The statements above can be proven by rigorous mathematical means!

CG in finite precision arithmetic can be seen as the exact arithmetic CG for the problem with the slightly modified distribution function with larger support, i.e., with single eigenvalues replaced by tight clusters.

Paige (1971-80), Greenbaum (1989),
Parlett (1990), S (1991), Greenbaum and S (1992), Notay (1993), ... , Druskin, Kniznermann, Zemke, Wülling, Meurant, ...
Recent reviews and updates in Meurant and S, Acta Numerica (2006); Meurant (2006); Liesen and S (2013).

One particular consequence is becoming very relevant: In FP computations, the composite convergence bounds eliminating large outlying eigenvalues at the cost of one iteration per eigenvalue (see Axelsson (1976), Jennings (1977)) are not valid.

5 Optimality in finite precision Lanczos (CG) computations?

- In exact arithmetic, local orthogonality properties of CG are equivalent to the global orthogonality properties and therefore also to the CG optimality recalled above.
- In finite precision arithmetic the local orthogonality properties are preserved proportionally to machine precision, but the global orthogonality and therefore the optimality wrt the underlying distribution function is lost.
- in finite precision arithmetic computations (or, more generally, in inexact Krylov subspace methods) the optimality property does not have any easily formulated meaning with respect to the subspaces generated by the computed residual (or direction) vectors.
- Using the results of Greenbaum from 1989, it does have, however, a well defined meaning with respect to the particular distribution functions defined by the original data and the rounding errors in the steps 1 trough n.

5 Optimality in finite precision Lanczos (CG) computations?

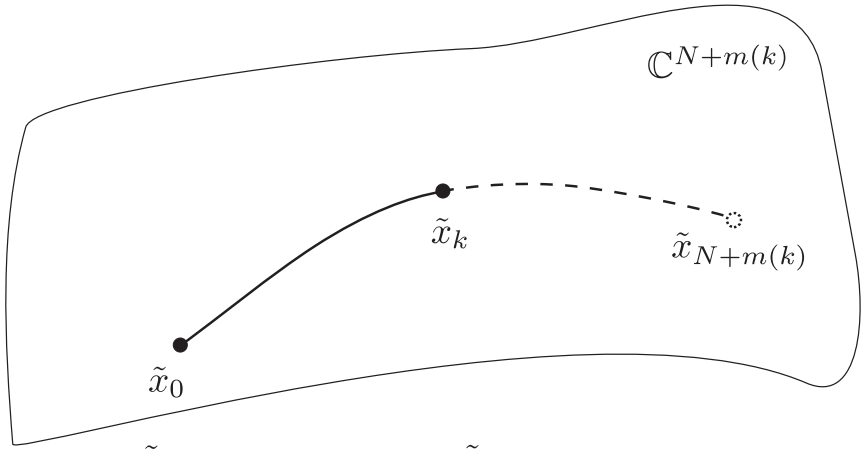
Consider the following mathematically equivalent formulation of CG

$$A\,W_n \;=\; W_n\,\mathbf{T_n} \;+\; \delta_{n+1} w_{n+1}\mathbf{e}_n^T\,, \quad \mathbf{T}_n \;=\; W_n^*(A, r_0)\,A\,W_n(A, r_0)\,,$$

and the CG approximation given by

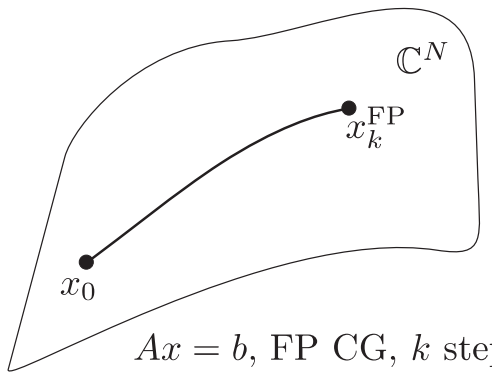
$$\mathbf{T}_n\,\mathbf{y}_n \;=\; \|r_0\|\mathbf{e}_1\,, \quad x_n \;=\; x_0 \;+\; W_n\,\mathbf{y}_n\,.$$

- Greenbaum proved that the Jacobi matrix computed in finite precision arithmetic can be considered a left principal submatrix of a certain larger Jacobi matrix having all its eigenvalues close to the eigenvalues of the original matrix A.
- This is equivalent to saying that convergence behavior in the first n steps of the given finite precision Lanczos computation can equivalently be described as the result of the exact Gauss quadrature for certain distribution function that depends on n having tight clusters of points of increase around the original eigenvalues of A.



$\tilde{A}_{N+m(k)}\tilde{x}_{N+m(k)} = \tilde{b}_{N+m(k)}$, EXACT CG

$\tilde{T}_k \equiv T_k^{(\mathrm{FP})}$

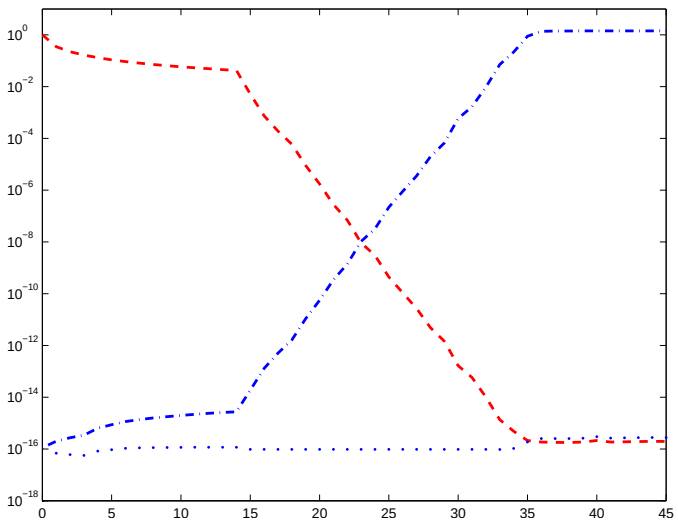


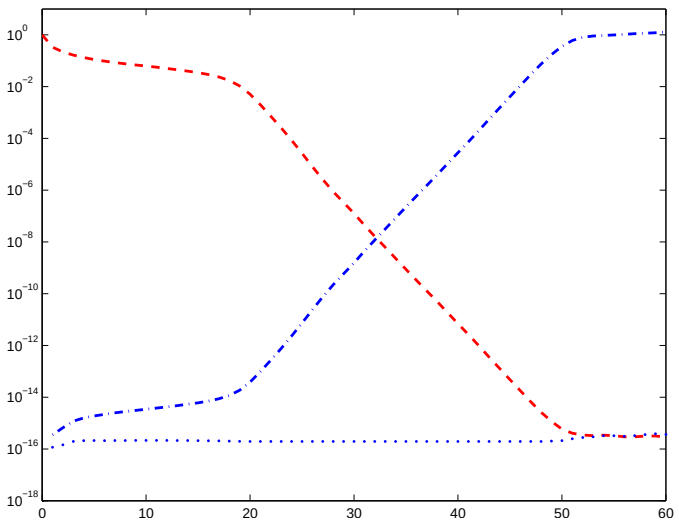
$Ax = b$, FP CG, k step

- In finite precision, the loss of orthogonality using the modified Gram-Schmidt GMRES is inversely proportional to the normwise relative backward error

$$\frac{\|b - Ax_n\|_2}{\|b\|_2 + \|A\|_2\|x_n\|_2}.$$

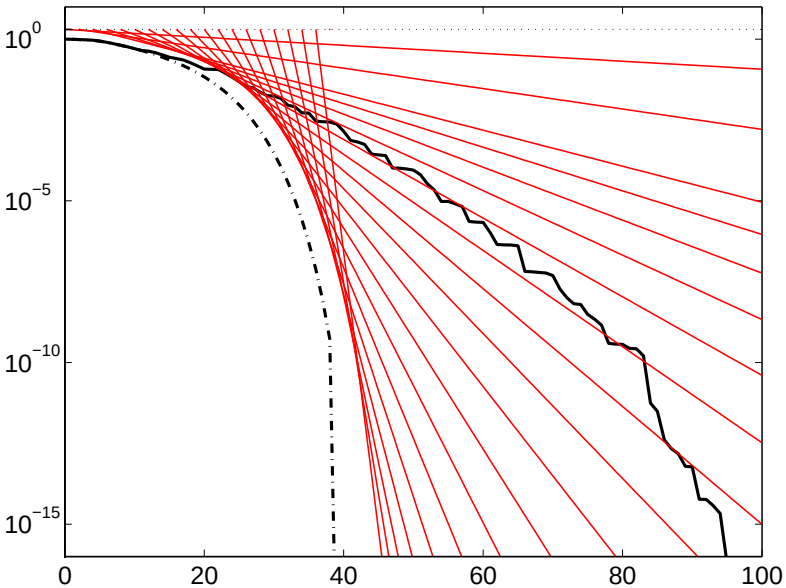
Loss of orthogonality (blue) and normwise relative backward error (red) for a convection-diffusion model problem with two different "winds":



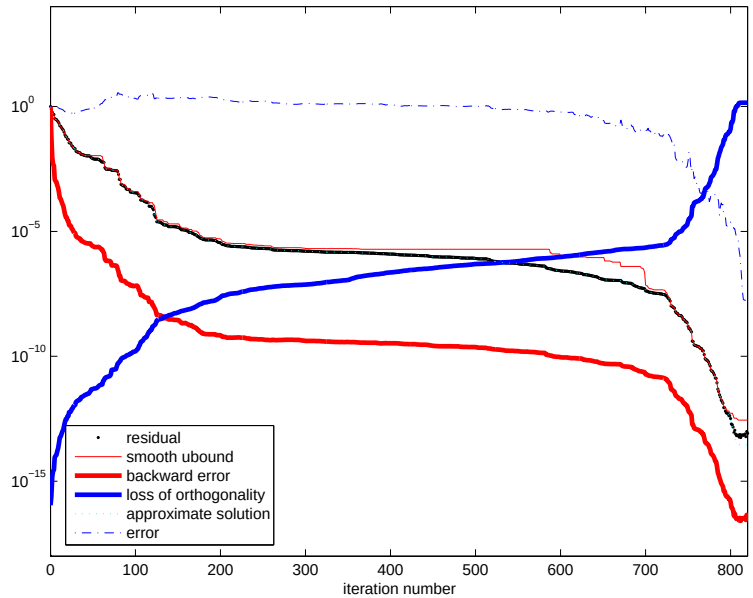


- It can be shown that the MGS-GMRES is normwise backward stable.

5 Delay of convergence due to inexactness

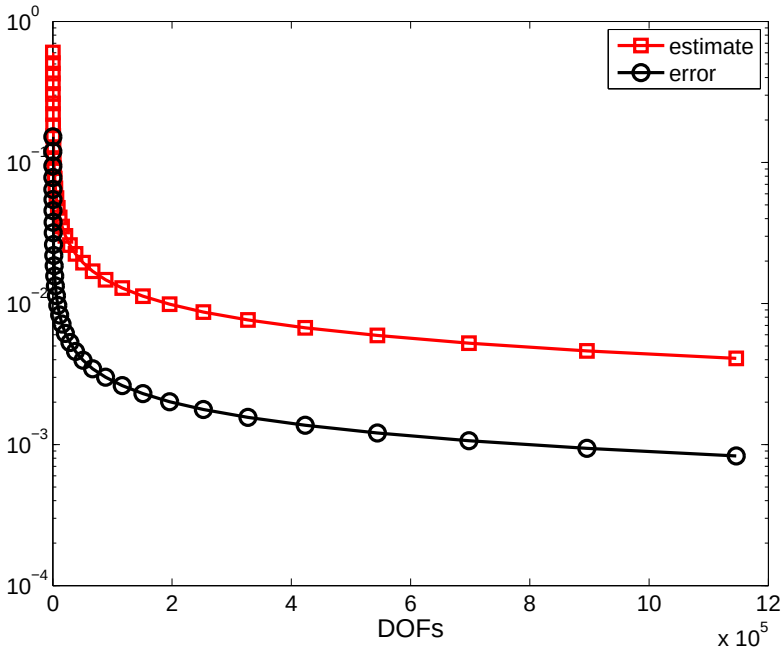


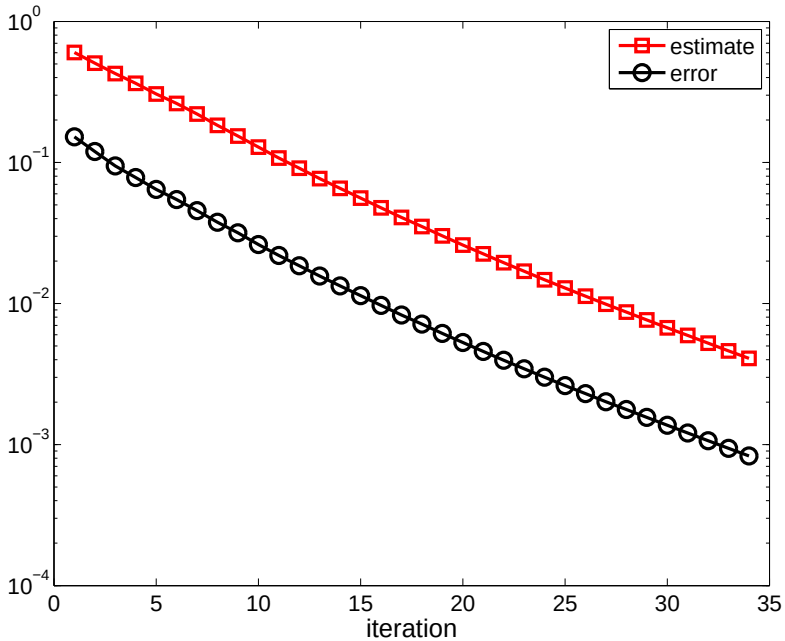
?



Here numerical inexactness due to roundoff. How much may we relax accuracy of the most costly operations without causing an unwanted delay and/or affecting the maximal attainable accuracy? That will be crucial in exascale computations.

5 Reaching an arbitrary accuracy in AFEM?





Inexactness and maximal attainable accuracy in matrix computations?

6. Functional analysis and infinite dimensional considerations

References

- J. Málek and Z.S., *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlight Series, SIAM (2015), Chapter 9

Let V be an infinite dimensional Hilbert space, $\mathcal{B}$ a bounded linear operator on V that has a bounded inversion. Consider the problem

$$\mathcal{B}\,u \;=\; f\,, \quad f \;\in\; V\,.$$

- The identity operator on an infinite dimensional Hilbert space is not compact.
- Since $\mathcal{B}\mathcal{B}^{-1} = \mathcal{I}\,,$ it follows that $\mathcal{B}$ can not be compact.
- Approximation of $\mathcal{B}$ by finite dimensional operators $\mathcal{B}_n : V \to V_n\,,$ V_n is finite dimensional?

- A uniform (in norm) limit of finite dimensional operators $\mathcal{B}_n$ is a compact operator.
- Every compact operator on a Hilbert space is a uniform limit of a sequence of finite dimensional operators.
- A uniform limit of compact operators is a compact operator.

Bounded invertible operators in Hilbert (holds also for Banach) spaces can not be approximated in norm to an arbitrary accuracy by neither compact nor finite dimensional operators! Approximation can be considered only in the sense of strong convergence (pointwise limit); for the method of moments see Vorobyev (1958, 1965)

$$\|\mathcal{B}_n\, w - \mathcal{B}\, w\| \to 0 \quad \forall w \in V\,.$$

Let $\mathcal{Z}_h$ be a numerical approximation of the bounded operator $\mathcal{Z}$ such that, with an appropriate extension, $\|\mathcal{Z} - \mathcal{Z}_h\| = \mathcal{O}(h)$.

Then we have $[(\lambda - \mathcal{Z})^{-1} - (\lambda - \mathcal{Z}_h)^{-1}] = \mathcal{O}(h)$ uniformly for $\lambda \in \Gamma$, where Γ surrounds the spectrum of $\mathcal{Z}$ with a distance of order $\mathcal{O}(h)$ or more. For any polynomial p

$$p(\mathcal{Z}) - p(\mathcal{Z}_h) = \frac{1}{2\pi i} \int_\Gamma p(\lambda)[(\lambda - \mathcal{Z})^{-1} - (\lambda - \mathcal{Z}_h)^{-1}]\, d\lambda\,,$$

and it seems that one can investigate $p(\mathcal{Z})$ instead of $p(\mathcal{Z}_h)$.

But the *assumption* $\|\mathcal{Z} - \mathcal{Z}_h\| = \mathcal{O}(h)\,,\ h \to 0$ does not hold for any bounded invertible infinite dimensional operator $\mathcal{Z}$.

7. Operator preconditioning, discretization and algebraic computation

References

- J. Málek and Z.S., *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlight Series, SIAM (2015)
- J. Papež, J.Liesen and Z.S., *Distribution of the discretization and algebraic error in numerical solution of partial differential equations,* Linear Alg. Appl. 449, 89-114 (2014)
- J. Papež, Z.S., and M. Vohralík *Estimating and localizing the algebraic and total numerical errors using flux reconstructions,* (2016, submitted for publication)
- J. Papež and Z.S., *Subtleties of the residual-based a posteriori error estimator for total error,* (2016, submitted for publication)

R. C. Kirby, SIREV (2010):

"We examine condition numbers, preconditioners and iterative methods for FEM discretization of coercive PDEs in the context of the solvability result, the Lax-Milgram lemma.

Moreover, useful insight is gained as to the relationship between Hilbert space and matrix condition numbers, and translating Hilbert space fixed point iterations into matrix computations provides new ways of motivating and explaining some classic iteration schemes. [...] This paper is [...] intending to bridge the functional analysis techniques common in finite elements and the linear algebra community."

K. A. Mardal and R. Winther, NLAA (2011):

"The main focus will be on an abstract approach to the construction of preconditioners for symmetric linear systems in a Hilbert space setting [...] The discussion of preconditioned Krylov space methods for the continuous systems will be a starting point for a corresponding discrete theory.

By using this characterization it can be established that the conjugate gradient method converges [...] with a rate which can be bounded by the condition number [...] However, if the operator has a few eigenvalues far away from the rest of the spectrum, then the estimate is not sharp. In fact, a few 'bad eigenvalues' will have almost no effect on the asymptotic convergence of the method."

O. Axelsson and J. Karátson, Numer. Alg. (2009):

"To preserve sparsity, the arising system is normally solved using an iterative solution method, commonly a preconditioned conjugate gradient method [...] the rate of convergence depends in general on a generalized condition number of the preconditioned operator [...]

- *if the two operators (original and preconditioner) are equivalent, then the corresponding PCG method provides mesh independent linear convergence [...]*
- *if the two operators (original and preconditioner) are compact-equivalent, then the corresponding PCG method provides mesh independent superlinear convergence."*

R. Hiptmair, CMA (2006):

"There is a continuous operator equation posed in infinite-dimensional spaces that underlines the linear system of equations [...] awareness of this connection is key to devising efficient solution strategies for the linear systems.

Operator preconditioning is a very general recipe [...]. It is simple to apply, but may not be particularly efficient, because in case of the [condition number] bound of Theorem 2.1 is too large, the operator preconditioning offers no hint how to improve the preconditioner. Hence, operator preconditioner may often achieve [...] the much-vaunted mesh independence of the preconditioner, but it may not perform satisfactorily on a given mesh."

7 Linear asymptotic behavior?

V. Faber, T. Manteuffel and S. V. Parter, Adv. in Appl. Math. (1990):

"For a fixed h, using a preconditioning strategy based on an equivalent operator may not be superior to classical methods [...] Equivalence alone is not sufficient for a good preconditioning strategy. One must also choose an equivalent operator for which the bound is small.

There is no flaw in the analysis, only a flaw in the conclusions drawn from the analysis [...] asymptotic estimates ignore the constant multiplier. Methods with similar asymptotic work estimates may behave quite differently in practice."

Let V be an infinite dimensional Hilbert space with the inner product

$$(\cdot,\cdot)_V : V \times V \to \mathbb{R}, \quad \text{the associated norm } \|\cdot\|_V ,$$

$V^\#$ be the dual space of bounded (continuous) linear functionals on V with the duality pairing

$$\langle \cdot,\cdot \rangle : V^\# \times V \to \mathbb{R} .$$

For each $f \in V^\#$ there exists a unique $\tau f \in V$ such that

$$\langle f, v \rangle = (\tau f, v)_V \quad \text{for all } v \in V .$$

In this way the inner product $(\cdot,\cdot)_V$ determines the Riesz map

$$\tau : V^\# \to V .$$

Let $a(\cdot,\cdot) = V \times V \to R$ be a bounded and coercive bilinear form. For $u \in V$ we can write the bounded linear functional $a(u,\cdot)$ on V as

$$\mathcal{A}u \equiv a(u,\cdot) \in V^{\#}\,, \quad \text{i.e.}\,,$$
$$\langle \mathcal{A}u, v\rangle = a(u,v) \quad \text{for all } v \in V\,.$$

This defines the bounded and coercive operator

$$\mathcal{A} : V \to V^{\#}, \qquad \inf_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle = \alpha > 0,\ \|\mathcal{A}\| = C\,.$$

The Lax-Milgram theorem ensures that for any $b \in V^{\#}$ there exists a unique solution $x \in V$ of the problem

$$a(x,v) = \langle b, v\rangle \quad \text{for all } v \in V\,.$$

Equivalently,

$$\langle \mathcal{A}x - b, v \rangle = 0 \quad \text{for all } v \in V \,,$$

which can be written as the equation in $V^{\#}$,

$$\mathcal{A}x = b \,, \qquad \mathcal{A} : V \to V^{\#}, \quad x \in V, \quad b \in V^{\#} \,.$$

We will consider $\mathcal{A}$ self-adjoint with respect to the duality pairing $\langle \cdot, \cdot \rangle$.

7 Discretization using $V_h \subset V$

Let $\Phi_h = (\phi_1^{(h)}, \ldots, \phi_N^{(h)})$ be a basis of the subspace $V_h \subset V$,
let $\Phi_h^{\#} = (\phi_1^{(h)\#}, \ldots, \phi_N^{(h)\#})$ be the canonical basis of its dual $V_h^{\#}$.

The Galerkin discretization then gives

$$\mathcal{A}_h x_h = b_h\,, \quad x_h \in V_h\,, \quad b_h \in V_h^{\#}\,, \quad \mathcal{A}_h : V_h \to V_h^{\#}\,.$$

Using the coordinates $x_h = \Phi_h \mathbf{x}$, $b_h = \Phi_h^{\#} \mathbf{b}$, the discretization results in the linear algebraic system

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,.$$

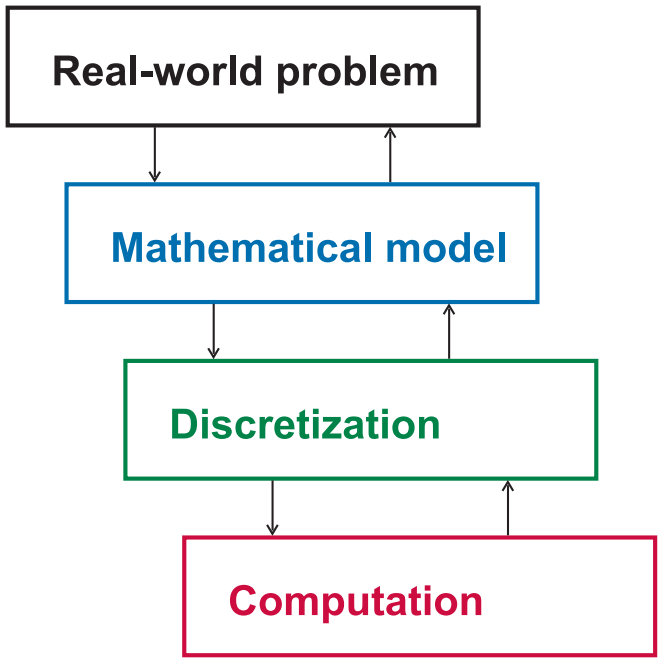
Preconditioning needed for accelerating the iterations is then often build up algebraically for the given matrix problem, giving (here illustrated as the left preconditioning)

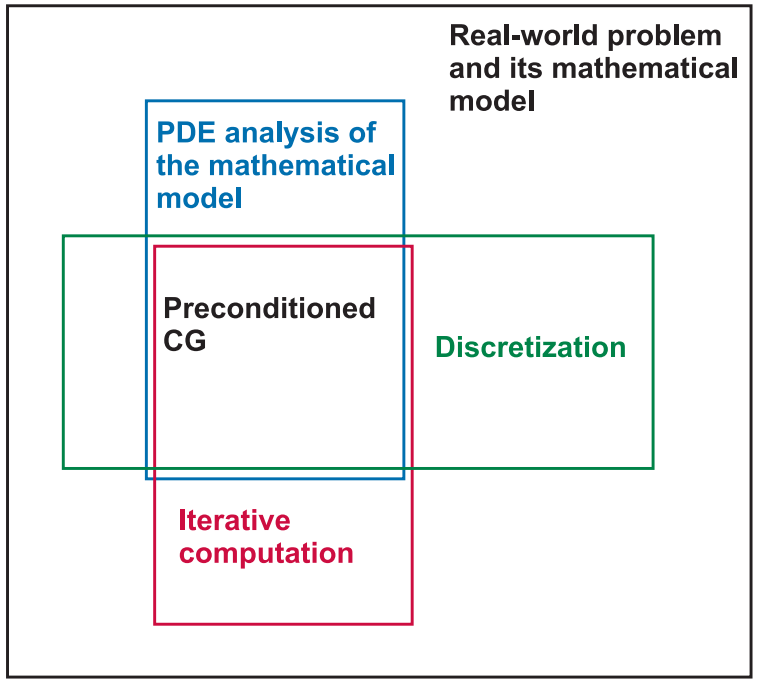
$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}\,.$$

Then the CG method is applied to the (symmetrized) preconditioned system, i.e., (PCG) (**M**-preconditioned CG) is applied to the unpreconditioned system. The schema of the solution process:

$$\mathcal{A},\, \langle b, \cdot \rangle \;\rightarrow\; \mathbf{A}, \mathbf{b} \rightarrow \;\; \text{preconditioning} \;\; \rightarrow \text{PCG applied to } \mathbf{A}\mathbf{x} = \mathbf{b}\,.$$

7 A bit different view





Formulation of the model, discretization and algebraic computation, including the evaluation of the error, stopping criteria for the algebraic solver, adaptivity etc. are very closely related to each other.

Recall that the inner product $(\cdot,\cdot)_V$ defines the Riesz map τ.
It can be used to transform the equation in $V^\#$

$$\mathcal{A}x = b\,, \qquad \mathcal{A} : V \to V^\#, \quad x \in V, \quad b \in V^\#\,.$$

into the equation in V

$$\tau\mathcal{A}x = \tau b, \qquad \tau\mathcal{A} : V \to V, \quad x \in V, \quad \tau b \in V\,,$$

This transformation is called operator preconditioning.

7 The mathematically best preconditioning?

With the choice of the inner product $(\cdot,\cdot)_V = a(\cdot,\cdot)$ we get

$$a(u,v) = \langle \mathcal{A}u, v\rangle = a(\tau\mathcal{A}\,u, v)$$

i.e.,

$$\tau = \mathcal{A}^{-1}\,, \quad \text{and the preconditioned system} \quad x = \mathcal{A}^{-1}b\,.$$

The inner product can be defined using an operator

$$\mathcal{B} \approx \mathcal{A}\,, \quad (\cdot,\cdot)_V \;=\; (\cdot,\cdot)_{\mathcal{B}} = \langle \mathcal{B}u, v\rangle\,.$$

Then

$$\tau = \mathcal{B}^{-1}\,, \quad \text{and the preconditioned system} \quad \mathcal{B}^{-1}\mathcal{A}x = \mathcal{B}^{-1}b\,.$$

What does it mean $\mathcal{B} \approx \mathcal{A}$?

Concept of norm equivalence and spectral equivalence of operators.

$r_0 = b - \mathcal{A}x_0 \in V^{\#}, \quad p_0 = \tau r_0 \in V$. For $n = 1, 2, \ldots, n_{\max}$

$$\alpha_{n-1} = \frac{\langle r_{n-1}, \tau r_{n-1} \rangle}{\langle \mathcal{A}p_{n-1}, p_{n-1} \rangle}$$

$x_n = x_{n-1} + \alpha_{n-1} p_{n-1}$, stop when the stopping criterion is satisfied

$r_n = r_{n-1} - \alpha_{n-1} \mathcal{A}p_{n-1}$

$$\beta_n = \frac{\langle r_n, \tau r_n \rangle}{\langle r_{n-1}, \tau r_{n-1} \rangle}$$

$p_n = \tau r_n + \beta_n p_{n-1}$

Hayes (1954); Vorobyev (1958, 1965); Karush (1952); Stesin (1954)
Superlinear convergence for (identity + compact) operators.
Here the Riesz map τ indeed serves as the preconditioner.

Using the coordinates in the bases Φ_h and $\Phi_h^\#$ of V_h and $V_h^\#$ respectively, $(\, V_h^\# = \mathcal{A}V_h \,)$,

$$
\begin{aligned}
\langle f, v\rangle &\to \mathbf{v}^*\mathbf{f}\,, \\
(u,v)_V &\to \mathbf{v}^*\mathbf{M}\mathbf{u}, \qquad (\mathbf{M}_{ij}) = ((\phi_j,\phi_i)_V)_{i,j=1,\dots,N}\,, \\
\mathcal{A}u &\to \mathbf{A}\mathbf{u}\,, \qquad \mathcal{A}u = \mathcal{A}\Phi_h\mathbf{u} = \Phi_h^\#\mathbf{A}\mathbf{u}\,; \quad (\mathbf{A}_{ij}) = (a(\phi_j,\phi_i))_{i,j=1,\dots,N}\,, \\
\tau f &\to \mathbf{M}^{-1}\mathbf{f}\,, \qquad \tau f = \tau\Phi_h^\#\mathbf{f} = \Phi_h\mathbf{M}^{-1}\mathbf{f}\,;
\end{aligned}
$$

we get with $b = \Phi_h^\#\mathbf{b}$, $x_n = \Phi_h\,\mathbf{x}_n$, $p_n = \Phi_h\,\mathbf{p}_n$, $r_n = \Phi_h^\#\,\mathbf{r}_n$ the algebraic CG formulation

7 Galerkin discretization gives matrix CG in V_h

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, solve $\mathbf{M}\mathbf{z}_0 = \mathbf{r}_0$, $\mathbf{p}_0 = \mathbf{z}_0$. For $n = 1, \ldots, n_{\max}$

$$
\begin{aligned}
\alpha_{n-1} &= \frac{\mathbf{z}_{n-1}^* \mathbf{r}_{n-1}}{\mathbf{p}_{n-1}^* \mathbf{A}\mathbf{p}_{n-1}} \\
\mathbf{x}_n &= \mathbf{x}_{n-1} + \alpha_{n-1}\mathbf{p}_{n-1}\,, \quad \text{stop when the stopping criterion is satisfied} \\
\mathbf{r}_n &= \mathbf{r}_{n-1} - \alpha_{n-1}\mathbf{A}\mathbf{p}_{n-1} \\
\mathbf{M}\mathbf{z}_n &= \mathbf{r}_n\,, \quad \text{solve for } \mathbf{z}_n \\
\beta_n &= \frac{\mathbf{z}_n^* \mathbf{r}_n}{\mathbf{z}_{n-1}^* \mathbf{r}_{n-1}} \\
\mathbf{p}_n &= \mathbf{z}_n + \beta_n \mathbf{p}_{n-1}
\end{aligned}
$$

Günnel, Herzog, Sachs (2014); Málek, S (2015)

The bound

$$\kappa(\mathbf{M}^{-1}\mathbf{A}) \leq \frac{\sup_{u,v\in V,\,\|u\|_V=1,\|v\|_V=1} |\langle \mathcal{A}u, v\rangle|}{\inf_{u\in V,\,\|u\|_V=1} \langle \mathcal{A}u, u\rangle}$$

is valid independently of the discretization, see, e.g., Hiptmair (2006). If the bound is small enough, then the matter about the rate of convergence and its monitoring is resolved.

- Unpreconditioned CG, i.e. $\mathbf{M} = \mathbf{I}$, corresponds to the discretization basis Φ orthonormal wrt $(\cdot, \cdot)_V$.

- Orthogonalization of the discretization basis with respect to the given inner product in V will result in the unpreconditioned CG that is applied to the transformed (preconditioned) algebraic system. The resulting orthogonal discretization basis functions do not have local support and the transformed matrix is not sparse.

- Orthogonalization is not unique. For the same inner product we can get different bases and different discretized systems with exactly the same convergence behaviour.

7 Algebraic preconditioning?

Consider an algebraic preconditioning with the (SPD) preconditioner

$$\widehat{\mathbf{M}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^* = \widehat{\mathbf{L}}\,(\mathbf{Q}\mathbf{Q}^*)\,\widehat{\mathbf{L}}^*$$

Where $\mathbf{Q}\mathbf{Q}^* = \mathbf{Q}^*\mathbf{Q} = \mathbf{I}$.

Question: Can any algebraic preconditioning be expressed in the operator preconditioning framework? How does it link with the discretization and the choice of the inner product in V ?

Transform the discretization bases

$$\widehat{\Phi} = \Phi\,((\widehat{\mathbf{L}}\mathbf{Q})^*)^{-1}, \quad \widehat{\Phi}^{\#} = \Phi^{\#}\,\widehat{\mathbf{L}}\mathbf{Q}\,.$$

with the change of the inner product in V (recall $(u,v)_V = \mathbf{v}^*\mathbf{M}\mathbf{u}$)

$$(u,v)_{\mathrm{new},V} = (\widehat{\Phi}\widehat{\mathbf{u}}, \widehat{\Phi}\widehat{\mathbf{v}})_{\mathrm{new},V} := \widehat{\mathbf{v}}^*\widehat{\mathbf{u}} = \mathbf{v}^*\widehat{\mathbf{L}}\mathbf{Q}\mathbf{Q}^*\widehat{\mathbf{L}}^*\mathbf{u} = \mathbf{v}^*\widehat{\mathbf{L}}\widehat{\mathbf{L}}^*\mathbf{u} = \mathbf{v}^*\widehat{\mathbf{M}}\mathbf{u}\,.$$

The discretized Hilbert space formulation of CG gives the algebraically preconditioned matrix formulation of CG with the preconditioner $\widehat{\mathbf{M}}$

(more specifically, it gives the unpreconditioned CG applied to the algebraically preconditioned discretized system).

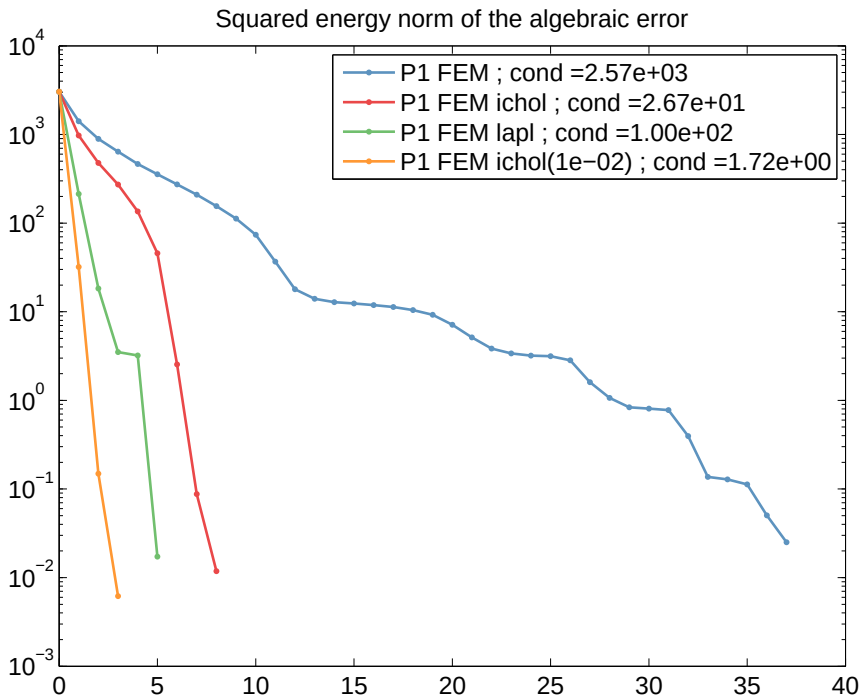
7 Sparsity, locality, global transfer of information

Sparsity of matrices of the algebraic systems is always presented as an advantage of the FEM discretizations.

Sparsity means locality of information in the individual matrix rows/columns. Getting a sufficiently accurate approximation to the solution may then require many matrix-vector multiplications (a large dimension of the Krylov space).

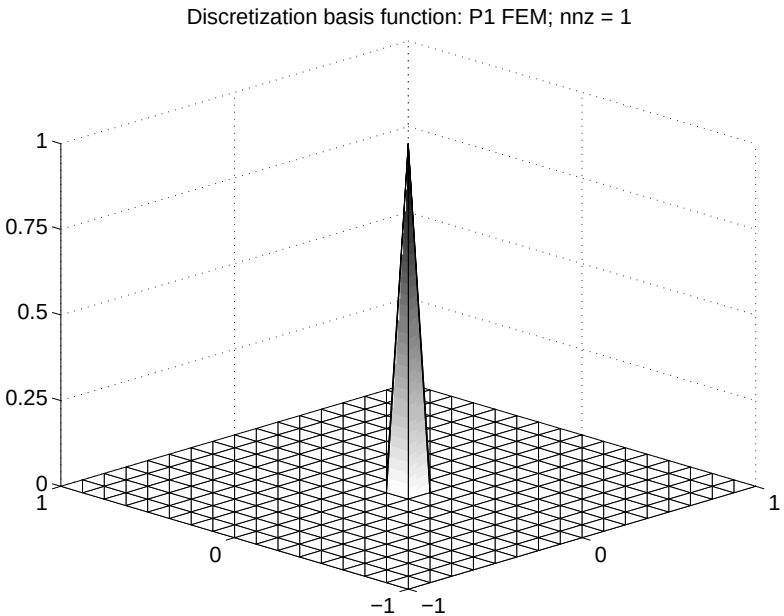
Preconditioning can be interpreted in part as addressing the unwanted consequence of sparsity (locality of the supports of the basis functions). Globally supported basis functions (hierarchical bases preconditioning, DD with coarse space components, multilevel methods, hierarchical grids etc.) can efficiently handle the transfer of global information.

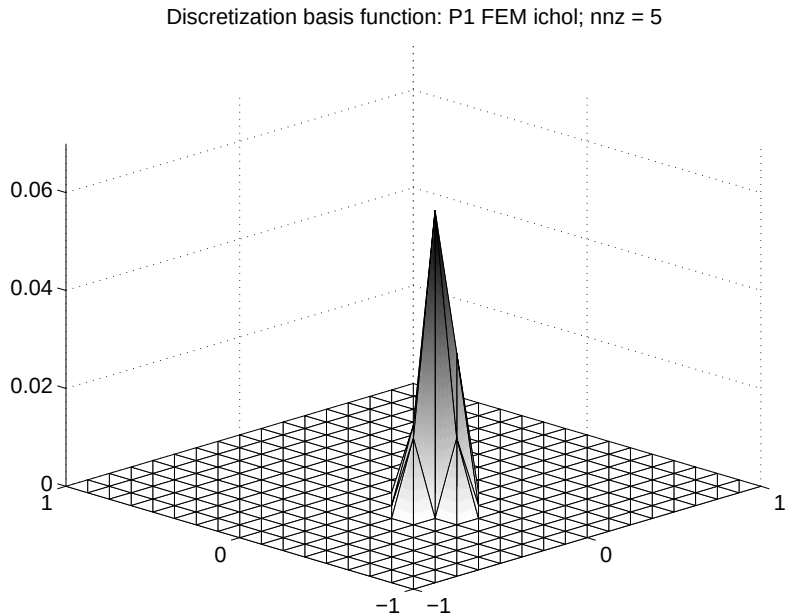
7 Example - Nonhomogeneous diffusion tensor



PCG convergence: unpreconditioned; ichol (no fill-in); Laplace operator preconditioning; ichol (drop-off tolerance 1e-02). Uniform mesh, condition numbers 2.5e03, 2.6e01, 1.0e02, 1.7e00.

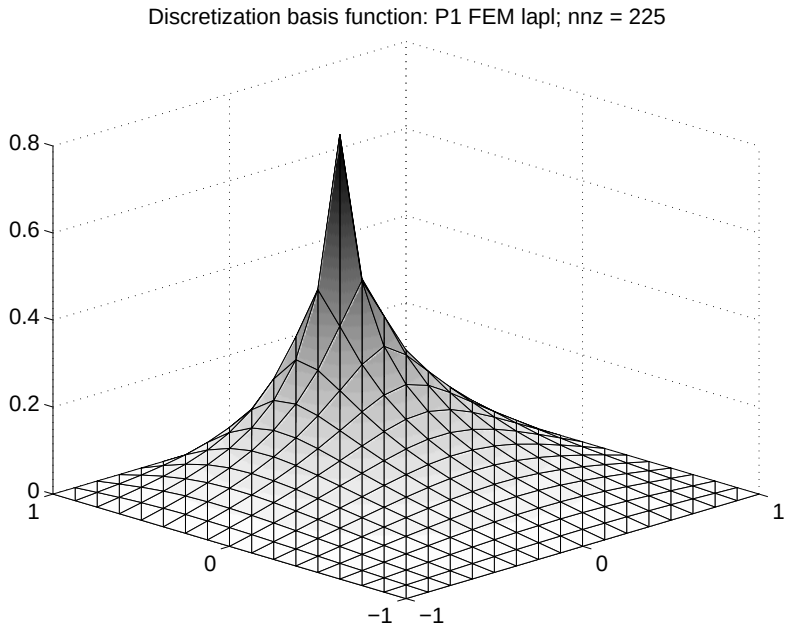
7 Transformed basis elements

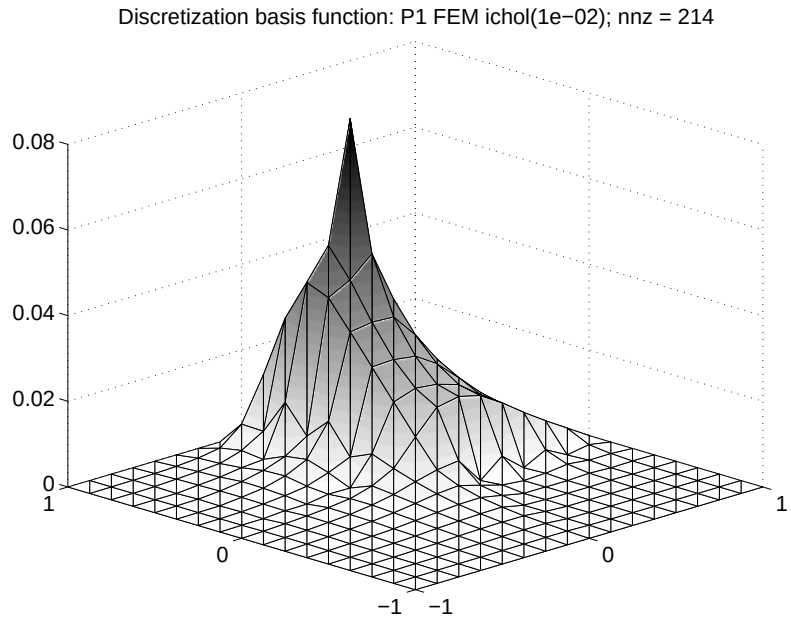




Original discretization basis element and its transformation corresponding to the ichol preconditioning.

7 Transformed basis elements





Transformed discretization basis elements corresponding o the lapl (left) and ichol(tol) preconditioning (right).

8. HPC computations with Krylov subspace methods?

References

- E. Carson, M. Rozložník, Z.S., P, Tichý, and M. Tůma, *On the numerical stability analysis of pipelined Krylov subspace methods,* (2016, submitted for publication).

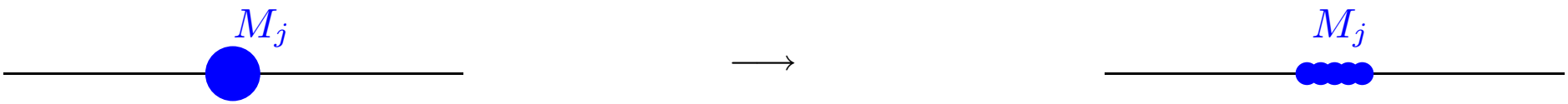
9. Myths about Krylov subspace methods

Myth: A belief given uncritical acceptance by the members of a group especially in support of existing or traditional practices and institutions.

Webster's Third New International Dictionary, Enc. Britannica Inc., Chicago (1986)

9 Widespread statements that are misleading or plainly incorrect

1. Minimal polynomials and finite termination property
2. Chebyshev bounds and CG
3. Spectral information and clustering of eigenvalues
4. Operator-based bounds and functional analysis arguments on convergence
5. Finite precision computations can not be seen as a minor modification of the exact considerations
6. Linearization of nonlinear phenomenon without noticing that this eliminates the main principle behind the phenomenon, i.e. the adaptation to the problem
7. Short term recurrences can not guarantee well conditioned basis due to rounding errors. This is true even for symmetric positive definite problems, and it remains true also for nonsymmetric problems
8. Sparsity can have positive as well as negative effects to computations



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).

If it does not, then it means that CG can not adapt to the problem, and it converges almost linearly. **In such cases - is it worth using?**

- It is not true that CG (or other Krylov subspace methods used for solving systems of linear algebraic equations with symmetric matrices) applied to a matrix with t distinct well separated tight clusters of eigenvalues produces in general a large error reduction after t steps; see Sections 5.6.5 and 5.9.1 of Liesen, S (2013). This myth has been disproved more than 20 years ago; see Greenbaum (1989); S (1991); Greenbaum, S (1992). Still it is persistently repeated in literature as an obvious fact.

- With no information on the structure of invariant subspaces it is not true that distribution of eigenvalues provides insight into the asymptotic behavior of Krylov subspace methods (such as GMRES) applied to systems with generally nonsymmetric matrices; see Sections 5.7.4, 5.7.6 and 5.11 of Liesen, S (2013). As before, the relevant results Greenbaum, S (1994); Greenbaum, Pták, S (1996) and Arioli, Pták, S (1998) are (almost) twenty years old.

9 How the mathematical myths are created?

- Rutishauser (1959) as well as Lanczos (1952) considered CG principally different in their nature from the method based on the Chebyshev polynomials.
- Daniel (1967) did not identify the CG convergence with the Chebyshev polynomials-based bound. He carefully writes (modifyling slightly his notation)

  "assuming only that the spectrum of the matrix A lies inside the interval $[\lambda_1, \lambda_N]$, we can do no better than Theorem 1.2.2."
- That means that the Chebyshev polynomials-based bound holds for any distribution of eigenvalues between λ_1 and λ_1 and for any distribution of the components of the initial residuals in the individual invariant subspaces.
- Why we do not read the original works? They are many times most valuable sources of insight, that can be gradually forgotten and can be overshadowed by commonly accepted myth ...

9 Analogy with a priori and a posteriori numerical PDE analysis

- Think of a priori and a posteriori numerical PDE analysis!
- The Chebyshev bound is a typical a priori bound; it uses no a posteriori information.
- A priori bounds are useful for the purpose they have been derived to. They can not take over the role of the a posteriori bounds.

9 Concluding remarks and outlook

- Krylov subspace methods adapt to the problem. Exploiting this adaptation is the key to their efficient use.
- Unlike in nonlinear problems and/or multilevel methods, analysis of Krylov subspace methods can not be based, in general, on contraction arguments.
- Individual steps modeling-analysis-discretization-computation should not be considered separately within isolated disciplines. They form a single problem. Operator preconditioning follows this philosophy.
- Fast HPC computations require handling all involved issues.
  A posteriori error analysis and stopping criteria are essential ...
- Assumptions must be honored
- Historia Magistra Vitae

Thank you very much for your kind patience!