O jednom starém článku a jeho mnohých souvislostech
On an old article and its many connections

Zdeněk Strakoš
Charles University, Prague
Jindřich Nečas Center for Mathematical Modelling

Czech Mathematical Society, Prague, February 2018

Many thanks

I am indebted to very many collaborators and friends at home and abroad. In this talk I will touch results of only some of them. I wish to particularly mention a friend whom I never wrote any paper with, but to whom I am greatly indebted for the invaluable moral support at the very beginning of my mathematical career.

Many thanks

I am indebted to very many collaborators and friends at home and abroad. In this talk I will touch results of only some of them. I wish to particularly mention a friend whom I never wrote any paper with, but to whom I am greatly indebted for the invaluable moral support at the very beginning of my mathematical career.

I would like to devote this lecture to the memory of

Professor Ivo Marek

Těžký je úděl dobrodruha

Tibor Dévényi,
Kariéra Dr. Gézy támhletoho aneb vědci a hlodavci,
Kapitola UMĚNÍ PŘEDNÁŠKY, MF, Praha (1981), p. 111

"Co neumíš povědět za třicet minut,
to necht' zůstane tajemstvím"

Not just one old paper but five. Old does not mean obsolete!

Not just one old paper but five. Old does not mean obsolete!

Even at our (post-)modern, advanced, progressive, excellent, post-factual times!

Not just one old paper but five. Old does not mean obsolete!

Even at our (post-)modern, advanced, progressive, excellent, post-factual times!

Cornelius Lanczos, *Why Mathematics,* 1966

"The naive optimist, who believes in progress and is convinced that today is better than yesterday and in ten years time the world will be infinitely better off than today, will come to the conclusion that mathematics (and more generally all the exact sciences) started only about twenty years ago, while all the predecessors must have walked in a kind of limbo of half-digested and improperly conceived ideas. { ... }

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."

Connections ... Why does science exist?

Connections ... Why does science exist?

- Humans must do science in order to survive.
  Question: How to make things work?

Connections ... Why does science exist?

- Humans must do science in order to survive.
  Question: How to make things work?

- Humans must do science because they are humans.
  Question: Why and how does the world work?

Connections ... Why does science exist?

- Humans must do science in order to survive.
  Question: How to make things work?

- Humans must do science because they are humans.
  Question: Why and how does the world work?

Success as a measure, avalanche of performance metrics, overspecialization, fragmentation, confusion .....

Pure against (!) applied mathematics,
basic research against (!) applied research, ...

Words have lost their meaning - the well known story!

Gen 11, 1-7

"The whole world spoke the same language, using the same words. { ... } They said to one another, "Come, let us mold bricks and harden them with fire. { ... } Then they said, "Come, let us build ourselves a city and a tower with its top in the sky, and so make a name for ourselves; otherwise we will be scattered all over the earth."

The Lord came down to see the city and the tower that men had built.Then the Lord said: { . . . } "Let us go down and confuse their language, so that one will not understand what another says." Thus the Lord scattered them from there all over the earth ... "

Matrix computations $A\,x = b\,, \quad A\,z(\lambda) = \lambda\,z(\lambda)$

Here we concentrate on solving systems of linear algebraic equations, which may seem at the first look a linear problem. Possible approaches:

- Direct methods - decomposition of A , e.g. $A = LU$
- Classical (fixed point) iterations - splitting $A = K - L$
- Any other possibility?

What does it mean linearity in solving the linearly formulated problem $A\,x = b\,?$

Moreover, the input finite data (such as the finite matrix A and vector b) typically represent a reduction of a problem formulated at infinite dimensional spaces.

What we will do for the remaining eighty five minutes

- Important and hard challenges of the contemporary computational mathematics are shown to be very closely related to some 19th century seminal works.

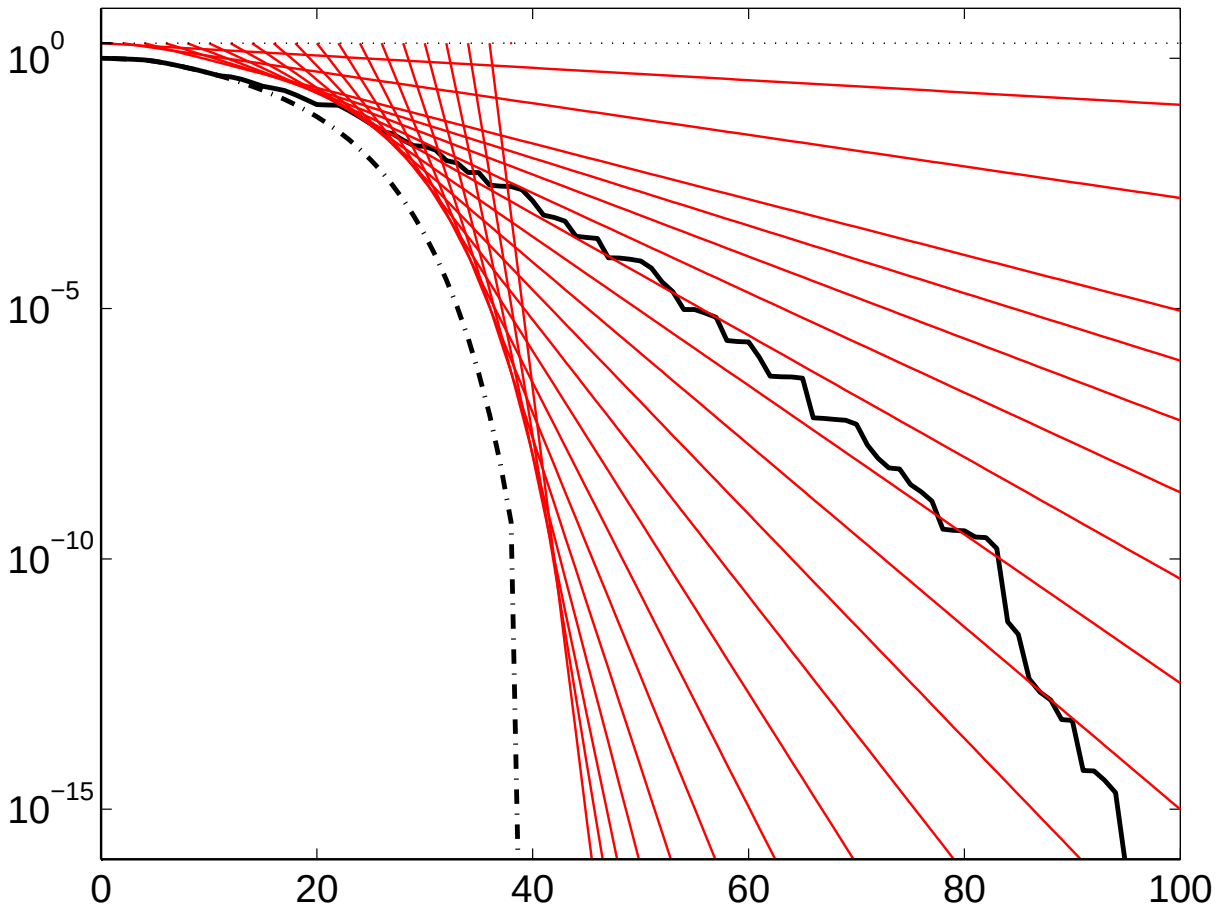
What we will do for the remaining eighty five minutes

- Important and hard challenges of the contemporary computational mathematics are shown to be very closely related to some 19th century seminal works.

- Revealing this historical link has a very deep impact to understanding the state-of-the-art computations.

  The work of Stieltjes (Gauss, Jacobi, ...) helps in solving the crucial problem of rounding error propagation in the Lanczos and conjugate gradient methods based on short recurrences.

Where are we heading to?



Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

1 Hierarchy of linear problems starting at infinite dimension

Problem with bounded invertible operator $\mathcal{G}$ on the infinite dim. Hilbert space S

$$\mathcal{G}\, u \;=\; f$$

is approximated on a finite dimensional subspace $S_h \subset S$ by the problem with the finite dimensional operator

$$\mathcal{G}_h\, u_h \;=\; f_h\,,$$

represented, using an appropriate basis of S_h, by the (sparse?) matrix problem

$$A\, x \;=\; b\,.$$

Bounded invertible operators in Hilbert spaces can be approximated by compact or finite dimensional operators only in the sense of strong convergence (pointwise limit)

$$\|\, \mathcal{G}_h\, w - \mathcal{G}\, w \,\| \;\to\; 0 \quad \text{as} \quad h \to 0 \quad \text{for all } w \in S\,;$$

The convergence $\mathcal{G}_h\, w \;\to\; \mathcal{G}\, w$ is not uniform w.r.t. w; the role of right hand sides.

1 Fundamental theorem of discretization of $\mathcal{G}\,u \;=\; f$

How closely $\mathcal{G}_h\,u_h \;=\; f_h$ approximates $\mathcal{G}\,u \;=\; f$? The residual measure

$$\mathcal{G}_h\,\pi_h u \;-\; f_h$$

gives

$$\pi_h u - u_h \;=\; \mathcal{G}_h^{-1}\,(\,\mathcal{G}_h\,\pi_h u \;-\; f_h\,).$$

If $\|\mathcal{G}_h^{-1}\|_h$ is bounded from above uniformly in h (the discretization is stable), then consistency

$$\|\,\mathcal{G}_h\,\pi_h u \;-\; f_h\,\|_h \;\to\; 0 \quad \text{as} \quad h \to 0$$

implies convergence of the discretization scheme

$$\|\,\pi_h u - u_h\,\|_h \;\to\; 0 \quad \text{as} \quad h \to 0\,.$$

Additional important point: In computations we only approximate u_h by $u_h^{(n)}$.

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

2 Polynomial (Krylov subspace) methods

Consider, as above, a linear invertible operator $\mathcal{G} : S \to S$ and the equation

$$\mathcal{G}\, u \;=\; f\,, \quad f \in S\,.$$

Krylov subspace methods at the step n implicitly construct a finite dimensional approximation $\mathcal{G}_n$ of $\mathcal{G}$ with the desired approximate solution u_n defined by $(u_0 = 0)$

$$u_n \;:=\; p_{n-1}(\mathcal{G}_n)\, f \;\approx\; u \;=\; \mathcal{G}^{-1} f\,,$$

where $p_{n-1}(\lambda)$ is a particular polynomial of degree at most $n-1$ and $\mathcal{G}_n$ is obtained by restricting and projecting $\mathcal{G}$ onto the nth Krylov subspace

$$\mathcal{K}_n(\mathcal{G}, f) \;:=\; \text{span}\left\{f, \mathcal{G}f, \ldots, \mathcal{G}^{n-1} f\right\}.$$

A.N. Krylov (1931), Gantmakher (1934)

2 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."

2 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."

2 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."

Nonlinear and globally optimal adaptation of the iterations in Krylov subspaces to the problem.

2 Four basic questions

1. How fast the iterations u_n, $n = 1, 2, \ldots$ approximate the desired solution u ?
   Nonlinear adaptation.
2. Which characteristics of $\mathcal{G}$ and f determine behaviour of the method?
   Inner nature of the problem.
3. How to handle efficiently discretization and computational issues?
   Provided that $\mathcal{K}_n(\mathcal{G}, f)$ can be computed, the projection provides discretization of the infinite dimensional problem $\mathcal{G}u = f$.
4. How to handle transformation of $\mathcal{G}u = f$ into
   an easier-to-solve problem? Preconditioning.

Vlastimil Pták: Finite dimensional nonlinearity is most difficult.

2 C. Lanczos (1950, 1952)

A computer, have to follow the notation!

Journal of Research of the National Bureau of Standards Vol. 49, No. 1, July 1952 Research Paper 2341

Solution of Systems of Linear Equations by Minimized Iterations[1]

Cornelius Lanczos

A simple algorithm is described which is well adapted to the effective solution of large systems of linear algebraic equations by a succession of well-convergent approximations.

1. Introduction

In an earlier publication [14][2] a method was described which generated the eigenvalues and eigenvectors of a matrix by a successive algorithm based on minimizations by least squares.[3] The advantage of this method consists in the fact that the successive iterations are constantly employed with maximum efficiency which guarantees fastest convergence for a given number of iterations. Moreover, with the proper care the accumulation of rounding errors can be avoided. The resulting high precision is of great advantage if the separation of closely bunched eigenvalues and eigenvectors is demanded [16].

It was pointed out in [14, p. 256] that the inversion of a matrix, and thus the solution of simultaneous systems of linear equations, is contained in the general procedure as a special case. However, in view of the great importance associated with the solution of large systems of linear equations, this problem deserved more than passing attention. It is the purpose of the present discussion to adopt the general principles of the previous investigation to the specific demands that arise if we are not interested in the complete analysis of a matrix but only in the more special problem of obtaining the solution of a given set of linear equations

$$Ay=b_0 \tag{1}$$

with a given matrix A and a given right side b_0. This is actually equivalent to the evaluation of one eigenvector only, of a symmetric, positive definite matrix. It is clear that this will require considerably less detailed analysis than the problem of constructing the *entire* set of eigenvectors and eigenvalues associated with an arbitrary matrix.

2. The Double Set of Vectors Associated With the Method of Minimized Iterations

The previous investigation [14] started out with an algorithm (see p. 261) which generated a double set of polynomials, later on denoted by $p_i(x)$ and $q_i(x)$ (see p. 274). Then a second algorithm was introduced, called "minimized iterations", which avoided the numerical difficulties of the first algorithm (see p. 287) and had, in addition, theoretically valuable properties for the solution of differential and integral equations (p. 272).

In this second algorithm, however, only *one-half* of the previous polynomials were represented, corresponding to the $p_i(x)$ polynomials whose coefficients appeared in the *full* columns of the original algorithm [14, (60)]. The polynomials $q_i(x)$, associated with the *half columns* of [14, (60)] did not come into evidence in the later procedure.

The vectors b_i, generated by minimized iterations, correspond to the polynomials $p_i(x)$ in the sense

$$b_k=p_k(A)b_0. \tag{2}$$

We should expect that the vectors generated by $q_k(A)b_0$ might also have some significance. We will see that this is actually the case. It is of considerable advantage to translate the *entire* scheme [14, (60)] into the language of minimized iterations, without omitting the half columns. We thus get a *double* set of vectors, instead of the single set considered before.

The additional work thus involved is not superfluous because the second set of polynomials can be put to good use. Moreover, the two sets of polynomials belong logically together and complement each other in a natural fashion. From the practical standpoint of adapting the resultant algorithm to the demands of large scale electronic computers, we gain in the simplicity of coding. The recurrence relations which exist between the polynomials $p_i(x)$; $q_i(x)$ are simpler in structure than the recurrence relation obtained by eliminating the second set of polynomials.

We want to simplify and systematize our notations. The vector obtained by letting the polynomial $p_k(A)$ operate on the original vector b_0, shall be called p_k:

$$p_k=p_k(A)b_0. \tag{3}$$

We thus distinguish between p_k as a *vector* and $p_k(A)$ as a *polynomial operator*. Hence the notation p_k will take the place of the previous b_k. Correspondingly we denote the associated second set of vectors by q_k:

$$q_k=q_k(A)b_0. \tag{4}$$

[1] The preparation of this paper was sponsored (in part) by the Office of Naval Research.
[2] Figures in brackets indicate the literature references at the end of this paper.
[3] The present paper is a natural sequel to the previous publication and depends on the previous findings. The reader's familiarity with the earlier development is assumed throughout this paper; the symbolism of the present paper is in harmony with that used before, in particular the notation pq, if applied to vectors, shall mean the scalar product of these two vectors.

1953

CHEBYSHEV POLYNOMIALS IN THE SOLUTION OF LARGE-SCALE LINEAR SYSTEMS*

(Toronto Symp. on Computing Techniques 1952, pp. 124–133)

By

Cornelius Lanczos
National Bureau of Standards, Los Angeles

An easily coded iterative routine is described which approximates the solution of a wide class of simultaneous linear algebraic equations by a succession of simple matrix multiplications.

The problem of finding the solution of the matrix equation

(1) $Ay = b$

demands the construction of the inverse matrix A^{-1} which in the case of very large matrices poses great numerical difficulties. On the other hand, the large scale machinery available in our days performs the mere substitution Ab with a given matrix A and a given vector b as a simple routine operation. Since this operation can be repeated any number of times, we have an arbitrary polynomial operation

(2) $G_m(A) \cdot b$

*The preparation of this paper was sponsored (in part) by the Office of Naval Research.

*The Editors wish to express their appreciation to the publisher for permission to reproduce this paper. A fuller statement of appreciation and permission appears in the Acknowledgements.

From Lanczos' Collected Works, Vol. VI (Paper + 2 commentaries)

2 M. Hestenes and E. Stiefel (1952)

Journal of Research of the National Bureau of Standards Vol. 49, No. 6, December 1952 Research Paper 2379

Methods of Conjugate Gradients for Solving Linear Systems[1]

Magnus R. Hestenes[2] and Eduard Stiefel[3]

An iterative algorithm is given for solving a system $Ax=k$ of n linear equations in n unknowns. The solution is given in n steps. It is shown that this method is a special case of a very general method which also includes Gaussian elimination. These general algorithms are essentially algorithms for finding an n dimensional ellipsoid. Connections are made with the theory of orthogonal polynomials and continued fractions.

1. Introduction

One of the major problems in machine computations is to find an effective method of solving a system of n simultaneous equations in n unknowns, particularly if n is large. There is, of course, no best method for all problems because the goodness of a method depends to some extent upon the particular system to be solved. In judging the goodness of a method for machine computations, one should bear in mind that criteria for a good machine method may be different from those for a hand method. By a hand method, we shall mean one in which a desk calculator may be used. By a machine method, we shall mean one in which sequence-controlled machines are used.

A machine method should have the following properties:

(1) The method should be simple, composed of a repetition of elementary routines requiring a minimum of storage space.

(2) The method should insure rapid convergence if the number of steps required for the solution is infinite. A method which—if no rounding-off errors occur—will yield the solution in a finite number of steps is to be preferred.

(3) The procedure should be stable with respect to rounding-off errors. If needed, a subroutine should be available to insure this stability. It should be possible to diminish rounding-off errors by a repetition of the same routine, starting with the previous result as the new estimate of the solution.

(4) Each step should give information about the solution and should yield a new and better estimate than the previous one.

(5) As many of the original data as possible should be used during each step of the routine. Special properties of the given linear system—such as having many vanishing coefficients—should be preserved. (For example, in the Gauss elimination special properties of this type may be destroyed.)

In our opinion there are two methods that best fit these criteria, namely, (a) the Gauss elimination method; (b) the conjugate gradient method presented in the present monograph.

There are many variations of the elimination method, just as there are many variations of the conjugate gradient method here presented. In the present paper it will be shown that both methods are special cases of a method that we call the method of conjugate directions. This enables one to compare the two methods from a theoretical point of view.

In our opinion, the conjugate gradient method is superior to the elimination method as a machine method. Our reasons can be stated as follows:

(a) Like the Gauss elimination method, the method of conjugate gradients gives the solution in n steps if no rounding-off error occurs.

(b) The conjugate gradient method is simpler to code and requires less storage space.

(c) The given matrix is unaltered during the process, so that a maximum of the original data is used. The advantage of having many zeros in the matrix is preserved. The method is, therefore, especially suited to handle linear systems arising from difference equations approximating boundary value problems.

(d) At each step an estimate of the solution is given, which is an improvement over the one given in the preceding step.

(e) At any step one can start anew by a very simple device, keeping the estimate last obtained as the initial estimate.

In the present paper, the conjugate gradient routines are developed for the symmetric and nonsymmetric cases. The principal results are described in section 3. For most of the theoretical considerations, we restrict ourselves to the positive definite symmetric case. No generality is lost thereby. We deal only with real matrices. The extension to complex matrices is simple.

The method of conjugate gradients was developed independently by E. Stiefel of the Institute of Applied Mathematics at Zurich and by M. R. Hestenes with the cooperation of J. B. Rosser, G. Forsythe, and L. Paige of the Institute for Numerical Analysis, National Bureau of Standards. The present account was prepared jointly by M. R. Hestenes and E. Stiefel during the latter's stay at the National Bureau of Standards. The first papers on this method were

[1] This work was performed on a National Bureau of Standards contract with the University of California at Los Angeles, and was sponsored (in part) by the Office of Naval Research, United States Navy.
[2] National Bureau of Standards and University of California at Los Angeles.
[3] University of California at Los Angeles, and Eidgenössische Technische Hochschule, Zurich, Switzerland.

given by E. Stiefel [4] and by M. R. Hestenes.[5] Reports on this method were given by E. Stiefel [6] and J. B. Rosser [7] at a Symposium [8] on August 23–25, 1951. Recently, C. Lanczos [9] developed a closely related routine based on his earlier paper on eigenvalue problem.[10] Examples and numerical tests of the method have been by R. Hayes, U. Hochstrasser, and M. Stein.

2. Notations and Terminology

Throughout the following pages we shall be concerned with the problem of solving a system of linear equations

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ \ldots\ +a_{1n}x_n&=k_1\\ a_{21}x_1+a_{22}x_2+\ \ldots\ +a_{2n}x_n&=k_2\\ \ldots\quad\quad\ \ldots\quad\quad\ \ldots&\\ a_{n1}x_1+a_{n2}x_2+\ \ldots\ +a_{nn}x_n&=k_n. \end{aligned} \tag{2:1}$$

These equations will be written in the vector form $Ax=k$. Here A is the matrix of coefficients (a_{ij}), x and k are the vectors $(x_1,\ldots,x_n)$ and $(k_1,\ldots,k_n)$. It is assumed that A is nonsingular. Its *inverse* A^{-1} therefore exists. We denote the *transpose* of A by A^*.

Given two vectors $x=(x_1,\ldots,x_n)$ and $y=(y_1,\ldots,y_n)$, their sum $x+y$ is the vector $(x_1+y_1,\ldots,x_n+y_n)$, and ax is the vector $(ax_1,\ldots,ax_n)$, where a is a scalar. The sum

$$(x,y)=x_1y_1+x_2y_2+\ldots+x_ny_n$$

is their *scalar product*. The *length* of x will be denoted by

$$|x|=(x_1^2+\ldots+x_n^2)^{\frac{1}{2}}=(x,x)^{\frac{1}{2}}.$$

The *Cauchy-Schwarz inequality* states that for all x,y:

$$(x,y)^2\leq(x,x)(y,y) \quad \text{or} \quad |(x,y)|\leq|x||y|. \tag{2:2}$$

The matrix A and its transpose A^* satisfy the relation

$$(x,Ay)=\sum_{i,j=1}^{n}a_{ij}x_iy_j=(A^*x,y).$$

If $a_{ij}=a_{ji}$, that is, if $A=A^*$, then A is said to be *symmetric*. A matrix A is said to be *positive definite* in case $(x,Ax)>0$ whenever $x\neq0$. If $(x,Ax)\geq0$ for all x, then A is said to be *nonnegative*. If A is symmetric, then two vectors x and y are said to be *conjugate* or *A-orthogonal* if the relation $(x,Ay)=(Ax,y)=0$ holds. This is an extension of the orthogonality relation $(x,y)=0$.

By an *eigenvalue* of a matrix A is meant a number λ such that $Ay=\lambda y$ has a solution $y\neq0$, and y is called a corresponding *eigenvector*.

Unless otherwise expressly stated the matrix A, with which we are concerned, will be *assumed to be symmetric and positive definite*. Clearly no loss of generality is caused thereby from a theoretical point of view, because the system $Ax=k$ is equivalent to the system $Bx=l$, where $B=A^*A$, $l=A^*k$. From a numerical point of view, the two systems are different, because of rounding-off errors that occur in joining the product A^*A. Our applications to the nonsymmetric case do not involve the computation of A^*A.

In the sequel we shall not have occasion to refer to a particular coordinate of a vector. Accordingly we may use subscripts to distinguish vectors instead of components. Thus x_0 will denote the vector $(x_{01},\ldots,x_{0n})$ and x_i the vector $(x_{i1},\ldots,x_{in})$. In case a symbol is to be interpreted as a component, we shall call attention to this fact unless the interpretation is evident from the context.

The *solution of the system $Ax=k$ will be denoted by h*; that is, $Ah=k$. If x is an estimate of h, the difference $r=k-Ax$ will be called the *residual* of x as an estimate of h. The quantity $|r|^2$ will be called the *squared residual*. The vector $h-x$ will be called the *error vector* of x, as an estimate of h.

3. Method of Conjugate Gradients (cg-Method)

The present section will be devoted to a description of a method of solving a system of linear equations $Ax=k$. This method will be called the *conjugate gradient method* or, more briefly, the cg-method, for reasons which will unfold from the theory developed in later sections. For the moment, we shall limit ourselves to collecting in one place the basic formulas upon which the method is based and to describing briefly how these formulas are used.

The cg-method is an iterative method which terminates in at most n steps if no rounding-off errors are encountered. Starting with an initial estimate x_0 of the solution h, one determines successively new estimates $x_0, x_1, x_2, \ldots$ of h, the estimate x_i being closer to h than x_{i+1}. At each step the residual $r_i=k-Ax_i$ is computed. Normally this vector can be used as a measure of the "goodness" of the estimate x_i. However, this measure is not a reliable one because, as will be seen in section 18, it is possible to construct cases in which the *squared residual* $|r_i|^2$ increases at each step (except for the last) while the length of the error vector $|h-x_i|$ decreases monotonically. If no rounding-off error is encountered, one will reach an estimate $x_m(m\leq n)$ at which $r_m=0$. This estimate is the desired solution h. Normally, $m=n$. However, since rounding-

[4] E. Stiefel, Uebersicht Methoden der Relaxationsrechnung, Z. angew. Math. Physik (3) (1952).
[5] M. R. Hestenes, Iterative methods for solving linear equations, NAML Report 52-9, National Bureau of Standards (1951).
[6] E. Stiefel, Some special methods of relaxation techniques, to appear in the Proceedings of the symposium (see footnote 8).
[7] J. B. Rosser, Rapidly converging iterative methods for solving linear equations, to appear in the Proceedings of the symposium (see footnote 8).
[8] Symposium on simultaneous linear equations and the determination of eigenvalues, Institute for Numerical Analysis, National Bureau of Standards, held on the campus of the University of California at Los Angeles (August 23–25, 1951).
[9] C. Lanczos, Solution of systems of linear equations by minimized iterations, NAML Report 52-13, National Bureau of Standards (1951).
[10] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research NBS 45, 255 (1950) RP2133; Proceedings of Second Symposium on Large-Scale Digital Calculating Machinery, Cambridge, Mass., 1951, pages 164–206.

2 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 the energy norm $\|x - x_n\|_A$ along the line

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

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

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

Indeed,

$$x - x_0 \;=\; \sum_{\ell=0}^{N-1} \alpha_\ell p_\ell \;=\; \sum_{\ell=0}^{n-1} \alpha_\ell p_\ell \;+\; x - x_n\,,$$

where

$$x - x_n \;\perp_A\; K_n(A, r_0)\,, \quad \text{or, equivalently,} \quad r_n \;\perp\; K_n(A, r_0)\,.$$

Mathematically, the orthogonality condition leads to short recurrences due to the relationship to the orthogonal polynomials that define the algebraic residuals and search vectors, see below.

Numerically, rounding errors can completely destroy the orthogonality and even linear independence of the computed search and residual vectors. As a consequence of experimental observations it was believed for several decades that the beautiful mathematical structure of the exact CG (Lanczos) was in practical computations inevitably lost and the finite precision behaviour would remain a mystery.

Crucial question: Is there any optimality of CG (Lanczos) left in the presence of rounding errors?

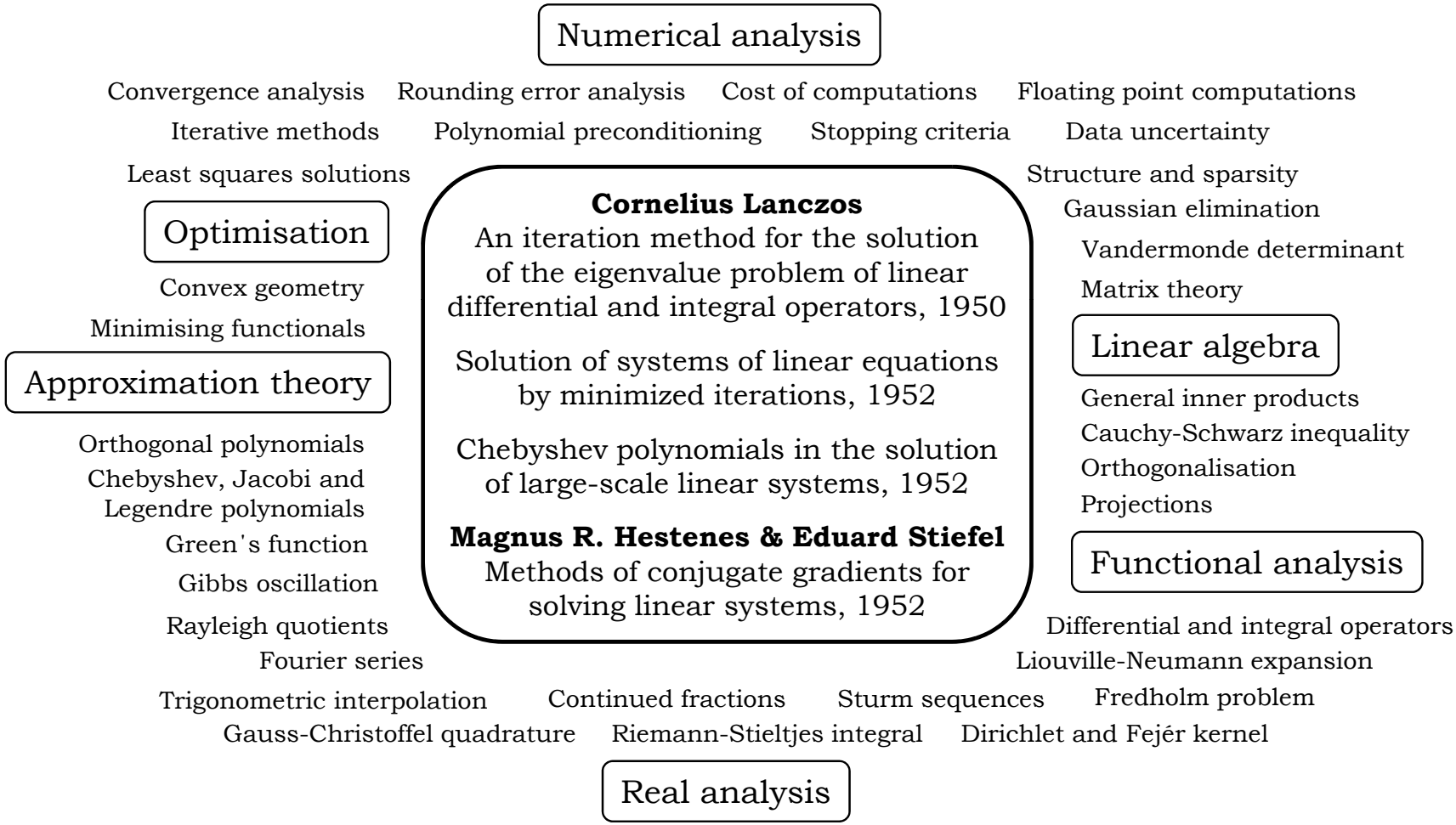
2 The depth of confusion and of unwillingness to listen

Referee report (2005): *"The only new items presented here have to do with analysis involving floating point operations (...). These are likely to bear very little interest to the audience of CMAME.*

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

It appears obvious to the authors that the A-norm is the quantity to measure to stop the iteration. In some case ... it is the residual norm (yes) that matters. For example, in nonlinear iterations, it is important to monitor the decrease of the residual norm - because the nonlinear iteration looks at the non-linear residual to build globally convergent strategies. This is known to practitioners, yet it is vehemently rejected by the authors."

2 Lanczos, Hestenes and Stiefel - phrases from the four papers



2 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. ...

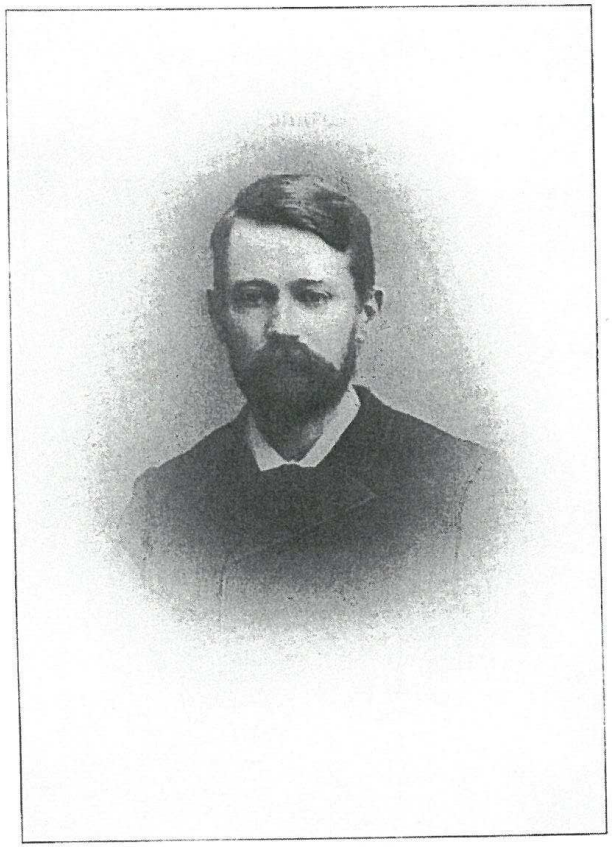
Little matter the actual things that are linked together; it is the links that I must begin by apprehending and interpreting."

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

3 Thomas Jan Stieltjes (1856 - 1894)



Thomas Jan Stieltjes
1856–1894

Investigations on Continued Fractions

T. J. Stieltjes

Ann. Fac. Sci. Toulouse 8 (1894) J.1–122; 9 (1895) A.1–47 (translation)

Introduction

The object of this work is the study of the continued fraction

$$\cfrac{1}{a_1 z + \cfrac{1}{a_2 + \cfrac{1}{a_3 z + \cdots + \cfrac{1}{a_{2n} + \cfrac{1}{a_{2n+1} z + \cdots,}}}}} \qquad \text{(I)}$$

in which the a_i are positive real numbers, while z is a variable which can take all real or complex values.

Denoting by $\frac{P_n(z)}{Q_n(z)}$ the nth convergent[1], which depends only on the first n coefficients a_i, 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 z.

We shall summarize the principal result of this study. There are two distinct cases.

First case. – The series $\sum_1^\infty a_n$ is convergent.

In this case we have for each finite value of z,

$$\lim P_{2n}(z) = p(z),$$
$$\lim Q_{2n}(z) = q(z),$$
$$\lim P_{2n+1}(z) = p_1(z),$$
$$\lim Q_{2n+1}(z) = q_1(z),$$

$p(z), q(z), p_1(z), q_1(z)$ being holomorphic functions in the whole plane which satisfy the relation

$$q(z)p_1(z) - q_1(z)p(z) = +1.$$

These functions are of genus zero and admit only simple zeros which are

$$1+\frac{1}{\boxed{2}} \qquad 1+\frac{1}{\boxed{2+\frac{1}{2}}} \qquad 1+\frac{1}{2+\frac{1}{2+\frac{1}{2}}} \qquad 1+\frac{1}{2+\frac{1}{2+\frac{1}{2+\frac{1}{2+\ddots}}}}$$

$$= 1.5 \qquad = 1.4 \qquad = 1.41666\overline{6} \qquad \longrightarrow \sqrt{2}$$

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 $\lambda \equiv -z$.

3 Remarkable mathematical object with remarkably long 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

- 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), Reid (1971), Greenbaum (1989), ...
- Magnus (1962a,b), Gragg (1974), Kalman (1979), Gragg, Lindquist (1983), Gallivan, Grimme, Van Dooren (1994), ...

Who is Yu. V. Vorobyev?

All what we know can be found in the monograph Liesen, S, *Krylov subspace methods, Principles and Analysis* , OUP, 2013, Section 3.7.

3 Book (1958, 1965), very few other traces ...

Russian Monographs and Texts on Advanced Mathematics and Physics

Volume I
A. I. KHINCHIN, *A Course of Mathematical Analysis*
Volume II
S. G. MIKHLIN, *Linear Integral Equations*
Volume III
P. P. KOROVKIN, *Linear Operators and Approximation Theory*
Volume IV
L. E. EL'SGOL'TS, *Differential Equations*
Volume V
L. A. LUSTERNIK and V. J. SOBOLEV, *Elements of Functional Analysis*
Volume VI
V. L. GINZBURG, *Propagation of Electromagnetic Waves in Plasma*
Volume VII
A. A. VLASOV, *Many-Particle Theory and its Application to Plasma*
Volume VIII
M. A. KRASNOSEL'SKI, *Convex Functions and Orlicz Spaces*
Volume IX
YU. V. VOROBYEV, *Method of Moments in Applied Mathematics*
Volume X
N. N. BOGOLIUBOV and Y. A. MITROPOLSKY, *Asymptotic Methods in the Theory of Non-Linear Oscillations*
Volume XI
S. D. VOLKOV, *Statistical Strength Theory*
Volume XII
Z. S. AGRANOVICH and V. A. MARCHENKO, *The Inverse Problem of Scattering Theory*
Volume XIII
V. F. NOZDREV, *Application of Ultrasonics in Molecular Physics*
Volume XIV
S. TARG, *Theoretical Mechanics: A Short Course*
Volume XV
A. S. KOMPANEYETS, *Theoretical Physics*
Volume XVI
P. STEPIN, *Strength of Materials*
Volume XVII
L. F. PONTRYAGIN, *Topological Groups*, Second Edition
Volume XVIII
S. B. PIKELNER, *Principles of Cosmic Electrodynamics*
Volume XIX
O. V. KOVALEV, *Irreducible Representation of Space Groups*

Additional Volumes in Preparation

Method of Moments in Applied Mathematics

by YU V. VOROBYEV

Translated from the Russian
by BERNARD SECKLER

GORDON AND BREACH SCIENCE PUBLISHERS
New York · London · Paris

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

Determine under which conditions the solution of the system of $2n$ 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 positive unknowns $\omega_j^{(n)} > 0,\ \theta_j^{(n)} > 0 \,, \quad j = 1, \ldots, n$
exists and is unique, and give the solution.

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

Determine under which conditions the solution of the system of $2n$ 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 positive unknowns $\omega_j^{(n)} > 0,\ \theta_j^{(n)} > 0 \,, \quad j = 1, \ldots, n$
exists and is unique, and give the solution.

Is this problem linear?
Mathematical description of the solution?
How to compute the solution?

3 Moment problem defined and solved by Stieltjes in (1894)

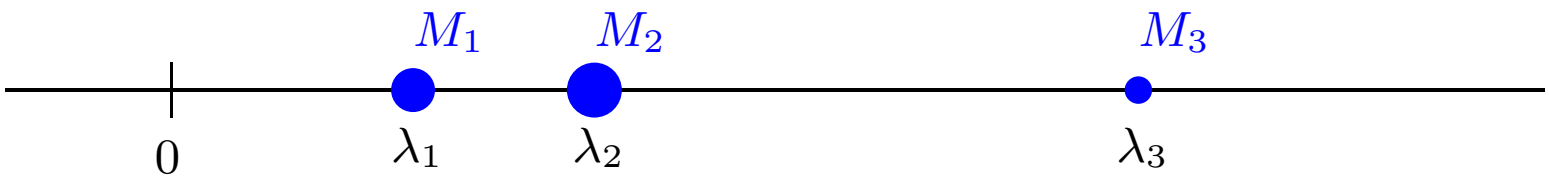
Consider an infinite sequence of real numbers $m_0, m_1, m_2, \ldots$.

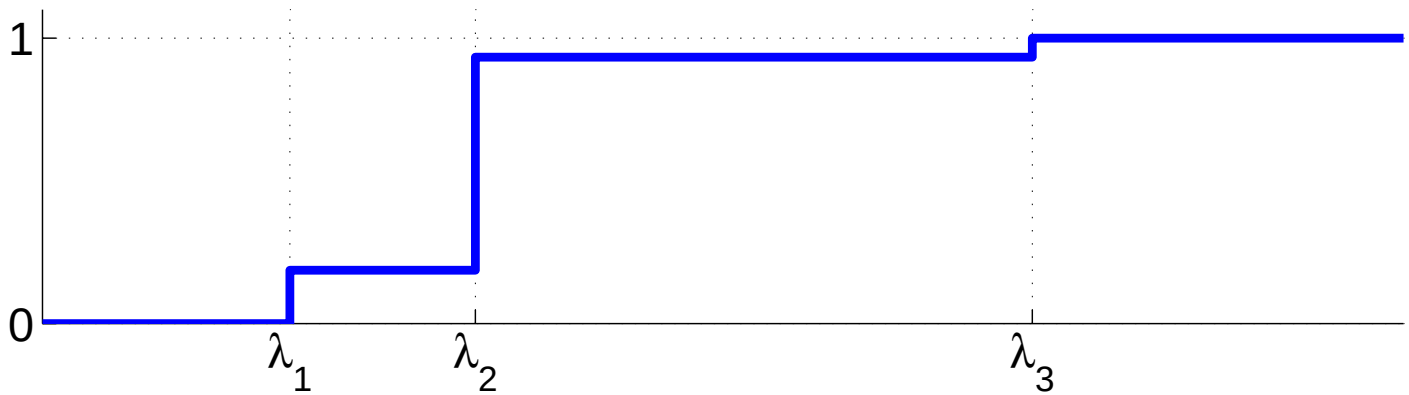
Find necessary and sufficient conditions for the existence of a Riemann-Stieltjes integral with the (positive nondecreasing) distribution function $\omega(\lambda)$ such that

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

and determine $\omega(\lambda)$.

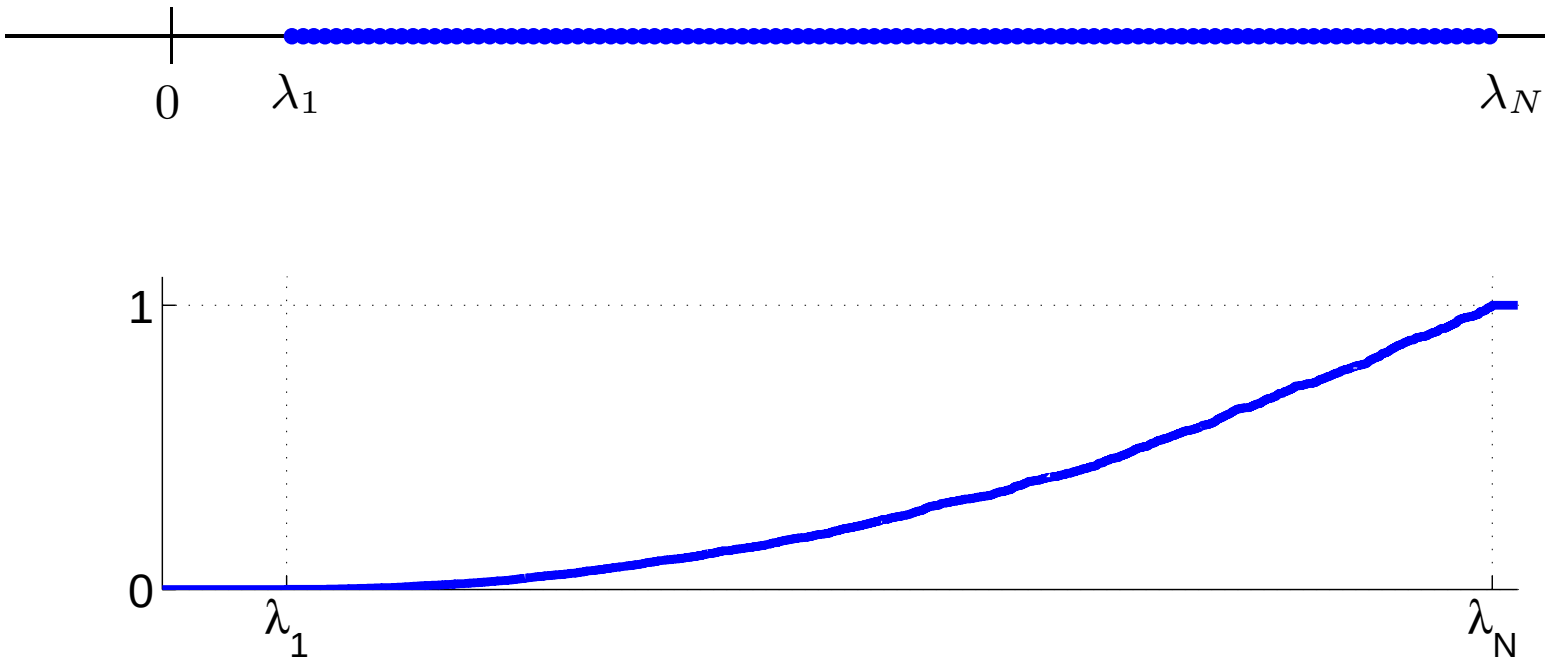
3 Distribution of mass points on the positive half real line





Meaning of $m_\ell = \int_0^\infty \lambda^\ell \, d\omega(\lambda)$ for $m = 0, 1, 2, \ldots$?

3 Many points of various weights between λ_1 and λ_N

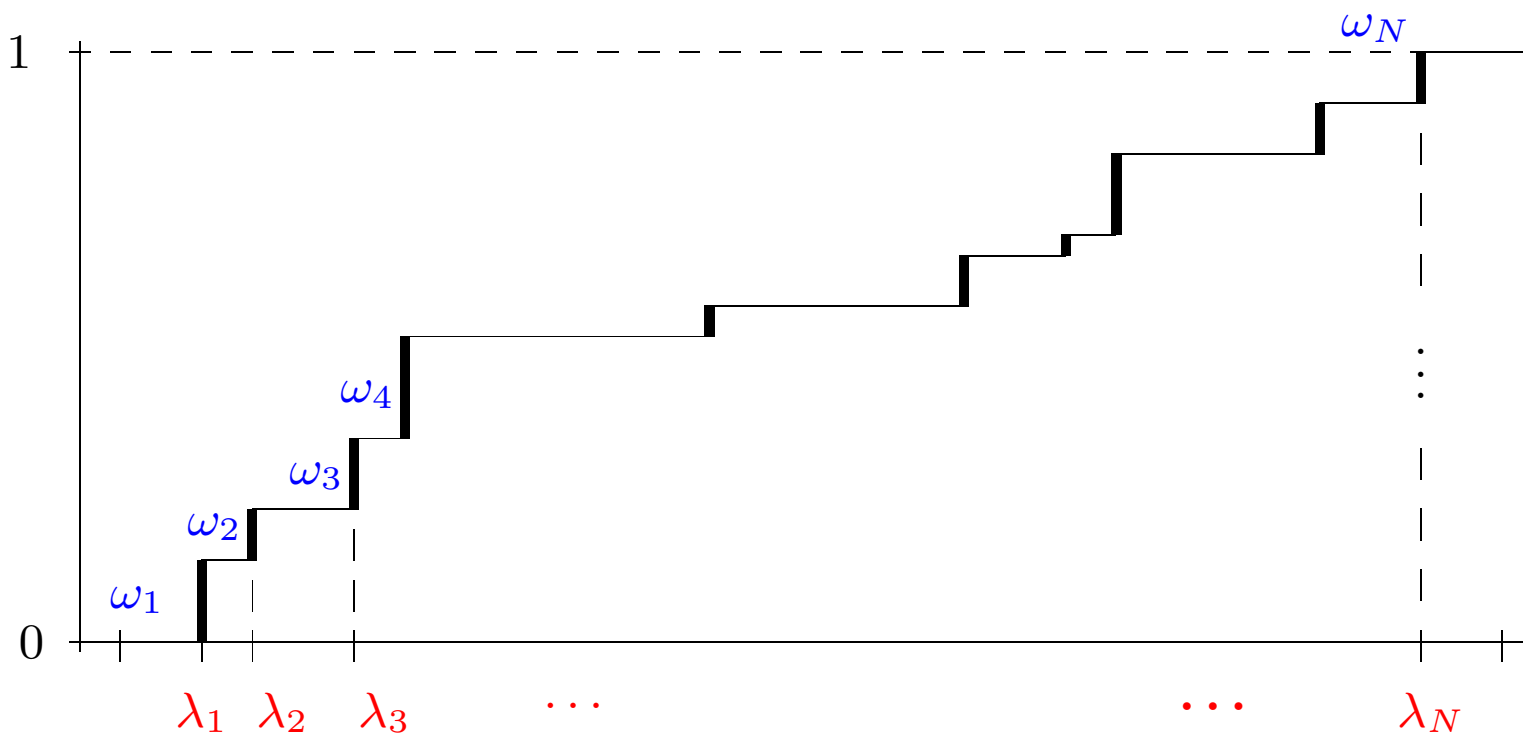


How is this related to approximating numerically the solution of a system of linear algebraic equations or the solution of a matrix eigenvalue problem?

3 CG and Lanczos give the solution of the homework problem!

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

λ_i, y_i are the eigenpairs of A, $\omega_i = |(y_i, w_1)|^2$, $(w_1 = r_0/\|r_0\|)$



Symbolically

$$w_1^* A w_1 = w_1^* \left(\sum_{\ell=1}^{N} \lambda_\ell \, y_\ell y_\ell^* \right) w_1 \equiv w_1^* \left(\int \lambda \, dE(\lambda) \right) w_1$$

$$= \sum_{\ell=1}^{N} \lambda_\ell \, |(y_\ell, w_1)|^2 = \sum_{\ell=1}^{N} \lambda_\ell \, \omega_\ell = \int \lambda \, d\omega(\lambda) \,,$$

where the spectral function $E(\lambda)$ of A is understood to be a nondecreasing family of projections with increasing λ, symbolically $dE(\lambda_\ell) \equiv y_\ell y_\ell^*$ and

$$I = \sum_{\ell=1}^{N} y_\ell y_\ell^* \equiv \int dE(\lambda) \,.$$

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

- Finite dimensional self-adjoint operators (finite Hermitian matrices)

$$\mathcal{G} = \frac{1}{2\pi\iota}\int_{\Gamma}\lambda\,(\lambda I_N - \mathcal{G})^{-1}\,d\lambda = \frac{1}{2\pi\iota}\sum_{j=1}^{N}\int_{\Gamma_j}\lambda\,(\lambda I_N - \mathcal{G})^{-1}\,d\lambda$$
$$= \sum_{j=1}^{N} Y\,\mathrm{diag}\left(\frac{1}{2\pi\iota}\int_{\Gamma_j}\frac{\lambda}{\lambda-\lambda_j}\,d\lambda\right)Y^* = \sum_{j=1}^{N}\lambda_j\,y_j y_j^*$$
$$= \int \lambda\,dE(\lambda)\,.$$

- Compact infinite dimensional self-adjoint operators
- Bounded infinite dimensional self-adjoint operators
- Generalization to bounded normal and non-normal operators

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

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

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

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

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

- Conjugate gradient (CG) method:
  - 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:
  - 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 Krylov subspace methods. Some of them are not well defined for each n (e.g. BiCGStab or QMR). They are not linear in A and b .

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

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

the CG approximations are given by

$$T_n\,t_n \;=\; \|r_0\| e_1\,, \quad x_n \;=\; x_0 \;+\; W_n\,t_n\,.$$

The nonlinearity of the CG projection process giving the CG optimality, wrt both A and b, is clearly visible.

4 (Petrov-) Galerkin framework

Projection idea in Krylov subspace methods is analogous to the (Petrov-) Galerkin framework, as, e.g., in numerical solution of PDEs.

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.

4 (Petrov-) Galerkin framework

Projection idea in Krylov subspace methods is analogous to the (Petrov-) Galerkin framework, as, e.g., in numerical solution of PDEs.

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:

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

4 Operator problem formulation (more in Appendix)

- 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.

4 CG in infinite dimensional Hilbert spaces (more in Appendix)

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

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

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

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

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

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

Superlinear convergence for (identity + compact) operators.

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

Here the Riesz map τ represents preconditioner.

4 Summary of this part and motivation

- Well defined Krylov subspace methods are based on orthogonality wrt nested Krylov-related subspaces, which can be equivalently formulated using the associated optimality property.
- Therefore the resulting methods are highly nonlinear in the data defining the problem to be solved.
- The nonlinearity allows to adapt to the problem as the iteration proceeds. This can be overlooked while using the derivation of CG based on the minimization of the quadratic functional. Therefore this fundamental nonlinearity of CG for solving system of linear equations is typically not presented in textbooks.
- The adaptation is obvious from the matching moments model reduction view presented next.

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

5 Recall the CG (Lanczos) relationship with Jacobi matrices

Let $W_n = [w_1, \ldots, w_n]$, $A W_n = W_n T_n + \delta_{n+1} w_{n+1} e_n^T$, form the Lanczos orthonormal basis of the Krylov subspace $K_n(A, r_0)$. Here the Jacobi matrix of the orthogonalization coefficients

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

represents, at the same time, the matrix of the restricted and orthogonally projected operator WW^*A on $K_n(A, r_0)$ in the basis W_n. The CG approximation is determined by

$$T_n t_n = \|r_0\| e_1 , \quad x_n = x_0 + W_n t_n .$$

5 CG as the matrix-formulated model reduction

- This can be viewed as a model reduction from a (large) system of order N to a (small) system of order n.
- In order to be efficient, the projection process should capture fast substantial part of information contained in the original data.
- Intuition: Repeated application of the operator tends to transfer into the projected system dominating information.
- Can this model reduction be defined by a matching moments property?

5 Recall the distribution functions and moments

- Let A be HPD with spectral decomposition $A = Y\Lambda Y^*$, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_N$ be the (distinct) eigenvalues (for simplicity).
- Let $\omega_k = |(w_1, y_k)|^2 > 0, \quad k = 1, \ldots, N,$ 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

$$\int_0^\infty \lambda^k d\omega(\lambda) = \sum_{\ell=1}^{N} \omega_\ell \{\lambda_\ell\}^k = w_1^* A^k w_1 \quad k = 0, 1, 2, \ldots$$

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

$$\int_0^\infty \lambda^k d\omega^{(n)}(\lambda) = \sum_{j=1}^{n} \omega_j^{(n)} \{\lambda_j^{(n)}\}^k = e_1^T T_n^k e_1, \quad k = 0, 1, 2, \ldots$$

5 Recall the distribution functions and moments

- Let A be HPD with spectral decomposition $A = Y\Lambda Y^*$, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_N$ be the (distinct) eigenvalues (for simplicity).
- Let $\omega_k = |(w_1, y_k)|^2 > 0, \quad k = 1, \ldots, N,$ 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

$$\int_0^\infty \lambda^k d\omega(\lambda) = \sum_{\ell=1}^{N} \omega_\ell \{\lambda_\ell\}^k = w_1^* A^k w_1 \quad k = 0, 1, 2, \ldots$$

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

$$\int_0^\infty \lambda^k d\omega^{(n)}(\lambda) = \sum_{j=1}^{n} \omega_j^{(n)} \{\lambda_j^{(n)}\}^k = e_1^T T_n^k e_1\,, \quad k = 0, 1, 2, \ldots$$

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 same Jacobi matrix as above

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

5 Relationship between CG (Lanczos) and Gauss quadrature

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

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

The quadrature nodes $\lambda_j^{(n)}$ are the eigenvalues $\theta_j^{(n)}$ of T_n and the weights $\omega_j^{(n)}$ are the squared first components of the associated normalized eigenvectors.

5 Continued fraction corresponding to $\omega(\lambda)$ defined by A, r_0

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

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

$$b^* (zI - A)^{-1} b = \int_0^\infty \frac{d\omega(\lambda)}{z-\lambda} = \sum_{j=1}^{N} \frac{\omega_j}{z-\lambda_j} = \frac{\mathcal{R}_N(z)}{\mathcal{P}_N(z)} = \mathcal{F}_N(z)$$

$$\approx \sum_{j=1}^{n} \frac{\omega_j^{(n)}}{z-\lambda_j^{(n)}} = \frac{\mathcal{R}_n(z)}{\mathcal{P}_n(z)} = \mathcal{F}_n(z)\,,$$

The denominator $\mathcal{P}_n(z)$ corresponding to the nth convergent $\mathcal{F}_n(z)$ of $\mathcal{F}_N(z)$, $n = 1, 2, \ldots$ is the nth monic orthogonal polynomial in the sequence determined by the distribution function ω and the numerator $\mathcal{R}_n(z)$ is determined by the same recurrence started instead of 1 and z with 0 and 1, see Chebyshev (1855).

5 Additional comments: History repeats

- 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

$$w_1^* A^k w_1 = e_1^T T_n^k e_1 \quad \text{for} \quad k = 0, 1, \ldots, 2n-1.$$

- 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).

We are in fact back (now using the language of matrix-vector algebra) to the investigations of Stieltjes (1894), or even Jacobi (1826) and Gauss (1814).

Let $z_0, z_1, \ldots, z_n$ be $n+1$ linearly independent elements of a 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 $z_1, z_2, \ldots, z_n, \ldots$ such that

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

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).

See CG in infinite dimensional Hilbert spaces given above in part 4.

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

- 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

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)| \leq \min_{p \in \mathcal{P}_n(0)} \max_{\lambda \in [\lambda_1, \lambda_N]} |p(\lambda)|$$

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

- 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

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq \min_{p \in \mathcal{P}_n(0)} \max_{1 \leq j \leq N} |p(\lambda_j)| \leq \min_{p \in \mathcal{P}_n(0)} \max_{\lambda \in [\lambda_1, \lambda_N]} |p(\lambda)|$$

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

- 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, $\kappa_s(A) \equiv \lambda_{N-s}/\lambda_1$. Then

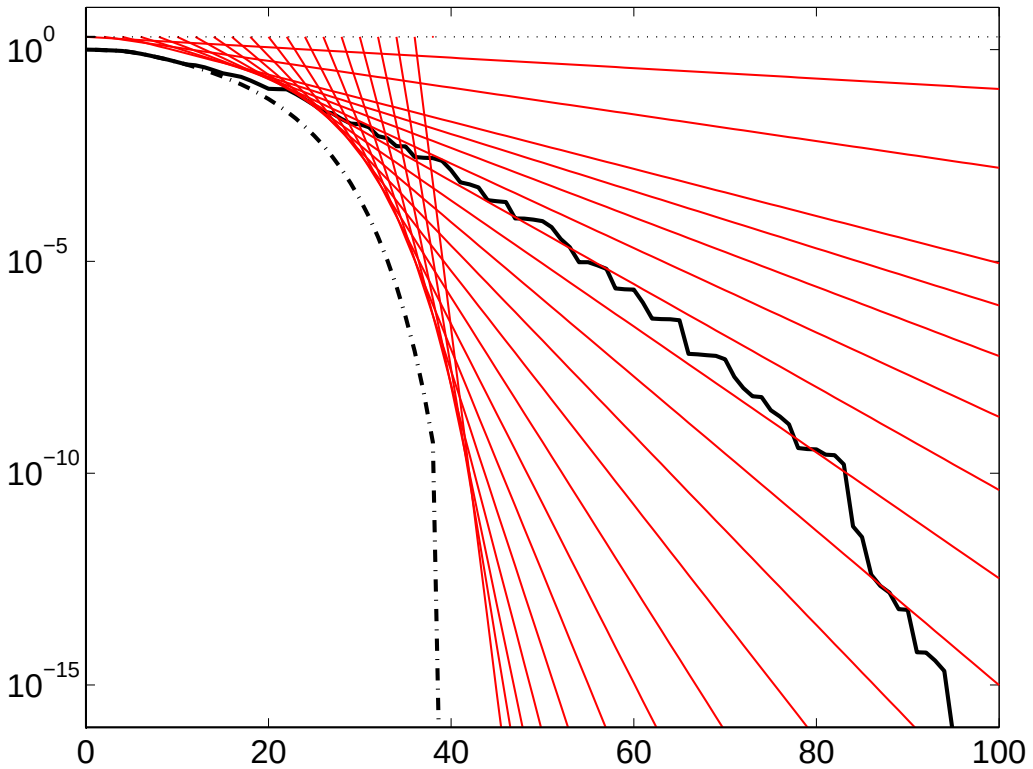
$$\mathbf{k} = \mathbf{s} + \left\lceil \frac{\ln(2/\epsilon)}{2} \sqrt{\kappa_s(A)} \right\rceil$$

CG steps will produce the approximate solution x_n satisfying

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

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

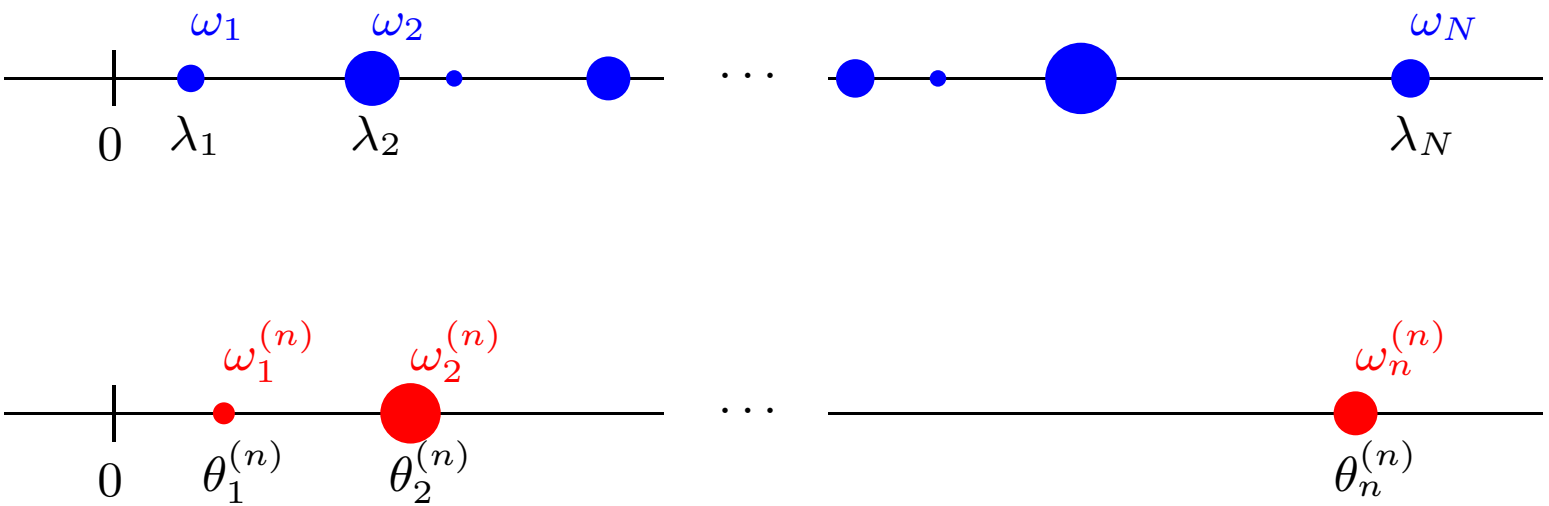
6 Adaptive Chebyshev bound fails to resolve the matter



The finite precision computation (the thick black line) is not captured quantitatively nor described qualitatively!

6 Recall the 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



6 Back to CG and Gauss quadrature - equivalence of errors!

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

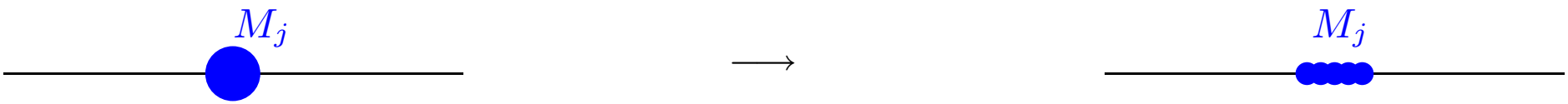
$$\int_0^\infty f(\lambda)\, d\omega(\lambda) \;=\; \sum_{i=1}^{n} \omega_i^{(n)} f(\theta_i^{(n)}) \;+\; R_n(f)\,.$$

For $f(\lambda) \equiv \lambda^{-1}$,

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

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

6 Clustering of eigenvalues does make a change! Even for A HPD!

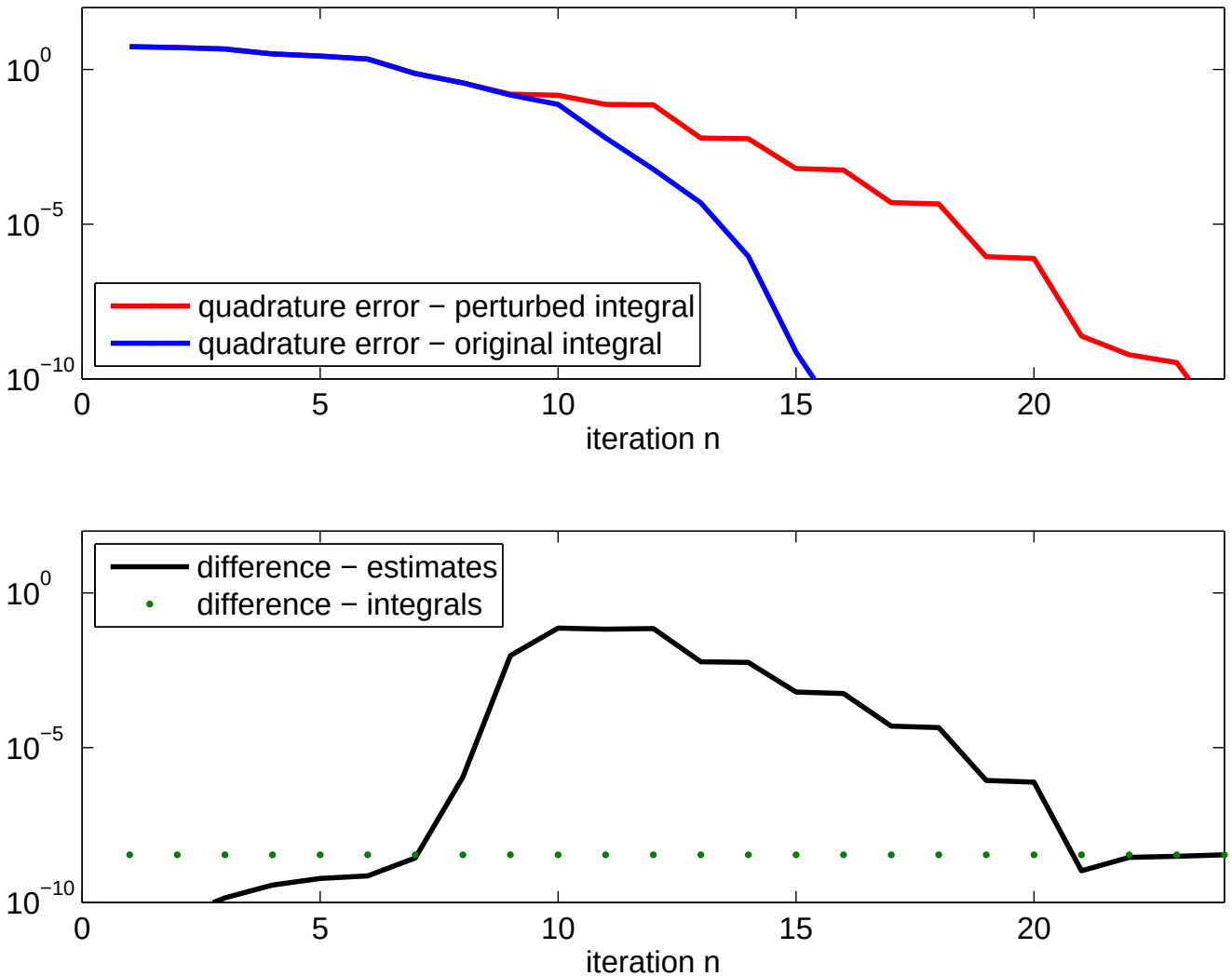


single eigenvalue λ_j $\longrightarrow$ many close eigenvalues $\widehat{\lambda}_{j_1}, \widehat{\lambda}_{j_2}, \ldots, \widehat{\lambda}_{j_\ell}$

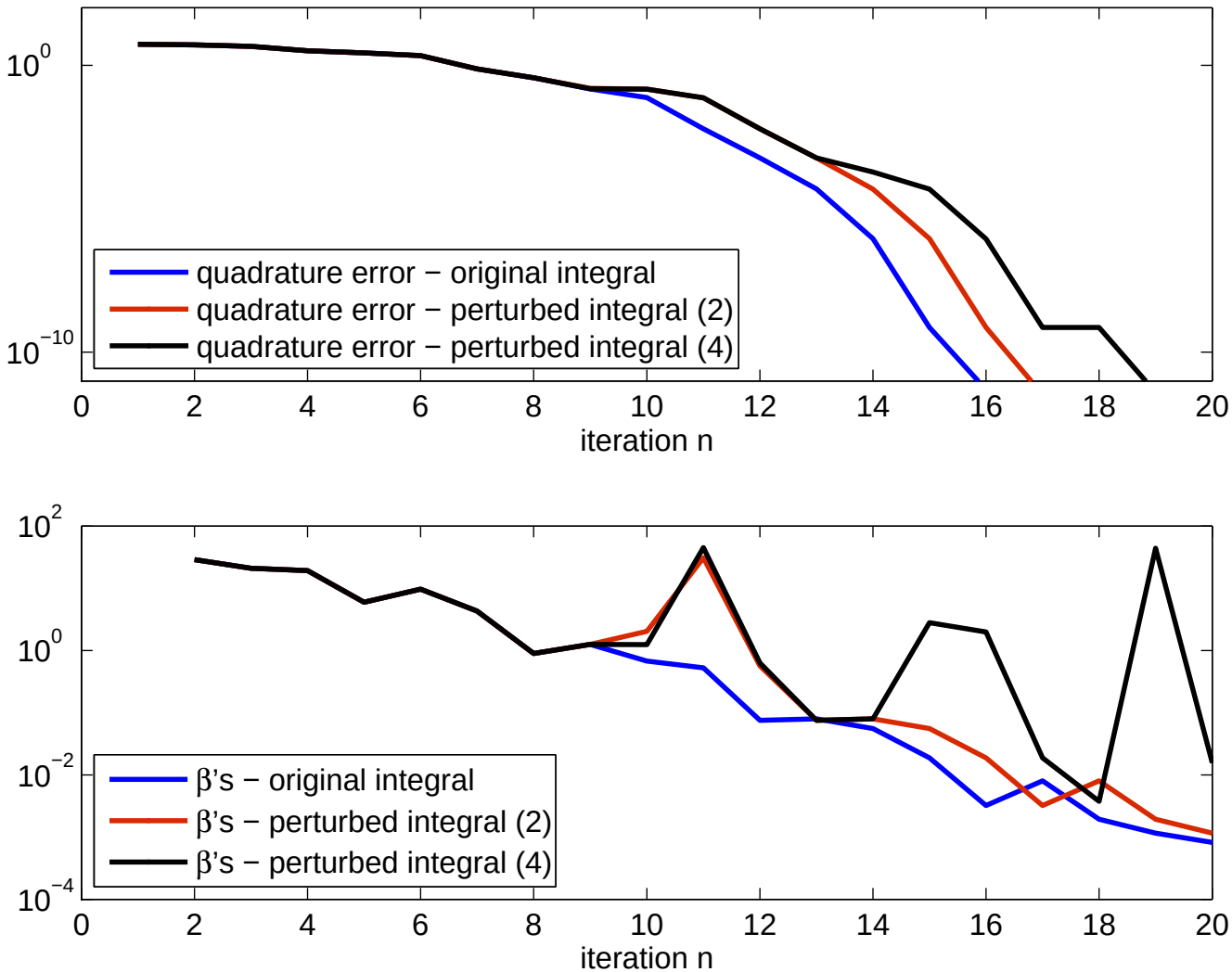
Replacing a single eigenvalue by a tight cluster can make a substantial difference; Greenbaum (1989); Greenbaum, S (1992); Golub, S (1994).

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?**

6 Gauss quadrature can be highly sensitive to small changes of $\omega(\lambda)$!



6 Illustration - only the largest eigenvalue is replaced by a cluster



6 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-node Gauss quadrature, is bounded as

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

6 Summary of the CG/Lanczos part in relation to the Gauss quadrature

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.

6 Another issue - convergence results for the GMRES method

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.

6 Another issue - convergence results for the GMRES method

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.

6 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.

6 Theorem: Any GMRES convergence with any spectrum

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

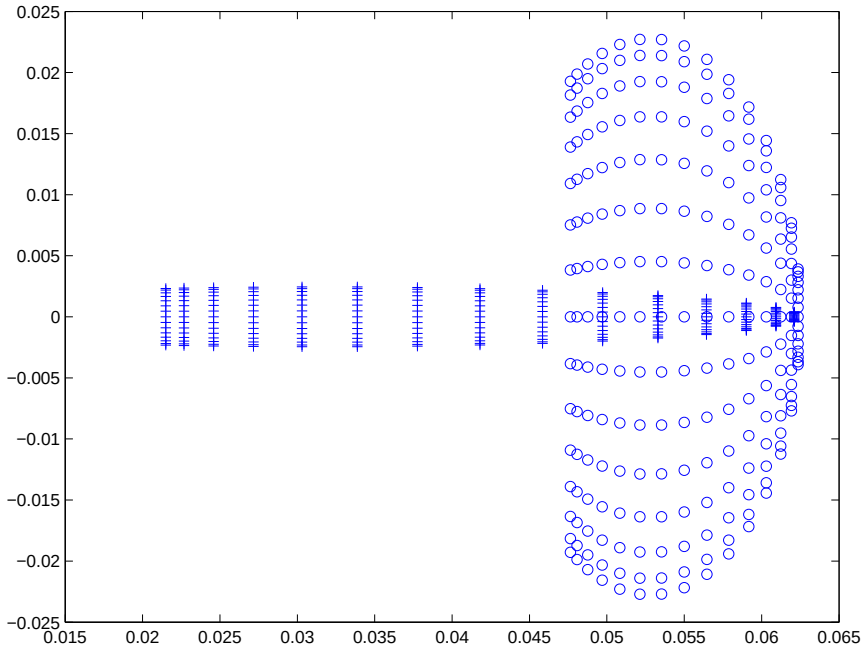
$$\|r_0\| \geq \|r_1\| \geq \cdots \geq \|r_{N-1}\| > \|r_N\| = 0\,.$$

2° Let C be the spectral companion matrix, $h = (h_1, \ldots, h_N)^T$, $h_i^2 = \|r_{i-1}\|^2 - \|r_i\|^2$, $i = 1, \ldots, N$ be the vector with its elements measuring the GMRES progress at the individual steps. Let R be a nonsingular upper triangular matrix such that $Rs = h$ with s being the first column of C^{-1}, and let W be unitary matrix. Then

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

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

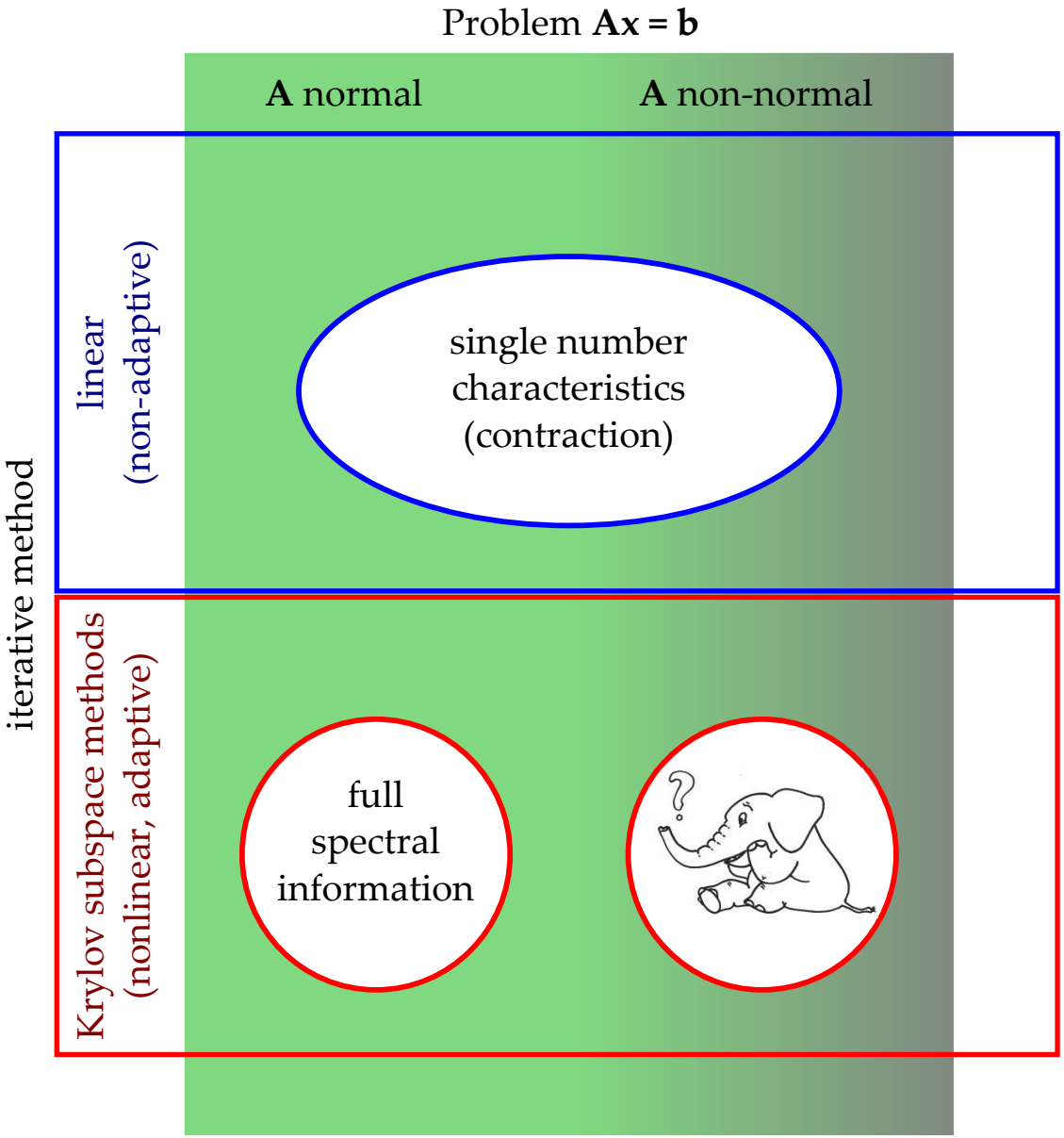
6 Quiz: Convection-diffusion model problem



$$-\nu\,\Delta u + w\cdot\nabla u \;=\; 0, \quad \text{in } \Omega = (0,1)\times(0,1), \qquad u = g \;\text{ on }\; \partial\Omega\,.$$

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

6 Adaptation to the inner nature of the problem?

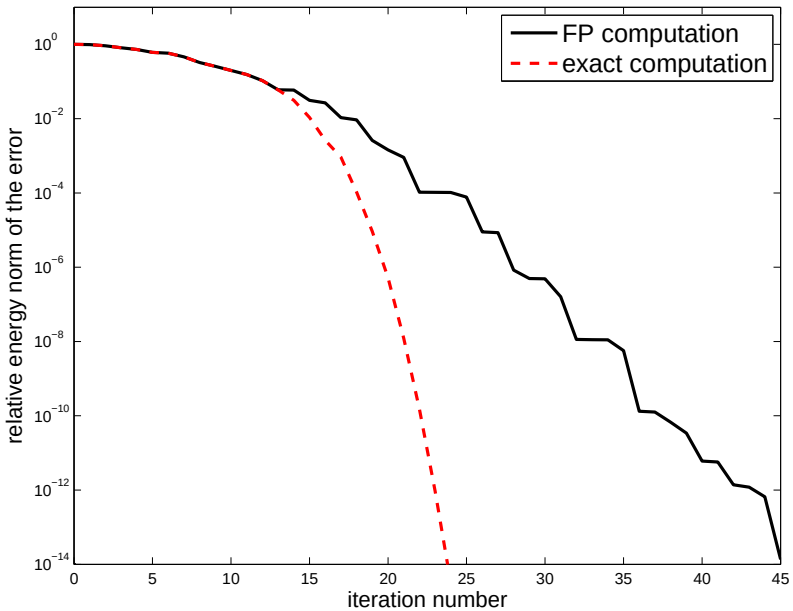


Outline

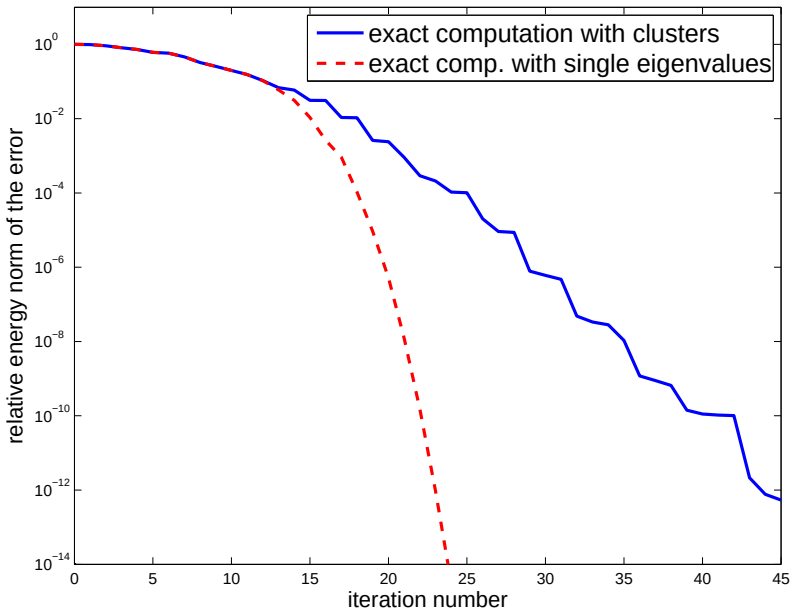
1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

7 CG convergence behavior in finite precision arithmetic

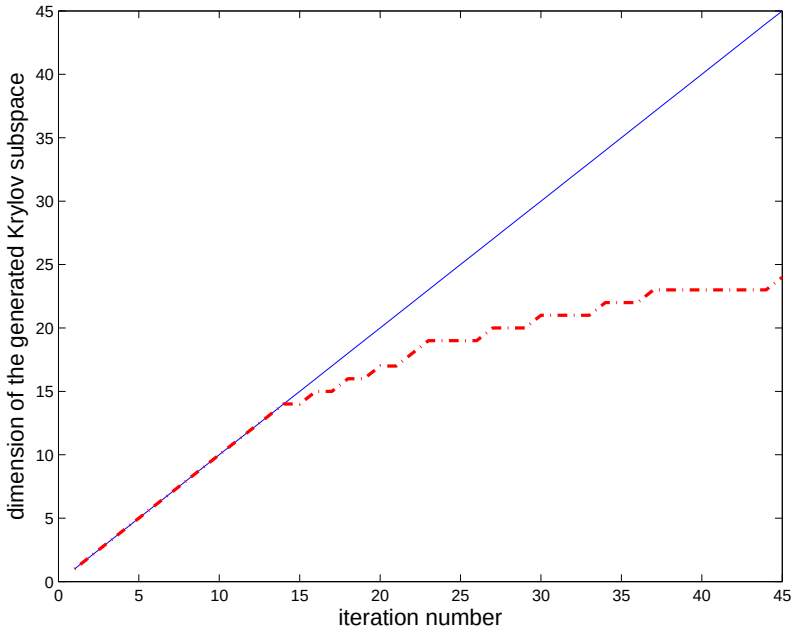


Rounding errors in finite precision CG computations cause a delay of convergence.

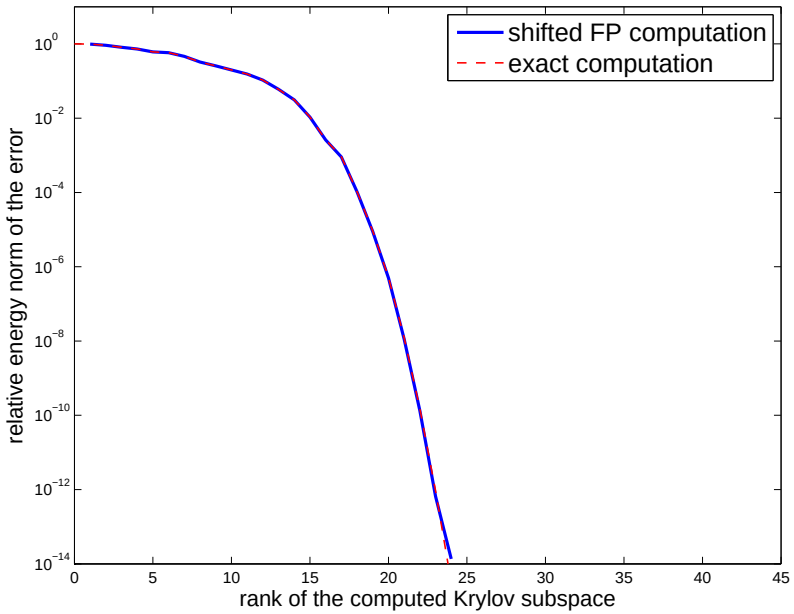


Exact CG computation for a matrix, where each eigenvalue is replaced by a tight cluster.

7 Delay of convergence and numerical rank of Krylov subspaces



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 observations given above have been proved by rigorous mathematical means!

Lanczos (with small inaccuracy also CG) in finite precision arithmetic can be seen as the exact arithmetic Lanczos (CG) for the problem with the slightly modified distribution function 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).

Now it is obvious why 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 applicable for description of the superlinear convergence behaviour of CG. They represent for methods with short recurrences a principally misleading concept.

7 Finite precision Lanczos (CG) computations are optimal!

- 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 through n.

7 Optimality in finite precision Lanczos (CG) computations?

Using the mathematically equivalent formulation of CG

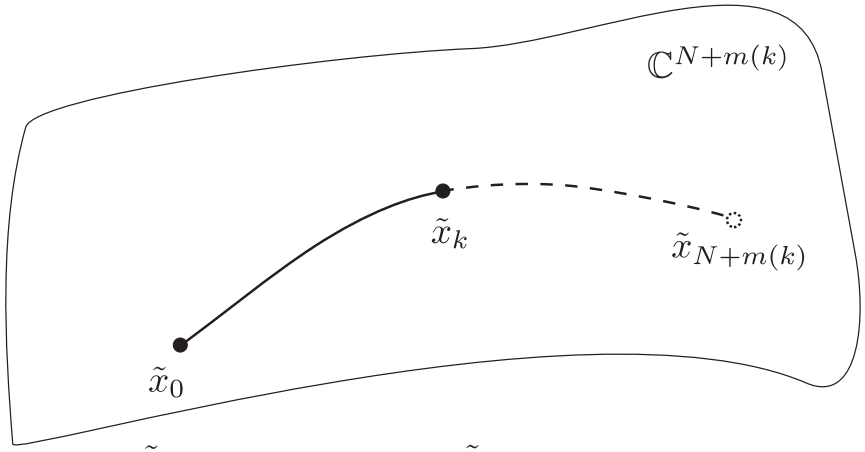
$$A\,W_n \;=\; W_n\,T_n \;+\; \delta_{n+1} w_{n+1} e_n^T\,, \quad T_n \;=\; W_n^*(A, r_0)\,A\,W_n(A, r_0)\,,$$

with the CG approximation given by

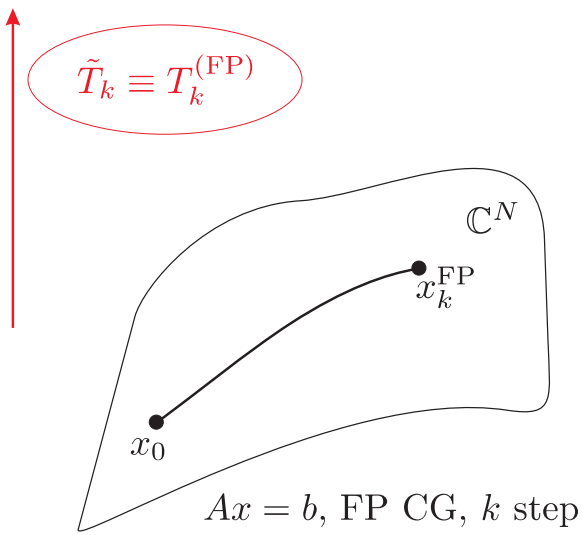
$$T_n\,t_n \;=\; \|r_0\| e_1\,, \quad x_n \;=\; x_0 \;+\; W_n\,t_n\,,$$

- 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 and has tight clusters of points of increase around the original eigenvalues of A.

7 Trajectories in spaces of different dimension



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

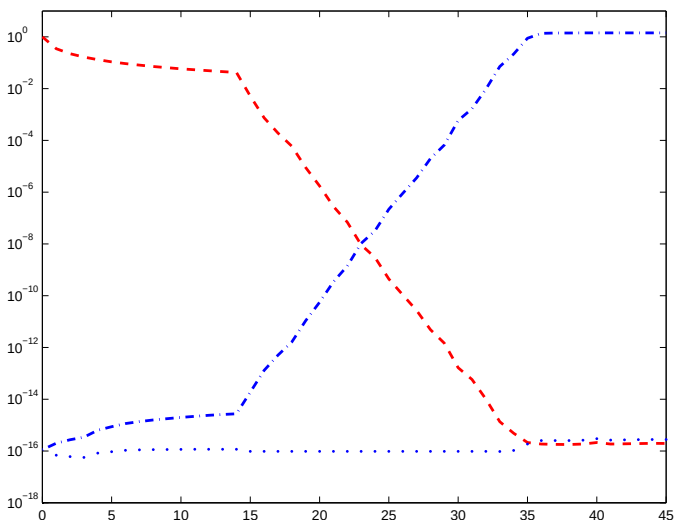


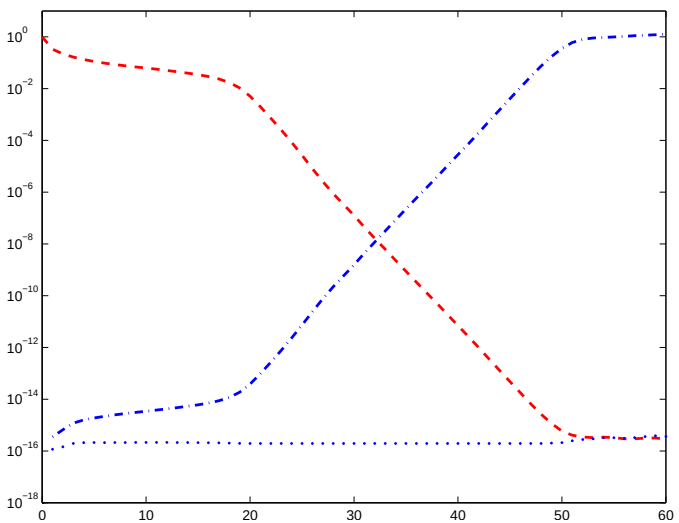
$Ax = b$, FP CG, k step

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

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

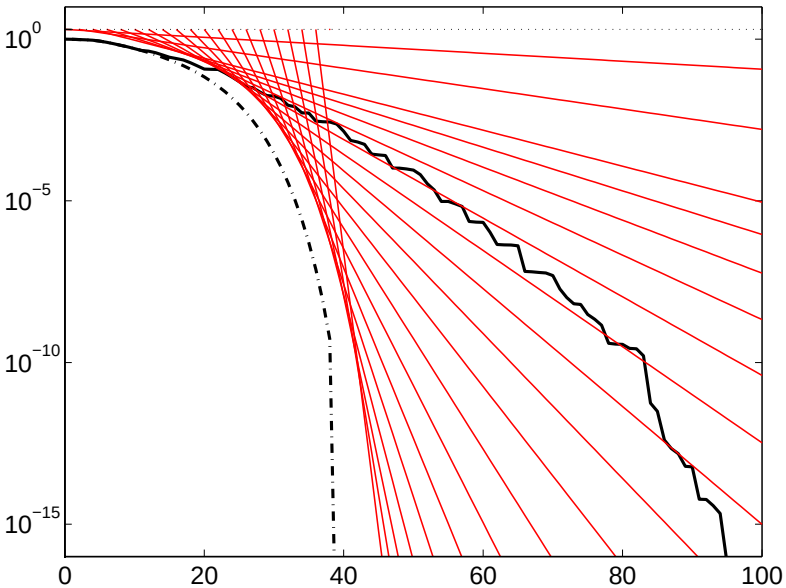
Loss of orthogonality (blue) and normwise relative backward error (red) for a convection-diffusion model problem with two different "winds":



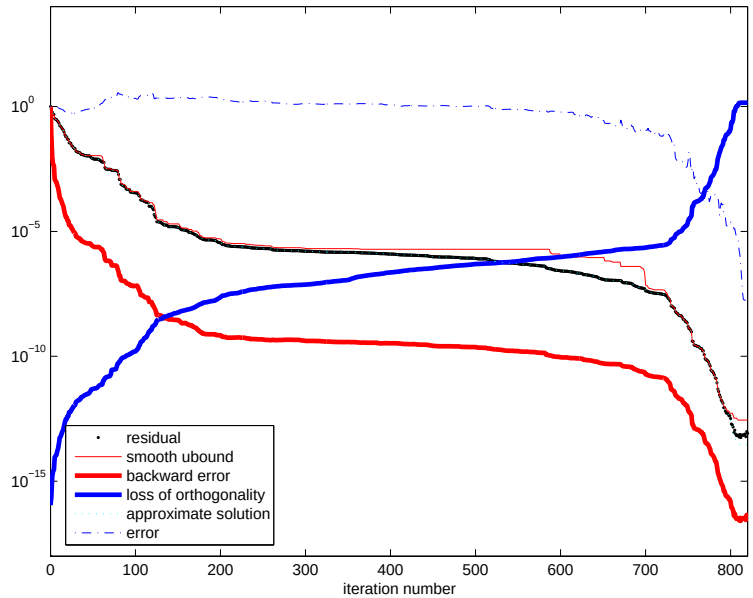


- It can be shown that the MGS-GMRES is normwise backward stable.

7 Major unresolved challenge - delay of convergence due to inexactness

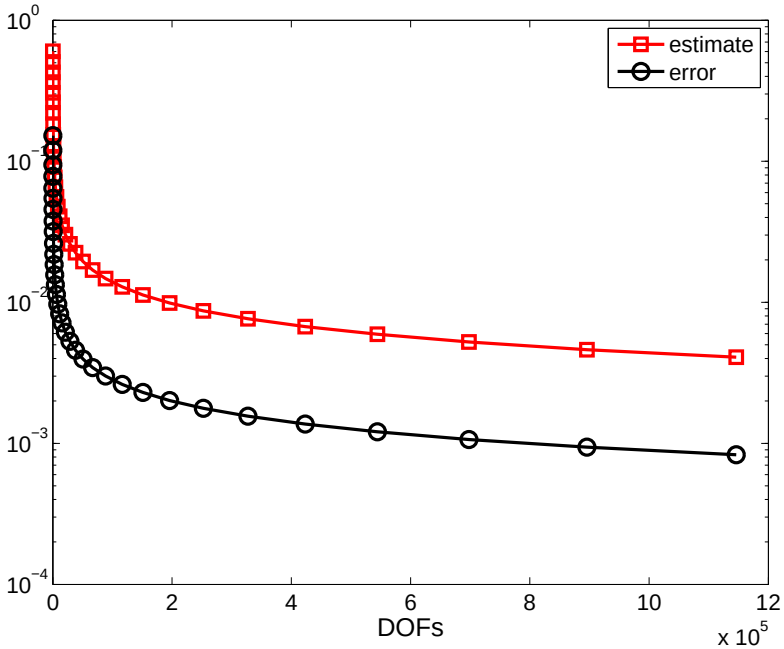


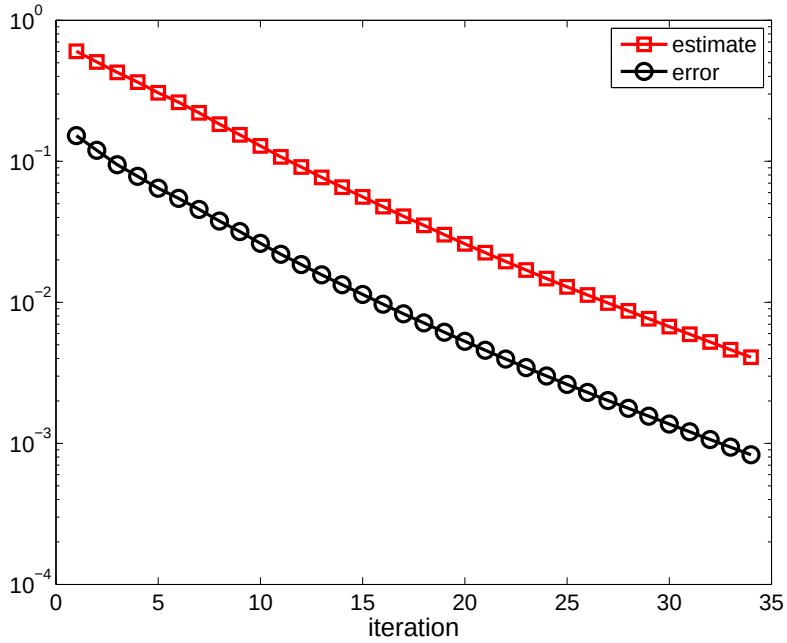
?



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.

7 A questionable conclusion - an arbitrary accuracy in AFEM?





Including inexactness and maximal attainable accuracy in matrix computations?

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

8 Myths in Mathematics?

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)

8 Myths in Mathematics?

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)

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

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

8 Examples of widespread myths concern

- Minimal polynomials and finite termination property
- Chebyshev bounds and CG
- Spectral information and clustering of eigenvalues
- Operator-based bounds and functional analysis arguments on convergence
- Finite precision computations seen as a minor modification of the exact considerations
- Linearization of nonlinear phenomenon
- Considering CG in matrix computations as a simplification of CG in general nonlinear optimization
- Well conditioned basis and short recurrences (look-ahead)
- Sparsity as an ultimate positive feature
- Discretization and algebraic errors in numerical PDEs

- 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). The associated myth has been proved false more than 25 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 can not be claimed 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 above, the relevant results Greenbaum, S (1994); Greenbaum, Pták, S (1996) and Arioli, Pták, S (1998) are more than 20 years old.

8 How the mathematical myths contradict the historical facts

- Lanczos, Hestenes and Stiefel did consider CG as iterative algorithm. In order to see that it is enough to read the titles and the abstracts of their papers published within 1950 - 52. Still, how many papers claim the opposite, even those published to celebrate the anniversaries of the original publications!

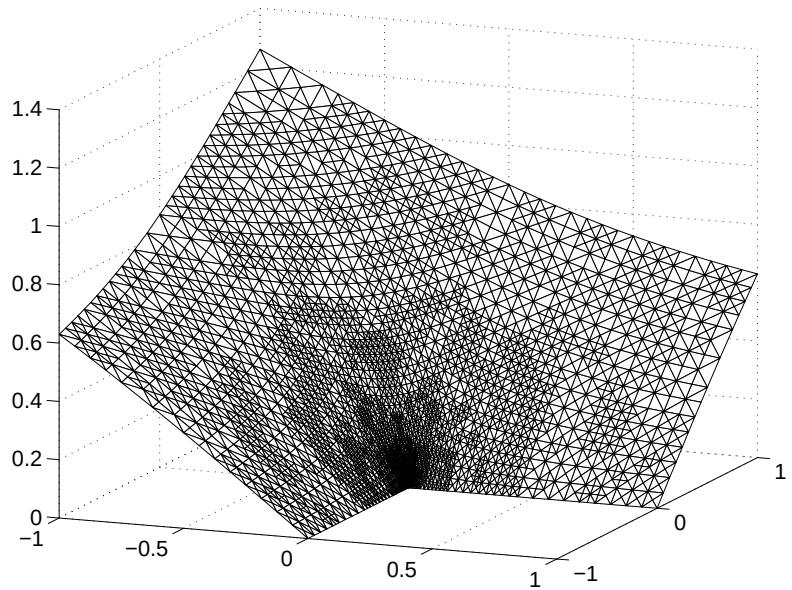
- Rutishauser (1959) as well as Lanczos (1952) considered CG principally different in their nature from the method based on Chebyshev polynomials.

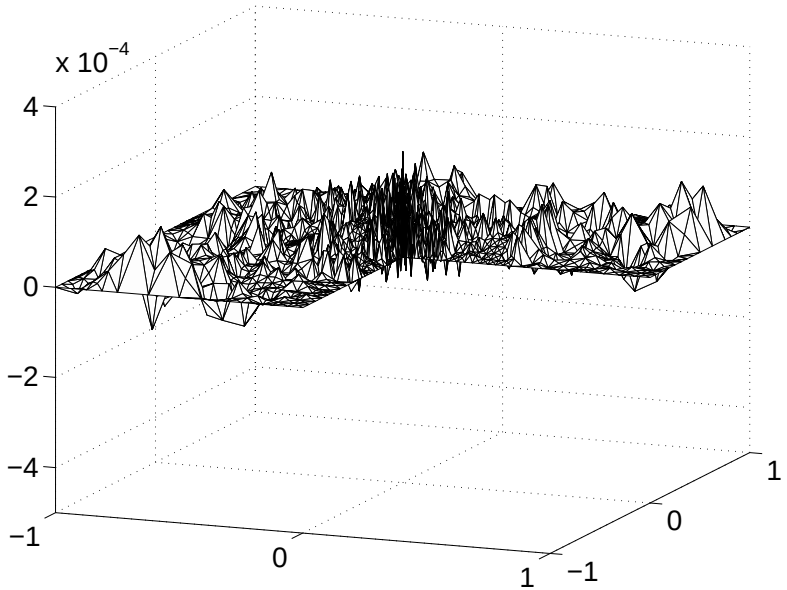
- Daniel (1967) did not identify the CG convergence with the Chebyshev polynomials-based bound. He carefully writes (modifyling slightly his notation)

  "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_N and for any distribution of the components of the initial residuals in the individual invariant subspaces.

8 Are algebraic errors in numerical PDEs easy to handle?

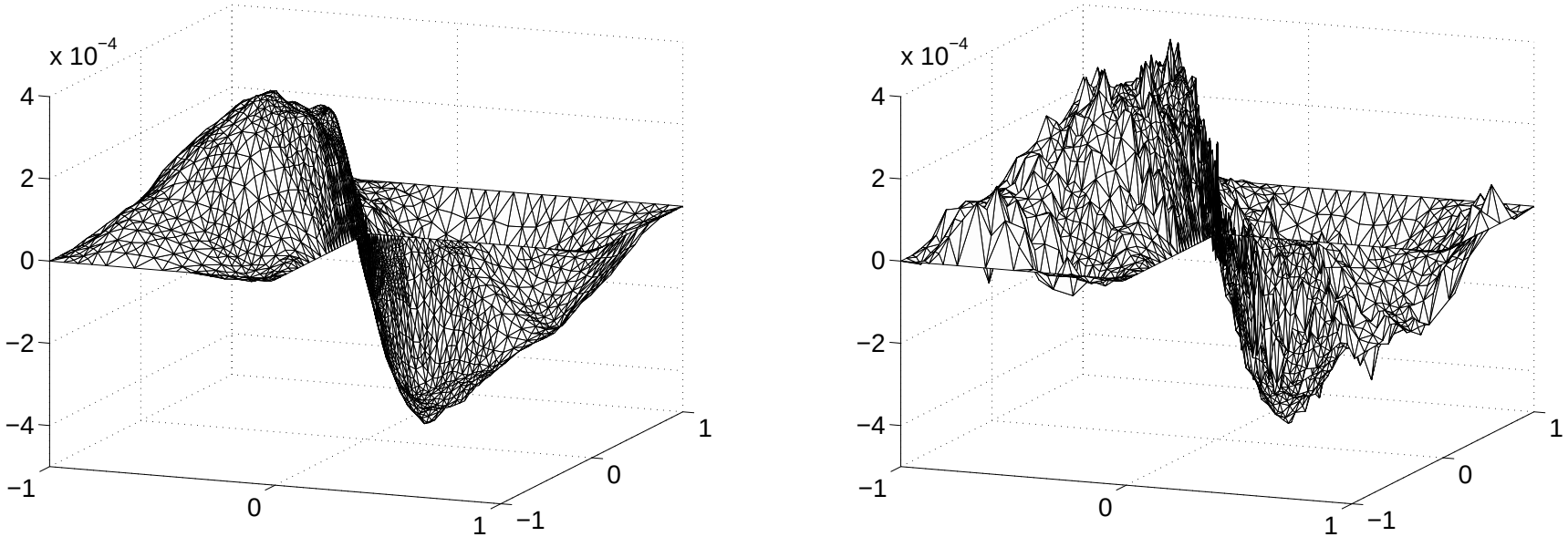




Exact solution u (left) and the discretization error $u - u_h$ (right) in the Poisson model problem, linear FEM, adaptive mesh refinement.

Quasi equilibrated discretization error over the domain.

8 L-shape domain, Papež, Liesen, S (2014)



Algebraic error $u_h - u_h^{(n)}$ (left) and the total error $u - u_h^{(n)}$ (right) after a number of CG iterations guaranteeing

$$\|\nabla(u - u_h)\| \gg \|x - x_n\|_A \,.$$

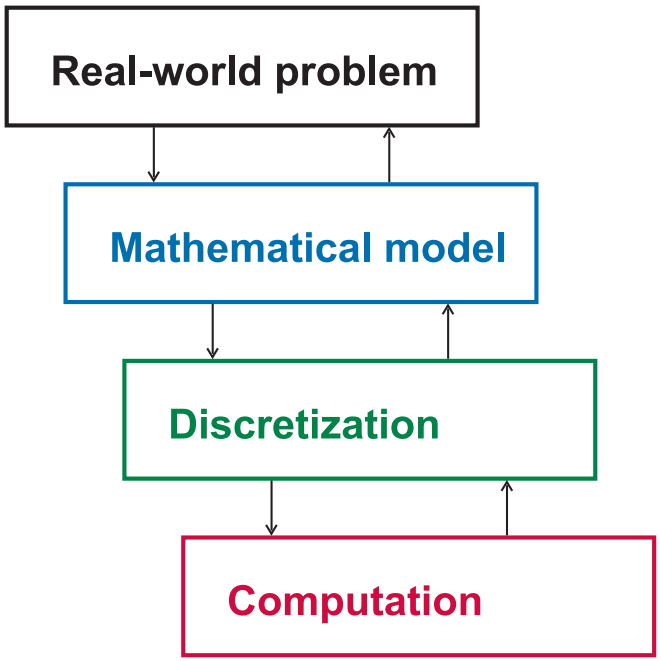
Outline

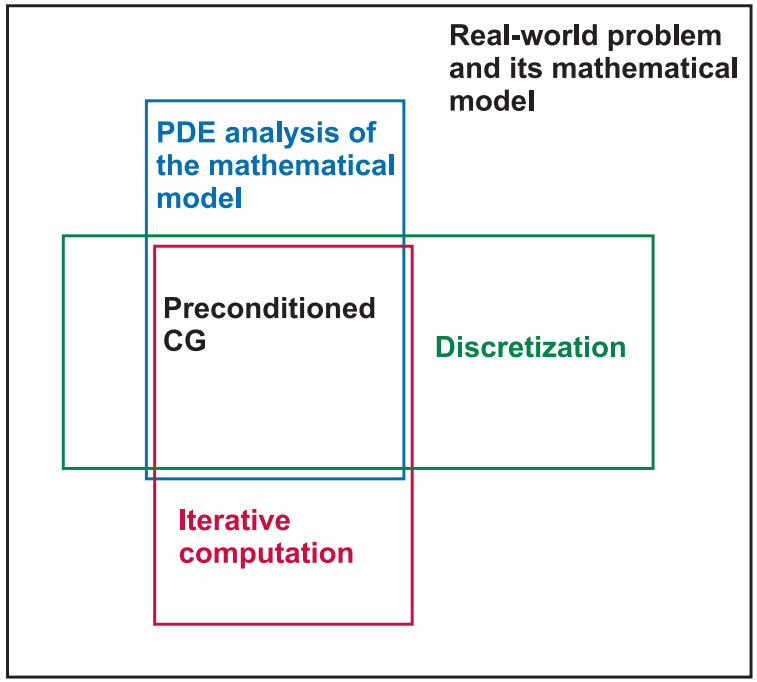
1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

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





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.

9 Lanczos, Why Mathematics (1966)

"We will go on pondering and meditating, the great mysteries still ahead of us, we will err and stumble on the way, and if we win a little victory, we will be jubilant and thankful, without claiming, however, that we have done something that can eliminate the contribution of all the millenia before us."

Thank you very much for your kind patience!

Outline

1. Infinite dimensional problems and finite dimensional computations
2. Krylov subspace methods: Hestenes, Stiefel, Lanczos (1950-52)
3. Problem of moments: Stieltjes (1894)
4. Projections onto highly nonlinear Krylov subspaces
5. Model reduction and moment matching
6. Convergence and spectral information
7. Inexact computations and numerical stability
8. Mathematical mythology
9. Optimistic outlook

Appendix: Operator preconditioning, discretization and algebraic computation

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."*

A Mesh independent condition number

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."

A 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."

A Literature on operator preconditioning

Gunn, D'yakonov, Faber, Manteuffel, Parter, Klawonn, Arnold, Falk, Winther, Axelsson, Karátson, Hiptmair, Vassilevski, Neytcheva, Notay, Elmann, Silvester, Wathen, Zulehner, Simoncini, Oswald, Griebel, Rüde, Steinbach, Wohlmuth, Bramble, Pasciak, Xu, Nyepomnyaschkikh, Dahmen, Kunoth, Yserentant, Mardal, Nordbotten, ...

Details, proofs and (incomplete) references can be found in

- J. Málek and Z.S., *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlight Series, SIAM (2015)
- J. Hrnčíř, I. Pultarová, Z.S., *Decomposition into subspaces and operator preconditioning* (submitted 2017)

A Basic setting

Hilbert space V with the inner product

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

dual space $V^\#$ of bounded linear functionals on V with the duality pairing and the associated Riesz map

$$\langle\cdot,\cdot\rangle : V^\# \times V \to \mathbb{R}\,, \quad \tau : V^\# \to V \quad \text{such that} \quad (\tau f, v)_V := \langle f, v\rangle \quad \text{for all } v \in V.$$

Equation in the functional space $V^\#$

$$\mathcal{A}u \;=\; b$$

with a linear, bounded, coercive, and self-adjoint operator

$$\mathcal{A} : V \to V^\#\,, \quad a(u,v) := \langle \mathcal{A}u, v\rangle\,,$$

$$\begin{aligned} C_{\mathcal{A}} \quad &:= \quad \sup_{v\in V,\,\|v\|_V=1} \|\mathcal{A}v\|_{V^\#} < \infty\,, \\ c_{\mathcal{A}} \quad &:= \quad \inf_{v\in V,\,\|v\|_V=1} \langle \mathcal{A}v, v\rangle > 0\,. \end{aligned}$$

A Operator preconditioning

Linear, bounded, coercive, and self-adjoint $\mathcal{B}$, $C_{\mathcal{B}}$, $c_{\mathcal{B}}$ defined analogously. Define

$$(\cdot,\cdot)_{\mathcal{B}} : V \times V \to \mathbb{R}, \qquad (w,v)_{\mathcal{B}} := \langle \mathcal{B}w, v\rangle \qquad \text{for all } w, v \in V\,,$$

$$\tau_{\mathcal{B}} : V^{\#} \to V, \qquad (\tau_{\mathcal{B}} f, v)_{\mathcal{B}} \;:=\; \langle f, v\rangle \qquad \text{for all } f \in V^{\#},\; v \in V\,.$$

Instead of the equation in the functional space $V^{\#}$

$$\mathcal{A}u \;=\; b$$

we solve the equation in the solution space V

$$\tau_{\mathcal{B}}\,\mathcal{A}\,u \;=\; \tau_{\mathcal{B}}\,b\,,$$

i.e.

$$\mathcal{B}^{-1}\,\mathcal{A}\,u \;=\; \mathcal{B}^{-1}b.$$

Theorem (norm equivalence and condition number)

Assuming that the linear, bounded, coercive and self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ are $V^{\#}$-*norm equivalent* on V, i.e. there exist $0 < \alpha \leq \beta < \infty$ such that

$$\alpha \leq \frac{\|\mathcal{A}w\|_{V^{\#}}}{\|\mathcal{B}w\|_{V^{\#}}} \leq \beta, \quad \text{for all } w \in V,\, w \neq 0\,.$$

Then

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) := \|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)}\|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)} \leq \frac{\beta}{\alpha}\,.$$

Theorem (spectral equivalence and spectral number)

Assuming that the linear, bounded, coercive and self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ are *spectrally equivalent* on V, i.e. there exist $0 < \gamma \leq \delta < \infty$ such that

$$\gamma \leq \frac{\langle \mathcal{A}w, w\rangle}{\langle \mathcal{B}w, w\rangle} \leq \delta, \quad \text{for all } w \in V,\, w \neq 0\,.$$

Then

$$\hat{\kappa}(\mathcal{A},\mathcal{B}) := \frac{\sup_{z\in V,\,\|z\|_V=1} \left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}\,(\tau\mathcal{B})^{-1/2}z, z\right)_V}{\inf_{v\in V,\,\|v\|_V=1} \left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}\,(\tau\mathcal{B})^{-1/2}v, v\right)_V} \leq \frac{\delta}{\gamma}\,.$$

A Galerkin discretization

N-dimensional subspace $V_h \subset V$; abstract Galerkin discretization gives $u_h \in V_h$, $u_h \approx u \in V$ satisfying Galerkin orthogonality

$$\langle \mathcal{A} u_h - b, v \rangle \;=\; 0 \qquad \text{for all } v \in V_h \,.$$

Restrictions $\mathcal{A}_h : V_h \to V_h^{\#}$, $b_h : V_h \to \mathbb{R}$ give the problem in $V_h^{\#}$

$$\mathcal{A}_h u_h \;=\; b_h, \qquad u_h \in V_h, \quad b_h \in V_h^{\#} \,.$$

With the inner product $(\cdot,\cdot)_{\mathcal{B}}$ and the associated restricted Riesz map

$$\tau_{\mathcal{B},h} : V_h^{\#} \to V_h$$

we get the abstract form of the preconditioned discretized problem in V_h

$$\tau_{\mathcal{B},h}\, \mathcal{A}_h\, u_h \;=\; \tau_{\mathcal{B},h}\, b_h \,.$$

A Preconditioning - straight consequence of the $V_h \longrightarrow V_h^{\#}$ setting

Using the discretization basis $\Phi_h = (\phi_1, \ldots, \phi_N)$ of V_h
and the canonical dual basis $\Phi_h^{\#} = (\phi_1^{\#}, \ldots, \phi_N^{\#})$ of $V_h^{\#}$, $(\Phi_h^{\#})^* \Phi_h = \mathbf{I}_N$,

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

where

$$\begin{aligned}
\mathbf{A}_h,\ \mathbf{M}_h \;&\in\; \mathbb{R}^{N\times N}, \quad \mathbf{x}_h, \mathbf{b}_h \in \mathbb{R}^N, \\
(\mathbf{x}_h)_i \;&=\; \langle \phi_i^{\#}, u_h \rangle, \quad (\mathbf{b}_h)_i \;=\; \langle b, \phi_i \rangle, \\
\mathbf{A}_h \;&=\; (a(\phi_j, \phi_i))_{i,j=1,\ldots,N} \;=\; (\langle \mathcal{A}\phi_j, \phi_i \rangle)_{i,j=1,\ldots,N}, \\
\mathbf{M}_h \;&=\; (\langle \mathcal{B}\phi_j, \phi_i \rangle)_{i,j=1,\ldots,N},
\end{aligned}$$

or

$$\mathbf{A}_h \;=\; (\mathcal{A}\Phi_h)^* \Phi_h, \qquad \mathbf{M}_h \;=\; (\mathcal{B}\Phi_h)^* \Phi_h\,.$$

A Matrix representation – symmetric form

Indeed, for the restricted Riesz map $\tau_{\mathcal{B},h}$ for $\mathbf{v}$ and $\mathbf{f}$, with $f = \Phi_h^{\#}\mathbf{f}$, $v = \Phi_h \mathbf{v}$

$$(\tau_{\mathcal{B},h} f, v)_{\mathcal{B}} = (\tau_{\mathcal{B},h}\Phi_h^{\#}\mathbf{f}, \Phi_h\mathbf{v})_{\mathcal{B}} = (\Phi_h \mathbf{M}_\tau \mathbf{f}, \Phi_h \mathbf{v})_{\mathcal{B}} = \langle \mathcal{B}\Phi_h \mathbf{M}_\tau \mathbf{f}, \Phi_h \mathbf{v}\rangle = \mathbf{v}^* \mathbf{M}_h \mathbf{M}_\tau \mathbf{f},$$
$$(\tau_{\mathcal{B},h} f, v)_{\mathcal{B}} = \langle f, v\rangle = \mathbf{v}^* \mathbf{f}$$

and therefore

$$\mathbf{M}_\tau \;=\; \mathbf{M}_h^{-1}\,.$$

Using (an arbitrary) decomposition $\mathbf{M}_h \;=\; \mathbf{L}_h \mathbf{L}_h^*$, the resulting preconditioned algebraic system can be transformed into

$$\left(\mathbf{L_h^{-1}} \mathbf{A}_h \mathbf{L_h^{*\,-1}}\right)\left(\mathbf{L_h^*} \mathbf{x}_h\right) \;=\; \mathbf{L}_h^{-1}\mathbf{b}_h\;,$$

i.e.,

$$\mathbf{A}_{t,h}\,\mathbf{x}_h^t \;=\; \mathbf{b}_h^t\;.$$

A Transformation of discretization basis and of the inner product

$$\Phi_h \to \tilde{\Phi}_{t,h} \quad \text{such that} \quad \mathbf{M}_{t,h} = (\mathcal{B}\tilde{\Phi}_{t,h})^* \tilde{\Phi}_{t,h} = \mathbf{I}\,,$$

i.e. orthogonalization of the basis with respect to the inner product $(\cdot,\cdot)_{\mathcal{B}}$,

$$\tilde{\Phi}_{t,h} = \Phi_h \mathbf{M_h}^{-1/2}, \quad \tilde{\Phi}^{\#}_{t,h} = \Phi^{\#}_h \mathbf{M_h}^{1/2}$$

gives immediately the preconditioned system $\tilde{\mathbf{A}}_{t,h}\,\tilde{\mathbf{x}}^t_h = \tilde{\mathbf{b}}^t_h$ with the reference choice $\mathbf{L_h} := \mathbf{M_h}^{1/2}$. Any other choice

$$\Phi_{t,h} = \Phi_h \mathbf{L_h}^{*\,-1}, \quad \Phi^{\#}_{t,h} = \Phi^{\#}_h \mathbf{L_h}$$

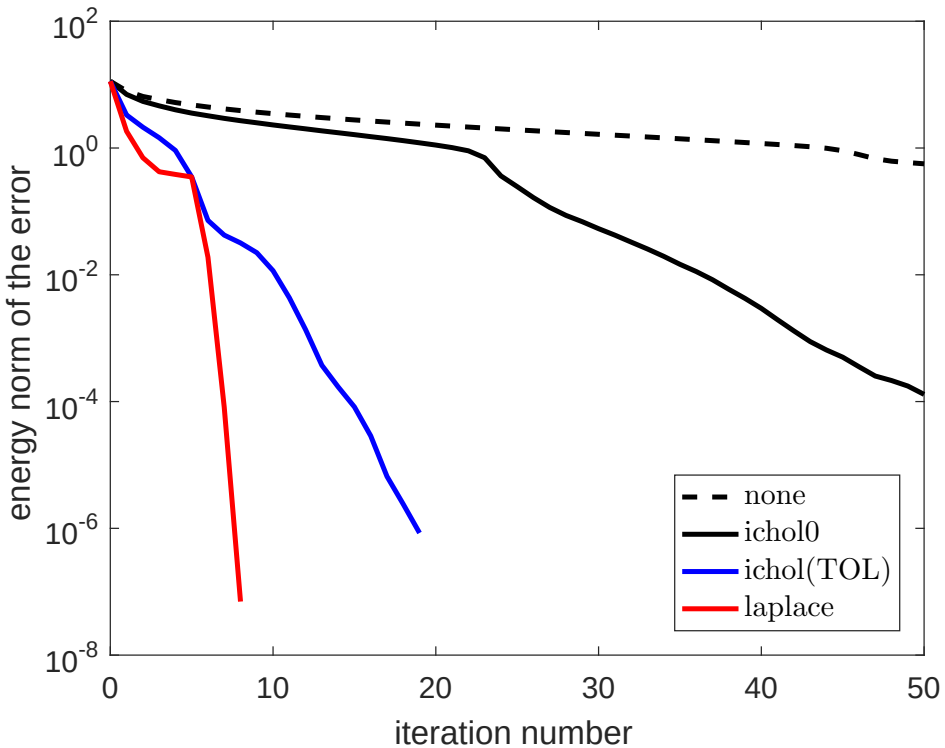
is given via an associated orthogonal transformation $\mathbf{Q}$,

$$\Phi_{t,h} = \tilde{\Phi}_{t,h}\mathbf{Q}^*\,, \quad \mathbf{Q}^* = \mathbf{M_h}^{1/2}\mathbf{L_h}^{*\,-1}\,.$$

A Points that are worth noticing

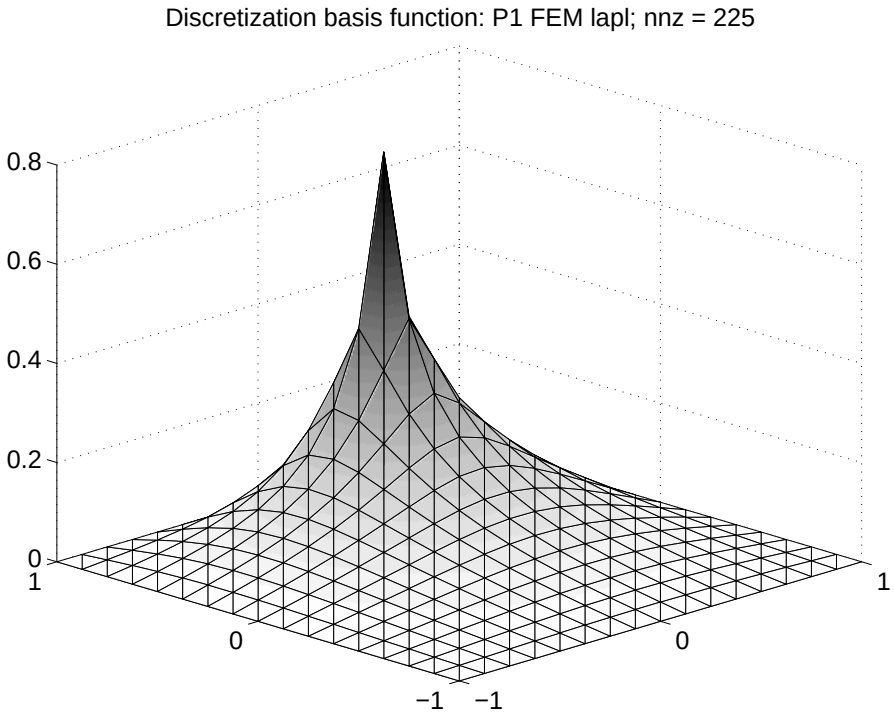
- Transformation of the discretization basis (preconditioning) is different from a change of the algebraic basis (similarity transformation).
- Any algebraic preconditioning can be put into the operator preconditioning framework by transformation of the discretization basis and the associated change of the inner product in the infinite dimensional Hilbert space V.

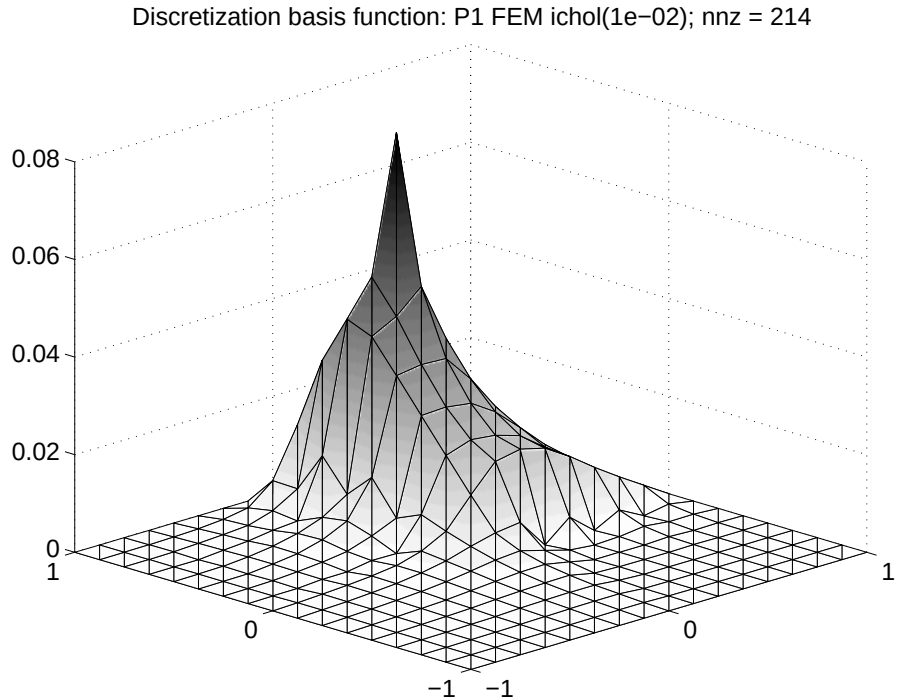
A Better conditioning does not necessarily mean faster convergence!



Nonhomogeneous diffusion tensor, uniform mesh. Unpreconditioned CG; ichol PCG (no fill-in); ichol PCG (drop-off tolerance 1e-02); Laplace operator PCG.
Condition numbers of $\mathbf{A}_{t,h}$: 6.75e04, 4.31e02, 1.6e01, 1.61e02.

A Transformed FEM nodal basis elements have global support





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

Theorem (norm equivalence and condition number)

Let the linear, bounded, coercive and self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ from V to $V^\#$ be $V^\#$-norm equivalent with the lower and upper bounds α and β, respectively, i.e.

$$\alpha \leq \frac{\|\mathcal{A}w\|_{V^\#}}{\|\mathcal{B}w\|_{V^\#}} \leq \beta \quad \text{for all } w \in V\,,\ w \neq 0, \quad 0 < \alpha \leq \beta < \infty\,.$$

Let $\mathbf{S}_h$ be the Gram matrix of the discretization basis $\Phi_h = (\phi_1, \ldots, \phi_N)$ of $V_h \subset V$, with $(\Phi_h^\#)^* \Phi_h = I$,

$$(\mathbf{S}_h)_{ij} = (\phi_i, \phi_j)_V\,.$$

Then the condition number of the matrix $\mathbf{M}_h^{-1}\mathbf{A}_h$ is bounded as

$$\kappa(\mathbf{M}_h^{-1}\mathbf{A}_h) := \|\mathbf{M}_h^{-1}\mathbf{A}_h\|\,\|\mathbf{A}_h^{-1}\mathbf{M}_h\| \leq \frac{\beta}{\alpha}\,\kappa(\mathbf{S}_h)\,.$$

Theorem (Spectral equivalence and spectral number)

Let the linear, bounded, coercive and self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ be spectrally equivalent with the lower and upper bounds γ and δ respectively, i.e.

$$\gamma \leq \frac{\langle \mathcal{A}w, w\rangle}{\langle \mathcal{B}w, w\rangle} \leq \delta \quad \text{for all } w \in V\,, \quad 0 < \gamma \leq \delta < \infty\,.$$

Then the spectral number $\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h)\,,$ which is equal to the condition number of the matrix $\mathbf{A}_{t,h} = \mathbf{L}_h^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}$ for any $\mathbf{L}_h$ such that $\mathbf{M}_h = \mathbf{L}_h\mathbf{L}_h^*$, is bounded as

$$\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) := \frac{\sup_{\mathbf{z}\in\mathbb{R}^N,\, \|\mathbf{z}\|=1} \left(\mathbf{M}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}\mathbf{z}, \mathbf{z}\right)}{\inf_{\mathbf{v}\in\mathbb{R}^N,\, \|\mathbf{v}\|=1} \left(\mathbf{M}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}\mathbf{v}, \mathbf{v}\right)} = \kappa(\mathbf{A}_{t,h}) \leq \frac{\delta}{\gamma}\,.$$