The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2044-D3613

Summer School and Advanced Workshop on Trends and Developments in Linear Algebra

22 June - 10 July, 2009

Metodi Iterativi - Iterative methods

Strada Costiera 11, 34014 Trieste, Italy - Tel. +39 040 2240 111; Fax +39 040 224 163 - sci_info@ictp.it, www.ictp.it

V. SIMONCINI

References:

"Iterative Methods for Sparse Linear Systems", Y. Saad, 2nd ed. SIAM 2003
"Iterative Methods for Solving Linear Systems", A. Greenbaum, SIAM 1997

METODI ITERATIVI - ITERATIVE METHODS

Problem: Solve $Ax = b$ with $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ known

$n >> 1000$ (e.g. $n = 10^8$)

Usually, but not always, A has a "sparsity" pattern, and it is sparse (= the # of nonzero entries per row is a low percentage of the total).

We look for a sequence $\{x_k^{(0)}, x_k^{(1)}, \dots x_k^{(n)}, \dots\}$, for a given $x_k^{(0)}$, that tends to approximate x, possibly at a reasonable cost. Moreover, we will content ourselves with a "good" approximation. In theory, $x_k^{(n)}$ is a "good" approx when $x_k^{(n)} \approx x$, that is $\|x - x_k^{(n)}\|$ small in some norm. In practice, since x is not known, instead of the error we shall look at the residual

$$r_k^{(n)} := b - A x_k^{(n)}$$

"how well the equation is satisfied" — "quality"

But first, let us consider a toy problem emphasizing where $Ax = b$ may come from:

Ex1
$$\begin{cases} -u'' = f(t) & 2^{\text{nd}} \text{ order, 1D} \\ u(0) = 0 & t \in [0,1] \\ u'(1) = 0 \end{cases} \qquad t_0 = 0 \dots t_i = ih \qquad t_{n+1} = 1 \qquad h = \frac{1.0}{n}$$

$$u''(t_i) \approx \frac{u(t_{i-1}) - 2u(t_i) + u(t_{i+1})}{h^2} \quad \left(\approx \frac{u'(t_{i+\frac{1}{2}}) - u'(t_{i-\frac{1}{2}})}{h} \right)$$

$x_i = u(t_i)$ i-th component of a vector x, $i = 1 \dots n$

$\forall\, t_i$ $i = 1 \dots n$ $\quad -u(t_{i-1}) - 2u(t_i) + u(t_{i+1}) = h^2 f(t_i)$

$$\left[-1, \; 2, \; -1\right] \begin{bmatrix} x_{i-1} \\ x_i \\ x_{i+1} \end{bmatrix} = h^2 f(t_i) \qquad i = n \quad -u(t_{n-1}) + 2u(t_n) - u(t_{n+1}) = 0$$

$$Ax \equiv \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} x = b \qquad b_i = h^2 f(t_i)$$

A tridiagonal (sparsity pattern)

Back to the general iterative procedure

KATHOLIEKE UNIVERSITEIT LEUVEN

Sequence $\{x^{(0)}, x^{(1)}, \dots, x^{(k)}, \dots\}$, $\lim_{k\to\infty} x^{(k)} = x$? Characterization:

- Convergence rate, in terms of iterations, in some cases, depending on h = mesh size
- Computational cost (floating point operations)
- Memory requirements

After k iterations, we can define $r^{(k)} := b - Ax^{(k)}$ $e^{(k)} = x - x^{(k)}$ error

A-priori estimates : estimate of convergence rate in terms of error
A-posteriori " : " " " " residuals
Relative estimates : $\frac{\|e^{(k)}\|}{\|x\|}$ $\frac{\|r^{(k)}\|}{\|b\|}$ in some norm

Remark : A small residual norm does not necessarily imply a small error norm

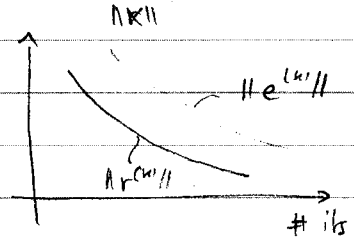
X HW Show that

$$\frac{\|e^{(k)}\|}{\|x\|} \le \kappa(A) \frac{\|r^{(k)}\|}{\|b\|} \qquad r^{(k)} = A e^{(k)} \qquad \kappa(A) := \|A\| \, \|A^{-1}\|$$

$\|\cdot\|$ induced

$\left(r^{(k)} = A e^{(k)}, \quad \frac{\|e^{(k)}\|}{\|x\|} \le \|A^{-1}\| \, \|r^{(k)}\| \, \frac{\|A\|}{\|b\|} \right)$

Typically,



In some cases, replaces $\|r^{(k)}\|$ with the "backward error" :

$$\frac{\|r^{(k)}\|}{\|A\| \, \|x^{(k)}\| + \|b\|}$$

Remark : how small is small ? $\|r^{(k)}\| \le$? tol $\quad tol = 10^{-16}$?
If the pb comes from a discretization (see previous example) then $\|r^{(k)}\| = O(h^p)$ where h^p is discretization error.

We start next with some recurrences. We shall see that properties of the methods may change (even dramatically !) depending on whether A is Hermitian, or normal.

Start with $Ax = b$ and assume you have $M \approx A$, in some sense. Then, given $x^{(k)}$

$$M^{-1} r^{(k)} \approx e^{(k)} \qquad \text{therefore}$$

$$x^{(k+1)} := x^{(k)} + M^{-1} r^{(k)}$$

which we expect to be a better approximation to x.

KATHOLIEKE UNIVERSITEIT LEUVEN

Recurrence : Simple iteration

$M = \text{diag}(A)$ = Jacobi iteration $\quad x^{(0)}$ given, $x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$

$M = \text{tril}(A)$ = Gauss-Seidel iteration

$M = \frac{1}{\omega} D - L$ ($D = \text{diag}(A)$, L strictly $\text{tril}(A)$) = successive over relaxation SOR method

Solving with M should be cheap !

$$e^{(k+1)} = (I - M^{-1}A)\, e^{(k)} = \dots = (I - M^{-1}A)^{k+1} e^{(0)} \Rightarrow \|e^{(k)}\| \le \|(I - M^{-1}A)^k\| \, \|e^{(0)}\|$$

(induced; sharp bound !)

LEMMA : $\|e^{(k)}\| \xrightarrow[k\to\infty]{} 0 \ \forall\, e^{(0)}$ iff $\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = 0$

Remark : $\|(I - M^{-1}A)^k\|^{1/k} \xrightarrow[k\to\infty]{} \rho(I - M^{-1}A)$ = spectral radius $\quad \rho(H) = \{\max |\lambda| : \lambda \in \sigma(H)\}$

THEOREM $x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$ converges to $x = A^{-1}b \ \forall\, e^{(0)}$ iff $\rho(I - M^{-1}A) < 1$

Rate of convergence ? $\quad \|e^{(k)}\| \le \|I - M^{-1}A\| \, \|e^{(0)}\|$

If $\|I - M^{-1}A\| < 1$, then the error is reduced at least by such factor at each iteration.

Therefore, $\dfrac{\|e^{(k)}\|}{\|e^{(0)}\|} \le \delta$ for $k \ge \dfrac{\ln \delta}{\ln \|I - M^{-1}A\|}$ HW

So, in this case, convergence is linear.

If $\rho(I - M^{-1}A) < 1$, and there $\exists$ norm s.t. $\|I - M^{-1}A\| < \rho(I - M^{-1}A) + \varepsilon$, then the whole norm reduction is approx $\rho(I - M^{-1}A)$. How to detect such a norm ? This is difficult !

Sometimes the norm given here is a strange one. We would like to work with the 2-norm or ∞-norm.

NOTE : for A normal $\rho(A) = \|A\|$ so in this case we know the reduction factor.

For non normal, we may have $\rho(I - M^{-1}A) < 1 < \|I - M^{-1}A\|$ 2-norm

and it is difficult to predict the number of its to achieve a tolerance, by only using $\rho(I - M^{-1}A)$.

EX : $A = \begin{pmatrix} 1 & -1.16 & \\ .16 & 1 & -1.16 \\ & .16 & 1 \end{pmatrix}$ $\quad M = \text{tril}(A)$ $\quad \rho(I - M^{-1}A) \approx 0.74$ Fig

Problems with early convergence history, but $\rho(\cdot)$ describes well the asymptotics

Fig.

KATHOLIEKE UNIVERSITEIT LEUVEN

Iterative methods for $Ax = b$ (Simultaneous displacements)

IMPROVEMENT: Introduce a parameter α_k: $x^{(k+1)} = x^{(k)} + \alpha_k (b - A x^{(k)})$

α_k chosen as $r^{(k+1)} = r^{(k)} - \alpha_k A r^{(k)}$

$$\alpha_k = \arg \min_{\alpha} \| r^{(k)} - \alpha A r^{(k)} \|_2 \qquad (*)$$

we find $\alpha_k = \dfrac{\langle r^{(k)}, A r^{(k)} \rangle}{\langle A r^{(k)}, A r^{(k)} \rangle}$ HW ($f(\alpha) = \| r - \alpha A r \|^2$ $f'(\alpha) = 0 \dots$)

minimizes the residual! (example of a class of methods),

If A is spd then we can minimize $\| e^{(k+1)} \|_A := \langle e^{(k+1)}, A e^{(k+1)} \rangle^{1/2}$:

$$e^{(k+1)} = e^{(k)} - \alpha_k r^{(k)} \Rightarrow \alpha_k = \frac{\langle e^{(k)}, A r^{(k)} \rangle}{\langle r^{(k)}, A r^{(k)} \rangle} = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k)}, A r^{(k)} \rangle}$$

("method of steepest descent": solving the linear system is equivalent to
min $x^T A x - 2 b^T x$, whose min is at $x = A^{-1} b$
negative gradient (or direction of steepest descent) at $x = x^{(k)}$ is $r^{(k)}$)

Going back to (*): $r^{(k+1)}$ is obtained by eliminating from $r^{(k)}$ its projection onto $A r^{(k)}$

$\Rightarrow \| r^{(k+1)} \| \leq \| r^{(k)} \|$

($\| r^{(k+1)} \| = \| r^{(k)} \|$ only if $r^{(k)} \perp A r^{(k)}$, that is, $\langle r^{(k)}, A r^{(k)} \rangle = 0$)

THEOREM $\| r^{(k)} \|$ ↘ strictly $\forall r^{(0)}$ iff $0 \notin F(A)$ (A real)

($\| r^{(k+1)} \| < \| r^{(k)} \|$ if $\langle r^{(k)}, A r^{(k)} \rangle \neq 0$

Convergence rate:

THEOREM (*) Converges to $A^{-1} b$ $\forall r^{(0)}$ iff $0 \notin F(A)$. In this case,

$$\| r^{(k+1)} \|_2 \leq \sqrt{1 - \frac{d^2}{\|A\|^2}} \, \| r^{(k)} \|_2 \qquad \forall k \qquad (**)$$

where $d = d(0, F(A))$.

Estimate is not sharp in general.

Proof. $\| r^{(k+1)} \|^2 = \| r^{(k)} \|^2 - \dfrac{\langle r^{(k)}, A r^{(k)} \rangle^2}{\langle A r^{(k)}, A r^{(k)} \rangle} = \| r^{(k)} \|^2 \left(1 - \left| \dfrac{\langle r^{(k)}, A r^{(k)} \rangle}{\langle r^{(k)}, r^{(k)} \rangle} \right|^2 \dfrac{\| r^{(k)} \|^2}{\langle A r^{(k)}, A r^{(k)} \rangle} \right)$

($\geq d^2$, $\leq \|A\|^2$)

$\leq \| r^{(k)} \|^2 \left(1 - \dfrac{d^2}{\dots} \right)$ #

KATHOLIEKE UNIVERSITEIT LEUVEN

the proof shows why the estimate may be non-sharp -

this is idea behind the proof of a famous convergence estimate for GMRES -

For A spd, noting that

$$\|r^{(k+1)}\| \le \|I - \alpha A\| \cdot \|r^{(k)}\| \quad \forall \alpha$$

we take $\alpha = \frac{2}{\lambda_1 + \lambda_n}$ $\quad \lambda_1 = \min \sigma(A)$, $\lambda_n = \max \sigma(A)$

$$\Rightarrow \|r^{(k+1)}\| \le \max_i \left| 1 - \frac{2\lambda_i}{\lambda_1 + \lambda_n} \right| \|r^{(k)}\|$$

$$\le \left| \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right| \|r^{(k)}\| \quad (\lambda_i \ge \lambda_1) \quad \boxplus$$

$$\le \left| \frac{k-1}{k+1} \right| \quad k \text{ second } (*)$$

Note, the bound in (KK) would give $\|r^{(k+1)}\| \le \sqrt{1 - \frac{1}{k^2}} \, \|r^{(k)}\|$

$$\approx \left(1 - \frac{1}{2}\frac{1}{k^2}\right) \|r^{(k)}\| \quad \text{weaker}$$

A bound similar to $\boxplus$ also for $\|e^{(k+1)}\|_A$ for A spd.

For A nonsymmetric, but with $F(A) \subset D = \{ |z - c| \le s \}$ with $0 \notin D$,

let $\alpha = \frac{1}{c}$. then

$$\|I - \alpha A\| \le \frac{s}{|c|} \qquad \text{HW} \left(\text{we have } F\left(I - \tfrac{1}{c}A\right) \subseteq \left\{ |z| \le \tfrac{s}{c} \right\} \right)$$

$$\|r^{(k+1)}\| \le \frac{2s}{|c|} \|r^{(k)}\|$$

which may be better, or a bound, than that in the theorem, depending on s and c.

Remark: ORTHOMIN(1) can be generalized to other norms $\|r^{(k+1)}\|_D$ different from A-norm

CONJUGATE GRADIENT METHOD (CG)

Assume we write now $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ (earlier we had $p^{(k)} = r^{(k)}$)

then we have

$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)} \qquad e^{(k+1)} = e^{(k)} - \alpha_k p^{(k)}$$

in addition to $r^{(k+1)} \perp Ap^{(k)}$, we subtract

the projection of $r^{(k)}$ in the direction of $Ar^{(k)}$ but orthogonal to $Ap^{(k-1)}$ by defining

$$p^{(k)} = r^{(k)} - \underbrace{\frac{\langle Ar^{(k)}, Ap^{(k-1)} \rangle}{\langle Ap^{(k-1)}, Ap^{(k-1)} \rangle}}_{\beta_{k-1}} p^{(k-1)} \quad \Rightarrow \quad \langle r^{(k-1)}, Ap^{(k-1)} \rangle = 0$$

$$\text{and } \langle r^{(k+1)}, Ap^{(k-1)} \rangle = 0$$

$$\Rightarrow \quad r^{(k+1)} = r^{(k)} - \alpha_k Ar^{(k)} + \alpha_k \beta_{k-1} Ap^{(k-1)} \quad , \quad r^{(k+1)} \perp \text{span}\{ Ar^{(k)}, Ap^{(k-1)} \}$$

if $0 \notin F(A)$ iteration does not fail $\qquad$ minimization of the residual

If A is symmetric, $r^{(k+1)}$ is minimized over a larger space:

THEOREM: A sym and $r^{(1)} \dots r^{(k+1)}$, $p^{(1)} \dots p^{(k+1)}$ are well defined. Then

$$\langle r^{(k+1)}, Ap^{(j)} \rangle = 0$$
$$\langle Ap^{(k+1)}, Ap^{(j)} \rangle = 0 \qquad \forall j \leq k$$

All vectors in

$$r^{(0)} + \text{span}\{Ar^{(0)}, A^2 r^{(0)}, \dots, A^{k+1} r^{(0)}\}$$

$r^{(k+1)}$ has the smallest 2-norm.

In particular, if all vectors are defined till $k = n$, then $r^{(n)} = 0$ ∎

MINRES provides a realization of this minimization, but with a lower implementation

For A spd, we want to eliminate the A-projection of the error in a direction that is already A-orthogonal to the previous direction vector, that is in the direction

$$p^{(k+1)} = r^{(k+1)} - \frac{\langle r^{(k+1)}, Ap^{(k-1)} \rangle}{\langle p^{(k-1)}, Ap^{(k-1)} \rangle} p^{(k-1)} \quad \leftarrow \hat{\beta}_{k-1}$$

so that we have $\langle e^{(k+1)}, Ap^{(k)} \rangle = \langle e^{(k+1)}, Ap^{(k-1)} \rangle = 0$

and $\|e^{(k+1)}\|_A$ is minimized over the affine space $e^{(k)} + \text{span}\{r^{(k)}, p^{(k-1)}\}$. This is CG

A result analogous to the previous theorem holds:

THEOREM: A spd. CG satisfies $\langle e^{(k+1)}, Ap^{(j)} \rangle = 0$

$$\langle p^{(k+1)}, Ap^{(j)} \rangle = 0 \qquad \forall j \leq k$$
$$\langle r^{(k+1)}, r^{(j)} \rangle = 0$$

and it follows $\|e^{(k+1)}\|_A$ smallest in $e_0 + \text{span}\{Ae^{(0)}, A^2 e^{(0)}, \dots, A^{k+1} e^{(0)}\}$

IMPLEMENTATION: Given $x^{(0)}$, $r^{(0)} = b - Ax^{(0)}$. Set $p^{(0)} = r^{(0)}$

For $k = 1, 2, \dots$

$$x^{(k)} = x^{(k-1)} + \alpha_{k-1} p^{(k-1)} \text{ with } \alpha_{k-1} = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle p^{(k-1)}, Ap^{(k-1)} \rangle}$$

$$r^{(k)} = r^{(k-1)} - \alpha_{k-1} Ap^{(k-1)}$$

$$p^{(k)} = r^{(k)} + \beta_{k-1} p^{(k-1)} \text{ with } \beta_{k-1} = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k-1)}, r^{(k-1)} \rangle}$$

Remarks: $\beta = \hat{\beta}$ and $\alpha = \hat{\alpha}$ ⟨r^(k+1), p^(k+1)⟩ / ... ...

this idea does not carry over to the case of A nonsym. Orthogonality and A-orthogonality would have to be enforced, at high computational cost.
We can still determine $x^{(k)}$ so that $r^{(k)}$ is minimized over span$\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\}$ without storing also the $Ap^{(k)}$ sequence, that is $x^{(k)} = \alpha_1 r^{(0)} + \alpha_2 Ar^{(0)} + \dots + \alpha_{k-1} A^{k-1} r^{(0)} = p(A) r^{(0)}$
A popular implementation of this idea is GMRES - this can be formulated in 3 steps

1. expand the space $\{r^{(0)}, Ar^{(0)}, \dots, A^{k-1} r^{(0)}\} \to \{r^{(0)}, \dots, A^{k-1} r^{(0)}, A^k r^{(0)}\}$
2. update an orthogonal basis for the new space
3. generate new approx. soln. $x^{(k)}$

we call

$K_k(A, r^{(0)}) = \text{span}\{r^{(0)}, \dots, A^{k-1} r^{(0)}\}$ Krylov subspace $\qquad V_k = [v^{(1)} \dots v^{(k)}]$ orth. basis for $K_k \qquad v^{(1)} = \frac{r^{(0)}}{\|r^{(0)}\|}$

steps 1. and 2. may be performed together by means of the Arnoldi method -
Given $\{v^{(1)}, \dots, v^{(k)}\}$ get $v^{(k+1)}$:

$$\hat{v} = A v^{(k)}$$

$$\tilde{v} = \hat{v} - \sum_{i=1}^{k} v^{(i)} \langle v^{(i)}, \hat{v} \rangle \quad \leftarrow \text{Gram-Schmidt}$$

$$v^{(k+1)} = \frac{\tilde{v}}{\|\tilde{v}\|}$$

Matrix form: set $h_{ik} = \langle v^{(i)}, \hat{v} \rangle \qquad h_{k+1,k} = \|\tilde{v}\| \qquad \underline{H}_k = (h_{ij}) \quad \begin{matrix} i = 1..k+1 \\ j = 1..k \end{matrix}$

$$(\boxplus) \qquad A V_k = V_{k+1} \underline{H}_k = V_k H_k + v^{(k+1)} h_{k+1,k} e_k^T \qquad e_k \text{ column of } I.$$

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix} \begin{pmatrix} | & & | \\ V_k & & \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ V_k & & \\ | & & | \end{pmatrix} \begin{pmatrix} & & \\ & H_k & \\ & & \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ | & & | \end{pmatrix}$$

GMRES step: $x^{(k)} \in x^{(0)} + K_k(A, r^{(0)}) \qquad x^{(k)} = V_k y^{(k)}$ for some $y^{(k)}$

$$y^{(k)} = \arg \min_y \|r^{(0)} - A V_k y\| \qquad \text{minimization of the residual}$$

$$= \arg \min_y \|V_{k+1} e_1 \|r^{(0)}\| - V_{k+1} \underline{H}_k y\| \qquad (\boxplus) \qquad r^{(k)} = b - A x^{(k)} = b - A x^{(0)} - A V_k y = r^{(0)} - A V_k y^{(k)}$$

$$= \arg \min_y \|e_1 \|r^{(0)}\| - \underline{H}_k y\|, \qquad \text{least squares pb of size } (k+1) \times k.$$

no implementation details, However
- H_k upper Hessenberg
- Least squares pb solved using QR decomposition with rotations (Q and R updated at each iteration
- $\|r^{(k)}\|$ computable without inverting R

Breakdown : the Arnoldi recurrence can continue as long as $Av^{(k)}$ is linearly independent wrt previous vectors $v^{(1)} \dots v^{(k)}$ - Otherwise, $\|\hat{v}\| = 0 = h_{k+1,k}$ - In this case
$A v^{(k)} \in \text{span}\{v^{(1)}, \dots, v^{(k)}\}$ and $AV_k = V_k H_k$ that is
V_k spans an invariant subspace of A - Since $\text{Range}(V_k) = k$, then $\text{Ran}(V_k) = k$!!
We call it a happy breakdown : the recurrence cannot continue but there is no need!
So, we can assume H_k to be full column rank (but not that H_k is nonsing)

Special properties : $H_k = V_k^T A V_k$ eigenvalues of the approximate of (A), especially the extreme ones.
$(\theta_j^{(k)}, z_j^{(k)})$ eigenpairs of H_k $\Rightarrow$ $\theta_j^{(k)}$ Ritz values, $V_k z_j^{(k)}$ Ritz vectors

NOTE : $u \in K_k(A, r^{(0)})$, $u = p_{k-1}(A) r^{(0)}$, p_{k-1} polynomials up to degree $k-1$
are exactly represented in $K_k(A, r^{(0)})$:

$$A r^{(0)} = A V_k e_1 \|r^{(0)}\| = V_k H_k e_1 \|r^{(0)}\| + v_{k+1} h_{k+1} e_k^T e_1 \|r^{(0)}\| \qquad k > 1$$
$$A^j r^{(0)} = \dots = V_k H_k^j e_1 \|r^{(0)}\| \qquad 1 \le j < k$$

From this it follows :

- Arnoldi iteration breaks down at step j iff the minimal polynomial of A wrt v_1 is of degree j (j = grade of v_1)
  - Hard ⇐ already seen - ⇒ $p_j(A) v_1 = 0$
    $A p_j(A) v_1 = 0 \Rightarrow \dots K_{j+1} \subseteq K_j$ invariant
    $\Rightarrow$ breakdown.
- If $\lambda \in \sigma(A)$ has geometric multiplicity greater than one (more than one eigvec. then only one copy of eigenvalue is approximated in $K_k(A, r^{(0)})$
  (in exact arithmetic)

KATHOLIEKE UNIVERSITEIT LEUVEN

CONVERGENCE : in Krylov subspace $K_k(A, r^{(0)})$

$r^{(k)} = r^{(0)} - A p_{k-1}(A) r^{(0)} = p_k(A) r^{(0)}$ with $p_k(0) = 1$ GMRES, MINRES

$e^{(k)} = q_k(A) e^{(0)}$ $q_k(0) = 1$ CG

p_k satisfies $\min_{\substack{p \in \mathbb{P}_k \\ p(0)=1}} \|p(A) r^{(0)}\|$ q_k satisfies $\min_{\substack{q \in \mathbb{P}_k \\ q(0)=1}} \|q(A) r^{(0)}\|_A$

Convergence estimates that do not depend on $x^{(0)}$:

A diagble $\Rightarrow$ $A = U \Lambda U^{-1}$ $r^{(k)} = p_k(A) r^{(0)} = U p_k(\Lambda) U^{-1} r^{(0)}$

$\Lambda = \mathrm{diag}(\lambda_1 \dots \lambda_n)$

$$\|r^{(k)}\| = \min_p \|p(A) r^{(0)}\| \le \|U\| \|U^{-1}\| \min_p \|p(\Lambda)\| \|r^{(0)}\|$$

A symmetric $\Rightarrow \|U\| = \|U^{-1}\| = 1$ $\Rightarrow$ $\dfrac{\|r^{(k)}\|}{\|r^{(0)}\|} \le \min_p \max_{\lambda_i} |p(\lambda_i)|$ MINRES

Aspd $\dfrac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A} \le \min_q \max_{\lambda_i} |q(\lambda_i)|$

Bound is sharp : $\exists\, e^{(0)}$ or $r^{(0)}$ s.t. equality is obtained.

the bound shows when spectrum of A is "good" or "bad" for convergence

If only limited information is available, we can still provide "worst case scenario"

THEOREM : Assume A spd. then CG satisfies

$$\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A} \le 2 \left(\frac{\sqrt{\mathrm{cond}(A)} - 1}{\sqrt{\mathrm{cond}(A)} + 1} \right)^k \qquad \mathrm{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

REMARK : Assume $\lambda_n \gg \lambda_i$ $i = 1 \dots n-1$ (λ_n much larger, "isolated")

we can take $q_k(\lambda) := \left(\dfrac{\lambda_n - \lambda}{\lambda_n}\right) S_{k-1}(\lambda)$

note that $\varphi_k(\lambda_n) = 0$ and $\left|\frac{\lambda_n - \lambda}{\lambda_n}\right| < 1$ for $\lambda = \lambda_i$ $i = 1 \dots n-1$

therefore $\max_{\lambda_i} |\varphi_k(\lambda_i)| < \max_{\lambda_i} |s_{k-1}(\lambda)|$

so that $\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A} \leq 2\left(\frac{\sqrt{\text{cond}(A)} - 1}{\sqrt{\text{cond}(A)} + 1}\right)^k$ $\qquad \widehat{\text{cond}}(A) = \frac{\lambda_{n-1}}{\lambda_1}$

the idea generalizes to a cluster.
A similar result holds for the 2-norm of MINRES residual (A symmetric)

Convergence of GMRES is much more treacherous, and still largely open pb, due to the non normality of A.

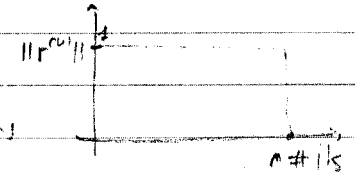
$$\frac{\|r^{(k)}\|}{\|r^{(0)}\|} \leq \text{cond}(U) \min_p \max_{\lambda_i} |p(\lambda_i)| \qquad \odot$$

is a good bound for U not ill-conditioned.
In general, the behavior of GMRES is not determined by the eigenvalues alone.

Example: A = companion matrix $A = \begin{pmatrix} & & & 0 \\ & & & \\ & & & \\ c_0 & c_1 & \cdots & c_{n-1} \end{pmatrix}$ $\qquad \sigma(A) = \text{roots}\left(t^n - \sum_{j=0}^{n-1} c_j t^j\right)$

$\sigma(A)$ any...

$r_0 = e_1 \xi \quad \xi \in \mathbb{R}$ multiple of first unit vector



no decrease of residual norm for the first n-2 iterations

For U ill-conditioned, then $\odot$ should not be used, not informative.
The convergence may still be good, it is just the bound that is not good.
For instance, if $0 \notin F(A)$ and $F(A) \subseteq D = \{|z - c| \leq s\}$
it can be shown that

$$\frac{\|r^{(k)}\|}{\|r^{(0)}\|} \leq 2\left(\frac{s}{|c|}\right)^k$$

(stronger than ORTHOMIN(1) bound because $\omega = 2^k$)

More sophisticated bounds may be obtained using the Cauchy integral, or also pseudo-spectrum recalling that

$$p(A) = \frac{1}{2\pi i} \int_\Gamma p(z)\,(zI - A)^{-1}\,dz$$

Γ = simple closed curve (or union thereof) enclosing $\sigma(A)$.

so that $\|p(A)\| \leq \frac{\mathcal{L}(\Gamma)}{2\pi} \max_{z \in \Gamma} |p(z)| \, \|(zI-A)^{-1}\|$.

$\mathcal{L}(\Gamma)$ length of Γ.

A more general perspective: Projection methods

$Ax = b$ $\quad x^{(k)} \in K$ approximation space

$\mathcal{L}$ trial space, constraints space

$r^{(k)} \perp \mathcal{L}$ Petrov-Galerkin condition $\Rightarrow$ as $\mathcal{L}$ expands, $r^{(k)}$ smaller...

- $\mathcal{L} = K \Rightarrow$ Galerkin condition $\Rightarrow$ orthogonal projection onto K

$x_k^{(k)} = V_k y$ $\quad$ Range $(V) = K$

$$V^T r^{(k)} = 0 \iff V^T (r^{(0)} - AVy) = 0$$

$$V^T A V y = V^T r^{(0)} \qquad y = \underbrace{(V^T A V)^{-1}}_{\text{working for A spd}} V^T r^{(0)}$$

$$r^{(k)} = r^{(0)} - AVy = r^{(0)} - AV(V^TAV)^{-1} V^T r^{(0)}$$

↳ CG in this class

Prop: $\|x - x^{(k)}\|_A = \min_{d \in x^{(0)} + K} \|x - d\|_A$

- $\mathcal{L} = AK$ $\quad (AV)^T r^{(k)} = 0$ $\quad V^T A^T r^{(0)} - V^T A^T A V y = 0$ $\quad$ oblique projector onto K orthogonal to $\mathcal{L}$

$$r^{(k)} = r^{(0)} - AV(V^TA^TAV)^{-1} V^T A^T r^{(0)}$$

↳ always working

↳ MINRES, GMRES in this class.

Prop: $\|b - Ax^{(k)}\|_2 = \min_{d \in x^{(0)} + K} \|b - Ad\|_2$

FOM (Full Orthogonalization method) $\mathcal{L} = K$ as in CG. A nonsing $\Rightarrow y = (V^TAV)^{-1} V^T r^{(0)}$

↑ no minimization (if A spd, FOM = CG with minimization)

For $K = K_k(A, r^{(0)})$, our already treated case, and A nonsing - Ortho

$$AV_k = V_{k+1} \underline{H}_k = V_k H_k + v^{(k+1)} h_{k+1,k} e_k^T \qquad H_k \text{ full rank}$$

we have:

GMRES: $x_G^{(k)} = x^{(0)} + V_k \underline{H}_k^+ e_1 \|r^{(0)}\|$ $\quad$ FOM: $x_F^{(k)} = x^{(0)} + V_k H_k^{-1} e_1 \|r^{(0)}\|$

(might not be defined!)

If they are both defined, it holds

$$\|r_k^F\| = \frac{\|r_k^G\|}{\sqrt{1 - \left(\frac{\|r_k^G\|}{\|r_{k-1}^G\|}\right)^2}}$$

Stagnation of GMRES at step k ($\|r_k^G\| = \|r_{k-1}^G\|$) corresponds to undefined FOM iterate - the same for quasi-stagnation $\rightsquigarrow$ so-called peak-and-plateau behavior.

KATHOLIEKE UNIVERSITEIT LEUVEN

Restarting

Both GMRES and FOM require storing V_k, and keep orthogonalizing the next vectors to maintain an orthogonal basis - High computational costs

Restarted procedure:
1. Select $x^{(0)}$,
2. Run k iterations of Arnoldi: $K_k(A, r_k^{(0)})$
3. Get FOM or GMRES solutions
4. If not converged, then $x^{(0)} \leftarrow x^{(k)}$ and restart from 2.

Crucial: How many iterations? (how to select k)
There is an unsolved issue, usually as large as possible (cope with the memory constraints)
GMRES is very popular - Is its restarted version trustable as well?

$$x_G^{(k)} = x^{(0)} + V_k H_k^{-1} e_1 \|r^{(0)}\| \quad \text{if } H_k^{-1} e_1 \approx 0$$

$$r_G^{(k)} = r^{(0)} - AV_k H_k^{-1} e_1 \|r^{(0)}\| = V_{k+1}\left(e_1 \|r^{(0)}\| - \underline{H}_k \underline{H}_k^{+} e_1 \|r^{(0)}\|\right)$$

$\Rightarrow r_G^{(k)} \approx r^{(0)}$ - Restarting will generate approximately the same space $\Rightarrow$ stagnation

On the other hand, FOM:

$$r_F^{(k)} = r^{(0)} - AV_k H_k^{-1} e_1 \|r^{(0)}\| = V_{k+1}\left(e_1 \|r^{(0)}\| - \underline{H}_k H_k^{-1} e_1 \|r^{(0)}\|\right)$$

$$= -v_{k+1} h_{k+1,k} e_k^T H_k^{-1} e_1 \|r^{(0)}\|$$

$$= -v_{k+1} h_{k+1,k} \left(e_k^T H_k^{-1} e_1 \|r^{(0)}\|\right)$$

$r_F^{(k)} \propto v_{k+1}$ next basis vector

If we restart with $r_F^{(k)}$, we continue building the space as if we had not restarted

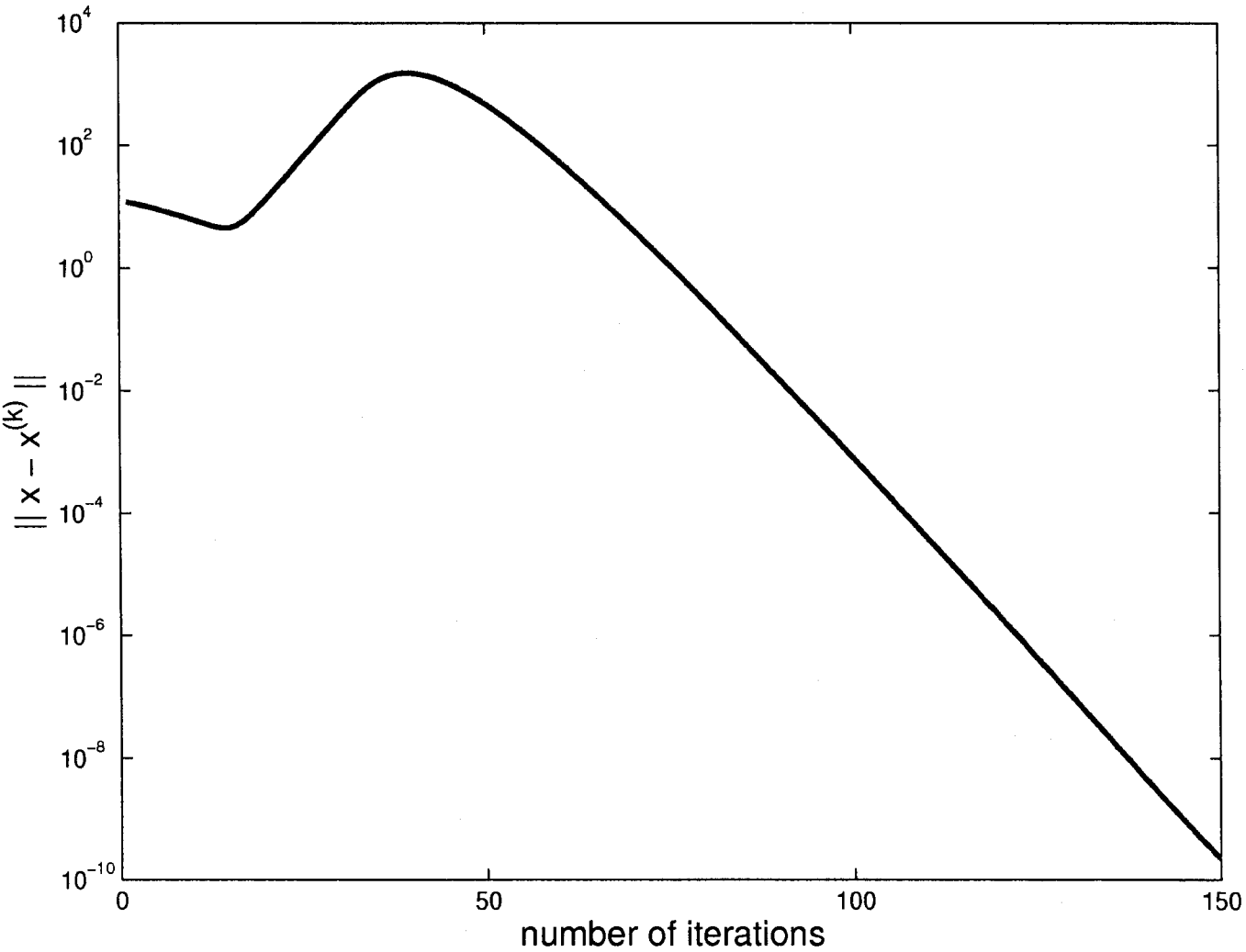
$$\underbrace{\{v^{(1)}\ v^{(2)} \cdots v^{(k)}\}}_{K_k(A, r^{(0)})} + \underbrace{\{v^{(k+1)}, v^{(k+2)}, \ldots\}}_{K_k(A, r_F^{(k)})}$$

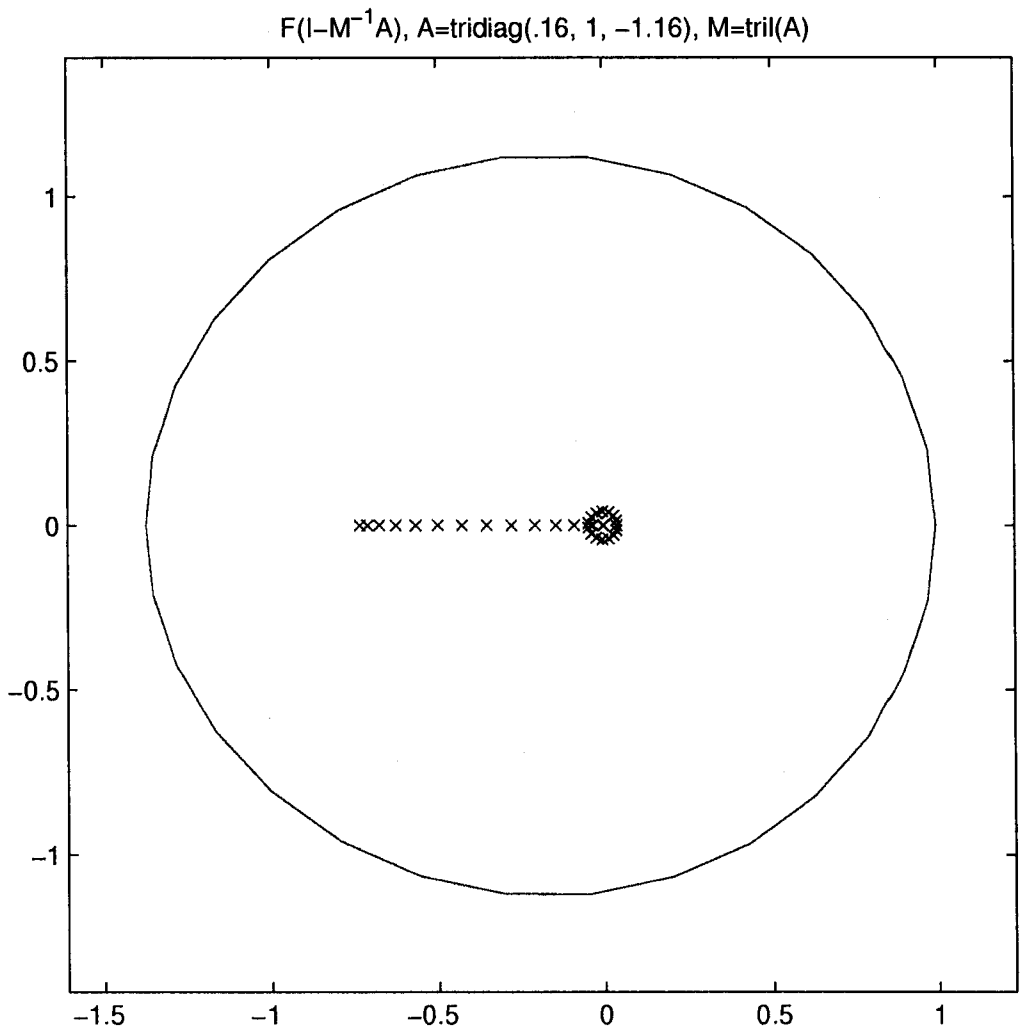
$r_F^{(k)}$: from first cycle.

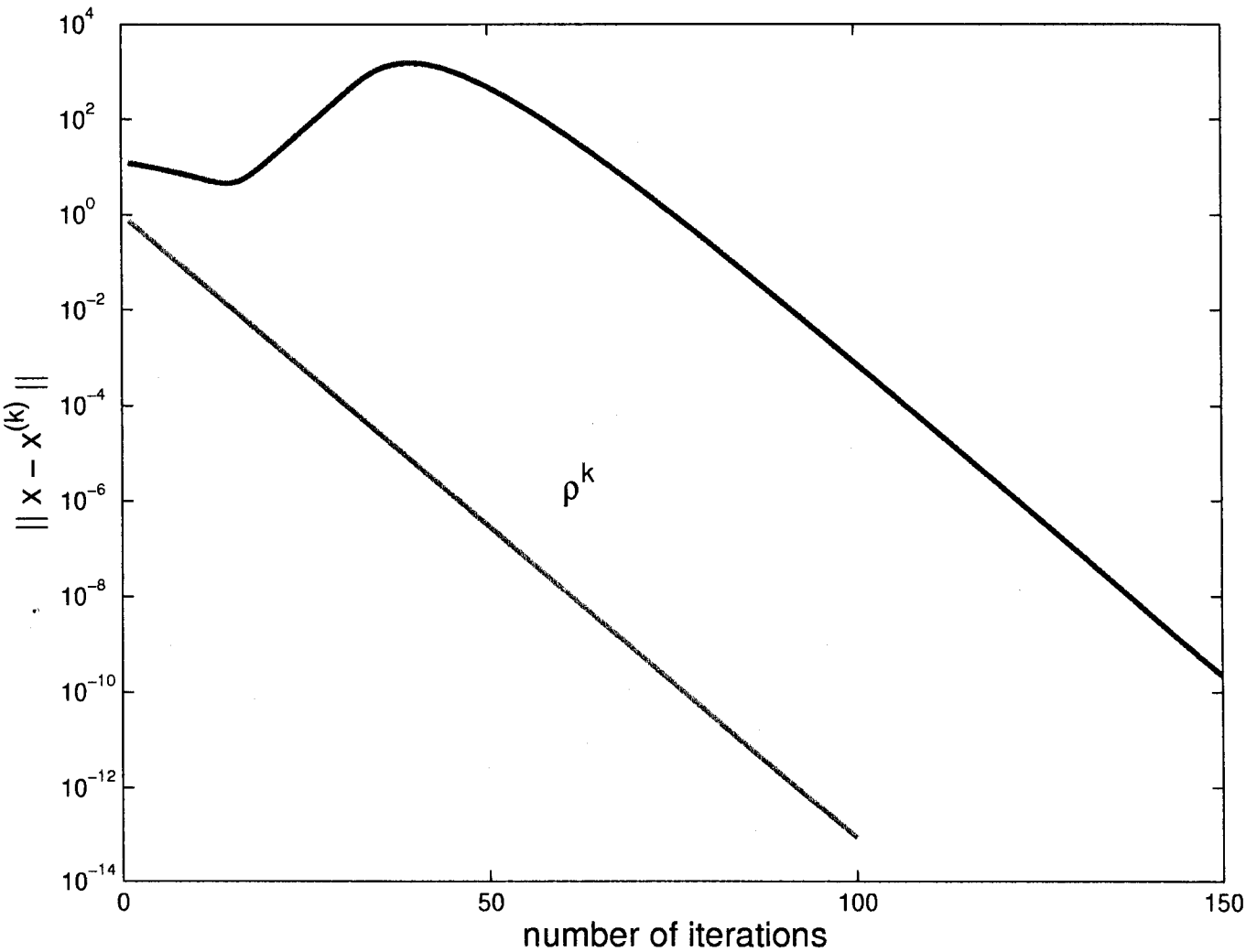
The major difference is now that the basis is only locally orthogonal, so we loose optimality of the projection -

Other methods have been devised, that do not require storing the whole basis (short-term recurrences) - In particular, "polynomial" methods, such as BiCGSTAB(ℓ) -

Very recently, development of a new algorithm, IDR(s), with performance similar to GMRES


10^4
10^2
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
$|| x - x^{(k)} ||$
0
50
100
150
number of iterations


F(I−M^−1A), A=tridiag(.16, 1, −1.16), M=tril(A)
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1


10^4
10^2
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
$\| x - x^{(k)} \|$
ρ^k
0
50
100
150
number of iterations