Utrecht, September–December 1996

Subspace methods

for

linear problems

and eigenproblems

Gerard Sleijpen
MI room 504
University Utrecht
Tel 030-253-1732
sleijpen@math.ruu.nl

Henk van der Vorst
MI room 521
University Utrecht
Tel 030-253-3732
vorst@math.ruu.nl

http://www.math.ruu.nl/people/sleijpen/

http://www.math.ruu.nl/people/vorst/

Program

1. Introduction, Conjugate Gradient Method
2. GCR — GCRO, GMRES
3. Convergence,

   relation to eigenvalue approximation
4. Bi-CG — Bi-CGSTAB, BiCGstab(ℓ)
5. Breakdown, look ahead, finite precision
6. Parallel aspects, preconditioning
7. Arnoldi, Rational Krylov Subspace, IRA
8. Jacobi-Davidson

Conclusions

1.1 Large systems of sparse linear equations can only be solved by iterative methods.
1.2 There is no best iterative method.
1.3 Krylov subspace methods subject to appropriate preconditioning are very attractive.

- *For a class of related problems for a given type of computer architecture one may find good methods by combining theoretical insight and experimental results.*
- *It is not easy to find a good method for a given situation.*
- *Numerical literature can easily be understood.*

2. Minimal residual methods
   as GCR, ..., GMRESR(m, ℓ)

and

3. efficient quasi-minimal residual methods
   as Bi-CG, ..., Bi-CGSTAB, BiCGstab(ℓ)

are very competitive.

Certain aspects of the reality

$\Downarrow$ Oceanography

Mathematical model (PDE)

$\Downarrow$ Finite elements
Finite differences
Finite volumes

Discretized model

$\Downarrow$

Computer model

$\Downarrow$ Computer Science

Simulation

Discretized model $\Rightarrow$ Computer model

Discretized PDE & discretized BC

Linear equation $\boxed{\mathbf{A}\mathbf{x} = \mathbf{b}}$ where

$\mathbf{A}$ $n \times n$–matrix, non-singular, n large, $\mathbf{A}$ sparse
$\mathbf{b}$ n–vector

Appear in linear problems and non-linear problems if
- the PDE problems are stationary
- implicit integration is used

(in non-linear problems, linear equations appear in the iteration steps of non-linear problem solvers, as the (damped) Newton method).

Solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ by
1. direct methods or
2. iterative methods as $\left\{ \begin{array}{l} \text{Multigrid techniques (smoothing)} \\ \text{Krylov subspace meth. (pre-conditioning)} \end{array} \right.$

Why iterative methods?

Direct method: $\mathbf{A} = \mathbf{L}\,\mathbf{U}$

Due to fill-in, $\mathbf{L}$ *and* $\mathbf{U}$ *can not be stored for realistic problems*

Comparison for $n \to \infty$

model problem	**Direct**	**Iterative**
Computational costs	$\mathcal{O}(n^\nu)$ $\nu \approx 2.0$ if $d = 2$ $\nu \approx 2.3$ if $d = 3$	$\mathcal{O}(n^\nu)$ $\nu \approx 1.4$ if $d = 2$ $\nu \approx 1.2$ if $d = 3$
Memory requirement	$\mathcal{O}(n^\nu)$ $\nu \approx 1.5$ if $d = 2$ $\nu \approx 1.7$ if $d = 3$	$\mathcal{O}(n)$

Is there a best iterative method?

If no further restrictions: no.

For a(ny) certain class of related problems, comp. arch.,...: yes.

Success of any iterative method depends on:

- properties of A
  (size, sparsity structure, spectral properties,...),
- computer architecture
  (scalar, vector, parallel, massive parallel, shared memory, distributed memory, ...),
- memory space that is available,
- accuracy that is required.

Iterative Methods to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$

step k
$\downarrow$

$\mathbf{x}_0 \rightsquigarrow \mathbf{x}_1 \rightsquigarrow \ldots \rightsquigarrow \mathbf{x}_{k-1} \rightsquigarrow \mathbf{x}_k$

$\uparrow$ initial guess $\qquad$ $\uparrow$ k-th approximate

- computational costs per step low
- low on memory
- $\|\mathbf{x} - \mathbf{x}_k\| \leq$ *tol* for some non-large k (hopefully)

k-th **error**
$\mathbf{x} - \mathbf{x}_k$
not computable

k-th **residual**
$\mathbf{r}_k = \mathbf{A}(\mathbf{x} - \mathbf{x}_k) = \mathbf{b} - \mathbf{A}\mathbf{x}_k$
computable

$$\begin{array}{ccccccccc} \mathbf{x}_0 & \rightsquigarrow & \mathbf{x}_1 & \rightsquigarrow & \ldots & \rightsquigarrow & \mathbf{x}_{k-1} & \rightsquigarrow & \mathbf{x}_k \\ \mathbf{r}_0 & & \mathbf{r}_1 & & & & \mathbf{r}_{k-1} & & \mathbf{r}_k \end{array}$$

stop if $\|\mathbf{r}_k\| \leq$ *tol*

In many iterative methods

compute **search direction** $\mathbf{u}_k$,
based on previously computed
$\mathbf{r}_k,\ \mathbf{u}_{k-1}, \mathbf{u}_{k-2}, \ldots,\ \mathbf{c}_{k-1}, \ldots$

$$\vdots$$
$$\mathbf{c}_k = \mathbf{A}\,\mathbf{u}_k$$
$$\vdots$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$$
$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$$

Objective: to get $\|\mathbf{r}_{k+1}\|$ "small"

Work on the residual, get the associated approximate as a side-product by updating the residual with a vector $\mathbf{c}_k$ of the form $\mathbf{c}_k = \mathbf{A}\mathbf{u}_k$ with $\mathbf{u}_k$ <u>explicitly</u> known.

Shifting the problem

All iterative methods that we consider are "linear" in the following sense.

Suppose $\mathbf{x}_0$ is the initial guess for $\mathbf{A}\mathbf{x} = \mathbf{b}$,
with approximates $\mathbf{x}_k$ and residuals $\mathbf{r}_k$.

Suppose $\mathbf{y}_0$ is the initial guess for
the "shifted problem" $\mathbf{A}\mathbf{y} = \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$,
with approximates $\mathbf{y}_k$.

Then $\quad \mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k \quad$ and $\quad \mathbf{r}_k = \mathbf{r}_0 - \mathbf{A}\mathbf{y}_k$.

Therefore, without loss of generality, we may take for the initial guess $\mathbf{x}_0 = 0$.

Stopping criteria I

If $\quad \dfrac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\| + \|\mathbf{b}\|} \le \epsilon \quad$ then

- $(\mathbf{A} + \Delta)\mathbf{x}_k = \mathbf{b} + \delta$ for some Δ and δ with $\quad \|\Delta\| \le \epsilon\|\mathbf{A}\|, \quad \|\delta\| \le \epsilon\|\mathbf{b}\|$
- $\|\mathbf{x} - \mathbf{x}_k\| \le \epsilon \left\|\mathbf{A}^{-1}\right\| (\|\mathbf{A}\|\|x_k\| + \|\mathbf{b}\|)$

(*ineq. & absolute values coordinate-wise*)

If $\quad \max_i \dfrac{(|\mathbf{r}_k|)_i}{(|\mathbf{A}||\mathbf{x}_k| + |\mathbf{b}|)_i} \le \epsilon \quad$ then

- $(\mathbf{A} + \Delta)\mathbf{x}_k = \mathbf{b} + \delta$ for some Δ and δ with $\quad |\Delta| \le \epsilon|\mathbf{A}|, \quad |\delta| \le \epsilon|\mathbf{b}|$
- $\|\mathbf{x} - \mathbf{x}_k\| \le \epsilon \left\||\mathbf{A}^{-1}|\ (|\mathbf{A}||x_k| + |\mathbf{b}|)\right\|$

Stopping criteria II

If $\quad \dfrac{\|\mathbf{r}_k\|}{\|\mathbf{b}\|} \le \epsilon \quad$ then

- $\mathbf{A}\mathbf{x}_k = \mathbf{b} + \delta$ for some δ with $\|\delta\| \le \epsilon\|\mathbf{b}\|$
- $\|\mathbf{x} - \mathbf{x}_k\| \le \epsilon \left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}\|$

However,

If $\|\,|\mathbf{A}|\,\|\,\|\mathbf{A}^{-1}\|$ is large, the iteration may never terminate with the above stopping criterion.

Suppose the exact solution $\mathbf{x}$ is known.
Then, in finite precision arithmetic,
$$\mathrm{fl}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{b} - (\mathbf{A} + \Delta_A)\mathbf{x} = \Delta_A\mathbf{x}$$
for some perturbation matrix Δ_A with
$$|\Delta_A| \le n_A\,\overline{\xi}\,|\mathbf{A}|,$$
where $\overline{\xi}$ is the relative machine precision,
and n_A max non-zeros of $\mathbb{A}$ per row. Hence,
$$\|\mathrm{fl}(\mathbf{b} - \mathbf{A}\mathbf{x})\| \le n_A\,\overline{\xi}\,\|\,|\mathbf{A}|\,\|\,\|\mathbf{x}\| \le n_A\,\overline{\xi}\,\|\,|\mathbf{A}|\,\|\,\|\mathbf{A}^{-1}\|\,\|\mathbf{b}\|$$

Consider a k-dimensional subspace $\mathcal{V}$ of $\mathbb{C}^n$.

<u>**Proposition.**</u> For a $\mathbf{y}_V \in \mathcal{V}$, $\mathbf{y}_V$ solves
$$\min\{\|\mathbf{r}_0 - \mathbf{y}\| \mid \mathbf{y} \in \mathcal{V}\},$$
that is $\quad \|\mathbf{r}_0 - \mathbf{y}_V\| = \min\{\|\mathbf{r}_0 - \mathbf{y}\| \mid \mathbf{y} \in \mathcal{V}\}$,
if and only if $\quad \mathbf{r}_0 - \mathbf{y}_V \perp \mathcal{V}$

Suppose $\mathcal{V}$ is spanned
by the columns of $\mathbf{V} = [\mathbf{v}_0 \mid \ldots \mid \mathbf{v}_{k-1}]$.

<u>**Proposition.**</u> Then, for $\mathbf{z} \in \mathbb{C}^n$,
$$\mathbf{z} \perp \mathcal{V} \quad \Leftrightarrow \quad \mathbf{V}^*\mathbf{z} = 0$$

Krylov subspace methods

$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k \quad$ where
$$\begin{aligned}\mathbf{y}_k &= \gamma_0\mathbf{r}_0 + \gamma_1\mathbf{A}\mathbf{r}_0 + \ldots + \gamma_{k-1}\mathbf{A}^{k-1}\mathbf{r}_0\\ &= \left(\gamma_0 + \gamma_1\mathbf{A} + \ldots + \gamma_{k-1}\mathbf{A}^{k-1}\right)\mathbf{r}_0\\ &= q_k(\mathbf{A})\mathbf{r}_0 \quad \text{where}\\ &\qquad q_k(t) = \gamma_0 + \gamma_1 t + \ldots + \gamma_{k-1}t^{k-1}.\end{aligned}$$

$$\begin{aligned}\mathbf{r}_k &= \mathbf{b} - \mathbf{A}\mathbf{x}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{y}_k\\ &= (\mathbf{I} - \mathbf{A}q_k(\mathbf{A}))\mathbf{r}_0\\ &= p_k(\mathbf{A})\mathbf{r}_0 \text{ where } p_k(t) = 1 - t q_k(t).\end{aligned}$$

Any <u>Krylov subspace method</u> looks for
a correction $\mathbf{y}_k$ of $\mathbf{x}_0$ in
the **Krylov subspace** $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ where

$$\begin{aligned}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) &= \mathrm{span}\left(\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{k-1}\mathbf{r}_0\right)\\ &= \{q(\mathbf{A})\mathbf{r}_0 \mid q \text{ polynomial deg. } < k\}\\ &= \{q(\mathbf{A})\mathbf{r}_0 \mid q \in \mathcal{P}_{k-1}\}\end{aligned}$$

Success depends on:

- how well it finds a good correction
  ($\|\mathbf{r}_0 - \mathbf{A}\mathbf{y}_k\|$ should be as small as possible);

- how efficiently it computes $\mathbf{y}_k$
  (how efficiently it computes the update $\mathbf{u}_{k-1}$ of $\mathbf{x}_{k-1}$: $\quad \mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_{k-1}$);

- how much memory space it requires.

$\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) := \mathrm{span}\left(\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{k-1}\mathbf{r}_0\right)$ is
the Krylov subspace of **order** k
generated by $\mathbf{A}$ and $\mathbf{r}_0$.

<u>**Proposition.**</u> $\mathbf{x} - \mathbf{x}_0 \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$
if $\quad k = \dim\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = \dim\mathcal{K}_{k+1}(\mathbf{A}; \mathbf{r}_0)$

(<u>*Proof.*</u> $\quad \alpha_0\mathbf{r}_0 + \alpha_1\mathbf{A}\mathbf{r}_0 + \ldots + \alpha_k\mathbf{A}^k\mathbf{r}_0 = 0$,
for some $\quad \alpha_i \in \mathbb{C}, \quad \alpha_0 \neq 0,\ \alpha_k \neq 0$. $\quad$ Hence
$$\mathbf{b} - \mathbf{A}\mathbf{x}_0 = \mathbf{r}_0 = -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\mathbf{A}^j\mathbf{r}_0 = \mathbf{A}\underbrace{\left(-\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\mathbf{A}^{j-1}\mathbf{r}_0\right)}_{=\,\mathbf{y}}$$
$$\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}_0 + \mathbf{y}) = \mathbf{b}. \qquad)$$

In particular $\mathbf{x}_n = \mathbf{x}$.
However, in general $\mathbf{x}_k \approx \mathbf{x}$ for $k \ll n$.

From now on we assume that k is such that $\dim\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = k$.

Why Krylov subspace methods?

- Uses only very elementary ingredients,

BLAS routines:

Ingredients:

AXPY (vector update): $\alpha\mathbf{x}+\mathbf{y}$

MV (Matrix vector multiplication): $\mathbf{M}\mathbf{v}$

DOT (inner product): $(\mathbf{x},\mathbf{y})=\mathbf{y}^T\mathbf{x}=\sum y_i x_i$

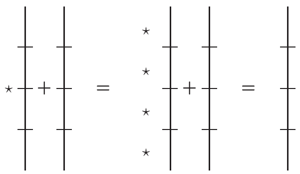
- Portable

- Need often only a non-large number of iteration steps.

- Flexible

◦ Simple algorithms (often only a few lines) that are easy implement, to add to an existing code and to read.

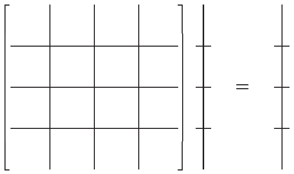
AXPY: $\alpha\mathbf{x}+\mathbf{y}$



Easy to vectorize parallelize.

Group expressions for chaining.

MV: $\mathbf{M}\mathbf{v}$

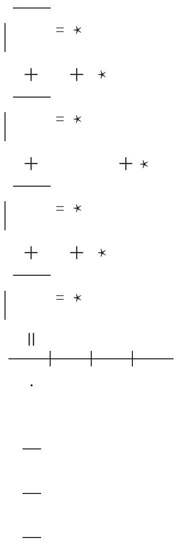


Store compressed row-wise

Easy to vectorize parallelize.

Only local communication.

DOT: $(\mathbf{x},\mathbf{y})=\mathbf{y}^T\mathbf{x}=\sum y_i x_i$



Can be vectorized, parallelized. **Global** communication, synchronization

Minimal residual methods

Optimal Krylov subspace methods

$\mathbf{y}_k=\mathbf{x}_k-\mathbf{x}_0\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ solves

$$\|\mathbf{r}_k\|=\min\{\|\mathbf{r}_0-\mathbf{A}\mathbf{y}\| \mid \mathbf{y}\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)\},$$

where $\mathbf{r}_k=\mathbf{b}-\mathbf{A}\mathbf{x}_k=\mathbf{r}_0-\mathbf{A}\mathbf{y}_k$.

step k: $\mathbf{u}_k$ "orthogonal",

that is $\mathbf{u}_k\in\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$

$\mathbf{u}_k$ "$\perp$" $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$

Costs of step k:

$\left.\begin{array}{l}1\text{ MV}\\ k\text{ AXPYs}\\ k\text{ DOTs}\end{array}\right\}$ computation

k n-vectors to store memory

(Bi–)orthogonal residual methods

Efficient Krylov subspace methods

Costs of step k: 1 MV, ℓ AXPYs, ℓ DOTs

ℓ n-vectors to be stored

$\mathbf{u}_k$ "bi-orthogonal",

that is $\mathbf{u}_k\in\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$

$\mathbf{u}_k$ "$\perp$" $\mathcal{K}_k(\mathbf{A}^T;\tilde{\mathbf{r}}_0)$

(o.k if $\mathbf{u}_k$ "$\perp$" $\tilde{\mathbf{s}}_{k-1}$ for $\tilde{\mathbf{s}}_{k-1}\in\mathcal{K}_k(\mathbb{A}^T;\tilde{\mathbf{r}}_0)\backslash\mathcal{K}_{k-1}(\mathbf{A}^T;\tilde{\mathbf{r}}_0)$)

$\mathbf{x}_k$ such that $\mathbf{r}_k=\mathbf{b}-\mathbf{A}\mathbf{x}_k$ "$\perp$" $\mathcal{K}_k(\mathbf{A}^T;\tilde{\mathbf{r}}_0)$

Theorem

- If $\mathbf{A}$ is positive definite $(\mathbf{A}^*=\mathbf{A})$ then CG is efficient and optimal.
- In general there is **no** efficient & optimal Krylov subspace method.

Krylov subspace methods

$\boxed{\text{Optimal methods}}$

GMRES, $\left\{\begin{array}{l}\text{GCR}\\ \text{ORTHODIR}\\ \text{ORTHOMIN}\\ \vdots\end{array}\right.$

GCRO(ℓ,m)

GMRESR(ℓ,m)

GMRES(m), ℓ–truncated GCR

⋮

some hybrid Bi-CG method??

⋮

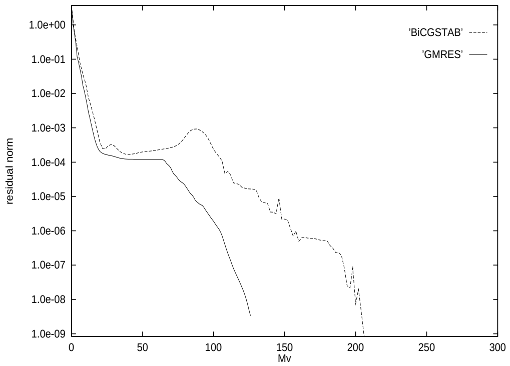
BiCGstab(ℓ),

Bi-CGSTAB

CGS

Bi-CG, QMR

$\boxed{\text{Efficient methods}}$



$-(D\,u_x)_x-(D\,u_y)_y=1$ on $[0,1]\times[0,1]$.

$D=1000$ for $0.1\le x,y\le 0.9$

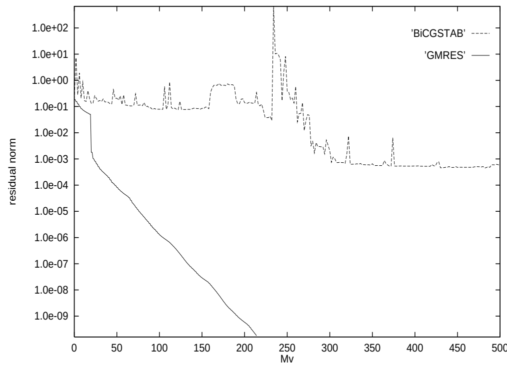
and $D=1$ elsewhere.

Dirichlet BC on $y=0$,

Neumann BC on other parts of Boundary.

82×82 discretization.

ILU Decomp. as preconditioner.



$u_{xx}+u_{yy}+u_{zz}+1000\,u_x=F.$

F is defined by the solution

$u(x,y,z)=xyz(1-x)(1-y)(1-z)$.

Discretized using $(22\times 22\times 22)$ volumes.

No preconditioning was used.

Convergence

How small can $\|\mathbf{r}_k\|$ be?

On what properties does the (minimal) size $\|\mathbf{r}_k\|$ depend?

How can we enforce these properties?

$$\|\mathbf{r}_k\| = \|p_k(\mathbf{A})\mathbf{r}_0\| \leq \|p_k(\mathbf{A})\| \, \|\mathbf{r}_0\|$$
$$\leq \mathcal{C}_E \max_{\lambda \text{ eigenv. } \mathbf{A}} |p_k(\lambda)| \, \|\mathbf{r}_0\|$$

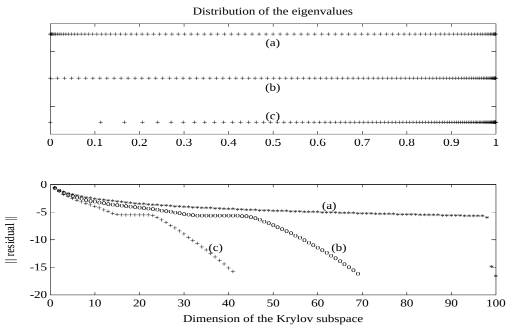
$$\min_{p \in \mathcal{P}_k^1} \max_{\lambda \text{ eigenv. } \mathbf{A}} |p_k(\lambda)|$$

$\mathcal{P}_k^1 := \{\, p \mid p \text{ pol. deg. } \leq k, \quad p(0) = 1\}$

Convergence behavior ($\|\mathbf{r}_k\|$) depends on:

- $\mathcal{C}(\mathbf{A}) = \|\mathbf{A}\| \, \|\mathbf{A}^{-1}\|$ **condition number** of $\mathbf{A}$,
- convex hull of the spectrum $\mathbf{A}$ (=set of eigenvalues of $\mathbf{A}$),
- distribution of the eigenvalues of $\mathbf{A}$ *(distribution of the eigenv. close to 0 and isolated extremal eigenvalues),*
- pseudo eigenvalues, distance from normality, ...,
- ..., $\mathbf{A}$,
- ◦ the initial residual $\mathbf{r}_0$.

Example



Example, continue

The lower frame shows the **convergence history** (the graph of $\|\mathbf{r}_k\|$ —on ^{10}log-scale— as a function of k) for an optimal Krylov subspace method for 3 symmetric $\mathbf{A}$ with

- the same condition number $\mathcal{C}(\mathbf{A}) \approx 4\,10^3$, (the same largest and smallest eigenvalue, 1, resp., $2.5\,10^{-4}$)
- but with a distribution of the eigenvalues that differs as indicated in the upper frame.

Preconditioning

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$
$$\mathbb{A}\mathbf{y} = \mathbf{A}\mathbf{K}^{-1}\mathbf{y} = \mathbf{b} \quad \text{and} \quad \mathbf{x} = \mathbf{K}^{-1}\mathbf{y}$$
(right preconditioning)

$\mathbf{u} = \mathbf{K}^{-1}\mathbf{r}$ means "solve $\mathbf{K}\mathbf{u} = \mathbf{r}$".

$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{K}^{-1}\mathbf{y}_k$ where $\mathbf{y}_k = q_k(\mathbb{A})\mathbf{r}_0$ with $q_k(t) = \gamma_0 + \gamma_1 t + \ldots + \gamma_{k-1} t^{k-1}$

$$\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k = \mathbf{r}_0 - \mathbb{A}\mathbf{y}_k$$
$$= p_k(\mathbb{A})\mathbf{r}_0 \quad \text{with} \quad p_k(t) = 1 - t q_k(t)$$

$$\min_{p \in \mathcal{P}_k^1} \|p(\mathbf{A})\| \rightsquigarrow \min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|$$

No preconditioning?

Example: discretized advection-diffusion PDE with large advection term (1-dim.)

$\mathbf{A}$ is an 99×99 matrix.

$$\mathbf{A} = \begin{bmatrix} -102 & 1 & 0 & 0 & \ldots \\ 101 & -102 & 1 & 0 & \\ 0 & 101 & -102 & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -101 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then $\mathbf{x} \approx (1, 1, \ldots, 1)^T$ is the exact solution.

With $\mathbf{x}_0 = 0$ we have that $\mathbf{r}_0 = \mathbf{b}$ and

$$\boxed{\mathcal{K}_k(A; \mathbf{r}_0) = \text{span}(e_1, \ldots, e_k)}$$

Hence $\|\mathbf{x} - \mathbf{x}_k\| \geq 1$ for every $k < 100$.

Any unpreconditioned Krylov subspace method needs at least 100 iteration steps in this example to "drag" the information to the other end of the grid. The preconditioner can "drag" the information over the whole grid in one step.

Incorporating a preconditioner

- **Explicit** use original algorithm, modify linear equation

  right: $\mathbf{A}\mathbf{K}^{-1}\mathbf{y} = \mathbf{b}$ and $\mathbf{x} = \mathbf{K}^{-1}\mathbf{y}$
  left: $\mathbf{K}^{-1}\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$ with $\tilde{\mathbf{b}} = \mathbf{K}^{-1}\mathbf{b}$
  two-sided: $\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_2^{-1}\mathbf{y} = \tilde{\mathbf{b}}$ with $\mathbf{K} = \mathbf{K}_1\mathbf{K}_2$

\+ *efficient MV possible (Eisenstat trick)*
− *only preconditioned residuals or preconditioned approximates available*

- **Implicit** use original linear equation, modify algorithm

  use preconditioned search directions: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{K}^{-1}\mathbf{u}_k$

\+ *no post-processing needed*

$$\mathbf{A} = \mathbf{L}_A + \mathbf{D}_A + \mathbf{U}_A$$

De Eisenstat trick

With Δ diagonal,
$\mathbf{K}_1 = \mathbf{L}_A + \Delta$, and $\mathbf{K}_2 = \Delta^{-1}(\mathbf{U}_A + \Delta)$
the MV $\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_2^{-1}\mathbf{u}$ can be as efficient as the MV $\mathbf{A}\mathbf{u}$

Compute
$\mathbf{D} = \Delta^{-1}\mathbf{D}_A - 2\mathbf{I}$, $\mathbf{L} = \Delta^{-1}\mathbf{L}_A$, $\mathbf{U} = \Delta^{-1}\mathbf{U}_A$,

solve $(\mathbf{U} + \mathbf{I})\mathbf{w}_1 = \mathbf{u}$
solve $(\mathbf{L} + \mathbf{I})\mathbf{w}_2 = \mathbf{D}\mathbf{w}_1 + \mathbf{u}$
Compute $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$

Then $\mathbf{w} = \mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_2^{-1}\mathbf{u}$

Explicit preconditioning

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$$

Proposition.

$\mathbf{r}_k = p(\mathbf{A}\mathbf{K}^{-1})\mathbf{r}_0$ determines convergence also for left and two-sided preconditioning.

Left: $\mathbf{K}^{-1}\mathbf{A}\mathbf{x} = \mathbf{K}^{-1}\mathbf{b}$.

Then $p(\mathbf{K}^{-1}\mathbf{A})\mathbf{K}^{-1}\mathbf{r}_0 = \mathbf{K}^{-1}p(\mathbf{A}\mathbf{K}^{-1})\mathbf{r}_0$

Two-sided: With $\mathbf{K} = \mathbf{K}_1\mathbf{K}_2$,

$\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_2^{-1}\mathbf{y} = \mathbf{K}_1^{-1}\mathbf{b}$ and $\mathbf{x} = \mathbf{K}_2^{-1}\mathbf{y}$.

Then $p(\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_2^{-1})\mathbf{K}_1^{-1}\mathbf{r}_0 = \mathbf{K}_1^{-1}p(\mathbf{A}\mathbf{K}^{-1})\mathbf{r}_0$

Good preconditioners

- $\mathbf{I} - \mathbf{A}\mathbf{K}^{-1}$ small in some sense
  $\mathbf{A}\mathbf{K}^{-1}$ has a more favorable eigenvalue-distribution, etc..
- It should be cheap to solve $\mathbf{K}\mathbf{u} = \mathbf{r}$.
- It should be cheap to store $\mathbf{K}$

Trade of:
number of iterations goes down
solving $\mathbf{K}\mathbf{u} = \mathbf{r}$ is more expensive

New architecture = new game:
"Good" on a vector machine does not imply "good" on parallel machine, ...

$$\mathbf{A} = \mathbf{L}_A + \mathbf{D}_A + \mathbf{U}_A$$

Examples

Classical preconditioned Krylov methods

Jacobi: $\mathbf{K} = \mathbf{D}_A$, $\mathbf{r}_k = (\mathbf{I} - \mathbf{A}_A\mathbf{D}_A^{-1})^k\mathbf{r}_0$

Gauss-Seidel: $\mathbf{K} = \mathbf{D}_A + \mathbf{L}_A$,
$$\mathbf{r}_k = \left(\mathbf{I} - \mathbf{A}(\mathbf{D}_A + \mathbf{L}_A)^{-1}\right)^k\mathbf{r}_0$$

Sym. GS (SSOR): $\mathbf{K} = (\mathbf{L}_A + \mathbf{D}_A)\mathbf{D}_A^{-1}(\mathbf{D}_A + \mathbf{U}_A)$

Polynomial preconditioning: $\mathbf{K} = q(\mathbf{A})$ where $p = 1 - tq$ is such that $\mathbf{I} - \mathbf{A}q(\mathbf{A})$ is "small"

Example: **Chebychev iteration**: $p = T_\ell$ where T_ℓ is the Chebychev polynomial of degree ℓ properly shifted and scaled.

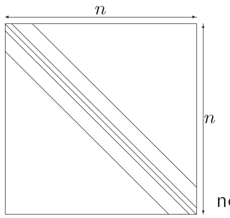
Modern preconditioned Krylov methods

Bi-CGSTAB with D-ILU:
$\mathbf{K} = (\mathbf{L}_A + \Delta)\Delta^{-1}(\Delta + \mathbf{U}_A)$ with
Δ diagonal such that $\mathrm{diag}(\mathbf{K}) = \mathbf{D}_A$
or Δ diagonal such that $\mathbf{K}\,\mathbf{1} = \mathbf{A}\,\mathbf{1}$

ILU, ...with GMRES, ..., ...

GMRES(ℓ) with GCR,
GCR with Chebychev, ...

Operators $n \times n$ matrices $H_0^2(\Omega) \xrightarrow{\mathcal{L}} L^2(\Omega)$

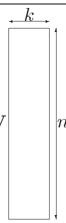


$\mathbf{A}$, $\mathbf{K}$
$\mathbb{A}$

$$\mathbb{R}^n = \mathbb{E} \underset{\mathbf{K}^{-1}}{\overset{\mathbf{A},\,\mathbf{K}}{\rightleftarrows}} \mathbb{F} = \mathbb{R}^n$$

$\mathbf{A}^k$, $\mathbf{A}^{-1}$
$\mathbf{K}^{-1}$, $\mathbf{A}$, $\mathbf{A}^T$, $\mathbb{V}$, Λ
only
for theory
not stored as matrices!

Large matrices and Large vectors $n \times k$, n

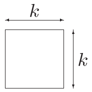


$\mathbf{U}$, $\mathbf{R}$, $\mathbf{V}$

$\mathbf{x}, \mathbf{x}_k$
$\mathbf{b}, \mathbf{r}_k$
$\mathbf{y}, \mathbf{u}_k$
$\mathbf{c}_k$

Small matrices and small vectors $k \times k$, k

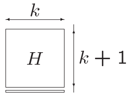
T, H, L, U



$\vec{\gamma}_k$, $\vec{\zeta}_k$

$$\vec{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{k-1} \end{bmatrix} = (\gamma_0, \ldots, \gamma_{k-1})^T$$

$\underline{U}$, $\underline{H} =$



Miscellaneous scalars: α_k, λ_k, ζ, $\mathcal{C}$;
integers: n, k, i, j, m, ℓ;
polynomials: p, p_k, q, T_k

Minimal residual methods

Optimality:

$\mathbf{y}_k \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ solves
$$\min\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

or, equivalently

$\mathbf{s}_k \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ solves
$$\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

$$\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

$\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ such that
$\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{k-1}$ is an orthogonal basis of $\mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$
where $\mathbf{c}_j = \mathbf{A}\mathbf{u}_j$ for $j = 0, 1, \ldots, k-1$

$$\begin{aligned} \text{Then}\quad \mathbf{r}_k &= \mathbf{r}_0 - \sum_{j=0}^{k-1} \frac{(\mathbf{r}_0, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)}\mathbf{c}_j \\ &= \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{c}_{k-1} \\ \text{and}\quad \mathbf{x}_k &= \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{u}_{k-1} \end{aligned}$$

$\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ such that

$\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{k-1}$ is an orthogonal basis of $\mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$

where $\quad \mathbf{c}_j = \mathbf{A}\mathbf{u}_j \quad$ for $\quad j = 0, 1, \ldots, k-1$

How to compute the $\mathbf{c}_j$?

Note that:

$$\mathbf{r}_k \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0) \quad \Rightarrow \quad \mathbb{A}\mathbf{r}_k \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0).$$

$$\mathbf{c}_k = \mathbb{A}\mathbf{r}_k - \sum_{j=0}^{k-1} \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)} \mathbf{c}_j \quad \text{and}$$

$$\mathbf{u}_k = \mathbf{K}^{-1}\mathbf{r}_k - \sum_{j=0}^{k-1} \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)} \mathbf{u}_j$$

Gram-Schmidt orthogonalization

[Vinsome '76]

ORTHOMIN

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\quad \|\mathbf{r}_k\| > tol \quad$ do
  solve $\mathbf{u}$ from $\quad \mathbf{K}\,\mathbf{u} = \mathbf{r}_k$
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $\quad i = 0, \ldots, k-1 \quad$ do
    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
  end for
  for $\quad i = 0, \ldots, k-1 \quad$ do
    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{u}_k = \mathbf{u}, \quad \mathbf{c}_k = \mathbf{c}$
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k + 1$
end while

Minimal residual methods

Advantages:

- Krylov subspace methods.
- Optimality.
- 1 MV needed to increase the Krylov subspace by one dimension
- No knowledge (in advance) on (spectral) properties of $\mathbb{A}$ is needed.
- ◦ Approximations of certain eigenvalues can cheaply be computed from β_{ij}, σ_j and α_j.

Drawbacks:

- Growing computational costs if the number of iteration steps increases, # DOTs & # AXPYs proportional to the step number.
- Growing memory requirement if the number of iteration steps increases, # large vectors is proportional to the step number.
- Any step requires DOTs.

CG: Conjugate Gradients $\quad \mathbb{A}^* = \mathbb{A} > 0$

If $\mathbb{A}$ is symmetric positive definite.

Savings in ORTHOMIN lead to CG

- $k-1$ DOTs & $k-1$ AXPYs in step k:
  $\beta_{ik} = 0$ for $i \leq k-2$.
  Put β_k instead of $\beta_{k-1\,k}$.

◦ 1 DOT by using $\quad \rho_k = (\mathbf{r}_k, \mathbb{A}\mathbf{r}_k)$
  and $\beta_k = -\rho_k/\rho_{k-1}$ & $\alpha_k = \rho_k/\sigma_k$

Replacing $\quad (\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbf{x} \quad$ by

$$(\mathbf{x}, \mathbf{y})_{\mathbb{A}^{-1}} = (\mathbb{A}^{-1}\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbb{A}^{-1}\mathbf{x}$$

saves

◦ 1 DOT: $\quad \rho_k = \mathbf{r}_k^T\mathbf{r}_k = \|\mathbf{r}_k\|^2$,
◦ 1 AXPY: $\quad \mathbf{c}_k = \mathbf{A}\mathbf{u}_k \quad$ (and $\sigma_k = \mathbf{c}_k^T\mathbf{r}_k$).

$$\tau_{ik} = (\mathbb{A}\mathbf{r}_k, \mathbf{c}_i) = (\mathbf{r}_k, \mathbb{A}^T\mathbf{c}_i) = (\mathbf{r}_k, \mathbb{A}\mathbf{c}_i)$$

$$\mathbb{A}\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0) = \mathrm{span}(\mathbf{c}_0, \ldots, \mathbf{c}_{k-2}) \subset \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$$

$$\mathbb{A}\mathbf{c}_i \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0) = \mathrm{span}(\mathbf{c}_0, \ldots \mathbf{c}_{k-1}) \perp \mathbf{r}_k \qquad i \leq k-2$$

$$\rho_k = (\mathbf{r}_k, \mathbf{c}_k) = (\mathbf{r}_k, \mathbb{A}\mathbf{r}_k - \beta_{k-1}\mathbf{c}_{k-1}) = (\mathbf{r}_k, \mathbb{A}\mathbf{r}_k)$$

Use $\quad \alpha_{k-1}\mathbf{c}_{k-1} = \mathbf{r}_{k-1} - \mathbf{r}_k$:

$$\beta_k = \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{c}_{k-1})}{(\mathbf{c}_{k-1}, \mathbf{c}_{k-1})} = \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{r}_{k-1} - \mathbf{r}_k)}{(\mathbf{c}_{k-1}, \mathbf{r}_{k-1} - \mathbf{r}_k)}$$

and note that

$\mathbf{r}_k \perp \mathbf{c}_{k-1} \quad$ and $\quad (\mathbb{A}\mathbf{r}_k, \mathbf{r}_{k-1}) = (\mathbf{r}_k, \mathbb{A}\mathbf{r}_{k-1}) = 0$.

Preconditioned CG $\quad \mathbb{A} = \mathbb{A}^* > 0$

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = 0, \quad \rho_{-1} = 1$
while $\quad \|\mathbf{r}_k\| > tol \quad$ do
  solve $\tilde{\mathbf{u}}$ from $\quad \mathbf{K}\tilde{\mathbf{u}} = \mathbf{r}_k$
  $\rho_k = (\mathbf{r}_k, \mathbf{r}_k), \quad \beta_k = -\rho_k/\rho_{k-1}$
  $\mathbf{u}_k = \tilde{\mathbf{u}} - \beta_k\mathbf{u}_{k-1}, \quad \mathbf{c}_k = \mathbf{A}\mathbf{u}_k$
  $\sigma_k = (\mathbf{c}_k, \mathbf{r}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k = k + 1$
end while

Properties CG

Advantages

- 1 MV, 2 DOT, 3 AXPY per step (to increase the Krylov subspace dimension by one).
- 4 large vectors have to be stored
- Optimal Krylov subspace method ($\|\mathbf{r}_k\|_{\mathbb{A}^{-1}}$ is minimal).
- No knowledge on properties of $\mathbb{A}$ is needed.
- ◦ Robust: CG (for pos.def.) always converges.

Drawback

◦ CG is sensitive to evaluation errors (often loss of super-linear convergence).

CGNE: CG for normal equations; precond. CG with $\mathbf{K} = \mathbf{A}^T$

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = 0, \;\; \rho_{-1} = 1$
while $\quad \|\mathbf{r}_k\| > tol \quad$ do
  $\mathbf{u} = \mathbf{A}^T\mathbf{r}_k$
  $\rho_k = (\mathbf{r}_k, \mathbf{r}_k), \;\; \beta_k = -\rho_k/\rho_{k-1}$
  $\mathbf{u}_k = \mathbf{u} - \beta_k\,\mathbf{u}_{k-1}$
  $\mathbf{c}_k = \mathbf{A}\mathbf{u}_k$
  $\sigma_k = (\mathbf{c}_k, \mathbf{r}_k), \;\; \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k + 1$
end while

Additional properties of CG for normal equations

Advantages

- 3 large vectors have to be stored
  (use: $\sigma_k = (\mathbf{u}_k, \mathbf{u}_k)$, $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{u}_k$).

Drawbacks

- CG for normal equations
usually convergences very very slowly
($\mathcal{C}(\mathbf{A})^2$ plays a role, rather than $\mathcal{C}(\mathbf{A})$).
- 2 MV in each step.
- 1 MV is by $\mathbf{A}^T$ in each step.

CGNE versus ORTHOMIN, I

$$\mathbf{A} = \text{blockdiag}\left(\ldots, \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}, \ldots\right)$$

$$\mathbf{A}\mathbf{x} = \mathbf{r}_0 = \mathbf{1}.$$

$$(\mathbf{A} - \mathbf{I})^2 = 0 \ \Rightarrow \dim \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = 2$$

ORTHOMIN: $\|\mathbf{r}_2\| = 0$.

CGNE: $\|\mathbf{r}_{n/2}\| \approx 1$

$$\mathbf{A}^T\mathbf{A} = \text{blockdiag}\left(\ldots, \begin{bmatrix} 1+k^2 & k \\ k & 1 \end{bmatrix}, \ldots\right)$$

CGNE versus ORTHOMIN, II

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ldots & 1 & 0 \end{bmatrix}$$

If $\mathbf{A}\mathbf{x} = \mathbf{r}_0 = e_1$ then
$\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = \text{span}(e_1, \ldots, e_k)$ and $\mathbf{x} = e_n$,
$\min\{\|e_1 - \mathbf{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)\} = 1$
$\mathbf{y}_k = 0$ if $k < n$

ORTHOMIN: $\|\mathbf{r}_k\| = 1$ for $k < n$.

CGNE: $\|\mathbf{r}_1\| = 0$ $(\mathbf{A}^T\mathbf{A} = \mathbf{I}!)$.

$<\cdot, \cdot>$ standard inner product: $<\mathbf{x}, \mathbf{y}> = \sum_j x_j \overline{y_j}$,
$|\cdot|$ Euclidean norm

CG for $\mathbf{A}$ and $\mathbf{K}$ pos.def.

$$\begin{aligned} \|\mathbf{r}_k\|_{\mathbb{A}^{-1}} &= \min\left\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\|_{\mathbb{A}^{-1}} \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\right\} \\ &= \min\left\{\|p(\mathbb{A})\mathbf{r}_0\|_{\mathbb{A}^{-1}} \mid p \in \mathcal{P}_k^1\right\} \end{aligned}$$

$\mathbf{A}^H = \mathbf{A} > 0$ and $\mathbf{K}^H = \mathbf{K} > 0$ (**Ex.** $\mathbf{K} = \mathbf{L}^T\mathbf{L}$),
$(\mathbf{x}, \mathbf{y}) := <\mathbf{K}^{-1}\mathbf{x}, \mathbf{y}>$ and $\|\mathbf{x}\| := \sqrt{<\mathbf{K}^{-1}\mathbf{x}, \mathbf{x}>}$

Then $\boxed{\begin{array}{l} \mathbb{A} = \mathbf{A}\mathbf{K}^{-1} \text{ pos. def. w.r.t. } (\cdot,\cdot) \\ \|\mathbf{z}\|_{\mathbb{A}^{-1}} = |\mathbf{z}|_{\mathbf{A}^{-1}} \end{array}}$

$$|\mathbf{r}_k|_{\mathbf{A}^{-1}} = \min\left\{|p(\mathbb{A})\mathbf{r}_0|_{\mathbf{A}^{-1}} \mid p \in \mathcal{P}_k^1\right\}$$

[Hestenes Stiefel '52]

CG $\quad \mathbf{A} > 0,\ \mathbf{K} > 0$

```
Choose x_0
k = 0
r_0 = b - A x_0
u_{-1} = 0,  rho_{-1} = 1
while  ||r_k|| > tol  do
    solve w from  K w = r_k
    rho_k = (r_k, w),  beta_k = -rho_k / rho_{k-1}
    u_k = w - beta_k u_{k-1}
    c_k = A u_k
    sigma_k = (c_k, u_k),  alpha_k = rho_k / sigma_k
    x_{k+1} = x_k + alpha_k u_k
    r_{k+1} = r_k - alpha_k c_k
    k <- k + 1
end while
```


$$\|\mathbf{r}_k\|_{\mathbb{A}^{-1}} = \min\left\{\|p(\mathbb{A})\mathbf{r}_0\|_{\mathbb{A}^{-1}} \mid p \in \mathcal{P}_k^1\right\}$$

Convergence CG

$\mathbb{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, $\mathbf{v}_i \perp \mathbf{v}_j$ $(i \neq j)$, $\|\mathbf{v}_i\| = 1$,

$$\mathbf{r}_0 = \sum_i \mu_i \mathbf{v}_i = \sum_i (\mathbf{r}_0, \mathbf{v}_i)\, \mathbf{v}_i$$

$$\begin{aligned} \|p(\mathbb{A})\mathbf{r}_0\|_{\mathbb{A}^{-1}} &= \sqrt{\sum_i |\mu_i|^2 \frac{p(\lambda_i)^2}{\lambda_i}} \\ &\leq \sqrt{\sum_i \frac{|\mu_i|^2}{\lambda_i}} \max_i |p(\lambda_i)| \\ &\leq \|\mathbf{r}_0\|_{\mathbb{A}^{-1}} \max_{\lambda \in \sigma(\mathbb{A})} |p(\lambda)| \end{aligned}$$

$$\|p(\mathbb{A})\mathbf{r}_0\|_{\mathbb{A}^{-1}} \leq \|\mathbf{r}_0\|_{\mathbb{A}^{-1}} \max_{\lambda \in \sigma(\mathbb{A})} |p(\lambda)|$$

$\sigma(\mathbb{A}) \subset [\lambda_1, \lambda_n] \subset (0, \infty)$, $\qquad \mathcal{C} := \lambda_n / \lambda_1$

$$\max_{\lambda \in \sigma(\mathbb{A})} |p(\lambda)| \leq 2\left(\frac{\sqrt{\mathcal{C}} - 1}{\sqrt{\mathcal{C}} + 1}\right)^k \leq 2 \exp\left(-2k \frac{1}{\sqrt{\mathcal{C}}}\right)$$

With $p(t) := \dfrac{T_k\left(\frac{\lambda_n - t}{\lambda_n - \lambda_1} - \frac{t - \lambda_1}{\lambda_n - \lambda_1}\right)}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)}$, we find

$$\max_{\lambda \in \sigma(\mathbb{A})} |p(\lambda)| \leq \frac{1}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)} \leq 2\left(\frac{\sqrt{\mathcal{C}} - 1}{\sqrt{\mathcal{C}} + 1}\right)^k$$

Theorem.
$\mathbf{r}_0, \ldots, \mathbf{r}_k$ form an orthogonal basis of $\mathcal{K}_{k+1}(\mathbb{A}; \mathbf{r}_0)$.

Proof. $\mathbf{r}_k \in \mathcal{K}_{k+1}(\mathbb{A}; \mathbf{r}_0)$.
$\mathbf{r}_k$ and $\mathbf{c}_k$ are orthogonal to
$\text{span}(\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}) = \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ w.r.t. $(\cdot,\cdot)_{\mathbb{A}^{-1}}$.
$\Rightarrow$ $\mathbf{r}_k$ and $\mathbf{c}_k$ orthogonal to $\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ w.r.t. $(\cdot,\cdot)$.
$\Rightarrow$ $\mathbf{r}_k$ orthogonal to $\mathbf{r}_0, \ldots, \mathbf{r}_{k-1}$.

Hence $\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0) = \{\, \mathbf{R}_k \vec{\gamma} \mid \vec{\gamma} \in \mathbb{R}^k\}$
where $\mathbf{R}_k := [\mathbf{r}_0 \,|\, \mathbf{r}_1 | \ldots | \mathbf{r}_{k-1}]$
orthogonal columns

For $m < k$:
$0 = (\mathbf{c}_m, \mathbf{c}_k)_{\mathbb{A}^{-1}} = (\mathbb{A}\mathbf{u}_m, \mathbb{A}\mathbf{u}_k)_{\mathbb{A}^{-1}} = (\mathbf{u}_m, \mathbb{A}\mathbf{u}_k)$
search directions are "Conjugate Gradients"

$\mathbf{r}_{k+1}=\mathbf{r}_k-\alpha_k\mathbf{c}_k \quad \& \quad \mathbf{u}_k=\mathbf{K}^{-1}\mathbf{r}_k-\beta_k\mathbf{u}_{k-1}$

Two coupled "2–term" CG-recursions can be combined to one "3–term" recursion for residuals:

$$\begin{aligned}\mathbb{A}\mathbf{r}_k &= \mathbf{A}\,(\mathbf{u}_k+\beta_k\mathbf{u}_{k-1})\\ &= \frac{1}{\alpha_k}(\mathbf{r}_k-\mathbf{r}_{k+1})+\frac{\beta_k}{\alpha_{k-1}}(\mathbf{r}_{k-1}-\mathbf{r}_k)\\ &= \underbrace{\frac{\beta_k}{\alpha_{k-1}}}_{\tilde\beta_k}\mathbf{r}_{k-1}+\underbrace{\left(\frac{1}{\alpha_k}-\frac{\beta_k}{\alpha_{k-1}}\right)}_{\tilde\alpha_k}\mathbf{r}_k\underbrace{-\frac{1}{\alpha_k}}_{\tilde\gamma_k}\mathbf{r}_{k+1}\end{aligned}$$

$$\tilde\gamma_k+\tilde\alpha_k+\tilde\beta_k=0$$

Hence, $\mathbb{A}\mathbf{R}_k=\mathbf{R}_{k+1}\underline{\widetilde{T}}_k$, with

$$\underbrace{\mathbf{R}_k=[\mathbf{r}_0\,|\,\mathbf{r}_1|\,\ldots\,|\mathbf{r}_{k-1}]}_{\text{orthogonal columns}},\quad \underbrace{\underline{\widetilde{T}}_k=\begin{bmatrix}\tilde\alpha_0 & \tilde\beta_0 & & \\ \tilde\gamma_0 & \tilde\alpha_1 & \ddots & \\ & \tilde\gamma_1 & \ddots & \tilde\beta_k\\ & & \ddots & \tilde\alpha_k\\ & & & \tilde\gamma_k\end{bmatrix}}_{(k+1)\times k \text{ tri-diagonal}}$$

Lanczos, $\mathbb{A}^*=\mathbb{A}$

$$\begin{array}{ll}\mathbf{w}=\mathbb{A}\,\mathbf{v}_k-\alpha\,\mathbf{v}_k, & \alpha_k \text{ s.t. } \mathbf{w}\perp\mathbf{v}_k\\ \mathbf{v}=\mathbf{w}-\beta_k, & \beta_k \text{ s.t. } \mathbf{v}\perp\mathbf{v}_{k-1}\\ \mathbf{v}_{k+1}=\mathbf{v}/\gamma_k & \text{with } \gamma_k=\|\mathbf{v}\|\end{array}$$

Theorem. With $\mathbf{V}_k=[\mathbf{v}_0\,|\,\ldots\,|\,\mathbf{v}_{k-1}]$ and

$$\underline{T}_k=\begin{bmatrix}\alpha_0 & \beta_1 & & & \\ \gamma_0 & \alpha_1 & \ddots & & \\ & \gamma_1 & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{k-1}\\ & & & \ddots & \alpha_{k-1}\\ & & & & \gamma_{k-1}\end{bmatrix}$$

$\mathbf{v}_0,\ldots,\mathbf{v}_{k-1}$ is an orthonormal basis of $\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$,

$\underline{T}_k$ is $(k+1)\times k$ tridiagonal, and $\mathbb{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$.

Let T_k be the $k\times k$ upper block T_k of $\underline{T}_k$.

T_k is real symmetric: $\beta_k=\gamma_{k-1}$.

$\mathbf{V}_k$ orthonormal, $\quad \mathbb{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$

Lanczos and CG, I

If $\quad \mathbb{A}^*=\mathbb{A}>0 \quad$ and $\quad \mathbf{v}_0=\rho\,\mathbf{r}_0 \quad$ then

$\mathbf{v}_k=\mathbf{r}_k/\|\mathbf{r}_k\|$ all k.

ORTHORES: an alternative for γ_k

With $\mathbf{v}_0=\mathbf{r}_0$ & $\mathbf{y}_0=\mathbf{x}_0$, generate $\mathbf{v}_k$ & $\mathbf{y}_k$ by

$$\begin{aligned}\gamma_k\mathbf{v}_{k+1} &= \mathbb{A}\mathbf{v}_k-\alpha_k\mathbf{v}_k-\beta_k\mathbf{v}_{k-1},\\ \mathbf{y}_{k+1} &= (\mathbf{K}^{-1}\mathbf{v}_k+\alpha_k\mathbf{y}_{k-1}+\beta_k\mathbf{y}_{k-1})/\gamma_k,\end{aligned}$$

with α_k & β_k such that

$$\mathbf{v}_{k+1}\perp\mathbf{v}_k,\mathbf{v}_{k-1} \quad \& \quad \gamma_k=-(\alpha_k+\beta_k).$$

Then $\quad \mathbf{v}_k=\mathbf{r}_k \quad$ and $\quad \mathbf{y}_k=\mathbf{x}_k \quad$ all k.

$\mathbf{V}_k$ orthonormal, $\quad \mathbb{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$

Lanczos and CG, II

$\mathbf{y}_k=\mathbf{V}_k\vec\gamma_k\in\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$ solves

$$\min\{\|\mathbf{r}_0-\mathbb{A}\mathbf{y}\|_{\mathbb{A}^{-1}}\mid \mathbf{y}\in\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)\}$$

if and only if $\quad T_k\vec\gamma_k=\rho\,e_1$,

with ρ such that $\mathbf{r}_0=\rho\,\mathbf{v}_0$

and T_k the $k\times k$ upper block of $\underline{T}_k$.

Proposition. If $\quad \mathbb{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$

is the Lanczos reduction with $\mathbf{v}_0=\rho\,\mathbf{r}_0$,

then for the CG approximate $\mathbf{x}_k$ we have

$$\mathbf{x}_k=\mathbf{x}_0+\mathbf{K}^{-1}\mathbf{V}_k\,\vec\gamma_k \quad \text{with} \quad T_k\vec\gamma_k=\rho\,e_1.$$

$\mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k,\ \ \mathbf{x}_k=\mathbf{V}_k\,\vec\gamma_k,\ \ T_k\,\vec\gamma_k=\rho\,e_1$

Lanczos and CG, III $\quad \mathbf{K}=\mathbf{I},\ \mathbf{x}_0=0$

$$T_k=L_k\,U_k$$

$$\mathbf{x}_k=\left(\mathbf{V}_k\,U_k^{-1}\right)\left(L_k^{-1}(\rho\,e_1)\right)$$

Expand $\mathbf{V}_k$, T_k and L_k and U_k

Expand $\mathbf{W}_k:=\mathbf{V}_k\,U_k^{-1}$

Expand $(\zeta_0,\ldots,\zeta_{k-1})^T:=L_k^{-1}(\rho\,e_1)$

Update $\mathbf{x}_k$

Note that $\mathbf{A}\mathbf{W}_k=\mathbf{A}\,\mathbf{V}_k\,U_k^{-1}=\mathbf{V}_{k+1}\,\underline{L}_k$

CG=Lanczos + "T=LU" + efficient impl.

$\mathbf{V}_k^*\mathbf{V}_k=I, \quad \mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$

MINRES, $\quad \mathbf{K}=\mathbf{I},\ \mathbf{x}_0=0$

With $\mathbf{r}_0=\rho\,\mathbf{v}_0,\ \mathbf{y}_k=\mathbf{V}_k\vec\gamma_k\in\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$ solves

$$\min\{\|\mathbf{r}_0-\mathbf{A}\mathbf{y}\|\mid \mathbf{y}\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)\}$$

if and only if $\vec\gamma_k\in\mathbb{R}^k$ solves

$$\min\{\|\rho\,e_1-\underline{T}_k\vec\gamma\|\mid\vec\gamma\in\mathbb{R}^k\}.$$

If $\quad \underline{T}_k=Q_{k+1}\,\underline{R}_k \quad$ with

Q_{k+1} product Givens rotations,

$\underline{R}_k$ $(k+1)\times k$ upper triangular,

then

$$\mathbf{x}_k=\left(\mathbf{V}_k\,R_k^{-1}\right)\vec\zeta_k \text{ with } \begin{bmatrix}\vec\zeta_k\\ \rho_k\end{bmatrix}:=Q_{k+1}^*(\rho\,e_1).$$

$\mathbf{V}_k^*\mathbf{V}_k=I, \quad \mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\underline{T}_k$

SYMMLQ, $\quad \mathbf{K}=\mathbf{I},\ \mathbf{x}_0=0$

With $\mathbf{r}_0=\rho\,\mathbf{v}_0,\ \mathbf{y}_k=\mathbf{V}_k\vec\gamma_k\in\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$ solves

$$\min\{\|\mathbf{x}-\mathbf{A}\mathbf{y}\|\mid \mathbf{y}\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)\}$$

if and only if $\quad (\underline{T}_k^*\,\underline{T}_k)\vec\gamma=\rho\,e_1$.

If $\quad \underline{T}_k=Q_{k+1}\,\underline{R}_k \quad$ with

Q_{k+1} product Givens rotations,

$\underline{R}_k$ $(k+1)\times k$ upper triangular,

then

$$\begin{aligned}\mathbf{x}_k &= \mathbf{A}\,\mathbf{V}_k\,\vec\gamma_k=\mathbf{V}_{k+1}\left(\underline{T}_k\,(\underline{T}_k^*\,\underline{T}_k)^{-1}\right)(\rho\,e_1)\\ &= \left(\mathbf{V}_{k+1}\,Q_{k+1}\right)\begin{bmatrix}R_k^{-*}(\rho\,e_1)\\ 0\end{bmatrix}.\end{aligned}$$

$\mathbf{A}\mathbf{x}=\mathbf{b}$ with $\mathbf{A}^*=\mathbf{A}\not>0$

CG [Orthogonal residuals] $\mathbf{y}_k\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ s.t.

$$\mathbf{r}_k=\mathbf{r}_0-\mathbf{A}\mathbf{y}_k\perp\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$$

– may break down, irregular convergence

MINRES [Minimal residuals] $\mathbf{y}_k\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ s.t.

$$\|\mathbf{r}_0-\mathbf{A}\mathbf{y}_k\|=\min\{\|\mathbf{r}_0-\mathbf{A}\mathbf{y}\|\mid\mathbf{y}\in\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)\}$$

SYMMLQ [Minimal error] $\mathbf{y}_k\in\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ s.t.

$$\|\mathbf{x}-\mathbf{A}\mathbf{y}_k\|=\min\{\|\mathbf{x}-\mathbf{A}\mathbf{y}\|\mid\mathbf{y}\in\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)\}$$

\+ MINRES and SYMMLQ do not break down

\+ SYMMLQ more stable than MINRES

Orthogonalization

Consider
$\mathbf{V}_k = [\mathbf{v}_1 \,|\, \dots \,|\, \mathbf{v}_k]$ and $\mathbf{W}_k = [\mathbf{w}_1 \,|\, \dots \,|\, \mathbf{w}_k]$,
matrices of full rang (of rank k).

An **expansion vector** $\tilde{\mathbf{v}}$ for $\mathbf{V}_k$
is **orthogonalized w.r.t.** $\mathbf{W}_k$ by

$$\tilde{\mathbf{u}} = \tilde{\mathbf{v}} - \mathbf{V}_k\,(\mathbf{W}_k^*\mathbf{V}_k)^{-1}\,(\mathbf{W}_k^*\tilde{\mathbf{v}}),$$

select a $\gamma_k \neq 0$, $\mathbf{v}_{k+1} = \tilde{\mathbf{u}}/\gamma_k$.

Then $\mathbf{v}_{k+1} \perp \mathbf{w}_j \quad \forall j \leq k$
or, equivalently, $\mathbf{W}_k^*\mathbf{v}_{k+1} = 0$.

Moreover, $\tilde{\mathbf{v}} = [\mathbf{V}_k \,|\, \mathbf{v}_{k+1}] \begin{bmatrix} \vec{h}_k \\ \gamma_k \end{bmatrix}$
with $\vec{h}_k := (\mathbf{W}_k^*\mathbf{V}_k)^{-1}\,(\mathbf{W}_k^*\tilde{\mathbf{v}})$

$\mathbf{V}_k$ & $\mathbf{W}_k$ matrices of full rank

$(\mathbf{V}_k, \mathbf{W}_k)$ are **bi-orthogonal** iff
$\mathbf{W}_k^*\mathbf{V}_k$ is a non-singular diagonal

$\mathbf{V}_k$ is **orthogonal** iff
$\mathbf{V}_k^*\mathbf{V}_k$ is a non-singular diagonal

$\mathbf{V}_k$ is ortho**normal** iff $\mathbf{V}_k^*\mathbf{V}_k = I$

How to evaluate $\tilde{\mathbf{v}} - \mathbf{V}_k\,(\mathbf{W}_k^*\mathbf{V}_k)^{-1}\,(\mathbf{W}_k^*\tilde{\mathbf{v}})$*?*

Important case: $\mathbf{V}_k$ orthogonal
How to evaluate $\tilde{\mathbf{v}} - \mathbf{V}_k\,D_k^{-1}(\mathbf{V}_k^*\tilde{\mathbf{v}})$*?*
where $D_k := \mathbf{V}_k^*\mathbf{V}_k$ is diagonal

$D_k = \mathbf{V}_k^*\mathbf{V}_k = \mathrm{diag}(\sigma_1, \dots, \sigma_k)$
How to evaluate $\tilde{\mathbf{v}} - \mathbf{V}_k\,D_k^{-1}(\mathbf{V}_k^*\tilde{\mathbf{v}})$*?*

Gram-Schmidt orthogonalization

```
for i = 1,...,k do
    τ_i = (ṽ, v_i),  β_i = τ_i/σ_i
end for
for i = 1,...,k do
    ṽ ← ṽ − β_i v_i
end for
```

$$\vec{\tau} = \mathbf{V}_k^*\tilde{\mathbf{v}}$$
$$\vec{\beta} = D_k^{-1}\vec{\tau}$$
$$\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec{\beta}$$

Advantage
- Additional parallelism

Drawback
- May be unstable

Gram–Schmidt is unstable (if the vectors to be orthogonalized make small angles)

[Walker '92]

Classical Gram-Schmidt

(Classical) Gram–Schmidt is unstable
(if the vectors to be orthogonalized make small angles)

$$\vec{\tau} = \mathbf{V}_k^*\tilde{\mathbf{v}}$$
$$\vec{\beta} = D_k^{-1}\vec{\tau}$$
$$\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec{\beta}$$

Compute all components, compute the projection on the span of all $\mathbf{v}_j$ *(sum of components) and subtract the projection.*

All errors are summed.

Modified Gram–Schmidt

```
for i = 1,...,k do
    τ_i = (ṽ, v_i),   β_i = τ_i/σ_i
    ṽ ← ṽ − β_i v_i
end for
```

Project and subtract the components one by one: compute one component and subtract that component, ...

Any error is projected and components are subtracted also.

Modified Gram-Schmidt is a less unstable than classical Gram-Schmidt. Its stability is satisfactory in practice for linear solvers.
As expensive as classical Gram-Schmidt (on non-parallel machines)

Repeated Gram-Schmidt

```
ρ = (ṽ, ṽ),  β = D_k^{-1}(V_k^* ṽ)
ṽ ← ṽ − V_k β
σ_{k+1} = (ṽ, ṽ),
while σ_k < κ ρ
    γ = D_k^{-1}(V_k^* ṽ)
    ṽ ← ṽ − V_k γ
    β ← β + γ
    ρ ← σ_{k+1},  σ_{k+1} = (ṽ, ṽ)
end while
```

e.g., $\kappa = 0.5$

In practice:
- Appears to be stable
- At most 1 refinement step needed (Only a little more expensive than Gram-Schmidt)

Theory:
Proof for "once is enough" only if $k = 1$.

Householder reflections

+ Are very stable
− Twice as expensive as
  classical or modified Gram-Schmidt.

[Gronopolous '92]

s-step approach

Work with blocks of s vectors:
Orthogonalize $\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_s$ w.r.t. $\mathbf{V}_k = [\mathbf{v}_1 \,|\, \dots \,|\, \mathbf{v}_k]$.

+ More parallelism can be exploited
− Can be very unstable

Krylov basis

$\mathbf{v}_1, \dots, \mathbf{v}_k$ is a **Krylov basis** of $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ if
$\mathcal{K}_j(\mathbf{A}; \mathbf{r}_0) = \mathrm{span}\big(\mathbf{v}_1, \dots, \mathbf{v}_j\big)$ for all $j \leq k$

Note that then $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = \mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$.
Therefore, when referring to $\mathbf{A}$*, we will not specify the Krylov subspace explicitly.*

Theorem. If $\mathbf{v}_1, \dots, \mathbf{v}_{k+1}$ is a Krylov basis
then $$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k}$$

where $\mathbf{V}_j := [\mathbf{v}_1 \,|\, \dots \,|\, \mathbf{v}_j]$ and
$\underline{H}_k$ is a $(k+1) \times k$ (upper) Hessenberg.
$\underline{H}_k$ is **irreducible**, i.e. $h_{j+1\,j} \neq 0 \quad \forall j$.

H_k is the $k \times k$ upper block of $\underline{H}_k$.

Arnoldi

Arnoldi expands recursively an orthonormal basis $\mathbf{v}_1,\ldots,\mathbf{v}_k$ of $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ $(\rho\mathbf{v}_1=\mathbf{r}_0)$ by orthogonalizing $\mathbf{A}\mathbf{v}_k$ w.r.t. $\mathbf{v}_1,\ldots,\mathbf{v}_k$:

$$\tilde{\mathbf{v}}=\mathbf{A}\mathbf{v}_k-\sum_{j=1}^{k}h_{jk}\,\mathbf{v}_j \quad\text{with}\quad h_{jk}=(\mathbf{A}\mathbf{v}_k,\mathbf{v}_j)$$

$$\mathbf{v}_{k+1}=\tilde{\mathbf{v}}/h_{k+1\,k} \quad\text{with}\quad h_{k+1\,k}=\|\tilde{\mathbf{v}}\|$$

Theorem

Then $\boxed{\mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\,\underline{H}_k}$ where

$$\mathbf{V}_k:=[\mathbf{v}_1\,|\,\ldots\,|\,\mathbf{v}_k] \quad\text{and}\quad \underline{H}_k:=\begin{bmatrix} h_{11} & \ldots & h_{1k}\\ h_{21} & & \\ & \ddots & \vdots\\ 0 & & h_{k+1\,k}\end{bmatrix}$$

$\mathbf{V}_k$ is orthonormal: $\mathbf{V}_k^*\mathbf{V}_k=I$,

$\underline{H}_k$ is $(k+1)\times k$ irreducible upper Hessenberg

[Arnoldi '51]

Arnoldi's expansion

```
Choose r_0
k = 1
ρ = ||r_0||,  v_1 = r_0/ρ
V_1 = [v_1],  H_0 = [ ]
while ... do
    ṽ ← A v_k
    h_k = V_k^* ṽ,  ṽ ← ṽ − V_k h_k
    h_{k+1 k} = ||ṽ||,  v_{k+1} = ṽ/h_{k+1 k}
    V_{k+1} = [V_k | v_{k+1}],  H_k = [ H_{k−1}   h_k ; 0...0 h_{k+1 k} ]
    k ← k + 1
end while
```

$$\vec{h}_k=\mathbf{V}_k^*\tilde{\mathbf{v}},\quad \tilde{\mathbf{v}}\leftarrow\tilde{\mathbf{v}}-\mathbf{V}_k\vec{h}_k,\qquad \mathbf{V}_{k+1}=[\mathbf{V}_k\,|\,\mathbf{v}_{k+1}],\ \underline{H}_k=\begin{bmatrix}\underline{H}_{k-1} & \vec{h}_k\\ 0\ldots0 & h_{k+1\,k}\end{bmatrix}$$

Arnoldi with Householder r.

Suppose $\mathbf{v}_1,\ldots,\mathbf{v}_k$ orthonormal,

$\mathbb{H}_2,\ldots,\mathbb{H}_k$ Householder reflections s.t.

$$\mathbb{H}_k\cdots\mathbb{H}_3\cdot\mathbb{H}_2\cdot[\,\mathbf{A}\mathbf{v}_1\,|\,\ldots\,|\,\mathbf{A}\mathbf{v}_k\,]=\begin{bmatrix}R_k\\ \mathbf{0}\end{bmatrix}$$

with R_k upper triangular

$$\boxed{\begin{aligned}&\tilde{\mathbf{v}}\leftarrow\mathbb{H}_k\cdots\mathbb{H}_2\cdot(\mathbf{A}\mathbf{v}_k)\\ &\mathbb{H}_{k+1}\text{ reflects }(\tilde{\mathbf{v}})_{j>k}\text{ on span}\big(e_{k+1}\big)\\ &\mathbf{v}_{k+1}=\mathbb{H}_2\cdots\mathbb{H}_{k+1}\,e_{k+1}\\ &R_{k+1}=\left[\begin{bmatrix}R_k\\ 0\ldots0\end{bmatrix}(\mathbb{H}_{k+1}\,\tilde{\mathbf{v}})_{j\leq k+1}\right]\end{aligned}}$$

Then $\mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\,\underline{H}_k$ with $\mathbf{V}_j=[\mathbf{v}_1\,|\,\ldots\,|\,\mathbf{v}_j]$

and $\underline{H}_k\,e_j=R_{k+1}e_{j+1}\qquad(j=1,\ldots,k)$

$\mathbf{y}=(y_1,\ldots,y_n)^T$. With $\mathbf{w}:=\|\mathbf{y}\|\,e_k+\mathrm{sign}(y_k)\,\mathbf{y}$,

$\mathbb{H}=\mathbf{I}-2\frac{\mathbf{w}\,\mathbf{w}^*}{\mathbf{w}^*\mathbf{w}}$, reflects $\mathbf{y}$ to span(e_k).

$(\mathbf{y})_{j\leq k}:=(y_1,\ldots,y_k)^T$ & $(\mathbf{y})_{j>k}:=(0,\ldots,0,y_{k+1},\ldots,y_n)^T$

s-step Arnoldi

Expand an orthonormal basis $\mathbf{v}_1,\ldots,\mathbf{v}_k$ (with $k=s\ell$) of $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ by

$$\tilde{\mathbf{v}}_1=\mathbf{A}\,\mathbf{v}_k,\quad \tilde{\mathbf{v}}_2=\mathbf{A}\tilde{\mathbf{v}}_1,\ \ldots,\quad \tilde{\mathbf{v}}_s=\mathbf{A}\tilde{\mathbf{v}}_{s-1}$$

Orthogonalize these s-vector w.r.t. $\mathbf{v}_1,\ldots,\mathbf{v}_k$

$$\mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\,\underline{H}_k$$

is an **Arnoldi expansion of order** k,

i.e., $\mathbf{V}_{k+1}^*\mathbf{V}_{k+1}=I$, $\underline{H}_k$ irreducible Hessenberg

---

Lanczos

H_k the upper $k\times k$ block of $\underline{H}_k$.

Theorem

H_k is symmetric tridiagonal if $\mathbf{A}^*=\mathbf{A}$

Lanczos exploits this fact: it expands recurs. an orthonormal basis $\mathbf{v}_1,\ldots,\mathbf{v}_k$ of $\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)$ by:

$$\tilde{\mathbf{v}}=\mathbf{A}\,\mathbf{v}_k-\alpha_k\,\mathbf{v}_k-\beta_{k-1}\,\mathbf{v}_{k-1}$$

$$\text{with }\alpha_k=(\mathbf{A}\mathbf{v}_k,\mathbf{v}_k)$$

$$\mathbf{v}_{k+1}=\tilde{\mathbf{v}}/\beta_k \quad\text{with}\quad \beta_k=\|\tilde{\mathbf{v}}\|.$$

Then $\boxed{\mathbf{A}\mathbf{V}_k=\mathbf{V}_{k+1}\,\underline{T}_k}$ where

$\underline{T}_k$ is $(k+1)\times k$ irreducible tridiagonal

[Lanczos '50]

Lanczos

```
Choose r_0
k = 1
ρ = ||r_0||,  v_1 = r_0/ρ
β_0 = 0,  v_0 = 0
while ... do
    ṽ ← A v_k − β_{k−1} v_{k−1}
    α_k = v_k^* ṽ,  ṽ ← ṽ − α_k v_k
    β_k = ||ṽ||,  v_{k+1} = ṽ/β_k
    k ← k + 1
end while
```


Lanczos versus Arnoldi: $\mathrm{A}^*=\mathrm{A}$

Arnoldi: stable (very stable with repeated GS)

Lanczos: very unstable

but, very surprisingly, still usable

Arnoldi: costs (storage & computation) grow with k

Lanczos: costs constant and small per step

Alternative approaches

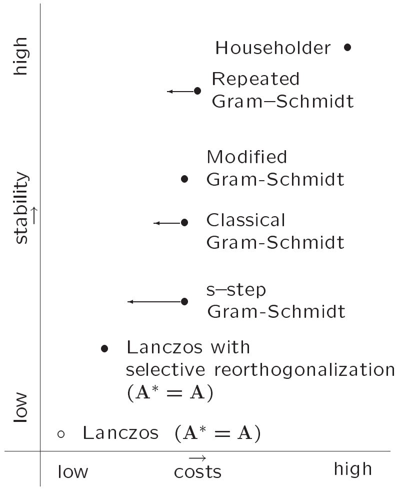
if Lanczos needs more stability

- **Selective Reorthogonalization**: orthogonalize also with respect to an appropriate part of the "old" space
- **Complete reorthogonalization**: Arnoldi but retain only tridiagonal part of $\underline{H}_k$.

The error in Lanczos process propagates in specific directions

plot fl$(\mathbf{V}_k)^*$fl$(\mathbf{V}_k)$ and $\mathbb{V}^*($fl$(\mathbf{V}_k)-\mathbf{V}_k)$

Properties orthogonalization



- • proportional with k, ∘ fixed per step

⟵ Improvements on parallel machines

The above figure shows the $\|I - \mathbf{V}_k^* \mathbf{V}_k\|$ (on ^{10}log-scale) for several orthogonalization techniques

Minimal residual methods

All minimal residual methods require the computation of an orthogonal basis.

for $\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$, $\mathbf{v}_1 = \mathbf{r}_0$
(GMRES, FOM, *Arnoldi*)

for $\mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$, $\mathbf{v}_1 = \mathbb{A}\mathbf{r}_0$
(ORTHOMIN, GCR, ORTHODIR)

ORTHOMIN, GCR expand with $\mathbb{A}\,\mathbf{r}_k$ (or $\mathbb{A}\mathbf{K}_k^{-1}\mathbf{r}_k$)
- − *may break down,*
- + *allow varying preconditioning*

GMRES, ORTHODIR expand with $\mathbb{A}\,\mathbf{c}_{k-1}$ ($\mathbb{A}\mathbf{v}_k$)
- + *robust,*
- − *allows only fixed preconditioning*

GMRES, FOM, *Arnoldi* normalize the orthogonalizes expansion vector

How to evaluate $\tilde{\mathbf{v}} - \mathbf{V}_k(\mathbf{W}_k^*\mathbf{V}_k)^{-1}(\mathbf{V}_k^*\tilde{\mathbf{v}})$*?*

Theorem. Modified Gram-Schmidt is possible
iff $M_k := \mathbf{W}_k^* \mathbf{V}_k$ is upper triangular.

Bi–Lanczos

If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a Krylov basis of $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$
and $\mathbf{w}_1, \ldots, \mathbf{w}_k$ is a Krylov bas. of $\mathcal{K}_k(\mathbf{A}^*; \tilde{\mathbf{r}}_0)$,
then **Bi-Lanczos** expands recursively

- $\mathbf{v}_1, \ldots, \mathbf{v}_k$ by $\mathbf{A}\mathbf{v}_k$ orthogonalized w.r.t. $\mathbf{w}_1, \ldots, \mathbf{w}_k$,
- and $\mathbf{w}_1, \ldots, \mathbf{w}_k$ by $\mathbf{A}^*\mathbf{w}_k$ orthogonalized w.r.t. $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Put $\mathbf{V}_j = [\mathbf{v}_1 \,|\, \ldots \,|\, \mathbf{v}_j]$ and $\mathbf{W}_j = [\mathbf{w}_1 \,|\, \ldots \,|\, \mathbf{w}_j]$.
The bi-Lanczos process does not **break down** if $D_{k+1} := \mathbf{W}_{k+1}^* \mathbf{V}_{k+1}$ is non-singular diagonal.
$\underline{H}_k$ and $\underline{F}_k$ are Hessenberg matrices such that

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k \;\&\; \mathbf{A}^*\,\mathbf{W}_k = \mathbf{W}_{k+1}\,\underline{F}_k$$

Theorem. $H_k = D_k^{-1} F_k^* D_k$ is tridiagonal, if the bi-Lanczos process does not break down.

Proof. $F_k^* D_k = \mathbf{W}_k^* \mathbf{A}\,\mathbf{V}_k = D_k H_k$

[Lanczos '50]

Bi-Lanczos

Select a $\mathbf{r}_0$, and a $\tilde{\mathbf{r}}_0$
$k = 1$
$\mathbf{v}_1 = \mathbf{r}_0/\|\mathbf{r}_0\|$, $\mathbf{v}_0 = \mathbf{0}$, $\mathbf{w}_1 = \tilde{\mathbf{r}}_0/\|\tilde{\mathbf{r}}_0\|$, $\mathbf{w}_0 = \mathbf{0}$
$\gamma_0 = 0$, $\delta_0 = 1$, $\tilde{\gamma}_0 = 0$, $\tilde{\delta}_0 = 1$
while ... do
  $\delta_k = \mathbf{w}_k^* \mathbf{v}_k$,
  $\tilde{\mathbf{v}} = \mathbf{A}\,\mathbf{v}_k$, $\tilde{\mathbf{w}} = \mathbf{A}^*\,\mathbf{w}_k$
  $\beta_k = \overline{\tilde{\gamma}}_{k-1}\delta_k/\delta_{k-1}$, $\tilde{\beta}_k = \overline{\gamma}_{k-1}\overline{\delta}_k/\overline{\delta}_{k-1}$
  $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \beta_k \mathbf{v}_{k-1}$, $\tilde{\mathbf{w}} \leftarrow \tilde{\mathbf{w}} - \tilde{\beta}_k \mathbf{w}_{k-1}$
  $\alpha_k = \mathbf{w}_k^* \tilde{\mathbf{v}}/\delta_k$, $\tilde{\alpha}_k = \overline{\alpha}_k$
  $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \alpha_k\,\mathbf{v}_k$, $\tilde{\mathbf{w}} \leftarrow \tilde{\mathbf{w}} - \tilde{\alpha}_k\,\mathbf{w}_k$
  Select a $\gamma_k \neq 0$ and a $\tilde{\gamma}_k \neq 0$
  $\mathbf{v}_{k+1} = \tilde{\mathbf{v}}/\gamma_k$, $\mathbf{w}_{k+1} = \tilde{\mathbf{w}}/\tilde{\gamma}_k$,
  $k \leftarrow k + 1$
  $\mathbf{V}_k = [\mathbf{V}_{k-1} \,|\, \mathbf{v}_k]$, $\mathbf{W}_k = [\mathbf{W}_{k-1} \,|\, \mathbf{w}_k]$
end while

Minimal residual methods

Optimality:

$\mathbf{y}_k \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ solves

$$\min\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

or, equivalently

$\mathbf{s}_k \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ solves

$$\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

• = **Minimal residual methods**

• ORTHOMIN *ORTHOgonal MINimization*
⇓ *stability*
Generalized Conjugate Residuals
• GCR ⇒ • efficient GCR ⟷ • GMRES
Breakdown ⇒ • ORTHODIR *ORTHOgonal DIRections*
⇓ *efficiency*
$\boxed{\text{Truncated GCR}}$ & $\boxed{\text{Restarted } \textit{GCR}\text{, GMRES}}$
⇓ *fast convergence*
$\boxed{\text{GMRESR}}$ *GMRES Repeated*
⇓ *fast convergence*
$\boxed{\text{GCRO}}$ *GCR Orthogonalized*

• $\boxed{\text{GMRES}}$ Generalized Minimal RESiduals ⟷ ∘ FOM Ful Orthogonalization Method

Minimal residual methods

Advantages:

- Krylov subspace methods.
- Optimality.
- 1 MV needed to increase the Krylov subspace by one dimension
- No knowledge (in advance) on (spectral) properties of $\mathbb{A}$ is needed.

∘ Approximations of certain eigenvalues can cheaply be computed from β_{ij}, σ_j and α_j.

Drawbacks:

- Growing computational costs if the number of iteration steps increases, # DOTs & # AXPYs proportional to the step number.
- Growing memory requirement if the number of iteration steps increases, # large vectors is proportional to the step number.
- Any step requires DOTs.

$$\|\mathbf{r}_k\| = \min\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

GMRES

If $\mathbb{A}\mathbf{V}_k = \mathbf{V}_{k+1}\, \underline{H}_k$ is the Arnoldi expansion of order k with $\mathbf{r}_0 = \rho\, \mathbf{v}_1$ then

$$\begin{aligned} \|\mathbf{r}_k\| &= \min_{\vec\gamma} \|\rho\, \mathbf{v}_1 - \mathbb{A}\, \mathbf{V}_k\, \vec\gamma\| \\ &= \min_{\vec\gamma} \|\rho\, \mathbf{v}_1 - \mathbf{V}_{k+1}\, \underline{H}_k\, \vec\gamma\| \\ &= \min_{\vec\gamma} \|\rho\, e_1 - \underline{H}_k\, \vec\gamma\| \quad (\star) \end{aligned}$$

If $\vec\gamma_k$ solves $(\star)$ then

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{K}^{-1}\, \mathbf{V}_k\, \vec\gamma_k$$

Per step (neglecting operations of dimension k)

work: 1 MV, $k+1$ DOT, $k+1$ AXPY

memory: $\mathbf{x}_0$, $(k+1) \times \mathbf{v}_j$

Only once: $k+1$ AXPY.

[Saad Schultz '86]

GMRES

Choose $\mathbf{x}_0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

$\rho_0 = \rho = \|\mathbf{r}_0\|$, $\mathbf{v}_1 = \mathbf{r}_0/\rho$

$k = 0$, $\mathbf{V}_1 = [\mathbf{v}_1]$, $\underline{H}_0 = [\,]$

while $\rho_{k-1} > tol$ do

  $k \leftarrow k+1$

  $\tilde{\mathbf{v}} \leftarrow \mathbf{K}^{-1}\mathbf{v}_k$, $\tilde{\mathbf{v}} \leftarrow \mathbf{A}\,\tilde{\mathbf{v}}$

  $\vec h_k = \mathbf{V}_k^*\,\tilde{\mathbf{v}}$, $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec h_k$

  $h_{k+1\,k} = \|\tilde{\mathbf{v}}\|$, $\mathbf{v}_{k+1} = \tilde{\mathbf{v}}/h_{k+1\,k}$

  $\mathbf{V}_{k+1} = [\mathbf{V}_k|\mathbf{v}_{k+1}]$, $\underline{H}_k = \begin{bmatrix} \underline{H}_{k-1} & \vec h_k \\ 0\ldots 0 & h_{k+1\,k} \end{bmatrix}$

  Compute $\rho_k = \min \|\rho\, e_1 - \underline{H}_k\, \vec\gamma\|$

end while

Solve $\min_{\vec\gamma} \|\rho\, e_1 - \underline{H}_k\, \vec\gamma\|$ $(\star)$

$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{K}^{-1}(\mathbf{V}_k\,\vec\gamma_k)$ with $\vec\gamma_k$ solution $(\star)$

Solve $\min_{\vec\gamma} \|\rho\, e_1 - \underline{H}_k\, \vec\gamma\|$ $(\star)$

Consider the solution $\vec\gamma_k$ of $(\star)$

If $\underline{H}_k = Q_{k+1}\, \underline{R}_k$ is the QR-decomposition, that is Q_{k+1} is $(k+1)\times(k+1)$ unitary, $\underline{R}_k$ is $(k+1)\times k$ upper triangular, then

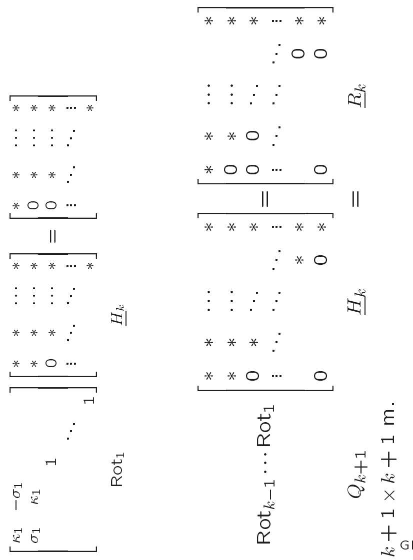
$$\|\rho\, e_1 - \underline{H}_k\, \vec\gamma_k\| = \|Q_{k+1}^*(\rho\, e_1) - \underline{R}_k\, \vec\gamma_k\|$$

Note that,

with $\vec\zeta_{k+1} := \begin{bmatrix} \tilde\zeta_k \\ \rho_k \end{bmatrix} := Q_{k+1}^*(\rho\, e_1)$

we have

$$\|\vec\zeta_{k+1} - \underline{R}_k\,\vec\gamma_k\| = |\rho_k| \quad \text{and} \quad \vec\gamma_k = R_k^{-1}\tilde\zeta_k$$



GMRES

While $\rho_k > tol$

  $k \leftarrow k+1$;

  Compute $\mathbf{v}_{k+1}$ and (h_{jk}) by one Arnoldi step;

  $\vec\eta_k := \text{Rot}_{k-1}\cdots\text{Rot}_1 \begin{bmatrix} h_{1k} \\ \vdots \\ h_{kk} \end{bmatrix}$, $\underline{R}_k = \left[\begin{array}{c|c} R_{k-1} & \vec\eta_k \\ 0\cdots 0 & \\ \hline 0\cdots 0 & h_{k+1\,k} \end{array}\right]$

  Compute κ_k and σ_k; $\text{Rot}_k\, \underline{R}_k = \begin{bmatrix} R_k \\ \hline 0\cdots 0 \end{bmatrix}$;

  $\vec\zeta_{k+1} = \text{Rot}_k \begin{bmatrix} \vec\zeta_k \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde\zeta_k \\ \rho_k \end{bmatrix}$ (then $\rho_k = \rho_{k-1}\,\sigma_k$);

end while

Solve $R_k\,\vec\gamma_k = \tilde\zeta_k$ by backward substitution

and compute $x_k = x_0 + \mathbf{V}_k\,\vec\gamma_k$

$$\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$

$\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ such that

$\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{k-1}$ is an orthogonal basis of $\mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$

where $\mathbf{c}_j = \mathbf{A}\mathbf{u}_j$ for $j = 0, 1, \ldots, k-1$

How to select the expansion vector for $\{\mathbf{c}_j\}$?
How to compute the $\mathbf{c}_k$?

Note that:

$$\mathbf{r}_k \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0) \quad \Rightarrow \quad \mathbb{A}\mathbf{r}_k \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0).$$

$$\mathbf{c}_k = \mathbb{A}\mathbf{r}_k - \sum_{j=0}^{k-1} \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)}\, \mathbf{c}_j \quad \text{and}$$

$$\mathbf{u}_k = \mathbf{K}^{-1}\mathbf{r}_k - \sum_{j=0}^{k-1} \frac{(\mathbb{A}\mathbf{r}_k, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)}\, \mathbf{u}_j$$

ORTHOMIN uses Classical Gram-Schmidt
GCR uses modified Gram-Schmidt

[Vinsome '76]

ORTHOMIN

Choose $\mathbf{x}_0$

$k = 0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

while $\|\mathbf{r}_k\| > tol$ do

  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}_k$

  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$

  for $i = 0, \ldots, k-1$ do

    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i)$, $\beta_{ik} = \tau_{ik}/\sigma_i$

  end for

  for $i = 0, \ldots, k-1$ do

    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$

    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$

  end for

  $\mathbf{u}_k = \mathbf{u}$, $\mathbf{c}_k = \mathbf{c}$

  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k)$, $\rho_k = (\mathbf{r}_k, \mathbf{c}_k)$, $\alpha_k = \rho_k/\sigma_k$

  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$

  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$

  $k \leftarrow k+1$

end while

[Elman '82]

GCR

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}_k$
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $i = 0, \ldots, k-1$ do
    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{u}_k = \mathbf{u}, \quad \mathbf{c}_k = \mathbf{c}$
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
end while

Savings in GCR

Proposition. $\mathbf{c}_k = \mathbf{A}\,\mathbf{u}_k$

- ∘ Useful for the k's for which k AXPYs are more expensive than 1 MV.

- **Efficient GCR**

  Compute only the $\mathbf{c}_j$ and $\mathbf{r}_j$
  Store only β_{ij}, α_j, $\mathbf{c}_j$.
  If $\|\mathbf{r}_k\|$ is small enough,
  Compute $\mathbf{x}_k$ from
  the **implicit** reconstruction of
  the $\mathbf{r}_j$ and the $\mathbf{u}_j$

solve $\mathbf{u}$ from $\mathbf{K}\mathbf{u} = \mathbf{r}_k$

$$\mathbf{u}_k = \mathbf{u} - \sum_{i=0}^{k-1} \beta_{ik}\,\mathbf{u}_i$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$$

Matrix representation

$\mathbf{K}^{-1}\mathbf{R}_{k+1} = \mathbf{U}_{k+1}\,B_{k+1}$, where

$\mathbf{R}_{k+1} := [\mathbf{r}_0 \,|\, \ldots \,|\, \mathbf{r}_k]$ & $\mathbf{U}_{k+1} := [\mathbf{u}_0 \,|\, \ldots \,|\, \mathbf{u}_k]$

$$B_{k+1} := \begin{bmatrix} 1 & \beta_{01} & \beta_{02} & \ldots & \beta_{0k} \\ 0 & 1 & \beta_{12} & \ldots & \beta_{1k} \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \end{bmatrix}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k \quad \Rightarrow \quad \mathbf{x}_{k+1} = \mathbf{x}_0 + \mathbf{U}_{k+1}\begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{bmatrix}$$

$\mathbf{x}_{k+1} = \mathbf{x}_0 + \mathbf{K}^{-1}\mathbf{y}_k$ where $\mathbf{y}_k = \mathbf{R}_{k+1}(B_{k+1}^{-1}\,\vec{\alpha}_{k+1})$

[Sleijpen van der Vorst '93]

An efficient form of GCR

Choose $\mathbf{x}_0$
$k = 0, \quad \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}_k$
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $i = 0, \ldots, k-1$ do
    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{c}_k = \mathbf{c}$,
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
end while
solve $\vec{\gamma}_k$ from $B_k\vec{\gamma}_k = \vec{\alpha}_k$
solve $\mathbf{y}$ from $\mathbf{K}\mathbf{y} = \mathbf{R}_k\vec{\gamma}_k$
$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k = \mathbf{r}_0 - \sum_{i=0}^{k} \alpha_i\,\mathbf{c}_i = \mathbf{r}_0 - \mathbf{C}_{k+1}\,\vec{\alpha}_{k+1}$$

Properties (efficient) GCR

Advantages:

- (One of) the most efficient minimal residual methods. 1 MV to increase the dimension of the Krylov subspace by 1.

In step k: $k+1$ DOT, k AXPY
$k+2$ large vectors to be stored

∘ Stability is satisfactory by the use of modified Gram-Schmidt to obtain orthogonal basis $\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}$ of $\mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$.

Drawbacks:

- Costs grow per iteration step.

∘ Not robust: iteration process may break down ($\rho_k = 0 \quad \Rightarrow \quad \mathbf{c}_{k+1} = 0$)

∘ The approximation $\mathbf{x}_k$ is only known at the end.

Breakdown in GCR

Example: $\mathbf{A} = \mathbb{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

If $\mathbf{x}_0 = 0$ then $\mathbf{r}_0 = \mathbf{b} = e_1$,
$\mathbf{c}_0 = \mathbf{A}\mathbf{r}_0 = e_2$ and $\rho_0 = (\mathbf{r}_0, \mathbf{c}_0) = 0$.

Remedy: Replace (in GCR)

solve $\mathbf{u}$ from $\mathbf{K}\mathbf{u} = \mathbf{r}_k$, $\quad \mathbf{c} = \mathbf{A}\mathbf{u}$

by (ORTHODIR)

solve $\mathbf{u}$ from $\mathbf{K}\mathbf{u} = \mathbf{c}_{k-1}$, $\quad \mathbf{c} = \mathbf{A}\mathbf{u}$

— If GCR does not break down then GCR $\mathbf{c}_j$ and ORTHODIR $\mathbf{c}_j$ are equal (except for a scalar multiple).

— ORTHODIR does not break down.

Explanation:

$\mathbf{r}_k = \mathbf{r}_0 + \gamma_1\,\mathbb{A}\mathbf{r}_0 + \ldots + \gamma_k\,\mathbb{A}^k\mathbf{r}_0$
γ_k may be 0; control on γ_0: $\gamma_0 = 1$

in ORTHODIR:
$\mathbf{c}_{k-1} = \gamma_1\,\mathbb{A}\mathbf{r}_0 + \ldots + \gamma_{k-1}\,\mathbb{A}^{k-1}\mathbf{r}_0 + \mathbb{A}^k\mathbf{r}_0$
control on γ_k: $\gamma_k = 1$

As GCR, ORTHODIR can be reformulated in an efficient form (no breakdown & optimal efficient in the class of minimal residual methods)

GMRES & efficient ORTHODIR

GMRES has the same properties (**+ and −**) as efficient ORTHODIR

Orthogonal basis for:
— $\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ in GMRES
— $\mathbb{A}\,\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ in ORTHODIR (GCR)
produced by (class.) Gram-Schmidt:
$\mathbf{c}_0 = \mathbf{r}_0$ (GMRES), $\mathbf{c}_0 = \mathbb{A}\mathbf{r}_0$ (ORTHODIR)

solve $\mathbf{u}$ from $\mathbf{K}\mathbf{u} = \mathbf{c}_{k-1}$,
$\mathbf{c} = \mathbf{A}\mathbf{u}$, orthogonalize

GMRES:
$\mathbf{y}_k$ solves $\min\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$
ORTHODIR (and GCR):
$\mathbf{s}_k$ solves $\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$

Keeping costs (work & storage) limited

Naive approaches

- Restart after ℓ steps (GCR(ℓ))
- Truncate:

  work with $\mathbf{u}_{k-\ell}, \ldots, \mathbf{u}_{k-1}$

  $\mathbf{c}_{k-\ell}, \ldots, \mathbf{c}_{k-1}$,

  orthogonalize only w.r.t. the last ℓ vectors.

– Lost of super-linear convergence;
  may lead to stagnation

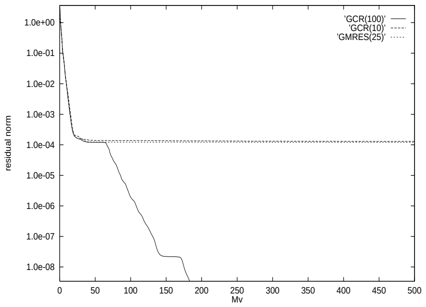
Advance approaches

Select
- implicitly or
- explicitly

an "interesting" part of the Krylov subspace

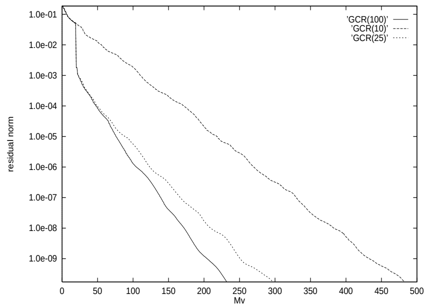
GCR(ℓ): restarted GCR

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}_k$
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $i = \lfloor \frac{k}{\ell} \rfloor \ell, \ldots, k-1$ do
    $\tau_{ik} = (c, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{u}_k = \mathbf{u}, \quad \mathbf{c}_k = \mathbf{c}$
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
end while



$-(Du_x)_x - (Du_y)_y = 1$ on $[0,1] \times [0,1]$.
$D = 1000$ for $0.1 \le x, y \le 0.9$ and $D = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
82×82 discretization.
ILU Decomp. as preconditioner.



$u_{xx} + u_{yy} + u_{zz} + 1000u_x = F.$

F is defined by the solution
$u(x,y,z) = xyz(1-x)(1-y)(1-z).$

Discretized using $(22 \times 22 \times 22)$ volumes.
No preconditioning was used.

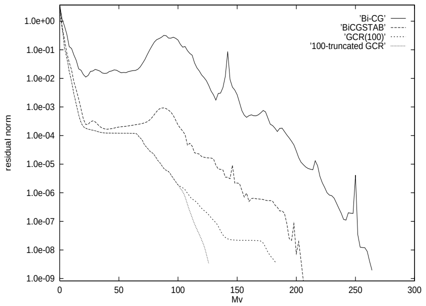
m–truncated GCR

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}_k$
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $i = \max(k-m, 0), \ldots, k-1$ do
    $\tau_{ik} = (c, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{u}_k = \mathbf{u}, \quad \mathbf{c}_k = \mathbf{c}$
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
end while

GMRES and GCR

In contrast to GCR,
GMRES and efficient GCR (ORTHODIR)
do **not** allow easy inclusion of

- truncation
- varying precondition



$-(Du_x)_x - (Du_y)_y = 1$ on $[0,1] \times [0,1]$.
$D = 1000$ for $0.1 \le x, y \le 0.9$ and $D = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
82×82 discretization.
ILU Decomp. as preconditioner.

$\min\{\|\mathbf{r}_0 - \mathbf{s}\| \mid \mathbf{s} \in \mathbb{A}\mathcal{V}\}$

$\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ such that
$\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{k-1}$ is an orthogonal basis $\mathbb{A}\mathcal{V}$
where $\mathbf{c}_j = \mathbf{A}\,\mathbf{u}_j$ for $j = 0, 1, \ldots, k-1$

How to select the expansion vector for $\{\mathbf{c}_j\}$?
get $\angle(\mathbf{r}_0, \mathbb{A}\mathcal{V})$ small

Select a $\mathbf{u}$ (e.g., such that $\mathbf{u} \approx \mathbf{A}^{-1}\mathbf{r}_k$)

$$\mathbf{c}_k = \mathbf{A}\,\mathbf{u} - \sum_{j=0}^{k-1} \frac{(\mathbf{A}\mathbf{u}, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)}\mathbf{c}_j \quad \text{and}$$

$$\mathbf{u}_k = \mathbf{u} - \sum_{j=0}^{k-1} \frac{(\mathbf{A}\mathbf{u}, \mathbf{c}_j)}{(\mathbf{c}_j, \mathbf{c}_j)}\mathbf{u}_j$$

[van der Vorst & Vuik '93]

GCR$\star$

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do
  Select a $\mathbf{u}$ (e.g. s.t. $\mathbf{u} \approx \mathbf{A}^{-1}\mathbf{r}_k$)
  $\mathbf{c} \leftarrow \mathbf{A}\,\mathbf{u}$
  for $i = 0, \ldots, k-1$ do
    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i)$, $\beta_{ik} = \tau_{ik}/\sigma_i$
    $\mathbf{u} \leftarrow \mathbf{u} - \beta_{ik}\,\mathbf{u}_i$
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{ik}\,\mathbf{c}_i$
  end for
  $\mathbf{u}_k = \mathbf{u}$, $\mathbf{c}_k = \mathbf{c}$
  $\sigma_k = (\mathbf{c}_k, \mathbf{c}_k)$, $\rho_k = (\mathbf{r}_k, \mathbf{c}_k)$, $\alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k + 1$
end while

GMRESR

GMRESR(ℓ): solve $\mathbf{A}\,\mathbf{u} = \mathbf{r}_k$ approximately by ℓ steps of GMRES (To precondition, do it explicitly)

Other iterative methods in the inner-loop

- LSQR switch [van der Vorst Vuik '93]
- Bi-CGSTAB [de Sturler Fokkema '93]
- A polynomial method using a fixed polynomial p (for which $\mathbf{I} - \mathbf{A}p(\mathbf{A})$ is "small" and take $\mathbf{u} = p(\mathbf{A})\mathbf{r}_k$) (thus avoiding inner products)

Example: Chebychev iteration [van Gijzen '93]

Deflating eigenvectors

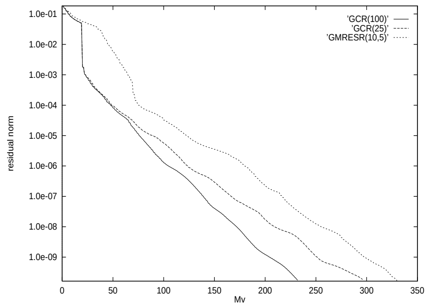
Suppose an (approximate) eigenvector $\tilde{\mathbf{u}}$ is explicitly known: $\mathbf{A}\tilde{\mathbf{u}} \approx \lambda\tilde{\mathbf{u}}$

Select in one (e.g. first) step $\mathbf{u} = \tilde{\mathbf{u}}$
select in other steps $\mathbf{u} \approx \mathbf{A}^{-1}\mathbf{r}_k$

In step "$\mathbf{u} = \tilde{\mathbf{u}}$" component of $\mathbf{x}$ (and $\mathbf{r}_0$) in the direction of $\tilde{\mathbf{u}}$ is computed. In subsequential steps, the search is in the space orthogonal to span($\tilde{\mathbf{u}}$) *(Deflation).*

Efficient for eigenvalues $\lambda \approx 0$,
no significant improvement if $\lambda \not\approx 0$.

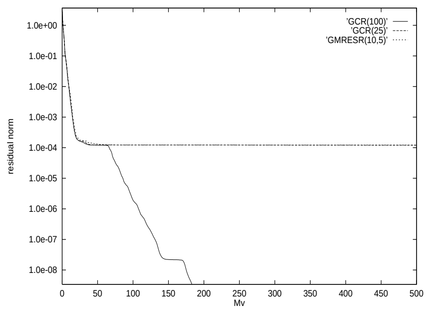
(Approximate) eigenvectors can be computed from iteration coeffients and vectors before, e.g., restart



$u_{xx} + u_{yy} + u_{zz} + 1000u_x = F.$

F is defined by the solution
$u(x, y, z) = xyz(1 - x)(1 - y)(1 - z).$

Discretized using $(22 \times 22 \times 22)$ volumes.
No preconditioning was used.



$-(Du_x)_x - (Du_y)_y = 1$ on $[0, 1] \times [0, 1]$.
$D = 1000$ for $0.1 \leq x, y \leq 0.9$ and $D = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
82×82 discretization.
ILU Decomp. as preconditioner.

Choosing ℓ in GMRESR(ℓ)

GMRES($m\ell$):
GMRES restarted every $m\ell$-th step.
GMRESR(m,ℓ):
GMRESR(ℓ) restarted every m-th step.

Experimentally: often GMRES($m\ell$) and GMRESR(m,ℓ) require the same number of MVs to obtain equally good approximations. One "sweep" requires $K := m\ell$ MV.

Per sweep:
AXPYs & DOTs GMRESR(m,ℓ) $\approx$ $(1/m + 2/\ell^2)\times$ # AXPYs & DOTs GMRES($m\ell$).

Optimal sweep length K depends on

- complexity per "step",
- accelerations

If optimal sweep length is K then choose ($m = K/\ell$ and) ℓ such that

$$m\,\ell = K \qquad \text{minimizing} \qquad 1/m + 2/\ell^2$$

$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k$: the correction $\mathbf{y}_k$ belongs to span($\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}$) and is optimal
in the sense that $\mathbf{r}_k \perp$ span($\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}$).

Waste of computational effort to let the next $\mathbf{u}$ —the approximate solution of $\mathbf{A}\mathbf{u} = \mathbf{r}_k$— have components in span($\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}$)*:* $\mathbf{u}$ *should be such that $\mathbf{c} = \mathbf{A}\mathbf{u}$ is orthogonal to* span($\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}$)

(Exact solution has this property.
For exact solution: $\mathbf{r}_{k+1} = \mathbf{r}_k - \mathbf{r}_k = 0$)

$\mathbf{P}_k = \sum_{i=0}^{k-1} \frac{\mathbf{c}_i\,\mathbf{c}_i^T}{\mathbf{c}_i^T\,\mathbf{c}_i}$ orthogonal projection onto span($\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}$).

$\mathbf{Q}_k = \sum_{i=0}^{k-1} \frac{\mathbf{u}_i\,\mathbf{c}_i^T\mathbf{A}}{\mathbf{c}_i^T\,\mathbf{c}_i}$ skew projection onto span($\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}$).

$$\mathbf{P}_k\mathbf{A} = \mathbf{A}\mathbf{Q}_k, \quad \mathbf{A}_k \equiv (\mathbf{I} - \mathbf{P}_k)\mathbf{A} = \mathbf{A}(\mathbf{I} - \mathbf{Q}_k)$$

$$\mathbf{A}_k\mathbf{u} = (\mathbf{I} - \mathbf{P}_k)\mathbf{A}\mathbf{u} = (\mathbf{I} - \mathbf{P}_k)\mathbf{r}_k = \mathbf{r}_k$$

GCRO$\star$:

Solve $\mathbf{u}$ approximately from $\mathbf{A}_k\,\mathbf{u} = \mathbf{r}_k$ such that $\mathbf{Q}_k\,\mathbf{u} = \mathbf{0}$.

Krylov subspace methods are very well suited: if $\mathbf{Q}_k\,\mathbf{u}^{(0)} = \mathbf{0}$ then $\mathbf{Q}_k\,\mathbf{u}^{(j)} = \mathbf{0}$.

Note. $\mathbf{A}\,\mathbf{u}^{(\ell)}$ orthogonal $\mathbf{c}_0, \ldots, \mathbf{c}_{k-1}$:
no orthogonalisation needed in outer loop.

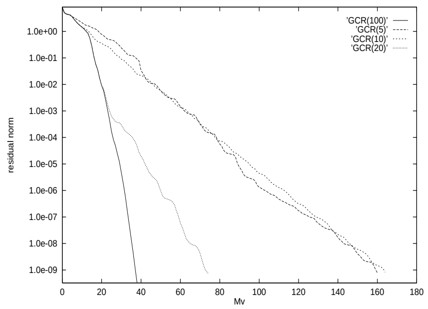
[de Sturler Fokkema '93]

GCRO⋆

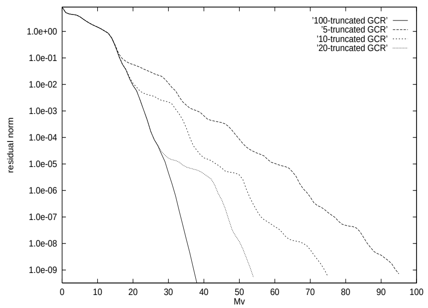
Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
while $\|\mathbf{r}_k\| > tol$ do

> solve $\mathbf{u}$ approximately from
> $\mathbf{A}_k\,\mathbf{u} = \mathbf{r}_k$ such that $\mathbf{Q}_k\,\mathbf{u} = \mathbf{0}$

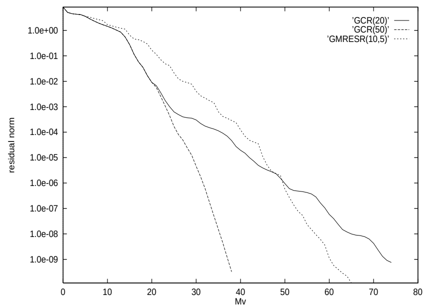
$\mathbf{u}_k = \mathbf{u}, \quad \mathbf{c}_k = \mathbf{A}\,\mathbf{u}$
$\sigma_k = (\mathbf{c}_k, \mathbf{c}_k), \quad \rho_k = (\mathbf{r}_k, \mathbf{c}_k), \quad \alpha_k = \rho_k/\sigma_k$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
$k \leftarrow k+1$
end while



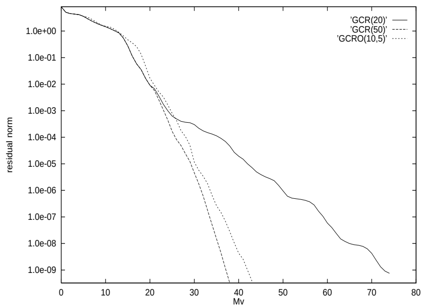
Sherman5 from the
Harwell-Boeing
Sparse Matrix Collection



Sherman5 from the
Harwell-Boeing
Sparse Matrix Collection



Sherman5 from the
Harwell-Boeing
Sparse Matrix Collection



Sherman5 from the
Harwell-Boeing
Sparse Matrix Collection

Other approaches

Implicit selection
There are other methods that allow varying preconditioners. They are not based on GCR.

- **F**lexible **GMRES** *based on GMRES* [Saad '93]
+ A little more efficient than GMRESR,
- no control on the residual

- **GEN**eralized **C**onjugate **G**radient [Axelsson '92]
Mathematically equivalent to GMRESR
- inefficient (2MV per step)
- less stable (uses classical Gram-Schmidt)

Explicit selection
Some methods try to identify converged eigenvector components and remove these eigenvectors from $\mathbf{A}$ by some deflation technique.

- Shift detected eigenvalues ≈ 0 to 1 [Kharchenko & Yeremin '94]
- Deflate eigenvector associated with eigenvalues ≈ 0 (take $\mathbf{u} =$ eigenvector) [Morgan '92]

A Flexible Arnoldi expansion

Expand recursively an orthonormal system $\mathbf{v}_1, \ldots, \mathbf{v}_k$ as follows

> Select a $\mathbf{u}_k$, (e.g., $\mathbf{u}_k \approx \mathbf{A}^{-1}\mathbf{v}_k$)
> $\tilde{\mathbf{v}} \leftarrow \mathbf{A}\,\mathbf{u}_k, \ \vec{h}_k = \mathbf{V}_k^*\,\tilde{\mathbf{v}}, \ \tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec{h}_k$
> $h_{k+1\,k} = \|\tilde{\mathbf{v}}\|, \ \mathbf{v}_{k+1} = \tilde{\mathbf{v}}/h_{k+1\,k}$

Theorem. Then, with $\mathbf{U}_k := [\mathbf{u}_1 \,|\, \ldots \,|\, \mathbf{u}_k]$,

$$\boxed{\mathbf{A}\,\mathbf{U}_k = \mathbf{V}_{k+1}\,\underline{H}_k \ \ \& \ \ \mathbf{V}_{k+1}^*\mathbf{V}_{k+1} = I}$$

Examples.

- $\mathbf{u}_1 = \rho\,\mathbf{r}_0, \ldots, \mathbf{u}_k = \mathbf{v}_k$ (then $\mathbf{U}_k = \mathbf{V}_k$) [Arnoldi '51]
- $\mathbf{u}_1 = \mathbf{v}$ with $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{u}_1 = \rho\,\mathbf{r}_0, \ldots, \mathbf{u}_k = \mathbf{v}_k$ (deflation of eigenvectors) [Morgan '92]
- $\mathbf{u}_k$ solves $\mathbf{K}_k\,\mathbf{u} = \mathbf{v}_k$ [Saad '93]
- mixtures of the above

[Saad '93]

Flexible GMRES

With $\mathbf{r}_0 = \rho\mathbf{v}_1$, use Flexible Arnoldi to get

$$\mathbf{A}\mathbf{U}_k = \mathbf{V}_{k+1}\,\underline{H}_k \text{ with } \mathbf{V}_{k+1}^*\mathbf{V}_{k+1} = I$$

$$\begin{aligned}\|\,\mathbf{r}_k\,\| &= \min_{\vec{\gamma}}\|\,\mathbf{r}_0 - \mathbf{A}\mathbf{U}_k\vec{\gamma}\,\| \\ &= \min_{\vec{\gamma}}\|\,\mathbf{r}_0 - \mathbf{V}_{k+1}\,\underline{H}_k\,\vec{\gamma}\,\| \\ &= \min_{\vec{\gamma}}\|\,\rho\,e_1 - \underline{H}_k\,\vec{\gamma}\,\| \quad (\star)\end{aligned}$$

If $\vec{\gamma}$ solves $(\star)$ then

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{U}_k\vec{\gamma}_k$$

If $\mathbf{K}_k = \mathbf{K}\ \forall k$ then $\mathbf{U}_k = \mathbf{K}^{-1}\mathbf{V}_k$ and

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{K}^{-1}(\mathbf{V}_k\vec{\gamma}_k)$$

Suffices to store the $\mathbf{v}_j$

Convergence minimal residual methods

- Upper bounds
- Accelerations
- Ritz values and eigenvalues
- Polynomial preconditioners

$\mathcal{P}_k = \{$polynomials of degree $\leq k\}$,
$\mathcal{P}_k^1 = \{p \in \mathcal{P}_k \mid p(0) = 1\}$

$\mathbb{A}$ is **diagonizable** if
for some non-singular $n \times n$ matrix $\mathbb{X}$

$$\mathbb{A}\mathbb{X} = \Lambda\mathbb{X} \quad \text{with} \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n). \qquad (1)$$

Then $\quad \mathcal{C}_E := \min \|\mathbb{X}\|\,\|\mathbb{X}^{-1}\|$,
where min over all $\mathbb{X}$ with (**??**),
is the **condition number**
of the **eigen problem** of $\mathbb{A}$.

If $\mathbb{A}$ is normal, i.e. $\mathbb{A}^*\mathbb{A} = \mathbb{A}\mathbb{A}^*$, then $\mathcal{C}_E = 1$.

$\sigma(\mathbb{A}) \;=\; \{\lambda_i\}$, the **spectrum** of $\mathbb{A}$,
$\qquad\;\;=\;$ the set of eigenvalues.

Exact arithmetic

Breakdown, stagnation

If GCR does not break down
and GCR $\mathbf{r}_0$ = GMRES $\mathbf{r}_0$
$\Rightarrow$ GCR $\mathbf{r}_k$ = GMRES $\mathbf{r}_k$

$\|\mathbf{r}_{k+1}\| = \|\mathbf{r}_k\|$ (**stagnation**)
$\Leftrightarrow \rho_k = (\mathbf{r}_k, \mathbb{A}\mathbf{r}_k) = 0$ (for GCR)
$\Leftrightarrow \sigma_{k+1} = (\mathbf{c}_{k+1}, \mathbf{c}_{k+1}) = 0$
(breakdown GCR)

$(\mathbb{A}\mathbf{y}, \mathbf{y}) \neq 0 \quad \forall\, \mathbf{y} \neq 0 \Rightarrow \|\mathbf{r}_{k+1}\| < \|\mathbf{r}_k\| \quad \forall k$
no stagnation

Minimal residuals [GMRES, MINRES, ...]
$\mathbf{r}_k^{\mathsf{MR}} := \mathbf{r}_0 - \mathbb{A}\mathbf{y}_k^{\mathsf{MR}}$ with $\mathbf{y}_k^{\mathsf{MR}} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$
such that
$$\|\mathbf{r}_k^{\mathsf{MR}}\| = \min\{\|\mathbf{r}_0 - \mathbb{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)\}$$
or, equivalently,
$$\mathbf{r}_k^{\mathsf{MR}} \perp \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$$

Orthogonal residuals [FOM, CG, ...]
$\mathbf{r}_k^{\mathsf{OR}} := \mathbf{r}_0 - \mathbb{A}\mathbf{y}_k^{\mathsf{OR}}$ with $\mathbf{y}_k^{\mathsf{OR}} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$
such that
$$\mathbf{r}_k^{\mathsf{OR}} \perp \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$$

[Brown '91]

Theorem. With $s := \|\mathbf{r}_k^{\mathsf{MR}}\|/\|\mathbf{r}_{k-1}^{\mathsf{MR}}\|$,
$$\boxed{\|\mathbf{r}_k^{\mathsf{MR}}\| = \sqrt{1 - s^2}\,\|\mathbf{r}_k^{\mathsf{OR}}\|}$$

Proof.
$\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0) \subset \mathbb{A}\,\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0) \subset \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$

$\mathbf{r}_{k-1}^{\mathsf{MR}} = \mathbf{r}_0 - \mathbf{w} \perp \mathbb{A}\,\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0)$
for some $\mathbf{w} \in \mathbb{A}\,\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0)$

Take $\mathbf{v} \in \mathbb{A}\,\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$
such that $\mathbf{v} \perp \mathbb{A}\,\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0)$, $\|\mathbf{v}\| = 1$.
Take $\mathbf{r} \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$
such that $\mathbf{r} \perp \mathbb{A}\,\mathcal{K}_{k-1}(\mathbb{A}; \mathbf{r}_0)$, $\mathbf{r} \neq 0$.
Observe that $\mathbf{r} = \mathbf{r}_{k-1}^{\mathsf{MR}}$

Then $\mathbf{r}_k^{\mathsf{MR}} = \mathbf{r}_{k-1}^{\mathsf{MR}} - \alpha\,\mathbf{v} \perp \mathbf{v}$: $\quad \alpha = (\mathbf{r}_{k-1}^{\mathsf{MR}}, \mathbf{v})$
$\mathbf{r}_k^{\mathsf{OR}} = \mathbf{r}_{k-1}^{\mathsf{MR}} - \beta\,\mathbf{v} \perp \mathbf{r}$: $\beta = \frac{(\mathbf{r},\mathbf{r})}{(\mathbf{r},\mathbf{v})} = \frac{\|\mathbf{r}_{k-1}^{\mathsf{MR}}\|^2}{\alpha}$

Hence $\quad \|\mathbf{r}_k^{\mathsf{MR}}\|^2 + \alpha^2 = \|\mathbf{r}_{k-1}\|^2$,
$\beta^2 = \|\mathbf{r}_{k-1}^{\mathsf{MR}}\|^2 + \|\mathbf{r}_k^{\mathsf{OR}}\|^2$

For $\mathbf{r}_0$ and GMRES (or GCR)
the residuals $\mathbf{r}_k$ are characterized by

$\mathbf{r}_k \in \mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ and $p_k \in \mathcal{P}_k^1$ are such that
$$\mathbf{r}_k = p_k(\mathbb{A})\mathbf{r}_0 \quad \text{and} \quad \|\mathbf{r}_k\| = \min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\mathbf{r}_0\|.$$

The following holds

$$\begin{aligned}
\|\mathbf{r}_k\| &\leq \left(\min_{p \in \mathcal{P}_{k-\ell}^1} \|p(\mathbb{A})\|\right) \|\mathbf{r}_\ell\| \\
&\leq \mathcal{C}_E \left(\min_{p \in \mathcal{P}_k^1} \max_{\lambda \in \sigma(\mathbb{A})} |p(\lambda)|\right) \|\mathbf{r}_0\| \qquad (*) \\
&\leq \mathcal{C}_E \left(\min_{p \in \mathcal{P}_k^1} \max_{\zeta \in \mathcal{G}} |p(\zeta)|\right) \|\mathbf{r}_0\| \qquad (**)
\end{aligned}$$

(*) if $\mathbb{A}$ diagonizable
(**) if $\sigma(\mathbb{A}) \subset \mathcal{G}$

Minimization problems

Is there a **unique** polynomial p in $\mathcal{P}_k^1$
that solves the minimization problems?

$\min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\mathbf{r}_0\|$	Existence! Uniqueness!	GMRES
$\min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|$	Existence! Uniqueness??	ideal GMRES
$\min_{p \in \mathcal{P}_k^1} \max_{\zeta \in \mathcal{G}} |p(\zeta)|$	Existence! Uniqueness!	

We are approximating the constant pol. $\mathbf{1}$
by polynomials $\zeta q(\zeta)$, where $q \in \mathcal{P}_{k-1}$,
with respect to some norm on the space of polynomials.

$$\min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\mathbf{r}_0\| \leq \left(\min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|\right) \|\mathbf{r}_0\|$$

Choice initial residual

Theorem
$\mathbb{A}$ normal $\Rightarrow \forall\, k\; \exists\, \mathbf{r}_0$ such that "=" [Greenbaum '79]
Counter examples for $\mathbb{A}$ non-normal ['95]

Numerical experiments
- $\forall\, k$, $\exists\, \mathbf{r}_0$ such that "$\approx$".
- $\mathbf{r}_k$ and $\mathbf{r}'_k$ minimal residuals
  with initial residual $\mathbf{r}_0$ and $\mathbf{r}'_0$, resp.
  $\Rightarrow$ for some "small" $\ell \in \mathbb{Z}$ $\quad \|\mathbf{r}_k\| \approx \|\mathbf{r}'_{k-\ell}\|$

(proof for pos. def. case) [S. & van der Sluis '96]

$\mathbb{A}$ pos. def. $\Rightarrow \min_{p \in \mathcal{P}_1^1} \|p(\mathbb{A})\| < 1 \Rightarrow \|\mathbf{r}_{k+1}\| < \|\mathbf{r}_k\|$
$(\|\mathbf{I} - \epsilon\mathbb{A}\| < 1$ if $0 < \varepsilon \ll 1)$

Suppose $\mathbb{A}$ diagonizable (& $\mathcal{C}_E$ not too large)

Then convergence is measured by

$$\min_{p\in\mathcal{P}_k^1}\ \max_{\lambda\in\sigma(\mathbb{A})} |p(\lambda)|$$

For $k\in\mathbb{N}$ and $\mathcal{G}\subset\mathbb{C}$ we put

$$\boxed{\varepsilon_k(\mathcal{G}) := \min\left\{\max_{\zeta\in\mathcal{G}}|p(\zeta)| \mid p\in\mathcal{P}_k^1\right\}}$$

$\mathbb{A}$ diagonizable $\Rightarrow$

$$\boxed{\sigma(\mathbb{A})\subset\mathcal{G} \ \Rightarrow\ \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \le \mathcal{C}_E\,\varepsilon_k(\sigma(\mathbb{A})) \le \mathcal{C}_E\,\varepsilon_k(\mathcal{G})}$$

Of interest if $\mathcal{C}_E$ is not too large.

Suppose $\mathbb{A}$ is normal. Then
$\mathbb{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$ where $\mathbf{x}_i\perp\mathbf{x}_j$, $\|\mathbf{x}_i\|=1 \quad (i\neq j)$.

$$\mathbf{r}_0 = \sum_i \mu_i\,\mathbf{x}_i \quad \text{where} \quad \mu_i = (\mathbf{r}_0,\mathbf{x}_i)$$

and

$$\|p(\mathbb{A})\mathbf{r}_0\| = \|\sum_i p(\lambda_i)\,\mu_i\,\mathbf{x}_i\| = \sqrt{\sum_i |p(\lambda_i)|^2\,|\mu_i|^2}$$

For some weight function ω we have that

$$\sum_i f(\lambda_i)\,|\mu_i|^2 \approx \int f(\lambda)\,\omega(\lambda)\,\mathrm{d}\lambda.$$

Therefore, alternatives

I. $\displaystyle \min_{p\in\mathcal{P}_k^1}\max_{\lambda\in\mathcal{G}}|p(\lambda)|$

II. $\displaystyle \min_{p\in\mathcal{P}_k^1}\sqrt{\int |p(\lambda)|^2\omega(\lambda)\,\mathrm{d}\lambda}$

$$\|\mathbf{r}_k\| \le \mathcal{C}_E\,\min_{p\in\mathcal{P}_k^1}\max_{\zeta\in\mathcal{G}}|p(\zeta)|\,\|\mathbf{r}_0\|$$

Example. Suppose $\sigma(\mathbb{A})\subset[\lambda_-,\lambda_+]\subset(0,\infty)$.

Then, with $\mathcal{C}:=\lambda_+/\lambda_-$, [Meinardus-Kaniel]

$$\boxed{\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \le 2\,\mathcal{C}_E\,\exp\left(-\frac{2k}{\sqrt{\mathcal{C}}}\right)}$$

$$k \ge \frac{1}{2}\sqrt{\mathcal{C}}\,\log\frac{2\,\mathcal{C}_E}{\epsilon} \ \Rightarrow\ \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|}\le\epsilon$$

On $\mathcal{G}:=[\lambda_-,\lambda_+]$,
the shifted and scaled Chebychev polynomial

$$\widetilde{p}_k(\zeta) := \frac{T_k\left(\frac{\lambda_+-\zeta}{\lambda_+-\lambda_-}-\frac{\zeta-\lambda_-}{\lambda_+-\lambda_-}\right)}{T_k\left(\frac{\lambda_++\lambda_-}{\lambda_+-\lambda_-}\right)}$$

solves

$$\min_{p\in\mathcal{P}_k^1}\ \max_{\zeta\in\mathcal{G}}|p(\zeta)| \le \frac{1}{T_k\left(\frac{\lambda_++\lambda_-}{\lambda_+-\lambda_-}\right)}$$

Chebychev polynomials

$T_k\in\mathcal{P}_k$ are defined by

$$\begin{cases} T_0(\zeta)=1, \quad T_1(\zeta)=\zeta \\ T_{k+1}(\zeta) = 2\,\zeta\,T_k(\zeta) - T_{k-1}(\zeta) \quad k=1,2,\ldots \end{cases}$$

or, equivalently,

$$T_k\left(\frac{1}{2}\left(\zeta+\frac{1}{\zeta}\right)\right) = \frac{1}{2}\left(\zeta^k+\frac{1}{\zeta^k}\right).$$

Properties.

- $|T_k|\le 1$ on $[-1,+1]$.
- $p\in\mathcal{P}_k$, $|p|\le 1$ on $[-1,+1]$
  $\Rightarrow$ $|T_k|\ge|p|$ on $\mathbb{R}\backslash[-1,+1]$

[Gauss-Seidel, SOR, ...]

Classical methods

with **relaxation parameter** $\alpha_k=\alpha\in\mathbb{C}$ $(\forall k)$

```
Choose x_0
k = 0
r_0 = b - A x_0
while ||r_k|| > tol do
   solve u_k from K u = r_k
   c_k = A u_k
   x_{k+1} = x_k + α_k u_k
   r_{k+1} = r_k - α_k c_k
   k = k + 1
end while
```

(locally) **M**inimal **R**esidual: $\alpha_k = (\mathbf{r}_k,\mathbf{c}_k)/(\mathbf{c}_k,\mathbf{c}_k)$

Classical methods II

Then $\quad \mathbf{r}_k = (\mathbf{I}-\alpha\,\mathbb{A})^k\mathbf{r}_0$

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \le \|(\mathbf{I}-\alpha\,\mathbb{A})^k\| \le \mathcal{C}_E\max_{\zeta\in\sigma(\mathbb{A})}|1-\alpha\,\zeta|^k$$

Convergence if $\displaystyle \max_{\zeta\in\sigma(\mathbb{A})}|1-\alpha\,\zeta|<1$

Suppose $\quad \sigma(\mathbb{A})\subset[\lambda_-,\lambda_+]. \quad \mathcal{C}=\lambda_+/\lambda_-$.

$\alpha = 2/\left(\lambda_++\lambda_-\right)$ minimizes "max ...". Then

$$\max\ldots = \frac{\lambda_+-\lambda_-}{\lambda_++\lambda_-}$$

$$\boxed{\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \le \mathcal{C}_E\left(\frac{\lambda_+-\lambda_-}{\lambda_++\lambda_-}\right)^k \le \mathcal{C}_E\,\exp\left(-\frac{2k}{\mathcal{C}}\right)}$$

$$k\ge\frac{1}{2}\mathcal{C}\,\log\frac{\mathcal{C}_E}{\epsilon} \ \Rightarrow\ \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|}\le\epsilon$$

$$\frac{\eta+\eta^{-1}}{2} = \tau = \frac{\lambda_+-\zeta}{\lambda_+-\lambda_-} - \frac{\zeta-\lambda_-}{\lambda_+-\lambda_-}$$

$$T_k(\tau) = \frac{\eta^k+\eta^{-k}}{2}$$

$$\tau_0 := \frac{\rho_++\rho_+^{-1}}{2} \ \Rightarrow\ |T_k(\tau_0)| = \frac{\rho_+^k+\rho_+^{-k}}{2}$$

$$|\eta|\le\rho_- \ \Rightarrow\ |T_k(\tau)| \le \frac{\rho_-^k+\rho_-^{-k}}{2}$$

If $\quad |\eta|\le\rho_- \quad$ and $\quad \widetilde{p}_k(\zeta) := \dfrac{T_k(\tau)}{T_k(\tau_0)} \quad$ then

$$|\widetilde{p}_k(\zeta)| \le \frac{\rho_-^k+\rho_-^{-k}}{\rho_+^k+\rho_+^{-k}}$$

Estimates on ellipses

$$\mathcal{E}_\rho := \left\{\zeta \;\middle|\; \frac{|\zeta-\lambda_-|+|\zeta-\lambda_+|}{|\lambda_+-\lambda_-|} \le \frac{\rho+\rho^{-1}}{2}\right\}$$

$$\boxed{\begin{array}{l} \sigma(\mathbb{A})\subset\mathcal{E}_{\rho_-} \quad \& \quad 0\in\partial(\mathcal{E}_{\rho_+}) \quad \Rightarrow \\ \varepsilon_k(\sigma(\mathbb{A})) \le \dfrac{\rho_-^k+\rho_-^{-k}}{\rho_+^k+\rho_+^{-k}} \le 2\left(\dfrac{\rho_-}{\rho_+}\right)^k \end{array}}$$

Observe that $\quad \dfrac{\rho_-}{\rho_+}<1 \ \Leftrightarrow\ 0\notin\mathcal{E}_{\rho_-}$

Examples.

- $\rho_-=1 \ \Rightarrow \mathcal{E}_{\rho_-}=[\lambda_-,\lambda_+]$ & $\quad \dfrac{\rho_-}{\rho_+}\le\dfrac{\sqrt{\mathcal{C}}-1}{\sqrt{\mathcal{C}}+1}$.
- $\lambda_-=\lambda_+=\mu \quad \& \quad \mathcal{E}_{\rho_-}=\{\zeta \mid |\zeta-\mu|\le\rho_0\}$
  $\Rightarrow \quad \dfrac{\rho_-}{\rho_+}=\dfrac{\rho_0}{|\mu|}$

Estimating accelerations

$$\boxed{\begin{aligned}&\sigma(\mathbb{A})\subset\{\lambda_0\}\cup\mathcal{E}_{\rho_-} \;\;\&\;\; 0\in\partial\mathcal{E}_{\rho_+} \;\Rightarrow\\ &\varepsilon_{k+1}(\sigma(\mathbb{A}))\leq 2\left(\frac{\rho_-}{\rho_+}\right)^k \max_{\zeta\in\mathcal{E}_{\rho_-}}\left|1-\frac{\zeta}{\lambda_0}\right|\end{aligned}}$$

Proof.
If $\sigma(\mathbb{A})\subset\{\lambda_0\}\cup\mathcal{E}_{\rho_-}$ and $0\notin\mathcal{E}_{\rho_-}$ then, with

$$\tilde{p}_{k+1}(\zeta):=\left(1-\frac{\zeta}{\lambda_0}\right)\frac{T_k(\tau)}{T_k(\tau_0)}$$

we have

$$|\tilde{p}_{k+1}|\leq 2\left(\frac{\rho_-}{\rho_+}\right)^k\max_{\zeta\in\mathcal{E}_{\rho_-}}\left|1-\frac{\zeta}{\lambda_0}\right| \quad \text{on} \quad \sigma(\mathbb{A})$$

$$\|p(\mathbb{A})\|\leq\mathcal{C}_E\max_{\lambda\in\sigma(\mathbb{A})}|p(\lambda)|$$

For cases where
- $\mathbb{A}$ is not diagonizable or
- $\mathcal{C}_E$ is extremely large:

For $\varepsilon>0$

$$\begin{aligned}\sigma_\varepsilon(\mathbb{A}) &:= \left\{\zeta\in\mathbb{C}\mid \|(\zeta\,\mathbf{I}-\mathbb{A})^{-1}\|^{-1}\leq\varepsilon\right\}\\ &= \bigcup_{\|\Delta\|\leq\varepsilon}\sigma(\mathbb{A}+\Delta)\end{aligned}$$

is the set of ε-**pseudo-eigenvalues** of $\mathbb{A}$.

[Trefethen '90]

Theorem. For arc $\mathcal{C}$ enclosing $\sigma_\varepsilon(\mathbb{A})$

$$\boxed{\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|}\leq\frac{\text{length}(\mathcal{C})}{2\pi\varepsilon}\,\varepsilon_k(\sigma_\varepsilon(\mathbb{A}))}$$

Proof. $p_k(\mathbb{A})\mathbf{r}_0=\frac{1}{2\pi i}\oint(\zeta\,\mathbf{I}-\mathbb{A})^{-1}\mathbf{r}_0\,p_k(\zeta)\,\mathrm{d}\zeta.$

Holomorphic functions

Consider a holomorphic function f on some open neighborhood $\mathcal{U}\subset\mathbb{C}$ of $\sigma(\mathbb{A})$.
$\mathcal{C}$ arc enclosing $\sigma(\mathbb{A})$, counterclockwise oriented
smooth simple
closed curve
in $\mathcal{U}$

$$f(\lambda)=\frac{1}{2\pi i}\oint_{\mathcal{C}}\frac{f(\zeta)}{\zeta-\lambda}\,\mathrm{d}\zeta.$$

$\sigma(\mathbb{A})=\{*\}$

Matrix functions

For f holomorphic on some open neighborhood $\mathcal{U}$ of $\sigma(\mathbb{A})$, and arc $\mathcal{C}$ such that ..., define

$$\boxed{f(\mathbb{A}):=\frac{1}{2\pi i}\oint_{\mathcal{C}}(\zeta\,\mathbf{I}-\mathbb{A})^{-1}f(\zeta)\,\mathrm{d}\zeta}$$

$$(f(\mathbb{A})\mathbf{x},\mathbf{y})=\frac{1}{2\pi i}\oint_{\mathcal{C}}((\zeta\,\mathbf{I}-\mathbb{A})^{-1}\mathbf{x},\mathbf{y})f(\zeta)\,\mathrm{d}\zeta$$

Properties. Linear in f.
$f(\zeta)=\zeta^j\Rightarrow f(\mathbb{A})=\mathbb{A}^j$.
$g(\mathbb{A})f(\mathbb{A})=f(\mathbb{A})g(\mathbb{A})=fg(\mathbb{A})$.
$(g\circ f)(\mathbb{A})=g(f(\mathbb{A}))$.

$$\|f(\mathbb{A})\|\leq\frac{\text{length}\,\mathcal{C}}{2\pi}\max_{\zeta\in\mathcal{C}}\|(\zeta\,\mathbf{I}-\mathbb{A})^{-1}\|\max_{\zeta\in\mathcal{C}}|f(\zeta)|$$

$\mathbb{A}$ normal $\quad \mathbb{A}=\int\zeta\,\mathrm{d}P_\eta$ spectral decomposition
$f(\mathbb{A})=\int f(\zeta)\,\mathrm{d}P_\zeta$. spectral measure

[Greenbaum, Strakos, Ptak '96]

$$\|\mathbf{r}_k\|\leq\mathcal{C}_E\min_{p\in\mathcal{P}_k^1}\max_{\lambda\in\sigma(\mathbb{A})}|p(\lambda)|\,\|\mathbf{r}_0\|$$

Do the eigenvalues tell it all?

Given $\quad \lambda_1,\ldots,\lambda_n\in\mathbb{C}$
Given $\quad \phi_1\geq\phi_2\geq\ldots\geq\phi_{n-1}$

Theorem. There is an $n\times n$ irreducible Hessenberg matrix $\mathbf{H}$ and an n-vector $\mathbf{r}_0$ such that for the GMRES residuals $\mathbf{r}_k$, with initial residual $\mathbf{r}_0$, we have

$$\sigma(\mathbf{H})=\{\lambda_1,\ldots,\lambda_n\} \quad \& \quad \|\mathbf{r}_k\|=\phi_k$$

Any monotonic decreasing graph can be the convergence history of a minimal residual method for a problem with any eigenvalue distribution

[Greenbaum, Strakos '93]

Stagnation for any eigenvalue distribution

$$\mathbf{H}=\begin{bmatrix}\alpha_1 & \alpha_2 & \ldots & \ldots & \alpha_{n-1} & \alpha_n\\ 1 & 0 & \ldots & \ldots & 0 & 0\\ & 1 & \ddots & & & \\ & & \ddots & \ddots & \vdots & \vdots\\ & & & \ddots & 0 & 0\\ 0 & & & & 1 & 0\end{bmatrix}$$

$\alpha_n\neq 0$

$\sigma(\mathbf{H})=\{\lambda\mid\lambda^n-\alpha_1\lambda^{k-1}-\ldots-\alpha_n=0\}$

Solve $\mathbf{H}\,\tilde{\mathbf{y}}=e_n$ with GMRES
and initial residual $\mathbf{r}_0:=e_n$

Note that $\mathbf{H}\,e_n=\alpha_n\,e_1$.
Therefore $\mathcal{K}_k(\mathbf{H};\mathbf{H}\,\mathbf{r}_0)=\text{span}(e_1,\ldots,e_k)$.
Hence $\mathbf{r}_k=e_n=\mathbf{r}_0 \quad \forall\,k<n$.

Hessenberg matrices tell it all

If $\rho\,\mathbf{A}\,\mathbf{r}_0=\mathbf{c}_0,\mathbf{c}_1,\ldots$ orthonormal such that

$$\mathcal{K}_k(\mathbf{A},\mathbf{A}\,\mathbf{r}_0)=\text{span}(\mathbf{c}_0,\ldots,\mathbf{c}_{k-1}) \quad \forall\,k$$

then, with $\mathbf{V}:=[\mathbf{c}_0|\ldots|\mathbf{c}_{n-1}]$,

$$\boxed{\mathbf{A}\,\mathbf{V}=\mathbf{H}\,\mathbf{V}^*,\quad \mathbf{V}^*\mathbf{V}=\mathbf{I},\quad \mathbf{H}\text{ Hessenberg}}$$

$\mathbf{A}\,\mathbf{y}=\mathbf{r}_0^{\mathbf{A}}=\mathbf{r}_0\Leftrightarrow$

$\mathbf{H}\,\mathbf{y}^{\mathbf{H}}=\mathbf{r}_0^{\mathbf{H}}:=\mathbf{V}^*\,\mathbf{r}_0^{\mathbf{A}} \quad \& \quad \mathbf{y}^{\mathbf{A}}=\mathbf{V}\,\mathbf{y}^{\mathbf{H}}$
then $\|\mathbf{r}_k^{\mathbf{A}}\|=\|\mathbf{r}_k^{\mathbf{H}}\|$, since $\mathbf{V}^*\mathbf{V}=\mathbf{I}$

Unitary and pos.def. tell it all

$\mathbf{H}$ Hessenberg, $\mathbf{r}_k^{\mathbf{H}}$ residuals of GMRES applied to $\mathbf{H}\,\tilde{\mathbf{y}}=\tilde{\mathbf{r}}_0$ with initial residual $\tilde{\mathbf{r}}_0$.
$\mathbf{B}$ Hessenberg, $\mathbf{r}_k^{\mathbf{B}}$ residuals of GMRES applied to $\mathbf{B}\,\tilde{\mathbf{y}}=\tilde{\mathbf{r}}_0$ with initial residual $\tilde{\mathbf{r}}_0$.

$$\boxed{\text{When is } \mathbf{r}_k^{\mathbf{H}}=\mathbf{r}_k^{\mathbf{B}} \text{ for all } k?}$$

Theorem. Iff $\quad \mathbf{B}=\mathbf{U}\mathbf{H}$
for some non-singular upper triangular $\mathbf{U}$.
Proof. Since $\mathbf{H}\,\tilde{\mathbf{r}}_0=e_1$ we have
$\mathcal{K}_k(\mathbf{H};\mathbf{H}\,\tilde{\mathbf{r}}_0)=\text{span}(e_1,\ldots,e_k)=\mathcal{K}_k(\mathbf{U}\mathbf{H};\mathbf{U}\mathbf{H}\,\tilde{\mathbf{r}}_0)$

Applications
- $\mathbf{H}=\mathbf{U}\mathbf{Q}$ for some $\mathbf{Q}$, $\mathbf{Q}^*\mathbf{Q}=\mathbf{I}$.
- $(\mathbf{H}\mathbf{y},\mathbf{y})\neq 0 \quad \forall\,\mathbf{y}\neq 0$
  $\Rightarrow \mathbf{H}=\mathbf{U}\mathbf{L}=(\mathbf{U}\mathbf{L}^{-*})\mathbf{L}^*\mathbf{L}$
  upper triang.   hermitean pos. def.

[van der Vorst, Vuik '93, vd V, van der Sluis '86, S. '90]

$\mathbf{r}_k$ GMRES residuals,

$\mathbb{A}$ diagonizable $\Rightarrow$ $\boxed{\dfrac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \leq \mathcal{C}_E\, \varepsilon_k(\sigma(\mathbb{A}))}$

Super-linear convergence I

p_k GMRES pol: $p_k \in \mathcal{P}_k^1$ s.t. $\mathbf{r}_k = p_k(\mathbb{A})\mathbf{r}_0$.

Choose $\lambda \in \sigma(\mathbb{A})$, $\ell \in \mathbb{N}$ and a zero θ of p_ℓ.

Put $$F := \max_{\zeta \in \sigma(\mathbb{A})} \left|\frac{1-\zeta/\lambda}{1-\zeta/\theta}\right|$$

Note that $\theta \approx \lambda \Rightarrow F \approx 1$

$\mathbb{A}$ diagonizable $\Rightarrow$

$$\boxed{\frac{\|\mathbf{r}_{k+\ell}\|}{\|\mathbf{r}_\ell\|} \leq \mathcal{C}_E\, F\, \varepsilon_k(\sigma(\mathbb{A})\backslash\{\lambda\})}$$

Projections

Consider a $\lambda \in \sigma(\mathbb{A})$.

Suppose λ is **semi-simple**, i.e.

$$\mathrm{Kern}(\mathbb{A} - \lambda\mathbf{I}) \cap \mathrm{Im}(\mathbb{A} - \lambda\mathbf{I}) = \{\mathbf{0}\}.$$

$$\underset{\mathbf{P}_\lambda}{\mathbb{C}^n = \mathrm{Kern}(\mathbb{A} - \lambda\mathbf{I})} \oplus \underset{\mathbf{I}-\mathbf{P}_\lambda}{\mathrm{Im}(\mathbb{A} - \lambda\mathbf{I})} \quad \text{projections}$$

$$\mathbf{P}_\lambda\,\mathbb{A} = \mathbb{A}\,\mathbf{P}_\lambda, \qquad \mathbb{A} - \lambda\,\mathbf{I} = (\mathbb{A} - \lambda\,\mathbf{I})(\mathbf{I} - \mathbf{P}_\lambda)$$

Suppose λ **simple**, i.e. dim Kern$(\mathbb{A} - \lambda\,\mathbf{I}) = 1$.

$\mathbf{x}$ and $\mathbf{y}$, $\neq \mathbf{0}$, s.t. $\mathbb{A}\mathbf{x} = \lambda\mathbf{x}$ & $\mathbf{y}^*\mathbb{A} = \lambda\,\mathbf{y}^*$

$$\Rightarrow \quad \mathbf{P}_\lambda = \frac{\mathbf{x}\,\mathbf{y}^*}{\mathbf{y}^*\mathbf{x}}$$

If $\mathbb{A}$ normal then $\mathbf{y} = \mathbf{x}$

$$\mathbf{P}_\lambda = \frac{1}{2\pi i}\oint_{\mathcal{C}} (\zeta\,\mathbf{I} - \mathbb{A})^{-1} f(\zeta)\,d\zeta \quad \text{with } f \text{ such that}$$

$f(\lambda) = 1$, $f(\zeta) = 0$ all $\zeta \in \sigma(\mathbb{A})$, $\zeta \neq \lambda$

$\mathbf{r}_k$ GMRES res. $\quad p_k \in \mathcal{P}_k^1$ s.t. $\mathbf{r}_k = p_k(\mathbb{A})\mathbf{r}_0$.

$\lambda \in \sigma(\mathbb{A})$, $\quad \ell \in \mathbb{N}$, $\quad p_\ell(\theta) = 0$,

$$F := \max_{\zeta \in \sigma(\mathbb{A})} \left|\frac{1-\zeta/\lambda}{1-\zeta/\theta}\right|$$

Super-linear convergence II

Suppose λ semi-simple.

$$\mathbb{A}' : \mathrm{Im}(\mathbb{A} - \lambda\,\mathbf{I}) \xrightarrow{\mathbb{A}} \mathrm{Im}(\mathbb{A} - \lambda\,\mathbf{I})$$

$$\mathbb{A}' := \mathbb{A}_{|\mathrm{Im}(\mathbb{A}-\lambda\,\mathbf{I})} = (\mathbf{I} - \mathbf{P}_\lambda)\mathbb{A}_{|\mathrm{Im}(\mathbb{A}-\lambda\,\mathbf{I})}$$

$\Rightarrow \sigma(\mathbb{A}') = \sigma(\mathbb{A})\backslash\{\lambda\}$ & $\mathcal{C}_E(\mathbb{A}') \leq \mathcal{C}_E(\mathbb{A}) := \mathcal{C}_E$

Comparison process

$\mathbf{r}'_k$ is the kth GMRES residual for $\mathbb{A}'$

with $\mathbf{r}'_0 := (\mathbf{I} - \mathbf{P}_\lambda)\mathbf{r}_\ell$

$$\Rightarrow \quad \boxed{\|\mathbf{r}_{k+\ell}\| \leq \mathcal{C}_E\, F\, \|\mathbf{r}'_k\|}$$

Proof. $\quad \widetilde{h}(\zeta) := \dfrac{1-\zeta/\lambda}{1-\zeta/\theta}, \qquad h := \widetilde{h}\, p_\ell$

Then

$$h \in \mathcal{P}_{k+\ell}^1, \quad \widetilde{h}(\lambda) = 0, \quad F = \max_{\zeta\in\sigma(\mathbb{A})} |\widetilde{h}(\zeta)|.$$

$$\mathbb{B} := \widetilde{h}(\mathbb{A}) = (\mathbf{I} - \mathbb{A}/\theta)^{-1}(\mathbf{I} - \mathbb{A}/\lambda)$$

$$\Rightarrow \mathbb{B} = \mathbb{B}\,(\mathbf{I} - \mathbf{P}_\lambda) = \widetilde{h}(\mathbb{A}')$$

For $\quad q_k \in \mathcal{P}_k^1$, $\quad$ we have $\quad q_k h \in \mathcal{P}_{k+\ell}^1 \quad$ and

$$\begin{aligned}\|\mathbf{r}_{k+l}\| \leq \|q_k(\mathbb{A})h(\mathbb{A})\mathbf{r}_0\| &= \|\mathbb{B}\,q_k(\mathbb{A})\,p_\ell(\mathbb{A})\,\mathbf{r}_0\| \\ &= \|\mathbb{B}\,q_k(\mathbb{A}')\,(\mathbf{I} - \mathbf{P}_\lambda)\,p_\ell(\mathbb{A})\,\mathbf{r}_0\|\end{aligned}$$

I. $$\begin{aligned}\|\mathbf{r}_{k+\ell}\| &\leq \|\mathbb{B}\,q_k(\mathbb{A})\|\,\|\mathbf{r}_\ell\| \\ &\leq \mathcal{C}_E \max\nolimits_{\zeta\in\sigma(\mathbb{A})} |\widetilde{h}(\zeta)q_l(\zeta)|\,\|\mathbf{r}_\ell\| \\ &\leq \mathcal{C}_E \max_{\zeta \in \sigma(\mathbb{A})\backslash\{\lambda\}} |\widetilde{h}(\zeta)q_l(\zeta)|\,\|\mathbf{r}_\ell\|\end{aligned}$$

Now, take the minimum over all $q_k \in \mathcal{P}_k^1$.

II. Take q_k such that $\mathbf{r}'_k = q_k(\mathbb{A}')\,\mathbf{r}'_0$:

$$\|\mathbf{r}_{k+\ell}\| \leq \|\mathbb{B}\|\,\|\mathbf{r}'_k\| \leq \mathcal{C}_E \max_{\zeta \in \sigma(\mathbb{A})} |\widetilde{h}(\zeta)|\,\|\mathbf{r}'_k\|$$

Accelerations and Ritz values

$\lambda \in \sigma(\mathbb{A})$, $\ell \in \mathbb{N}$ and θ Ritz value order ℓ.

Put $$F := \max_{\zeta \in \sigma(\mathbb{A})} \left|\frac{1-\zeta/\lambda}{1-\zeta/\theta}\right|$$

$\mathbb{A}$ diagonizable, $s = \|\mathbf{r}_\ell\|/\|\mathbf{r}_{\ell-1}\| \Rightarrow$

$$\boxed{\frac{\|\mathbf{r}_{k+\ell}\|}{\|\mathbf{r}_\ell\|} \leq \mathcal{C}_E\, \frac{F}{\sqrt{1-s^2}}\, \varepsilon_k(\sigma(\mathbb{A})\backslash\{\lambda\})}$$

Comparison process. Suppose λ semi-simple.

$\mathbf{r}'_k$ is the kth GMRES residual for $\mathbb{A}'$ with

$\mathbf{r}'_0 := (\mathbf{I} - \mathbf{P}_\lambda)\,\mathbf{r}_\ell^{\mathrm{OR}}$, where $\mathbf{r}_\ell^{\mathrm{OR}}$ FOM residual

$$\Rightarrow \quad \boxed{\|\mathbf{r}_{k+\ell}\| \leq \mathcal{C}_E\, F\, \|\mathbf{r}'_k\|}$$

Proof. $\widetilde{h}(\zeta) := \frac{1-\zeta/\lambda}{1-\zeta/\theta}$, $\quad h := \widetilde{h}p_\ell^{\mathrm{OR}}$, $\quad \|\mathbf{r}_\ell^{\mathrm{OR}}\| \leq \frac{1}{\kappa_\ell}\|\mathbf{r}_\ell\|$

Why approximating eigenvalues?

- If the physical problem requires the computation of eigenvalues.

- For linear problems, approximate eigenvalues provide information on the spectrum of $\mathbf{A}$ that can be used
  - to assess the effect of the preconditioner,
  - to construct an effective preconditioner,

- In the context of PDEs, approximate eigenvalues can be used
  - to analyze stability of integration method,
  - ...

Ritz-values (approximate eigenvalues) can cheaply be computed as a side-product of certain Krylov subspace methods as GCR, GMRES, ..., (Bi-CG, Bi-CGSTAB,...).

Approximate eigenvalues

If $(\theta, \mathbf{u})$ is an **approximate eigenpair**,

i.e. $\theta \in \mathbb{C}$ and $\mathbf{u} \in \mathbb{C}^n$,

then $\mathbf{r} := \mathbf{A}\mathbf{u} - \theta\mathbf{u}$ is the **residual**.

Theorem. If $\mathbf{A}$ is diagonalizable $\qquad$ [Bauer-Fike]

then there is an eigenvalue λ of $\mathbf{A}$ such that

$$\boxed{|\lambda - \theta| \leq \mathcal{C}_E\, \frac{\|\mathbf{r}\|}{\|\mathbf{u}\|}}$$

Theorem. If $(\mathbf{A} + \Delta_A)\mathbf{u} = \theta\,\mathbf{u}$ then

for some eigenvalue λ with

right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}$,

i.e. $\qquad \mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}$ and $\mathbf{y}^*\,\mathbf{A} = \lambda\,\mathbf{y}^*$

Then $$\boxed{|\lambda - \theta| \lesssim \|\Delta_A\|\, \frac{\|\mathbf{x}\|\,\|\mathbf{y}\|}{(\mathbf{x}, \mathbf{y})}}$$

Petrov values

Consider a (skew) projection Π on a subspace $\mathcal{V}$ of $\mathbb{C}^n$ (typically $\dim(\mathcal{V}) \ll n$):

$$\Pi(\mathbf{z}) \in \mathcal{V} \quad (\mathbf{z} \in \mathbb{C}^n), \qquad \Pi(\mathbf{w}) = \mathbf{w} \quad (\mathbf{w} \in \mathcal{V})$$

$\theta \in \mathbb{C}$ is a **Petrov value** of $\mathbf{A}$ w.r.t. Π

with **Petrov vector** $\mathbf{u} \in \mathbb{C}^n$ if

$$\mathbf{u} \in \mathcal{V} \quad \text{and} \quad \Pi(\mathbf{A}\mathbf{u} - \theta\mathbf{u}) = 0,$$

or, equivalently,

$$(\theta, \mathbf{u}) \text{ is an eigenpair of } \mathbf{A}_\Pi := \Pi\mathbf{A}_{|\mathcal{V}}$$

If the columns of $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ span $\mathcal{V}$

and H_Π is the matrix of $\mathbf{A}_\Pi$ w.r.t. $\mathbf{v}_1, \ldots, \mathbf{v}_k$

then $(\theta, \mathbf{V}\,\vec{\gamma})$ is a **Petrov pair** $\Leftrightarrow H_\Pi\,\vec{\gamma} = \theta\,\vec{\gamma}$

Practical computation.

Compute H_Π, solve the eigenproblem for H_Π with, for instance, the QR algorithm.

Improved eigenvector approximation

For θ Petrov value, find $\tilde{\mathbf{u}} \in \mathcal{V}$ that minimizes

$$\|\mathbf{A}\tilde{\mathbf{u}} - \theta\,\tilde{\mathbf{u}}\|.$$

$\tilde{\mathbf{u}}$ is the smallest singular vector of $\mathbf{A}_{|\mathcal{V}}$.

The quality of Petrov values

Consider $\qquad \mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}$

Is there a Petrov value θ close to λ?

Depends on

- $\angle(\mathbf{x}, \mathcal{V})$
- $\mathcal{C}_E$
- $\|\Pi\|$

If Π is **orthogonal**,

i.e. $\qquad \mathbf{z} - \Pi(\mathbf{z}) \perp \mathcal{V} \qquad (\mathbf{z} \in \mathbb{C}^n)$,

then $\|\Pi\| = 1$ and the Petrov pair is a **Ritz pair**

Examples

For the circular matrix,

$$\text{for } n = 4: \qquad \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$\mathcal{V} = \text{span}(e_1, \ldots, e_k)$ gives bad approximation for $k < n$.

For any matrix,

$$\Pi(\textstyle\sum_j \mu_j\,\mathbf{v}_j) := \sum_{j<k} \mu_j\,\mathbf{v}_j$$

with $\mathbf{v}_1, \mathbf{v}_2 := \mathbf{A}\mathbf{v}_1, \mathbf{v}_3 := \mathbf{A}\mathbf{v}_2, \ldots$,

gives bad approximations for $k < n$:

$$H_\Pi(e_i) = e_{i+1}$$

Ritz values

Consider a subspace $\mathcal{V}$ of $\mathbb{C}^n$ ($\dim(\mathcal{V}) \ll n$).

Proposition. For a $\theta \in \mathbb{C}$ and a $\mathbf{u} \in \mathcal{V}$ TFAE

(a) $\mathbf{A}\mathbf{u} - \theta\,\mathbf{u} \perp \mathcal{V}$;

(b) $(\theta, \mathbf{u})$ is a Petrov pair w.r.t. Π, where

Π is the **orthogonal** projection on $\mathcal{V}$,

that is, $\Pi(z) \in \mathcal{V}$ and $\mathbf{z} - \Pi(\mathbf{z}) \perp \mathcal{V}$ $(\mathbf{z} \in \mathbb{C}^n)$.

A **Ritz pair** of $\mathbf{A}$ w.r.t. $\mathcal{V}$ is a pair $(\theta, \mathbf{u})$ with one of the above properties.

If the columns of $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ form an orthonormal basis of $\mathcal{V}$ then

$$\Pi = \mathbf{V}\mathbf{V}^* \quad \text{and} \quad H_\Pi := H_\mathcal{V} := \mathbf{V}^*\mathbf{A}\mathbf{V}$$

$(\theta, \mathbf{u})$ Ritz-pair $\quad \Leftrightarrow \quad H_\mathcal{V}\,\vec{\gamma} = \theta\,\vec{\gamma}$ & $\mathbf{u} = \mathbf{V}\,\vec{\gamma}$

The quality of Ritz values

Consider $\qquad \mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x} \quad$ with $\quad \|\mathbf{x}\| = 1$

Write

$\mathbf{x} = \mathbf{u} + \mathbf{z}$ with $\mathbf{u} \in \mathcal{V}$, $\mathbf{z} \perp \mathcal{V}$.

Now, $\sin\phi_x = \text{sinus}\,\angle\,(\mathbf{x}, \mathcal{V}) = \|\mathbf{z}\|$.

With $\mathbf{A}_\mathcal{V} := \Pi\mathbf{A}_{|\mathcal{V}}$,

$$\begin{aligned} \|(\mathbf{A}_\mathcal{V} - \lambda)\mathbf{u}\| &= \|\mathbf{A}\,\mathbf{x} - \mathbf{A}\,\mathbf{z} - \lambda\,\mathbf{u}\| \\ &= \|(\lambda - \mathbf{A})\mathbf{z}\| \le 2\|\mathbf{A}\|\sin\phi_x. \end{aligned}$$

For certain Ritz value θ w.r.t. $\mathcal{V}$ we have

$$\boxed{|\theta - \lambda| \le \mathcal{C}_E\, 2\,\|\mathbf{A}\|\,\tan\phi_x}$$

Ritz values on Krylov subspaces

Consider Ritz values on $\mathcal{V} = \mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$.

Also the power method takes approximate eigenpairs from $\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$. For this method we know that

$\tan\phi_x \ll 1$ for "non-large" k

and for dominant $\mathbf{x}$.

However, for all $\zeta \in \mathbb{C}$,

$\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1) = \text{span}\big(\mathbf{v}_1, (\mathbf{A} - \zeta)\mathbf{v}_1, \ldots, (\mathbf{A} - \zeta)^{k-1}\mathbf{v}_1\big)$.

Therefore,

$$\boxed{\tan\phi_x \ll 1 \text{ for "non-large" } k \text{ and extremal } \lambda}$$

For normal matrices

Suppose $\mathbf{A}$ is diagonizable:

$\mathbf{A}\,\mathbf{x}_i = \lambda_i\,\mathbf{x}_i, \quad \mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, \ldots \quad$ basis

Write $\qquad \mathbf{v}_1 = \sum_{i=1}^n \mu_i\,\mathbf{x}_i \quad$ for certain $\mu_i \in \mathbb{C}$.

Then $\qquad \mathcal{K}_k(\mathbf{A}; \mathbf{v}_1) = \left\{ \sum_{i=1}^n \mu_i\,p(\lambda_i)\,\mathbf{x}_i \mid p \in \mathcal{P}_k \right\}$

Consider $\quad p \in \mathcal{P}_k, \quad \mu_1\,p(\lambda) = 1 \quad (\lambda = \lambda_1)$.

Then $\qquad \sin\angle(\mathbf{x}, \mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)) \le \|\mathbf{r}\|$

where $\quad \mathbf{r} := \sum_{i=2}^n \mu_i\,p(\lambda_i)\,\mathbf{v}_i$.

Therefore,
an upper bound of $\sin\angle$ is proportional to

$$\boxed{\min\{\max_{\substack{\zeta \in \sigma(\mathbf{A}) \\ \zeta \neq \lambda}} |p(\zeta)| \mid p \in \mathcal{P}_k, \;\; \mu_1\,p(\lambda) = 1\}}$$

For symmetric matrices

$$\mathbf{A}^* = \mathbf{A}, \qquad \lambda = \lambda_1 < \lambda_2 < \ldots < \lambda_n.$$

$$\tau := \tan\angle(\mathbf{x}, \mathbf{v}_1), \qquad \beta := \frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_1}$$

Gap ratio

[Kaniel-Paige-Saad]

Then, for the smallest Ritz value θ of order k,

$$\boxed{0 \le \frac{\theta - \lambda}{\lambda_n - \lambda_1} \le \tau^2 \exp(-[4k + 2]\sqrt{\beta})}$$

Projections: examples

Example I

If the columns of $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_k]$ span $\mathcal{V}$
& the columns of $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$ span $\mathcal{W}$
then $\Pi := \mathbf{V}(\mathbf{W}^*\mathbf{V})^{-1}\mathbf{W}^*$
is a projection on $\mathcal{V}$ orthogonal to $\mathcal{W}$:

$$\Pi(\mathbf{z}) \in \mathcal{V} \quad \& \quad \mathbf{z} - \Pi(\mathbf{z}) \perp \mathcal{W} \quad (\mathbf{z} \in \mathbb{C}^n)$$

Then $H_\Pi = (\mathbf{W}^*\mathbf{V})^{-1}\mathbf{W}^*\mathbf{A}\mathbf{V}$.

Example II

If $\mathbf{v}_1, \mathbf{v}_2, \dots$ span $\mathbb{C}^n$ then

$$\Pi_k(\mathbf{z}) := \sum_{i=1}^{k} \mu_i \mathbf{v}_i \quad \text{for all} \quad \mathbf{z} = \sum_{i=1}^{n} \mu_i \mathbf{v}_i \in \mathbb{C}^n,$$

is a projection onto $\mathcal{V} := \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$.
The $k \times k$ left upper block
of the matrix of $\mathbf{A}$ w.r.t. the $\mathbf{v}_j$ is
H_Π, the matrix of $\mathbf{A}_\Pi$ w.r.t. $\mathbf{v}_1, \dots, \mathbf{v}_k$.

Krylov basis

$\mathbf{v}_1, \mathbf{v}_2, \dots$ Krylov basis, i.e. for all k

$$\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1) = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$$

Then

$$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\dots|\mathbf{v}_k]}$$

$\underline{H}_k$ is an $(k+1) \times k$ upper Hessenberg matrix:

$$\underline{H}_k = \begin{bmatrix} \star & \star & \star & & \\ \star & \star & \star & & \\ 0 & \star & \star & & \\ \vdots & \vdots & \ddots & \ddots & \\ & & \ddots & \star & \star \\ & & & 0 & \star \end{bmatrix}$$

H_k is the $k \times k$ upper block of $\underline{H}_k$;
H_k is upper Hessenberg.

Proposition.

H_k irreducible $\Leftrightarrow$ $\dim \mathcal{K}_k(\mathbf{A}; s_0) = k$.

$\mathbf{v}_1, \mathbf{v}_2, \dots$ Krylov basis

$$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\dots|\mathbf{v}_k]}$$

H_k is the $k \times k$ upper block of $\underline{H}_k$.

Projections on Krylov subspaces

Π_k is the projection onto $\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$ given by

$$\Pi_k(\mathbf{z}) := \sum_{i=1}^{k} \mu_i \mathbf{v}_i \quad \text{for all } \mathbf{z} = \sum_{i=1}^{n} \mu_i \mathbf{v}_i \in \mathbb{C}^n.$$

Put $\mathbf{A}_k := \Pi_k \mathbf{A}_{|\mathcal{K}_k(\mathbf{A};\mathbf{v}_1)}$

Proposition. H_k is the matrix of $\mathbf{A}_k$ with respect to the basis $\mathbf{v}_1, \dots, \mathbf{v}_k$.
$(\theta, \mathbf{V}_k\vec{\gamma})$ is a Petrov pair of $\mathbf{A}$ w.r.t. Π
$\Leftrightarrow$ $(\theta, \vec{\gamma})$ is an eigenpair of H_k.

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

Petrov values on Krylov subspaces

Consider a Petrov pair $(\theta, \mathbf{u} := \mathbf{V}_k\vec{\gamma})$:

$$H_k\,\vec{\gamma} = \theta\,\vec{\gamma}.$$

Then, with $\vec{\gamma} = (\gamma_1, \dots, \gamma_k)^T$, $\underline{H}_k = (h_{ij})$,

$$\boxed{\begin{aligned} \mathbf{A}\mathbf{u} - \theta\,\mathbf{u} &= \mathbf{V}_{k+1}\,(\underline{H}_k\vec{\gamma} - \theta\,\vec{\gamma}) \\ &= h_{k+1\,k}\,\mathbf{v}_{k+1}\,\gamma_k \end{aligned}}$$

For certain eigenvalue λ of $\mathbf{A}$ we have

$$\boxed{|\,\theta - \lambda\,| \le \mathcal{C}_E\,|\,h_{k+1\,k}|\,|\gamma_k|\,\frac{\|\mathbf{v}_{k+1}\|}{\|\mathbf{u}\|}}$$

The quantities $h_{k+1\,k}$, γ_k, $\|\mathbf{v}_k\|$ are explicitly known and a lower bound for $\|\mathbf{u}\|$ can (easily) be estimated.

$\mathbf{v}_1, \mathbf{v}_2, \dots$ Krylov basis

$$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\dots|\mathbf{v}_k]}$$

H_k is the $k \times k$ upper block of $\underline{H}_k$.

Π_k is the projection onto $\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$ given by

$$\Pi_k(\mathbf{z}) := \sum_{i=1}^{k} \mu_i \mathbf{v}_i \quad \text{for all} \quad \mathbf{z} = \sum_{i=1}^{n} \mu_i \mathbf{v}_i \in \mathbb{C}^n.$$

Ritz values on Krylov subspaces

Proposition. If $\mathbf{v}_1, \dots$ are orthogonal then Π_k is orthogonal and $(\theta, \mathbf{V}_k\vec{\gamma})$ is a Ritz pair of $\mathbf{A}$ w.r.t. $\mathcal{K}_k(\mathbf{A}, \mathbf{v}_1)$ $\Leftrightarrow$ $(\theta, \vec{\gamma})$ is an eigenpair of H_k.

Examples. From Arnoldi, GMRES, GCR.

[Arnoldi '54]

Arnoldi for eigenvalue computation

Select a $\mathbf{v}_1$.

For $k = 1, 2, \dots$

- use Arnoldi's expansion to get
  $\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$ with $\mathbf{V}_{k+1}^*\mathbf{V}_{k+1} = I$
- compute all eigenpairs $(\theta, \vec{\gamma})$ of H_k:
  $H_k\,\vec{\gamma} = \theta\,\vec{\gamma}$ with $\|\vec{\gamma}\| = 1$,
- accept $(\theta, \mathbf{u})$, with $\mathbf{u} := \mathbf{V}_k\,\vec{\gamma}$,
  as an approximate eigenpair if
  $|h_{k+1\,k}|\,|e_k^*\vec{\gamma}| < \mathit{tol}$

end

Proposition. If $(\theta, \mathbf{u})$, with $\mathbf{u} := \mathbf{V}_k\,\vec{\gamma}$, is accepted, then, for some eigenvalue λ of $\mathbf{A}$,

$$|\theta - \lambda| \le \mathcal{C}_E\,\mathit{tol}$$

Matrix representation of GCR

Take $\mathbf{v}_1 = \mathbf{A}\mathbf{r}_0$.

$\mathbf{K}^{-1}\mathbf{R}_{k+1} = \mathbf{U}_{k+1}B_{k+1} \;\Rightarrow\; \mathbf{A}\mathbf{R}_{k+1} = \mathbf{C}_{k+1}B_{k+1}$
where $\mathbf{C}_{k+1} := [\mathbf{c}_0\,|\,\dots\,|\,\mathbf{c}_k] = \mathbf{A}\mathbf{U}_{k+1}$.
The $\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{k-1}$ form an orthogonal basis of $\mathcal{K}_k(\mathbf{A}; \mathbf{A}\mathbf{r}_0)$.

$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k \;\Rightarrow\; \mathbf{R}_{k+1}\,\underline{S}_k = \mathbf{C}_k\,D_k$, where

$$\underline{S}_k = \begin{bmatrix} 1 & 0 & \dots & & \\ -1 & 1 & \ddots & & \\ 0 & -1 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & 0 & -1 \end{bmatrix},$$

& $D_k = \text{diag}(\alpha_0, \dots, \alpha_{k-1})$.

$\mathbf{K}^{-1}\mathbf{R}_{k+1} = \mathbf{U}_{k+1}B_{k+1} \;\Rightarrow\; \mathbf{A}\mathbf{R}_{k+1} = \mathbf{C}_{k+1}B_{k+1}$
where $\mathbf{C}_{k+1} = [\mathbf{c}_0\,|\,\dots\,|\,\mathbf{c}_k] = \mathbf{A}\mathbf{U}_{k+1}$.
The $\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{k-1}$ are orthogonal.

$$\mathbf{R}_{k+1}\,\underline{S}_k = \mathbf{C}_k D_k,$$

GCR and Ritz values

$$\mathbf{A}\,\mathbf{C}_k D_k = \mathbf{C}_{k+1}\,B_{k+1}\,\underline{S}_k \quad \Rightarrow$$

$$\boxed{\mathbf{A}\,\mathbf{C}_k = \mathbf{C}_{k+1}\,\underline{H}_k \text{ where } \underline{H}_k = B_{k+1}\,\underline{S}_{k+1}\,D_k^{-1}}$$

$\underline{H}_k$ is an $(k+1) \times k$ upper Hessenberg matrix. Its matrix entries can easily be computed from β_{ij}, α_j and σ_j.

With $\Delta_k = \text{diag}(\sigma_0, \dots, \sigma_{k-1})$, we have that

$$\mathbf{C}_k^T\mathbf{C}_k = \Delta_k.$$

Therefore $\mathbf{C}_k\Delta_k^{-\frac{1}{2}}$ forms an orthonormal basis.

[Morgan '92, Paige Parlett van der vorst '95]

Harmonic Ritz values

Consider a subspace $\mathcal{V}$ of $\mathbb{C}^n$ ($\dim(\mathcal{V}) \ll n$).

Proposition. For a $\theta \in \mathbb{C}$ and a $\mathbf{u} \in \mathcal{V}$ TFAE

(a) $\mathbf{A}\mathbf{u} - \theta\,\mathbf{u} \perp \mathbf{A}\mathcal{V}$;

(b) $(\theta^{-1}, \mathbf{A}\mathbf{u})$ is a Ritz pair of $\mathbf{A}^{-1}$ w.r.t. $\mathbf{A}\mathcal{V}$.

(c) $(\theta, \mathbf{u})$ is a Petrov pair w.r.t. the projection Π on $\mathcal{V}$ that is orthogonal to $\mathcal{W} := \mathbf{A}\mathcal{V}$:

$$\Pi(z) \in \mathcal{V}, \quad \mathbf{z} - \Pi(\mathbf{z}) \perp \mathcal{W} \quad (\mathbf{z} \in \mathbb{C}^n).$$

A **harmonic Ritz pair** of $\mathbf{A}$ w.r.t. $\mathcal{V}$ is a pair $(\theta, \mathbf{u})$ with one of the above properties.

Why harmonic Ritz values?

Consider a Ritz pair $(\theta, \mathbf{u})$ of $\mathbf{A}$ w.r.t. $\mathcal{V}$:

$$\mathbf{A}\mathbf{u} - \theta\,\mathbf{u} \perp \mathcal{V}.$$

Note that $\qquad \theta = (\mathbf{A}\mathbf{u}, \mathbf{u})/(\mathbf{u}, \mathbf{u})$

Suppose $\mathbf{A}$ is normal. Write $\mathbf{u} = \sum \mu_j\,\mathbf{x}_j$.

Then $\qquad \theta = \left(\sum |\mu_j|^2\,\lambda_j\right) / \left(\sum |\mu_j|^2\right)$:

θ is a convex combination of the λ_j.

$$\boxed{\theta \approx \lambda_\ell \ \Rightarrow\ \mathbf{u} \approx \mathbf{x}_\ell ?}$$

$\theta \approx \lambda_\ell$ & λ_ℓ extremal $\quad \Rightarrow \quad |\mu_j| \ll |\mu_\ell| \quad (j \neq \ell)$

$\theta \approx \lambda_\ell$ & λ_ℓ interior $\quad \not\Rightarrow \quad |\mu_j| \ll |\mu_\ell| \quad (j \neq \ell)$

Note $\lambda_\ell \approx 0$ then λ_ℓ^{-1} extremal eigenvalue $\mathbf{A}^{-1}$.

$\mathbf{v}_1, \mathbf{v}_2, \ldots$ Krylov basis,

$$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\ldots|\mathbf{v}_k]}$$

Π_k is the projection onto $\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$ given by

$$\Pi_k(\mathbf{z}) := \sum_{i=1}^{k} \mu_i\,\mathbf{v}_i \ \text{ for all } \ \mathbf{z} = \sum_{i=1}^{n} \mu_i\,\mathbf{v}_i \in \mathbb{C}^n.$$

Harmonic Ritz values on Krylov subspaces

Proposition. If $\mathbf{v}_1, \ldots$ are $\mathbf{A}$**–orthogonal**, that is $\ \mathbf{v}_i \perp \mathbf{A}\mathbf{v}_j, \quad (\mathbf{v}_j, \mathbf{A}\mathbf{v}_j) \neq 0 \quad (j < i)$,
then Π_k is orthogonal to $\mathbf{A}\,\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ and $(\theta, \mathbf{V}_k\vec{\gamma})$ is a harmonic Ritz pair of $\mathbf{A}$ w.r.t. $\mathcal{K}_k(\mathbf{A}, \mathbf{v}_1)$
$\Leftrightarrow$ $(\theta, \vec{\gamma})$ is an eigenpair of H_k.

Examples. From GCR *(and GMRES)*.

GMRES and GCR residuals

The GCR (and GMRES) residuals $\mathbf{r}_0, \ldots, \mathbf{r}_{k-1}$ form an $\mathbf{A}$-orthogonal basis of $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ $\forall\, k$.

Proposition. $\mathbf{r}_0, \mathbf{r}_1, \ldots$ are $\mathbf{A}$-orthogonal and $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = \text{span}(\mathbf{r}_0, \ldots, \mathbf{r}_{k-1})$ $\quad$ for all k.

Proof. $\mathbf{r}_i \perp \text{span}(\mathbf{c}_0, \ldots, \mathbf{c}_{j-1}) = \mathcal{K}_j(\mathbf{A}; \mathbf{A}\mathbf{r}_0) = \mathbf{A}\mathcal{K}_j(\mathbf{A}; \mathbf{r}_0)$

$\mathcal{K}_j(\mathbf{A}; \mathbf{r}_0) \ni \mathbf{r}_i \quad$ for $\quad i < j$.

and $\qquad \rho_j = (\mathbf{r}_j, \mathbf{c}_j) = (\mathbf{r}_j, \mathbf{A}\mathbf{r}_j - \sum_{i<j} \beta_{ij}\mathbf{c}_i) = (\mathbf{r}_j, \mathbf{A}\mathbf{r}_j)$.

$\rho_j \neq 0 \quad$ all j $\ \Leftrightarrow$ no breakdown.

$\mathbf{K}^{-1}\mathbf{R}_{k+1} = \mathbf{U}_{k+1}B_{k+1} \ \Rightarrow\ \mathbf{A}\mathbf{R}_{k+1} = \mathbf{C}_{k+1}B_{k+1}$

where $\quad \mathbf{C}_{k+1} = [\mathbf{c}_0\,|\,\ldots\,|\,\mathbf{c}_k] = \mathbf{A}\mathbf{U}_{k+1}$

$$\mathbf{R}_{k+1}\,\underline{S}_k = \mathbf{C}_k\,D_k$$

GCR and harmonic Ritz values

$$\mathbf{A}\,\mathbf{R}_k = \mathbf{C}_k\,B_k = \mathbf{R}_{k+1}\,\underline{S}_k\,D_k^{-1}\,B_k$$

$$\boxed{\mathbf{A}\mathbf{R}_k = \mathbf{R}_{k+1}\,\underline{\widetilde{H}}_k \quad \text{where} \quad \underline{\widetilde{H}}_k = \underline{S}_k\,D_k^{-1}\,B_k.}$$

$\underline{\widetilde{H}}_k$ is an $(k+1) \times k$ upper Hessenberg matrix. Its matrix entries can easily be computed from β_{ij}, α_j and σ_j.

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots$ be an orthonormal Krylov basis,

$$\boxed{\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\ldots|\mathbf{v}_k]}$$

GMRES and harmonic Ritz values

Theorem. With $\beta := h_{k+1\,k}$, the harmonic Ritz values of $\mathbf{A}$ are the eigenvalues of

$$H_k^{-*}(\underline{H}_k^*\,\underline{H}_k) = H_k + |\beta|^2(H_k^{-*}e_k)e_k^*.$$

Proof. $(\theta, \mathbf{u} := \mathbf{V}_k\,\vec{\gamma})$ is an harmonic Ritz pair

$\Leftrightarrow \quad \mathbf{A}\mathbf{u} - \theta\,\mathbf{u} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$

$\Leftrightarrow \quad \mathbf{A}\mathbf{V}_k\,\vec{\gamma} - \theta\,\mathbf{V}_k\,\vec{\gamma} \perp \mathbf{A}\,\mathbf{V}_k$

$\Leftrightarrow \quad \mathbf{V}_{k+1}\,\underline{H}_k\,\vec{\gamma} - \theta\mathbf{V}_k\,\vec{\gamma} \perp \mathbf{V}_{k+1}\,\underline{H}_k$

$\Leftrightarrow \quad \underline{H}_k^*\,\underline{H}_k\,\vec{\gamma} - \theta\,H_k^*\,\vec{\gamma} = 0.$

In the last "$\Leftrightarrow$" we used $\underline{H}_k^*\,\mathbf{V}_{k+1}^*\,\mathbf{V}_k = H_k^*$.

Finally, note that $\underline{H}_k = H_k + \beta e_{k+1}e_k^*$. Therefore

$$H_k^{-*}(\underline{H}_k^*\underline{H}_k) = H_k^{-*}(H_k^*H_k + |\beta|^2 e_k e_k^*) = H_k + |\beta|^2 H_k^{-*}e_k e_k^*$$

Projections on Krylov subspaces

$\mathbf{v}_1, \mathbf{v}_2, \ldots$ Krylov basis

$$\boxed{\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k, \quad \mathbf{V}_k := [\mathbf{v}_1|\ldots|\mathbf{v}_k]}$$

H_k is the $k \times k$ upper block of $\underline{H}_k$.

Π_k is the projection onto $\mathcal{K}_k(\mathbf{A}; \mathbf{v}_1)$ given by

$$\Pi_k(\mathbf{z}) := \sum_{i=1}^{k} \mu_i\,\mathbf{v}_i \quad \text{for all } \mathbf{z} = \sum_{i=1}^{n} \mu_i\,\mathbf{v}_i \in \mathbb{C}^n.$$

Put $\qquad \mathbf{A}_k := \Pi_k\mathbf{A}_{|\mathcal{K}_k(\mathbf{A};\mathbf{v}_1)}$

Proposition. H_k is the matrix of $\mathbf{A}_k$ with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

$(\theta, \mathbf{V}_k\vec{\gamma})$ is a Petrov pair of $\mathbf{A}$ w.r.t. Π
$\Leftrightarrow$ $(\theta, \vec{\gamma})$ is an eigenpair of H_k.

Hessenberg matrices and polynomials

Theorem. If $q \in \mathcal{P}_k^1$ then

$\Pi_k(q(\mathbf{A})\mathbf{r}_0) = 0 \quad \Leftrightarrow$

$$\{\theta \mid q(\theta) = 0\} = \sigma(H_k) = \sigma(\mathbf{A}_k).$$

Lemma. $\quad \sigma(H_k) = \{\theta_i\}$. For $q \in \mathcal{P}_k^1$:

(a) $\Pi_k q(\mathbf{A})\mathbf{r}_0 = q(\mathbf{A}_k)\mathbf{v}_1 = \rho\,\mathbf{V}_k\,q(H_k)e_1$

where $\quad \mathbf{r}_0 = \rho\,\mathbf{v}_1$.

(b) $\Pi_k\,(q(\mathbf{A})\mathbf{r}_0) = 0 \ \Leftrightarrow\ q(\zeta) = \prod_{j=1}^{k} \left(1 - \zeta/\theta_j\right)$.

Proof. (b) $\quad \Pi_k(q(\mathbf{A})\mathbf{r}_0) = 0 \ \Leftrightarrow\ q(H_k)e_1 = 0$.

(a) $\quad \Pi_k\mathbf{A}^j\mathbf{r}_0 = \Pi_k\mathbf{A}\,\underbrace{(\mathbf{A}^{j-1}\mathbf{r}_0)}_{\in\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)} = \mathbf{A}_k(\mathbf{A}^{j-1}\mathbf{r}_0) \underset{\text{ind.}}{=} \mathbf{A}_k^j\mathbf{r}_0 \qquad j \leq k$

$$\mathbf{A}_k^j\,\mathbf{V}_k = \mathbf{A}_k^{j-1}\,\mathbf{A}_k\,\mathbf{V}_k = \mathbf{A}_k^{j-1}\,\mathbf{V}_k\,H_k = \mathbf{V}_k\,H_k^j,$$

$$\mathbf{A}_k^j\,\mathbf{r}_0 = \mathbf{A}_k^j\,\mathbf{V}_k\,\rho\,e_1 = \rho\,\mathbf{V}_k\,H_k^j e_1.$$

Irreducible Hessenberg matrices

$$H = \begin{bmatrix} * & * & \cdots & * & * \\ h_2 & * & \cdots & * & * \\ & \ddots & \ddots & \vdots & \vdots \\ & & \ddots & * & * \\ 0 & & & h_k & * \end{bmatrix}$$

$k \times k$ Hessenberg

irreducible iff $h_i \neq 0 \ \forall i$

Proposition.

$e_1, \quad He_1, \quad H^2e_1, \quad \ldots, \quad H^{k-1}e_1$ linearly independent

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} * \\ h_2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} * \\ * \\ h_2h_3 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \ldots \begin{bmatrix} * \\ * \\ * \\ * \\ \vdots \\ h_2 \cdots h_k \end{bmatrix}$$

Irreducible Hessenberg matrices

Proposition. $q \in \mathcal{P}_{k-1}$.

$$q(H)e_1 = 0 \ \Leftrightarrow q = 0$$

Proof. $0 = q(H)e_1 = \alpha_0 e_1 + \alpha_1 He_1 + \ldots + \alpha_{k-1}H^{k-1}e_1$

$\Rightarrow \alpha_0 = \alpha_1 = \ldots = \alpha_{k-1} = 0 \Rightarrow q = 0$.

Proposition. $\sigma(H) = \{\theta_i\}$. $\quad p \in \mathcal{P}_k^1$.

$$p(\zeta) = \prod_{i=1}^{k} (1 - \zeta/\theta_i) \ \Leftrightarrow p(H)e_1 = 0$$

Proof. Cayley-Hamilton: "$\Rightarrow$".

$p, q \in \mathcal{P}_k^1$. $\quad p(H)e_1 = q(H)e_1 = 0 \quad \& \quad q \notin \mathcal{P}_{k-1}$

$$\left.\begin{array}{r} \Rightarrow p - \alpha q \in \mathcal{P}_{k-1} \text{ for some } \alpha \in \mathbb{C}. \\ (p(H) - \alpha q(H))e_1 = 0 \end{array}\right\} \begin{array}{l} \Rightarrow p - \alpha q = 0 \\ \Rightarrow 1 = p(0) = \alpha q(0) = \alpha. \end{array}$$

Zeros of FOM polynomials

Orthogonal (=FOM) residual polynomial:

$$p_k^{\mathsf{OR}} \in \mathcal{P}_k^1 \ \ \& \ \ p_k^{\mathsf{OR}}(\mathbf{A})\mathbf{r}_0 \perp \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0).$$

Theorem. For a $\theta \in \mathbb{C}$ TFAE:

(a) $p_k^{\mathsf{OR}}(\theta) = 0$;

(b) there is a $\mathbf{u} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, $\mathbf{u} \neq 0$ for which

$$\mathbf{Au} - \theta\mathbf{u} \perp \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0).$$

Proof. $\mathbf{v}_1, \mathbf{v}_2 \ldots$ orthonormal system as produced by FOM ($\mathbf{v}_1 = \mathbf{r}_0$) with associated orthogonal projection $\Pi_k \ \Rightarrow \Pi_k(p_k^{\mathsf{OR}}(\mathbf{A})\mathbf{r}_0) = 0 \ \Rightarrow$

$$\sigma(\mathbf{A}_k) = \sigma(H_k) = \{\theta \mid p_k^{\mathsf{OR}}(\theta) = 0\}$$

Zeros of GMRES polynomials

Minimal (=GMRES) residual polynomial:

$$p_k^{\mathsf{MR}} \in \mathcal{P}_k^1 \ \ \& \ \ p_k^{\mathsf{MR}}(\mathbf{A})\mathbf{r}_0 \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0).$$

Theorem. For a $\theta \in \mathbb{C}$ TFAE:

(a) $p_k^{\mathsf{MR}}(\theta) = 0$;

(b) there is a $\mathbf{u} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, $\mathbf{u} \neq 0$ for which

$$\mathbf{Au} - \theta\mathbf{u} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0).$$

Remark. The GMRES residuals $\mathbf{r}_0^{\mathsf{MR}}, \ldots, \mathbf{r}_{k-1}^{\mathsf{MR}}$ form an $\mathbf{A}$-orthogonal basis of $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) \ \ \forall k$.

Proof. Take $\mathbf{v}_i = \mathbf{r}_i^{\mathsf{MR}} \quad \forall\, i$, with associated projection Π_k. Then Π_k is orthogonal to $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$

$\Rightarrow \Pi_k(p_k^{\mathsf{MR}}(\mathbf{A})\mathbf{r}_0) = 0 \ \Rightarrow p_k(H_k)e_1 = 0 \ \Rightarrow$

$$\sigma(\mathbf{A}_k) = \sigma(H_k) = \{\theta \mid p_k^{G}(\theta) = 0\}$$

Polynomial preconditioning

Suppose, $\quad \|\tilde{p}(\mathbb{A})\| \approx \min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|$,

$$\tilde{p}(\zeta) = 1 - \alpha_1\zeta - \ldots - \alpha_k\zeta^k = 1 - \zeta\tilde{q}(\zeta)$$

for some k and some explicitly known $\tilde{p} \in \mathcal{P}_k^1$

To be used

- to approximate the solution of $\mathbf{A}\,\mathbf{x} = \tilde{\mathbf{b}}$:
With $\tilde{\mathbf{r}}_0 = \tilde{\mathbf{b}} - \mathbf{A}\,\tilde{\mathbf{x}}_0$, $\ \tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_0 + \mathbf{K}^{-1}\tilde{q}(\mathbb{A})\tilde{\mathbf{r}}_0$
corresponds to $\ \tilde{\mathbf{r}}_k = \tilde{\mathbf{b}} - \mathbf{A}\,\tilde{\mathbf{x}}_k$
- as a preconditioner:
$\mathbf{K}^{-1}\tilde{q}(\mathbb{A})\tilde{\mathbf{b}}$ approx. the solution of $\mathbf{A}\,\mathbf{x} = \tilde{\mathbf{b}}$.
i.e., replace "Solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{r}$" by

Solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \tilde{q}(\mathbb{A})\mathbf{r}$

Advantage. No DOTs are needed.

Drawback. Sensitive to choice $\tilde{p}$.

$$\tilde{p}(\zeta) = 1 - \alpha_1\zeta - \ldots - \alpha_k\zeta^k = 1 - \zeta\tilde{q}(\zeta)$$

and $\quad \|\tilde{p}(\mathbb{A})\| \approx \min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|$

$\tilde{\mathbf{x}}_k := \tilde{\mathbf{x}}_0 + \mathbf{K}^{-1}\tilde{q}(\mathbb{A})\tilde{\mathbf{r}}_0$ approx. solution $\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$

Horner's scheme

$\mathbf{w} = \alpha_k\,\tilde{\mathbf{r}}_0$
for $\quad i = 1, \ldots, k-1 \quad$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{w}$
  $\mathbf{w} \leftarrow \mathbf{Au} + \alpha_{k-i}\,\tilde{\mathbf{r}}_0$
end for
solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \mathbf{w}$
$\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_0 + \mathbf{u}$
$\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_0 - \mathbf{A}\,\mathbf{u}$

$$\tilde{p}(\zeta) = 1 - \alpha_1\zeta - \ldots - \alpha_k\zeta^k = 1 - \zeta\tilde{q}(\zeta)$$

and $\quad \|\tilde{p}(\mathbb{A})\| \approx \min_{p \in \mathcal{P}_k^1} \|p(\mathbb{A})\|$

$\tilde{\mathbf{x}}_k := \tilde{\mathbf{x}}_0 + \mathbf{K}^{-1}\tilde{q}(\mathbb{A})\tilde{\mathbf{r}}_0$ approx. solution $\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$

$$\tilde{p}(\zeta) = \prod_{j=1}^{k} \left(1 - \frac{\zeta}{\theta_j}\right).$$

Smolariski & Saylor scheme

for $\quad i = 1, \ldots, k \quad$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \tilde{\mathbf{r}}_{i-1}$
  $\mathbf{c} \leftarrow \mathbf{Au}$
  $\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_{i-1} + (1/\theta_i)\,\mathbf{u}$
  $\tilde{\mathbf{r}}_i = \tilde{\mathbf{r}}_{i-1} - (1/\theta_i)\,\mathbf{c}$
end for

Leja ordering

For stability,

order zero's $\tilde{p}$ according to **Leja ordering**:

$$|\theta_1| = \max_{1 \le j \le k} |\theta_j|,$$

$$|\theta_j| \prod_{i=1}^{j-1} |\theta_j - \theta_i| \le \max_{j \le \ell \le k} \left(|\theta_\ell| \prod_{i=1}^{j-1} |\theta_\ell - \theta_i| \right)$$

Real arithmetic

For real problems,

use polynomial factors of degree 2:

$\theta_2 = \overline{\theta}_1 \ \Rightarrow$

$$\left(1 - \frac{\zeta}{\theta_1}\right)\left(1 - \frac{\zeta}{\theta_2}\right) = 1 - 2\,\mathrm{Re}\left(\frac{1}{\theta_1}\right)\zeta + \frac{\zeta^2}{|\theta_1|^2}$$

Using (harmonic) Ritz values

Harmonic Ritz values are zero's of p_k, where p_k is some kth GMRES polynomial.
Apply Leja, etc..

I. Use (harmonic) Ritz values to estimate a polygonial area $\mathcal{G}$ in $\mathbb{C}$ that contains $\sigma(\mathbb{A})$.

A. Estimate an ellipse $\mathcal{E}$ ($0 \notin \mathcal{E}$) that contains $\mathcal{G}$. Take for $\tilde{p}$ the shifted and scaled Chebychev polynomial. [Manteuffel '79]

B. Choose a weight function and compute $\tilde{p}$ as the polynomial in $\mathcal{P}_k^1$ that minimizes on $\partial\mathcal{G}$ with respect to the weighted 2-norm. [van Gijzen '93]

II. (If $\mathbb{A}$ is positive definite.) Use the Ritz values values to estimate a weight function ω for which $\int f\omega \approx \sum f(\lambda_i)|\mu_i|^2$. Compute $\tilde{p}$ as the polynomial in $\mathcal{P}_k^1$ (a Jacobi-type of polynomial) that minimizes with respect to the weighted 2-norm. [Fischer '93]

Using the GCR polynomial

Consider the GCR (=GMRES) polynomial p_k (with respect to some $\mathbf{r}_0$), that is
$p_k \in \mathcal{P}_k^1$ & $\mathbf{r}_k = p_k(\mathbb{A})\mathbf{r}_0$ the kth GCR residual

I. The harmonic Ritz values are the zero's of p_k and can cheaply be computed from GCR. (Apply Leja, etc..) [Nachtigal Reichel Trefethen '92]

II. Mimic GCR:

Store the coefficients β_{ij}, α_j, $\vec{\gamma}_k$ as computed in k steps GCR with initial residual $\mathbf{r}_0$.

Now apply k steps of "(efficient) GCR" with the new initial approximation, but using the **old** GCR coefficients.
[Sleijpen van der Vorst Fokkema '94]

Drawback. Depends on the initial residual.

$\tilde{\mathbf{x}}_k := \tilde{\mathbf{x}}_0 + \mathbf{K}^{-1}q(\mathbb{A})\tilde{\mathbf{r}}_0$ approx. sol. $\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$
$\tilde{\mathbf{r}}_k = \tilde{\mathbf{b}} - \mathbf{A}\tilde{\mathbf{x}}_k = p_k(\mathbb{A})\tilde{\mathbf{r}}_0$ where $p_k(\zeta) = 1 - \zeta q(\zeta)$

Mimic of efficient GCR

Choose $\tilde{\mathbf{x}}_0$
$\tilde{\mathbf{r}}_0 = \tilde{\mathbf{b}} - \mathbf{A}\,\tilde{\mathbf{x}}_0$
for $\ell = 0, \ldots, k-1$ do
  solve $\mathbf{u}$ from $\mathbf{K}\,\mathbf{u} = \tilde{\mathbf{r}}_\ell$
  $\mathbf{c} \leftarrow \mathbf{A}\mathbf{u}$
  for $i = 0, \ldots, \ell-1$ do
    $\mathbf{c} \leftarrow \mathbf{c} - \beta_{i\ell}\,\tilde{\mathbf{c}}_i$
  end for
  $\tilde{\mathbf{c}}_\ell = \mathbf{c}$,
  $\tilde{\mathbf{r}}_{\ell+1} = \tilde{\mathbf{r}}_\ell - \alpha_\ell\,\tilde{\mathbf{c}}_\ell$
end for
solve $\mathbf{y}$ from $\mathbf{K}\,\mathbf{y} = \widetilde{\mathbf{R}}_k\,\vec{\gamma}_k$
$\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_0 + \mathbf{y}$

Efficient methods

CG — Conjugate Gradients
if $\mathbb{A}^* = \mathbb{A} > 0$ $\Rightarrow$ MINRES, SYMMLQ if $\mathbb{A}^* = \mathbb{A}$

$\Downarrow$ *non-symmetric problems*

Bi-CG — Bi-Conjugate Gradients $\Rightarrow$ Quasi Minimal Res. simple QMR
smoothing convergence
avoiding breakdown

$\Downarrow$ *transpose free*
faster convergence

CGS — Conjugate Gradients Squared

$\Downarrow$ *more accuracy*
faster convergence

Bi-CGSTAB — Bi-CG STABilized

$\Downarrow$ *avoiding breakdown*
faster convergence

BiCGstab(ℓ)

"New" aspects

- **Breakdown**

- **Effects of finite precision arithmetic**

— Loss of speed of convergence
  due to ◦ **near breakdown**
  ◦ **loss of orthogonality**

— Loss of accuracy
  i.e., for computed $\mathbf{r}_k$ and computed $\mathbf{x}_k$

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|}{\|\mathbf{r}_k\|} \not\approx 1$$

  due to ◦ **non-monotone convergence**

Orthogonal residual methods Bi-CG and related methods

Program

- Bi-CG

- Miscellaneous

  Smoothing the convergence (fake)
  irregular convergence
  maintaining accuracy

- Hybrid Bi-CG

  CGS, GCGS, Bi-CGSTAB, BiCGstab(ℓ)

Conjugate Gradients $\mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{u}_k \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$$

Theorem. If $\mathbf{A}^* = \mathbf{A}$ and
$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \mathbf{r}_{k-1}$ then $\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$

Proof.
$\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathbf{r}_{k-1}$ by construction
$\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$ by induction

$\mathbf{A}\,\mathbf{u}_k = \mathbf{A}\,\mathbf{r}_k - \beta_k\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathbf{r}_k$ by construction
$\mathbf{A}\,\mathbf{u}_k = \mathbf{A}\,\mathbf{r}_k - \beta_k\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$ by induction:

$\mathbf{A}\mathbf{r}_k \perp \mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0) \quad\Leftrightarrow\quad \mathbf{r}_k \perp \mathbf{A}\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0) \subset \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$

Conjugate Gradients $\mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$\mathbf{r}_k,\ \mathbf{u}_k \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$
$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \mathbf{r}_{k-1}$

With $\rho_k := (\mathbf{r}_k, \mathbf{r}_k)$ & $\sigma_k := (\mathbf{A}\mathbf{u}_k, \mathbf{r}_k)$

and, since $\mathbf{r}_k + \alpha_{k-1}\,\mathbf{A}\,\mathbf{r}_{k-1} \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$,

we have $\alpha_k = \dfrac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{A}\mathbf{u}_k, \mathbf{r}_k)} = \dfrac{\rho_k}{\sigma_k}$ and

$$\beta_k = \frac{(\mathbf{A}\mathbf{r}_k, \mathbf{r}_{k-1})}{(\mathbf{A}\mathbf{u}_{k-1}, \mathbf{r}_{k-1})} = \frac{(\mathbf{r}_k, \mathbf{A}\mathbf{r}_{k-1})}{\sigma_{k-1}} = \frac{-\rho_k}{\alpha_{k-1}\sigma_{k-1}}$$

Conjugate Gradients

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{K}^{-1}\mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{K}\mathbf{u}_k \in \mathcal{K}_{k+1}(\mathbb{A};\mathbf{r}_0)$$

Theorem. If $\mathbf{A}^* = \mathbf{A}$ & $\mathbf{K}^* = \mathbf{K}$ and

$\mathbf{r}_k, \mathbf{A}\mathbf{u}_k \perp \mathbf{K}^{-1}\mathbf{r}_{k-1}$ then $\mathbf{r}_k, \mathbf{A}\mathbf{u}_k \perp \mathbf{K}^{-1}\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$

[Hestenes Stiefel '52]

CG $\quad \mathbf{A} > 0,\ \mathbf{K} > 0$

Choose $\mathbf{x}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = 0,\ \rho_{-1} = 1$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{w}$ from $\mathbf{K}\mathbf{w} = \mathbf{r}_k$
  $\rho_k = (\mathbf{r}_k, \mathbf{w}),\ \beta_k = -\rho_k/\rho_{k-1}$
  $\mathbf{u}_k = \mathbf{w} - \beta_k\,\mathbf{u}_{k-1}$
  $\mathbf{c}_k = \mathbf{A}\mathbf{u}_k$
  $\sigma_k = (\mathbf{c}_k, \mathbf{u}_k),\ \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
end while

[Hestenes Stiefel '52]

CG $\quad \mathbf{A} > 0,\ \mathbf{K} > 0$

Choose $\mathbf{x}_0$
$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$
$\mathbf{u} = 0,\ \rho = 1$
while $\|\mathbf{r}\| > tol$ do
  $\sigma \leftarrow -\rho$
  solve $\mathbf{w}$ from $\mathbf{K}\mathbf{w} = \mathbf{r}$
  $\rho \leftarrow (\mathbf{r}, \mathbf{w}),\ \beta \leftarrow \rho/\sigma$
  $\mathbf{u} \leftarrow \mathbf{w} - \beta\,\mathbf{u}$
  $\mathbf{c} \leftarrow \mathbf{A}\mathbf{u}$
  $\sigma \leftarrow (\mathbf{c}, \mathbf{u}),\ \alpha \leftarrow \rho/\sigma$
  $\mathbf{x} \leftarrow \mathbf{x} + \alpha\,\mathbf{u}$
  $\mathbf{r} \leftarrow \mathbf{r} - \alpha\,\mathbf{c}$
end while

Properties CG

Advantages

- 1 MV, 2 DOT, 3 AXPY per step to increase dimension Krylov subspace by one.
- Storage: 5 large vectors (incl. $\mathbf{b}$) (6 if $\mathbf{K} \neq \mathbf{I}$).
- Minimal residual method if $\mathbf{A}$, $\mathbf{K}$ pos.def.: $\|\mathbf{r}_k\|_{\mathbf{A}^{-1}}$ is minimal.
- Orth. residual method if $\mathbf{A}^* = \mathbf{A}$, $\mathbf{K}^* = \mathbf{K}$: $\mathbf{r}_k \perp \mathbf{K}^{-1}\mathcal{K}_k(\mathbb{A};\mathbf{r}_0)$.
- No knowledge on properties of $\mathbb{A}$ is needed.
- ◦ Robust: CG (for pos.def.) always converges.

Drawback

- Does not work if $\mathbf{A} \neq \mathbf{A}^*$.
- ◦ CG is sensitive to evaluation errors (often loss of super-linear convergence).

Bi-Conjugate Gradients $\quad \mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{u}_k \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$$

Theorem. $\widetilde{\mathbf{r}}_0, \ldots, \widetilde{\mathbf{r}}_{k-1}$ Krylov basis $\mathcal{K}_k(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$

$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \widetilde{\mathbf{r}}_{k-1}$ then $\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \mathcal{K}_k(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$

Proof.

$\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \widetilde{\mathbf{r}}_{k-1}$ by construction
$\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathcal{K}_{k-1}(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$ by induction

$\mathbf{A}\,\mathbf{u}_k = \mathbf{A}\,\mathbf{r}_k - \beta_k\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \widetilde{\mathbf{r}}_{k-1}$ by construction
$\mathbf{A}\,\mathbf{u}_k = \mathbf{A}\,\mathbf{r}_k - \beta_k\,\mathbf{A}\,\mathbf{u}_{k-1} \perp \mathcal{K}_{k-1}(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$ by induction:

$$\mathbf{A}\,\mathbf{r}_k \perp \mathcal{K}_{k-1}(\mathbf{A}^*;\widetilde{\mathbf{r}}_0) \ \Leftrightarrow\ \mathbf{r}_k \perp \mathbf{A}^*\mathcal{K}_{k-1}(\mathbf{A}^*;\widetilde{\mathbf{r}}_0) \subset \mathcal{K}_k(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$$

Bi-Conjugate Gradients $\quad \mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \widetilde{\mathbf{r}}_{k-1}$$

with $\widetilde{\mathbf{r}}_0, \ldots, \widetilde{\mathbf{r}}_{k-1}$ Krylov basis $\mathcal{K}_k(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$

With $\rho_k := (\mathbf{r}_k, \widetilde{\mathbf{r}}_k)$ & $\sigma_k := (\mathbf{A}\mathbf{u}_k, \widetilde{\mathbf{r}}_k)$

and θ_k such that $\widetilde{\mathbf{r}}_k - \bar{\theta}_k\,\mathbf{A}^*\,\widetilde{\mathbf{r}}_{k-1} \in \mathcal{K}_k(\mathbf{A}^*;\widetilde{\mathbf{r}}_0)$,

we have $\alpha_k = \dfrac{(\mathbf{r}_k, \widetilde{\mathbf{r}}_k)}{(\mathbf{A}\mathbf{u}_k, \widetilde{\mathbf{r}}_k)} = \dfrac{\rho_k}{\sigma_k}$ and

$$\beta_k = \frac{(\mathbf{A}\mathbf{r}_k, \widetilde{\mathbf{r}}_{k-1})}{(\mathbf{A}\mathbf{u}_{k-1}, \widetilde{\mathbf{r}}_{k-1})} = \frac{(\mathbf{r}_k, \mathbf{A}^*\,\widetilde{\mathbf{r}}_{k-1})}{\sigma_{k-1}} = \frac{\rho_k}{\theta_k\,\sigma_{k-1}}$$

Fletcher '76: $\widetilde{\mathbf{u}}_k = \widetilde{\mathbf{r}}_k - \bar{\beta}_k\,\widetilde{\mathbf{u}}_{k-1}$ & $\widetilde{\mathbf{r}}_{k+1} = \widetilde{\mathbf{r}}_k - \bar{\alpha}_k\,\mathbf{A}^*\,\widetilde{\mathbf{u}}_k$

Bi-Conjugate Gradients $\quad \mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned}\mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}\\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k\end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \bar{q}_{k-1}(\mathbf{A}^*)\widetilde{\mathbf{r}}_0$$

with q_{k-1} a polynomial of exact degree $k-1$.

With

$\rho_k := (\mathbf{r}_k, \bar{q}_k(\mathbf{A}^*)\widetilde{\mathbf{r}}_0)$ & $\sigma_k := (\mathbf{A}\mathbf{u}_k, \bar{q}_k(\mathbf{A}^*)\widetilde{\mathbf{r}}_0)$

and θ_k such that

$$q_k(\zeta) - \theta_k\,\zeta\,q_{k-1}(\zeta) \in \mathcal{P}_{k-1}$$

we have

$$\alpha_k = \frac{\rho_k}{\sigma_k} \quad \& \quad \beta_k = \frac{\rho_k}{\theta_k\,\sigma_{k-1}}$$

[Sonneveld '89]

Generalized Bi-CG

Choose $\mathbf{x}_0$ and $\widetilde{\mathbf{r}}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = 0,\ \theta_0 = 1,\ \sigma_{-1} = 1$
while $\|\mathbf{r}_k\| > tol$ do
  solve $\mathbf{w}$ from $\mathbf{K}\mathbf{w} = \mathbf{r}_k$
  $\rho_k = (\mathbf{r}_k, \widetilde{\mathbf{r}}_k),\ \beta_k = \rho_k/(\theta_k\,\sigma_{k-1})$
  $\mathbf{u}_k = \mathbf{w} - \beta_k\,\mathbf{u}_{k-1}$
  $\mathbf{c}_k = \mathbf{A}\mathbf{u}_k$
  $\sigma_k = (\mathbf{c}_k, \widetilde{\mathbf{r}}_k),\ \alpha_k = \rho_k/\sigma_k$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$
  $k \leftarrow k+1$
    select a $\widetilde{\mathbf{r}}_k \in \mathcal{K}_{k+1}(\mathbb{A}^*;\widetilde{\mathbf{r}}_0)$
    such that, for some $\theta_k \neq 0$,
    $\widetilde{\mathbf{r}}_k - \bar{\theta}_k\,\mathbb{A}^*\,\widetilde{\mathbf{r}}_{k-1} \in \mathcal{K}_k(\mathbb{A}^*;\widetilde{\mathbf{r}}_0)$
end while

[Fletcher '76]

Bi-CG

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$

$k = 0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$

$\mathbf{u}_{-1} = 0,\ \rho_{-1} = 1,\qquad \tilde{\mathbf{c}}_{-1} = 0$

while $\|\mathbf{r}_k\| > tol$ do

  solve $\mathbf{w}$ from $\mathbf{K}\mathbf{w} = \mathbf{r}_k$

  $\rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_k),\ \beta_k = -\rho_k/\rho_{k-1}$

  $\mathbf{u}_k = \mathbf{w} - \beta_k\,\mathbf{u}_{k-1}$

  $\mathbf{c}_k = \mathbf{A}\mathbf{u}_k$

  $\sigma_k = (\mathbf{c}_k, \tilde{\mathbf{r}}_k),\ \alpha_k = \rho_k/\sigma_k$

  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$

  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{c}_k$

    $\tilde{\mathbf{c}}_k = \mathbb{A}^*\,\tilde{\mathbf{r}}_k - \bar{\beta}_k\,\tilde{\mathbf{c}}_{k-1}$

    $\tilde{\mathbf{r}}_{k+1} = \tilde{\mathbf{r}}_k - \bar{\alpha}_k\,\tilde{\mathbf{c}}_k$

  $k \leftarrow k+1$

end while

[Fletcher '76]

Bi-CG

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}$

$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$

$\mathbf{u} = 0,\ \rho = 1,\qquad \tilde{\mathbf{c}} = 0$

while $\|\mathbf{r}\| > tol$ do

  $\sigma \leftarrow -\rho$

  solve $\mathbf{w}$ from $\mathbf{K}\mathbf{w} = \mathbf{r}$

  $\rho \leftarrow (\mathbf{r}, \tilde{\mathbf{r}}),\ \beta \leftarrow \rho/\sigma$

  $\mathbf{u} \leftarrow \mathbf{w} - \beta\,\mathbf{u}$

  $\mathbf{c} \leftarrow \mathbf{A}\mathbf{u}$

  $\sigma \leftarrow (\mathbf{c}, \tilde{\mathbf{r}}),\ \alpha \leftarrow \rho/\sigma$

  $\mathbf{x} \leftarrow \mathbf{x} + \alpha\,\mathbf{u}$

  $\mathbf{r} \leftarrow \mathbf{r} - \alpha\,\mathbf{c}$

    $\tilde{\mathbf{c}} \leftarrow \mathbb{A}^*\,\tilde{\mathbf{r}} - \bar{\beta}\,\tilde{\mathbf{c}}$

    $\tilde{\mathbf{r}} \leftarrow \tilde{\mathbf{r}} - \bar{\alpha}\,\tilde{\mathbf{c}}$

end while

Bi-CG = CG

Instead of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ we may as well solve

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^* \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \tilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \tilde{\mathbf{b}} \end{bmatrix} \qquad (\star)$$

$$2\left[\begin{bmatrix} \mathbf{x} \\ \tilde{\mathbf{x}} \end{bmatrix}, \begin{bmatrix} \mathbf{y} \\ \tilde{\mathbf{y}} \end{bmatrix}\right] := \begin{bmatrix} \mathbf{y} \\ \tilde{\mathbf{y}} \end{bmatrix}^* \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \tilde{\mathbf{x}} \end{bmatrix}$$
$$= \mathbf{y}^*\tilde{\mathbf{x}} + \tilde{\mathbf{y}}^*\mathbf{x} = (\mathbf{y}, \tilde{\mathbf{x}}) + (\mathbf{x}, \tilde{\mathbf{y}})$$

defines an inner product except for the positivity requirement (we now do not have

$$[\vec{x}, \vec{x}] \geq 0 \quad \text{nor} \quad [\vec{x}, \vec{x}] = 0 \Leftrightarrow \vec{x} = 0).$$

$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^* \end{bmatrix}$ is symmetric with respect to this "quarter inner product".

Bi-CG = CG applied to the **augmented problem** $(\star)$ with respect to this "inner product".

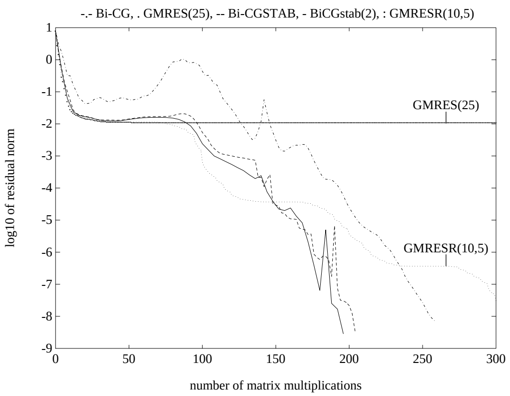
Properties Bi-CG

Advantages

- Usually selects good approximations from the search subspaces (Krylov subspaces).
- 2 DOT, 5 AXPY per step.
- Storage: 8 large vectors.
- No knowledge on properties of $\mathbb{A}$ is needed.

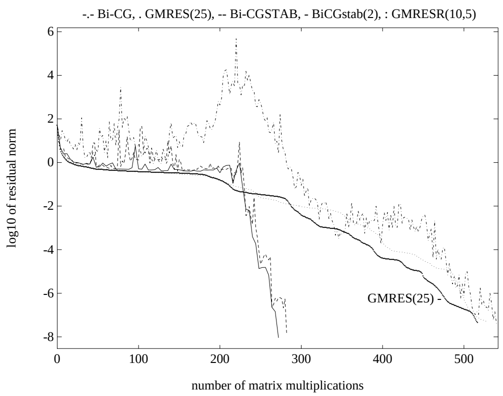
Drawbacks

- ◦ Non-optimal Krylov subspace method.
- Not robust: Bi-CG may break down.
- Bi-CG is sensitive to evaluation errors (often loss of super-linear convergence).
- ◦ Convergence depends on shadow residual $\tilde{\mathbf{r}}_0$.
- 2 MV needed to expand search subspace.
- 1 MV is by $\mathbf{A}^*$.



$-(a\,u_x)_x - (a\,u_y)_y = 1$ on $[0,1] \times [0,1]$.
$a = 1000$ for $0.1 \leq x, y \leq 0.9$
and $a = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
82×82 volumes. ILU Decomp.



$-u_{xx} - u_{yy} + a\,u_x + 100\,u_y = f$ on $[0,1] \times [0,1]$.
$a = 100$ if $x \in [0, \frac{1}{4}] \cup [\frac{1}{2}, \frac{3}{4}]$, $a = -100$ elsewhere

f such that $u(x,y) = sin(\pi x)\sin(\pi y)$.
Dirichlet BC. 81×81 volumes. No precond..

Expanding a Krylov basis

Suppose $\mathbf{w}_1, \mathbf{w}_2, \ldots$ are linearly independent.

Expand $\mathbf{V}_k = [\mathbf{v}_1|\ldots|\mathbf{v}_k]$ (recursively) with $\mathbf{A}\mathbf{v}_k$ orthogonalized w.r.t. $\mathbf{W}_k = [\mathbf{w}_1|\ldots|\mathbf{w}_k]$:

  $\tilde{\mathbf{v}} = \mathbf{A}\mathbf{v}_k,\ M_k = \mathbf{W}_k^*\mathbf{V}_k$

  $\vec{h}_k = M_k^{-1}(\mathbf{W}_k^*\tilde{\mathbf{v}}),\ \tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec{h}_k$

  Select a $h_{k+1\,k} \neq 0$, $\mathbf{v}_{k+1} = \tilde{\mathbf{v}}/h_{k+1\,k}$

Theorem. Then $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$ is a Krylov basis and $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{H}_k$, $\underline{H}_k = \begin{bmatrix} \underline{H}_{k-1} & \vec{h}_k \\ 0\ldots0 & h_{k+1\,k} \end{bmatrix}$

Examples. In all examples, $\forall k$

- $\mathbf{w}_k = \mathbf{v}_k$ [Arnoldi '51]
- $\mathbf{w}_1, \ldots, \mathbf{w}_k$ Krylov basis $\mathcal{K}_k(\mathbf{A}^*; \mathbf{w}_1)$ [Lanczos '52]
- π "selects" from $\{1, \ldots, n\}$: $\mathbf{w}_k = e_{\pi(k)}$ [Sadok '96]

$$\rho_k := \min\{\|\mathbf{r}_0 - \mathbf{A}\mathbf{y}\| \mid \mathbf{y} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)\}$$

$\mathbf{r}_0/\rho = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$ Krylov basis $\mathcal{K}_{k+1}(\mathbf{A}; \mathbf{r}_0)$

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

Then $\quad \rho_k = \min_{\vec{\gamma}} \|\mathbf{r}_0 - \mathbf{A}\mathbf{V}_k\,\vec{\gamma}\|$
$\quad\quad = \min_{\vec{\gamma}} \|\mathbf{V}_{k+1}(\rho e_1 - \underline{H}_k\vec{\gamma})\|$

Quasi minimalization

Solve $\quad \tilde{\rho}_k := \min_{\vec{\gamma}} \|\rho e_1 - \underline{H}_k\vec{\gamma}\|$

With solution $\vec{\gamma}_k$, take $\quad \mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\,\vec{\gamma}_k$

Theorem. $\sigma_1 \leq \ldots \leq \sigma_k$ singular values $\mathbf{V}_k$

$\Rightarrow \quad \sigma_1\,\tilde{\rho}_k \leq \rho_k \leq \sigma_k\,\tilde{\rho}_k,$

$$\rho_k \leq \|\mathbf{r}_k\| = \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\| \leq \sigma_k\,\tilde{\rho}_k \leq \mathcal{C}(\mathbf{V}_k)\,\rho_k.$$

$$\sigma_k \leq \sqrt{k}\ \max_j \|\mathbf{v}_j\|$$

$\mathbf{r}_0/\rho = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$ Krylov basis $\mathcal{K}_{k+1}(\mathbf{A}; \mathbf{r}_0)$

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

Orthogonalization

Solve $H_k\vec{\gamma} = \rho\, e_1$. With solution $\vec{\gamma}_k$, need not exist!
take $\mathbf{y}_k = \mathbf{V}_k\,\vec{\gamma}_k$ and $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k$

Theorem. $\mathbf{y}_k \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ and

$$\mathbf{r}_k = \mathbf{r}_0 - \mathbf{A}\mathbf{y}_k = \mathbf{v}_{k+1}(h_{k+1\,k}\, e_k^*\vec{\gamma}_k) =: \tau_k\,\mathbf{v}_{k+1}$$

If $\mathbf{W}_k^*\mathbf{v}_{k+1} = 0$ then $\mathbf{r}_k \perp \mathrm{span}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$

Proof. $\mathbf{r}_0 - \mathbf{A}\,\mathbf{y}_k = \mathbf{V}_{k+1}\,(\rho\, e_1 - \underline{H}_k\,\vec{\gamma}_k)$

Theorem. $|\tau_k| = \widetilde{\rho}_k/\sqrt{1 - \widetilde{\rho}_k^2/\widetilde{\rho}_{k-1}^2}$

with $\widetilde{\rho}_k := \min_{\vec{\gamma}} \|\rho\, e_1 - \underline{H}_k\,\vec{\gamma}\|$

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$ Krylov basis, $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$

Theorem. H_k is tri-diagonal $\forall k$

if, for some $\mathbf{w}_1$,

$\left.\begin{array}{l}\mathbf{v}_{k+1} \text{ is orthogonal with respect to the} \\ \textbf{shadow Krylov subspace } \mathcal{K}_k(\mathbf{A}^*; \mathbf{w}_1)\end{array}\right\}\ \forall k\ (\star)$

$\mathbf{w}_1, \ldots, \mathbf{w}_{k+1}$ Krylov basis $\mathcal{K}_{k+1}(\mathbf{A}^*; \mathbf{w}_1)$,

$$\mathbf{A}^*\,\mathbf{W}_k = \mathbf{W}_{k+1}\,\underline{F}_k.$$

Then, with $M_k := \mathbf{W}_k^*\mathbf{V}_k$:

$(\star)$ holds iff M_k non-singular lower triangular;

$(\star)$ implies $\quad H_k = M_k^{-1}\,((\underline{F}_k)^*\,\underline{M}_k)$

where $\underline{M}_k := \mathbf{W}_{k+1}^*\mathbf{V}_k$ is lower triangular.

Proof. $\quad \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k = (\mathbf{W}_k^*\mathbf{V}_{k+1})\underline{H}_k = M_k\,H_k$

$(\underline{F}_k)^*\underline{M}_k = (\underline{F}_k)^*\mathbf{W}_{k+1}^*\mathbf{V}_k = \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k$

[Lanczos '50]

Bi–Lanczos

$\mathbf{v}_1, \ldots, \mathbf{v}_k$ Krylov basis $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$.
$\mathbf{w}_1, \ldots, \mathbf{w}_k$ Krylov basis $\mathcal{K}_k(\mathbf{A}^*; \widetilde{\mathbf{r}}_0)$.

Bi-Lanczos expands recursively

- $\mathbf{V}_k = [\mathbf{v}_1\,|\,\ldots\,|\,\mathbf{v}_k]$ by $\mathbf{A}\mathbf{v}_k$ orthogonalized w.r.t. $\mathbf{w}_1, \ldots, \mathbf{w}_k$,

&
- $\mathbf{W}_k = [\mathbf{w}_1\,|\,\ldots\,|\,\mathbf{w}_k]$ by $\mathbf{A}^*\mathbf{w}_k$ orthogonalized w.r.t. $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Bi-Lanczos does not **break down**
iff $D_{k+1} := \mathbf{W}_{k+1}^*\mathbf{V}_{k+1}$ non-singular diagonal.

$\underline{H}_k$ and $\underline{F}_k$ Hessenberg such that

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k \quad \& \quad \mathbf{A}^*\,\mathbf{W}_k = \mathbf{W}_{k+1}\,\underline{F}_k$$

Theorem. $H_k = D_k^{-1}\,F_k^*\,D_k$ is tridiagonal,
if bi-Lanczos does not break down.

Proof. $\quad F_k^*D_k = \mathbf{W}_k^*\,\mathbf{A}\,\mathbf{V}_k = D_k H_k$

[Lanczos '50]

Bi-Lanczos

Select a $\mathbf{r}_0$,	and a $\widetilde{\mathbf{r}}_0$
$k = 1$	
$\mathbf{v}_1 = \mathbf{r}_0/\|\mathbf{r}_0\|$, $\mathbf{v}_0 = \mathbf{0}$,	$\mathbf{w}_1 = \widetilde{\mathbf{r}}_0/\|\widetilde{\mathbf{r}}_0\|$, $\mathbf{w}_0 = \mathbf{0}$
$\gamma_0 = 0$, $\delta_0 = 1$,	$\widetilde{\gamma}_0 = 0$, $\widetilde{\delta}_0 = 1$
while ... do	
$\quad\delta_k = \mathbf{w}_k^*\mathbf{v}_k$,	
$\quad\widetilde{\mathbf{v}} = \mathbf{A}\,\mathbf{v}_k$,	$\widetilde{\mathbf{w}} = \mathbf{A}^*\,\mathbf{w}_k$
$\quad\beta_k = \overline{\widetilde{\gamma}}_{k-1}\delta_k/\delta_{k-1}$,	$\widetilde{\beta}_k = \overline{\gamma}_{k-1}\overline{\delta}_k/\overline{\delta}_{k-1}$
$\quad\widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}} - \beta_k\mathbf{v}_{k-1}$,	$\widetilde{\mathbf{w}} \leftarrow \widetilde{\mathbf{w}} - \widetilde{\beta}_k\mathbf{w}_{k-1}$
$\quad\alpha_k = \mathbf{w}_k^*\widetilde{\mathbf{v}}/\delta_k$,	$\widetilde{\alpha}_k = \overline{\alpha}_k$
$\quad\widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}} - \alpha_k\,\mathbf{v}_k$,	$\widetilde{\mathbf{w}} \leftarrow \widetilde{\mathbf{w}} - \widetilde{\alpha}_k\,\mathbf{w}_k$
$\quad$Select a $\gamma_k \neq 0$	and a $\widetilde{\gamma}_k \neq 0$
$\quad\mathbf{v}_{k+1} = \widetilde{\mathbf{v}}/\gamma_k$,	$\mathbf{w}_{k+1} = \widetilde{\mathbf{w}}/\widetilde{\gamma}_k$,
$\quad k \leftarrow k + 1$	
$\quad\mathbf{V}_k = [\mathbf{V}_{k-1}\,|\,\mathbf{v}_k]$,	$\mathbf{W}_k = [\mathbf{W}_{k-1}\,|\,\mathbf{w}_k]$
end while	

$\mathbf{v}_1, \ldots, \mathbf{v}_k$ Krylov basis $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$.
$\mathbf{w}_1, \ldots, \mathbf{w}_k$ Krylov basis $\mathcal{K}_k(\mathbf{A}^*; \widetilde{\mathbf{r}}_0)$.

Bi-Lanczos expands recursively

- $\mathbf{V}_k = [\mathbf{v}_1\,|\,\ldots\,|\,\mathbf{v}_k]$ by $\mathbf{A}\mathbf{v}_k$ orthogonalized w.r.t. $\mathbf{w}_1, \ldots, \mathbf{w}_k$,

&
- $\mathbf{W}_k = [\mathbf{w}_1\,|\,\ldots\,|\,\mathbf{w}_k]$ by $\mathbf{A}^*\mathbf{w}_k$ orthogonalized w.r.t. $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Lanczos breakdowns

Breakdown iff $\mathbf{w}_{k+1}^*\mathbf{v}_{k+1} = 0$

- **Lucky** breakdown if $\mathbf{v}_{k+1} = 0$.
Then $\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_k\,T_k$.
- **Serious** or **Lanczos** breakdown:
$\mathbf{v}_{k+1} \neq 0$, $\mathbf{w}_{k+1} \neq 0$, $\mathbf{w}_{k+1}^*\mathbf{v}_{k+1} = 0$

Cure: **Look ahead**
i.e. form a non-singular block diagonal $D_{k+\ell}$

$$D_{k+\ell} = \mathbf{W}_{k+\ell}^*\mathbf{V}_{k+\ell}:$$

Orth. $\mathbf{A}\,\mathbf{v}_{k+1}$ w.r.t. $\mathbf{W}_k$ and $\mathbf{A}^*\,\mathbf{w}_{k+1}$ w.r.t. $\mathbf{V}_k$.
Check non-singularity $[\mathbf{w}_{k+1}|\mathbf{w}_{k+2}]^*[\mathbf{v}_{k+1}|\mathbf{v}_{k+2}]$
...

$\mathbf{r}_0 = \rho\,\mathbf{v}_1, \quad \mathbf{W}_k^*\mathbf{V}_k = D_k, \quad \mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{T}_k$
with D_k non-sing. diag., $\underline{T}_k$ tri-diagonal

Solve $T_k\vec{\gamma} = \rho\, e_1$. With solution $\vec{\gamma}_k$,
take $\mathbf{y}_k = \mathbf{V}_k\,\vec{\gamma}_k$ and $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k$

Bi-Conjugate Gradients

$$T_k = L_k\,U_k \quad \text{need not exist!}$$

$$\mathbf{y}_k = (\mathbf{V}_k\,U_k^{-1})(L_k^{-1}(\rho\, e_1))$$

Expand $\mathbf{V}_k$, $\mathbf{W}_k$, T_k
Expand L_k and U_k
Expand $\mathbf{U}_k := \mathbf{V}_k\,U_k^{-1}$
Expand $(\zeta_0, \ldots, \zeta_{k-1})^T := L_k^{-1}(\rho\, e_1)$
Update $\mathbf{x}_k$ $\quad (= \mathbf{x}_{k-1} + \zeta_{k-1}\,\mathbf{U}_k\,e_k)$

Bi-CG = efficient (in computational costs and storage) implementation of this, using

$$\mathbf{A}\,\mathbf{U}_k = \mathbf{A}\,\mathbf{V}_k\,U_k^{-1} = \mathbf{V}_{k+1}\,\underline{L}_k$$

for expanding $\mathbf{V}_k$.
$(\mathbf{V}_k = [\mathbf{r}_0|\ldots|\mathbf{r}_{k-1}],\ \mathbf{U}_k = [\mathbf{u}_0|\ldots|\mathbf{u}_{k-1}],\ \mathbf{W}_k = [\widetilde{\mathbf{r}}_0|\ldots|\widetilde{\mathbf{r}}_{k-1}])$

[Freund Nachtigal '91]

$\mathbf{r}_0 = \rho\,\mathbf{v}_1, \quad \mathbf{W}_k^*\mathbf{V}_k = D_k, \quad \mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{T}_k$
with D_k non-sing. diag., $\underline{T}_k$ tri-diagonal

Solve $\min_{\vec{\gamma}} \|\rho e_1 - \underline{T}_k\,\vec{\gamma}\|$. With sol. $\vec{\gamma}_k$,
take $\mathbf{y}_k = \mathbf{V}_k\,\vec{\gamma}_k$ and $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{y}_k$.

Quasi Minimal Residual

$$\underline{T}_k = Q_{k+1}\,\underline{R}_k$$

with Q_{k+1} product Givens rotations
$\underline{R}_k$ $(k+1) \times k$ upper triangular

$$\mathbf{y}_k = (\mathbf{V}_k\,R_k^{-1})\vec{\zeta}_k \ \text{ with } \begin{bmatrix}\vec{\zeta}_k \\ \widetilde{\rho}_k\end{bmatrix} := Q_{k+1}^*(\rho\, e_1)$$

QMR = efficient (in computational costs and storage) implementation of this.

QMR versus Bi-CG

– Both methods may suffer from
Lanczos breakdown $\quad (\rho_k = 0)$.

+ Look ahead strategies can easily be incorporated in QMR (much harder in Bi-CG).
In fact, look ahead forms part of QMR.

+ In contrast to QMR, Bi-CG may suffer from
LU-decomposition breakdown
or **pivot breakdown** $\quad (\sigma_k = 0)$.
($T_k = L_k\,U_k$ need not exist: pivot σ_k can be 0)

Cure: composite step Bi-CG [Bank Chan '94]
(uses 2×2 pivots if $\sigma_k = 0$ or σ_k is small).

+? Convergence history QMR tends to be smoother. No faster convergence!!

– QMR is a little less stable
Coupled 2-term recurrences (Bi-CG) appear to be more stable than 3-term recurrence (QMR).

– QMR algorithm is a little more complicated.

Near Breakdowns

Bi-CG breaks down if:

- $\rho_k = 0$ − $\mathbf{r}_k = 0$ **lucky breakdown**
  - − $\tilde{\mathbf{r}}_k = 0$
  - − $\mathbf{r}_k \neq 0$, $\tilde{\mathbf{r}}_k \neq 0$ and $\mathbf{r}_k \perp \tilde{\mathbf{r}}_k$ **Lanczos breakdown**
- $\sigma_k = 0$ **pivot breakdown**

Breakdown hardly ever occurs in practice.
Cure: restart with another $\tilde{\mathbf{r}}_0$.

However, **near-breakdown** may lead to
- inaccurate coefficients (ρ_k or σ_k)
- irregular convergence

Cure
- $\rho_k \approx 0$: look-ahead Lanczos
- $\sigma_k \approx 0$: composite step Bi-CG, QMR

Smoothing the convergence: cosmetic approaches I

$\mathbf{x}_0, \ldots, \mathbf{x}_k$ approximations with residuals $\mathbf{r}_0, \ldots, \mathbf{r}_k$ produced by some method.

[Neumaier '94]

- To "**smooth**" the convergence,

take $\mathbf{r}_k^{\text{best}} = \mathbf{r}_j$ with $\|\mathbf{r}_j\| = \min_{i \le k} \|\mathbf{r}_i\|$
$\mathbf{x}_k^{\text{best}} = \mathbf{x}_j$.

◦ *additional storage required*

[Weiss '90]

- To "**smooth**" the convergence,

take $\mathbf{r}_k^{\text{MR}} = (1-\gamma_k)\,\mathbf{r}_{k-1}^{\text{MR}} + \gamma_k\,\mathbf{r}_k$
$\mathbf{x}_k^{\text{MR}} = (1-\gamma_k)\,\mathbf{x}_{k-1}^{\text{MR}} + \gamma_k\,\mathbf{x}_k$

with γ_k minimizing $\|(1-\gamma)\,\mathbf{r}_{k-1}^{\text{MR}} + \gamma\,\mathbf{r}_k\|$.

◦ *additional storage & DOTs required*

Smoothing the convergence cosmetic approaches II

$\mathbf{x}_0, \ldots, \mathbf{x}_k$ approximations with residuals $\mathbf{r}_0, \ldots, \mathbf{r}_k$ produced by some method.

- To "**smooth**" the convergence, "**quasi-minimize**" $\|\sum \gamma_j\,\mathbf{r}_j\|$, that is:

$$\mathbf{r}_k^{\text{QMR}} = \sum_{j=0}^{k} \gamma_j\,\mathbf{r}_j \quad \text{with} \quad \sum_{j=0}^{k} \gamma_j^2 = 1$$

minimizing $\sum_{j=0}^{k} \gamma_j^2\,\|\mathbf{r}_k\|^2 \qquad (= \tau_k^2)$

Inductively: $\mathbf{r}_k^{\text{QMR}} = (1-\gamma_k)\,\mathbf{r}_{k-1}^{\text{QMR}} + \gamma_k\,\mathbf{r}_k$
$\mathbf{x}_k^{\text{QMR}} = (1-\gamma_k)\,\mathbf{x}_{k-1}^{\text{QMR}} + \gamma_k\,\mathbf{x}_k$

with
$\gamma_k^{-1} = 1 + \tau_{k-1}^{-2}\|\mathbf{r}_k\|^2$ & $\tau_k^{-2} = \tau_{k-1}^{-2} + \|\mathbf{r}_k\|^{-2}$

◦ *no additional DOTs.* Stable?

Theorem. $\|\mathbf{r}_k\| = \tau_k / \sqrt{1 - \tau_k^2/\tau_{k-1}^2}$.

QMR as smoothed Bi-CG

In exact arithmetic, if no breakdown occurs:

Theorem. QMR = producing $\mathbf{x}_k^{\text{QMR}}$ and $\mathbf{r}_k^{\text{QMR}}$ from the Bi-CG approximates and residuals by quasi-minimization.

If QMR reduces well, i.e. $\left\|\mathbf{r}_k^{\text{QMR}}\right\| \ll \left\|\mathbf{r}_{k-1}^{\text{QMR}}\right\|$,
then ($\tau_k \ll \tau_{k-1}$ and) $\left\|\mathbf{r}_k^{\text{QMR}}\right\| \approx \left\|\mathbf{r}_k^{\text{Bi-CG}}\right\|$

− *No real improvement on Bi-CG (in exact arithmetic, if no breakdown occurs).*

- *QMR allows look-ahead strategies.*

Inaccurate coefficients

- Loss of speed of convergence

Irregular convergence

- Loss of speed of convergence
- Loss of accuracy

$$\mathbf{u}_k = \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}$$
$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k$$

Rounded arithmetic

In rounded arithmetic $\cdot^\star$ computed quantity

$\mathbf{r}_{k+1}^\star = \mathbf{r}_k^\star - \alpha_k\,\mathbf{A}\,\mathbf{u}_k^\star + \mathbf{f}_k$
$\mathbf{f}_k$ is the **local** rounding error

Theorem. assuming MV main source of error

$$\boxed{\|\mathbf{f}_k\| \lesssim \overline{\xi}\,\Gamma\left(\|\mathbf{r}_{k+1}^\star\| + \|\mathbf{r}_k^\star\|\right)}$$

where $\overline{\xi}$ is the **relative machine precision**
and $\Gamma := n_A\,\left\||\mathbf{A}|\right\|\,\left\|\mathbf{A}^{-1}\right\|$
with $n_A :=$ max # non-zeros per row of $\mathbf{A}$

Proof. $\|\mathbf{f}_k\| \lesssim \overline{\xi}\,n_A\,\||\mathbf{A}|\|\,\|\alpha_k\mathbf{u}^\star\| \le \overline{\xi}\,\Gamma\,\|\mathbf{A}(\alpha_k\mathbf{u}_k^\star)\|$

$$\mathbf{u}_k = \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1}$$
$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k$$

In rounded arithmetic
$$\mathbf{r}_{k+1}^\star = \mathbf{r}_k^\star - \alpha_k\,\mathbf{A}\,\mathbf{u}_k^\star + \mathbf{f}_k$$
$$\|\mathbf{f}_k\| \lesssim \overline{\xi}\,\Gamma\left(\|\mathbf{r}_{k+1}^\star\| + \|\mathbf{r}_k^\star\|\right)$$

Perturbed convergence

Experimental result

Convergence depends very sensitively on $\mathbf{f}_k$.

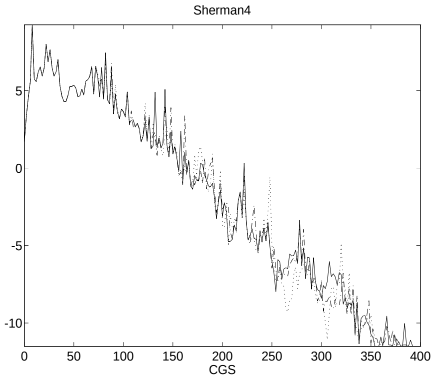
Speed of convergence does **not** depend heavily on $\mathbf{f}_k$ if $\|\mathbf{f}_k\| \ll \|\mathbf{r}_{k+1}^\star\|$

Speed of convergence may be disturbed seriously if $\|\mathbf{f}_k\| \not\ll \|\mathbf{r}_{k+1}^\star\|$

Therefore,
irregularities harmless if $\overline{\xi}\,\Gamma\,\|\mathbf{r}_k^\star\| \ll \|\mathbf{r}_{k+1}^\star\|$

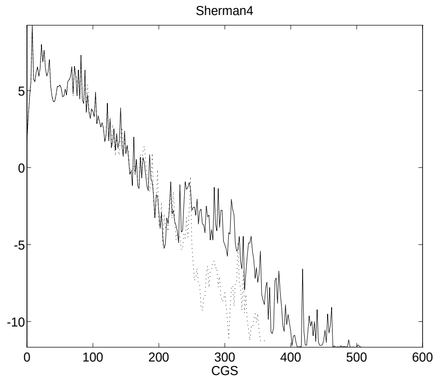
$$\mathbf{r}_{k+1}^\star = \mathbf{r}_k^\star - \alpha_k\,\mathbf{A}\,\mathbf{u}_k^\star + \mathbf{f}_k, \qquad \|\mathbf{f}_k\| \ll \|\mathbf{r}_{k+1}^\star\|$$

Small pert. in (hybrid) Bi-CG



$$\mathbf{r}^\star_{k+1} = \mathbf{r}^\star_k - \alpha_k\,\mathbf{A}\,\mathbf{u}^\star_k + \mathbf{f}_k, \qquad \|\mathbf{f}_k\| \not\ll \|\mathbf{r}^\star_{k+1}\|$$

Large pert. in (hybrid) Bi-CG



$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\,\mathbf{u}_k$$

In rounded arithmetic

$$\mathbf{r}^\star_{k+1} = \mathbf{r}^\star_k - \alpha_k\,\mathbf{A}\,\mathbf{u}^\star_k + \mathbf{f}_k$$
$$\|\mathbf{f}_k\| \lesssim \bar{\xi}\,\Gamma\left(\|\mathbf{r}^\star_{k+1}\| + \|\mathbf{r}^\star_k\|\right)$$

Accuracy

True residual

$$\mathbf{b} - \mathbf{A}\mathbf{x}^\star_{j+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^\star_j - \mathbf{A}(\alpha_j\,\mathbf{u}^\star_j)$$

and recursively updated residual

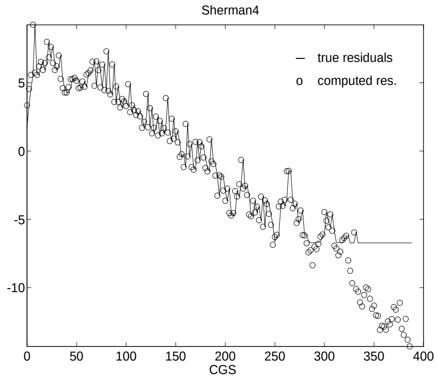
$$\mathbf{r}^\star_{j+1} \quad = \quad \mathbf{r}^\star_j \quad - \alpha_j\,\mathbf{A}\mathbf{u}^\star_j + \mathbf{f}_j$$

drift apart.

Theorem.

$$\boxed{\|(\mathbf{b} - \mathbf{A}\mathbf{x}^\star_k) - \mathbf{r}^\star_k\| \le 2\,\bar{\xi}\,\Gamma \sum_{j\le k} \|\mathbf{r}^\star_j\| \quad \text{where} \quad \Gamma := n_A\,\big\|\,|\mathbf{A}|\,\big\|\,\big\|\,\mathbf{A}^{-1}\big\|}$$

Large residuals in (hybrid) Bi-CG



Avoiding inaccurate solutions

Use $\quad \mathbf{r}_{k+1} \leftarrow \mathbf{b} - \mathbf{A}\,\mathbf{x}_{k+1} \quad$ in each step?

$$\begin{aligned} \mathbf{r}^\star_{k+1} &= \mathbf{b} - (\mathbf{A}\mathbf{x}^\star_{k+1})^\star = \mathbf{b} - \mathbf{A}\mathbf{x}^\star_{k+1} + \mathbf{f}'_k \\ &= \mathbf{b} - \mathbf{A}\mathbf{x}^\star_k - \mathbf{A}(\alpha_k\mathbf{u}^\star_k) + \mathbf{f}'_k \\ &= \mathbf{b} - (\mathbf{A}\mathbf{x}^\star_k)^\star - \mathbf{A}(\alpha_k\mathbf{u}^\star_k) + \mathbf{f}''_k + \mathbf{f}'_k \end{aligned}$$

Additional recursion error:

$$\|\mathbf{f}_k\| \le \bar{\xi}\,\Gamma\,\left(2\,\|\,\mathbf{b}\,\| + \|\,\mathbf{r}^\star_k\,\| + \|\,\mathbf{r}^\star_{k+1}\,\|\right)$$

Disadvantages:

- disturbs convergence if $\|\,\mathbf{r}^\star_{k+1}\,\| \ll \|\,\mathbf{b}\,\|$,
- (may) cost (many) additional MVs.

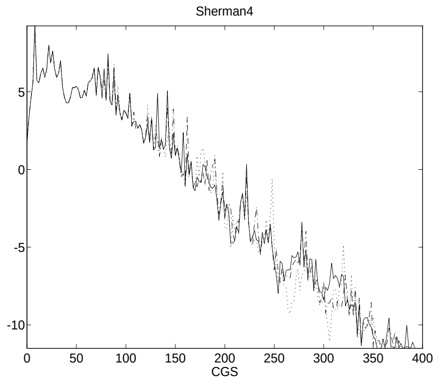
Use $\quad \mathbf{r}_{k+1} \leftarrow \mathbf{b} - \mathbf{A}\,\mathbf{x}_{k+1} \quad$ occasionally!

Additional recursion error:

$$\|\mathbf{f}\| \le \bar{\xi}\,\Gamma\,\left(2\,\|\,\mathbf{b}\,\| + \sum \|\,\mathbf{r}^\star_j\,\|\right),$$

summing over all res. since previous use of true res.

Use $\quad \mathbf{r}_{k+1} \leftarrow \mathbf{b} - \mathbf{A}\,\mathbf{x}_{k+1} \quad$ only if
the residual-norm decreases significantly
and is not (much) smaller than $\|\,\mathbf{b}\,\|$.



7 true residual computations

Bi-Conjugate Gradients $\quad \mathbf{K} = \mathbf{I}$

$$\boxed{\begin{aligned} \mathbf{u}_k &= \mathbf{r}_k - \beta_k\,\mathbf{u}_{k-1} \\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\mathbf{u}_k \end{aligned}}$$

$$\mathbf{r}_k,\ \mathbf{A}\mathbf{u}_k \perp \bar{q}_{k-1}(\mathbf{A}^*)\tilde{\mathbf{r}}_0$$

with q_{k-1} a polynomial of exact degree $k-1$.

With

$$\rho_k := (\mathbf{r}_k, \bar{q}_k(\mathbf{A}^*)\tilde{\mathbf{r}}_0) \quad \& \quad \sigma_k := (\mathbf{A}\mathbf{u}_k, \bar{q}_k(\mathbf{A}^*)\tilde{\mathbf{r}}_0)$$

and θ_k such that

$$q_k(\zeta) - \theta_k\,\zeta\,q_{k-1}(\zeta) \in \mathcal{P}_{k-1}$$

we have

$$\alpha_k = \frac{\rho_k}{\sigma_k} \qquad \& \qquad \beta_k = \frac{\rho_k}{\theta_k\,\sigma_{k-1}}$$

Transpose-free Bi-CG

$$\rho_k = (\mathbf{r}_k, \bar{q}_k(\mathbf{A}^*)\,\tilde{\mathbf{r}}_0) = (q_k(\mathbf{A})\mathbf{r}_k, \tilde{\mathbf{r}}_0),$$
$$\sigma_k = (\mathbf{A}\mathbf{u}_k, \bar{q}_k(\mathbf{A}^*)\,\tilde{\mathbf{r}}_0) = (\mathbf{A}q_k(\mathbf{A})\mathbf{u}_k, \tilde{\mathbf{r}}_0)$$

$$\mathbf{Q}_k := q_k(\mathbf{A})$$

$$\text{(Bi-CG)} \quad \begin{cases} \rho_k, & \mathbf{Q}_k\mathbf{u}_k = \mathbf{Q}_k\mathbf{r}_k - \beta_k\,\mathbf{Q}_k\mathbf{u}_{k-1}, \\ \sigma_k, & \mathbf{Q}_k\mathbf{r}_{k+1} = \mathbf{Q}_k\mathbf{r}_k - \alpha_k\,\mathbf{A}\mathbf{Q}_k\mathbf{u}_k, \end{cases}$$

(Pol) Compute q_{k+1} of degree $k+1$
such that $q_{k+1}(0) = q_k(0) = 1$.
Compute

$$\hat{\hat{\mathbf{u}}}_k := \mathbf{Q}_{k+1}\mathbf{u}_k, \quad \hat{\hat{\mathbf{r}}}_{k+1} := \mathbf{Q}_{k+1}\mathbf{r}_{k+1}$$

Transpose-free Bi-CG; Practice

Work with $\hat{\mathbf{u}}_k = \mathbf{Q}_k\mathbf{u}_k$ and $\hat{\mathbf{r}}_k = \mathbf{Q}_k\mathbf{r}_{k+1}$
Write $\mathbf{u}_k$, $\mathbf{r}_k$, $\mathbf{x}_k$ instead of $\hat{\hat{\mathbf{u}}}_k$, $\hat{\hat{\mathbf{r}}}_k$, $\hat{\hat{\mathbf{x}}}_k$, resp.

$$\rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_0),$$
$$\sigma_k = (\mathbf{A}\,\hat{\mathbf{u}}_k, \tilde{\mathbf{r}}_0)$$

$$\text{(Bi-CG)} \quad \begin{cases} \rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_0), & \hat{\mathbf{u}}_k = \mathbf{r}_k - \beta_k\mathbf{u}_{k-1}, \\ \sigma_k = (\mathbf{A}\,\hat{\mathbf{u}}_k, \tilde{\mathbf{r}}_0), & \hat{\mathbf{r}}_k = \mathbf{r}_k - \alpha_k\,\mathbf{A}\,\hat{\mathbf{u}}_k, \\ & \hat{\mathbf{x}} = \mathbf{x}_k + \alpha_k\hat{\mathbf{u}}_k \end{cases}$$

(Pol) Compute updating coefficients for q_{k+1}.
Compute

$$\mathbf{u}_k, \qquad \mathbf{r}_{k+1}, \qquad \mathbf{x}_{k+1}$$

Hybrid Bi-CG

$$\boxed{\mathbf{r}_k = q_k(\mathbf{A})\,\phi_k^{\text{Bi-CG}}(\mathbf{A})\,\mathbf{r}_0}$$

$r_k := \mathbf{r}_k^{\text{Bi-CG}} := \phi_k^{\text{Bi-CG}}(\mathbf{A})\mathbf{r}_0$:

$\phi_k^{\text{Bi-CG}}$ is the "Bi-CG polynomial"

of degree k

q_k??

q_k for efficient steps & fast convergence.

Fast convergence by

- reducing the residual
- stabilizing the Bi-CG part
- other — when used as linear solver for the Jacobian system in a Newton scheme for non-linear equations, by reducing the number of Newton steps

Properties hybrid Bi-CG

Advantages

- Converges often twice as fast as Bi-CG w.r.t. # MVs:

each MV expands the search subspace

Bi-CG: $\mathbf{x}_k - \mathbf{x}_0 \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ á 2k MV.

Hybrid Bi-CG: $\mathbf{x}_k - \mathbf{x}_0 \in \mathcal{K}_{2k}(\mathbf{A};\mathbf{r}_0)$ á 2k MV.

- Work/MV and storage similar to Bi-CG.
- Transpose free.
- ◦ Explicit computation of Bi-CG scalars.

Drawbacks

- Non-optimal Krylov subspace method.
- Peaks in the convergence history.
- Large intermediate residuals.
- ◦ Breakdown possibilities.

[Sonneveld '89]

CGS: Conjugate Gradients Squared

$$q_k = \phi_k^{\text{Bi-CG}}, \quad \mathbf{r}_k = \phi_k^{\text{Bi-CG}}(\mathbf{A})\,\phi_k^{\text{Bi-CG}}(\mathbf{A})\,\mathbf{r}_0$$

CGS exploits recurrence relations for the Bi-CG polynomials to design a very efficient algorithm.

Properties

\+ Hybrid Bi-CG.

\+ A very efficient algorithm:
1 DOT/MV, 3.25 AXPY/MV;
storage: 7 large vectors.

− Often high peaks in its convergence history

− Often large intermediate residuals

\+ Seems to do well as linear solver in a Newton scheme

CGS

Put $\quad r_k := \mathbf{r}_k^{\text{Bi-CG}}, \quad u_k := \mathbf{u}_k^{\text{Bi-CG}}$

$$\boxed{r_k = \phi_k(\mathbf{A})\mathbf{r}_0 = \Phi_k\mathbf{r}_0,\ u_k = \psi_k(\mathbf{A})\mathbf{r}_0 = \Psi_k\mathbf{r}_0}$$

$$\text{(Bi–CG)} \quad \left\{ \begin{array}{rcll} u_k &=& r_k - \beta_k u_{k-1} & \text{(U)} \\ r_{k+1} &=& r_k - \alpha_k \mathbf{A} u_k & \text{(R)} \end{array} \right.$$

$$\text{(Pol)} \quad \left\{ \begin{array}{rcll} \Psi_k &=& \Phi_k - \beta_k \Psi_{k-1} & \text{(U)} \\ \Phi_{k+1} &=& \Phi_k - \alpha_k \mathbf{A} \Psi_k & \text{(R)} \end{array} \right.$$

$$\theta_k = -\alpha_{k-1} \quad \Rightarrow \quad \beta_k = -\frac{\rho_k}{\rho_{k-1}}$$

$\psi_k - \phi_k$ degree $< k$

$$\Rightarrow \left\{ \begin{array}{l} \rho_k = (\Phi_k r_k, \tilde{\mathbf{r}}_0) = (\Psi_k r_k, \tilde{\mathbf{r}}_0) \\ \sigma_k = (\mathbf{A}\Phi_k u_k, \tilde{\mathbf{r}}_0) = (\mathbf{A}\Psi_k u_k, \tilde{\mathbf{r}}_0) \end{array} \right.$$

Bi-CG(1)	$\Phi_k u_k = \Phi_k r_k - \beta_k \Phi_k u_{k-1}$	$\Phi_k * \text{(U)}$
Bi-CG(2)	$\Psi_k u_k = \Psi_k r_k - \beta_k \Psi_k u_{k-1}$	$\Psi_k * \text{(U)}$
Bi-CG(3)	$\Psi_k r_{k+1} = \Psi_k r_k - \alpha_k \mathbf{A}\Psi_k u_k$	$\Psi_k * \text{(R)}$
Bi-CG(4)	$\Phi_k r_{k+1} = \Phi_k r_k - \alpha_k \mathbf{A}\Phi_k u_k$	$\Phi_k * \text{(R)}$
Pol(5)	$\Phi_{k+1} r_{k+1} = \Phi_k r_{k+1} - \alpha_k \mathbf{A}\Psi_k r_{k+1}$	$\text{(R)} * r_{k+1}$
Pol(6)	$\Phi_{k+1} u_k = \Phi_k u_k - \alpha_k \mathbf{A}\Psi_k u_k$	$\text{(R)} * u_k$
Pol(7)	$\Psi_{k+1} u_k = \Phi_{k+1} u_k - \beta_{k+1} \Psi_k u_k$	$\text{(U)} * u_k$
Pol(8)	$\Psi_{k+1} r_{k+1} = \Phi_{k+1} r_{k+1} - \beta_{k+1} \Psi_k r_{k+1}$	$\text{(U)} * r_{k+1}$

Since $\quad \Psi_k r_k = \Psi_k \Phi_k \mathbf{r}_0 = \Phi_k u_k, \quad \Psi_k r_{k+1} = \Phi_{k+1} u_k$
we have $\quad (6) = (3) \quad$ and $\quad (8) = (1) \quad$ next sweep.

Combine (4) & (5) to save one MV.
Skip (6) & (8). Start at (7). Put
$\mathbf{r}_k := \Phi_k r_k, \quad \mathbf{u}_k := \Phi_{k+1} u_k = \Psi_k r_{k+1}, \quad \mathbf{w}_k := \Psi_k u_k$

[Sonneveld '89]

CGS

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = \mathbf{w}_{-1} = 0,\ \rho_{-1} = 1$
while $\|\mathbf{r}_k\| > \mathit{tol}$ do
$\quad \rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_0), \quad \beta_k = -\rho_k/\rho_{k-1}$
$\quad \mathbf{w} = \mathbf{u}_{k-1} - \beta_k \mathbf{w}_{k-1}$
$\quad \tilde{\mathbf{u}} = \mathbf{r}_k - \beta_k \mathbf{u}_{k-1}$
$\quad \mathbf{w}_k = \tilde{\mathbf{u}} - \beta_k \mathbf{w}$
$\quad \mathbf{c} = \mathbf{A}\mathbf{w}_k$
$\quad \sigma_k = (\mathbf{c}, \tilde{\mathbf{r}}_0), \quad \alpha_k = \rho_k/\sigma_k$
$\quad \mathbf{u}_k = \tilde{\mathbf{u}} - \alpha_k \mathbf{c}$
$\quad \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{A}(\tilde{\mathbf{u}} + \mathbf{u}_k)$
$\quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k(\tilde{\mathbf{u}} + \mathbf{u}_k)$
$\quad k \leftarrow k + 1$
end while

[Sonneveld '89]

CGS

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$.
$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$.
$\mathbf{u} = \mathbf{w} = 0,\ \rho = 1$.
While $\|\mathbf{r}\| > \mathit{tol}$ do
$\quad \sigma \leftarrow -\rho$
$\quad \rho \leftarrow (\mathbf{r}, \tilde{\mathbf{r}}_0), \quad \beta \leftarrow \rho/\sigma$
$\quad \mathbf{w} \leftarrow \mathbf{u} - \beta\,\mathbf{w}$
$\quad \mathbf{v} \leftarrow \mathbf{r} - \beta\,\mathbf{u}$
$\quad \mathbf{w} \leftarrow \mathbf{v} - \beta\,\mathbf{w}$
$\quad \mathbf{c} \leftarrow \mathbf{A}\mathbf{w}$
$\quad \sigma \leftarrow (\mathbf{c}, \tilde{\mathbf{r}}_0), \quad \alpha \leftarrow \rho/\sigma$
$\quad \mathbf{u} \leftarrow \mathbf{v} - \alpha\,\mathbf{c}$
$\quad \mathbf{r} \leftarrow \mathbf{r} - \alpha\,\mathbf{A}(\mathbf{v} + \mathbf{u})$
$\quad \mathbf{x} \leftarrow \mathbf{x} + \alpha\,(\mathbf{v} + \mathbf{u})$
end while

Why high peaks and large res.?

Model. $\quad \mathbf{A}\mathbf{v}_m = \lambda_m \mathbf{v}_m, \qquad \mathbf{r}_0 = \sum \gamma_m \mathbf{v}_m$

in Bi-CG: $\quad r_k = \sum \gamma_m \phi_k(\lambda_m)\mathbf{v}_m$

$|\gamma_k \phi_k(\lambda_m)|$ *moderate or small*

in **CGS**: $\quad \mathbf{r}_k = \sum \gamma_m (\phi_k(\lambda_m))^2 \mathbf{v}_m$

is $|\gamma_k(\phi_k(\lambda_m))^2|$ *moderate or small?*

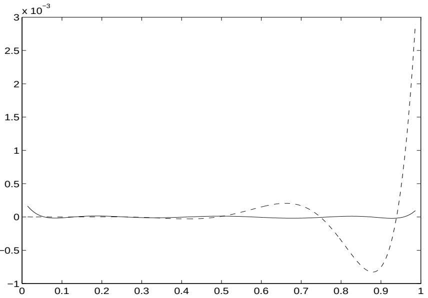
Why nice in a Newton scheme?

For extremal λ_m we may expect
$|\phi_k(\lambda_m)|$ small also for non-large m.
Hence $|\gamma_k(\phi_k(\lambda_m))^2|$ very small.

Inexact Newton with **CGS**
$\approx$ exact Newton w.r.t. extremal components.

How
to preserve the reducing effect of squaring
and to get rid of the amplifying effects?

Idea. Take, for instance, $q_k = \widetilde{\phi}_k$ with
$\widetilde{\phi}_k$ Bi-CG polynomial for $\mathbf{A}$,
but with another $\tilde{\mathbf{r}}_0$, say, $\tilde{\mathbf{s}}_0 \neq \tilde{\mathbf{r}}_0$.



[Fokkema, S, van der Vorst '96]

Generalized CGS

Choose $\mathbf{x}_0$, $\tilde{\mathbf{r}}_0$ and $\tilde{\mathbf{s}}_0$.
$\mathbf{x} = \mathbf{x}_0$, $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}_0$.
$\mathbf{v} = \mathbf{w} = \mathbf{u} = 0$, $\alpha = \tilde{\sigma} = \tilde{\alpha} = \sigma = 1$.
While $\|\mathbf{r}\| > tol$ do
$\quad \sigma \leftarrow -\tilde{\alpha}\sigma, \quad \rho \leftarrow (\mathbf{r}, \tilde{\mathbf{r}}_0), \quad \beta \leftarrow \rho/\sigma$
$\quad \mathbf{w} \leftarrow \mathbf{u} - \beta\,\mathbf{w}$
$\quad \mathbf{v} \leftarrow \mathbf{r} - \beta\,\mathbf{v}$
$\quad \tilde{\sigma} \leftarrow -\alpha\tilde{\sigma}, \quad \tilde{\rho} \leftarrow (\mathbf{r}, \tilde{\mathbf{r}}_0), \quad \tilde{\beta} \leftarrow \tilde{\rho}/\tilde{\sigma}$
$\quad \mathbf{u} \leftarrow \mathbf{r} - \tilde{\beta}\,\mathbf{u}$
$\quad \mathbf{w} \leftarrow \mathbf{v} - \tilde{\beta}\mathbf{w}, \quad \mathbf{c} \leftarrow \mathbf{A}\mathbf{w}$
$\quad \sigma \leftarrow (\mathbf{c}, \tilde{\mathbf{r}}_0),\ \alpha \leftarrow \rho/\sigma,\ \tilde{\sigma} \leftarrow (\mathbf{c}, \tilde{\mathbf{s}}_0),\ \tilde{\alpha} \leftarrow \tilde{\rho}/\tilde{\sigma}$
$\quad \mathbf{v} \leftarrow \mathbf{v} - \tilde{\alpha}\,\mathbf{c}$
$\quad \mathbf{r} \leftarrow \mathbf{r} - \mathbf{A}(\alpha\,\mathbf{v} + \tilde{\alpha}\,\mathbf{u})$
$\quad \mathbf{x} \leftarrow \mathbf{x} + \alpha\,\mathbf{v} + \tilde{\alpha}\,\mathbf{u}$
$\quad \mathbf{u} \leftarrow \mathbf{u} - \alpha\,\mathbf{c}$
end while

[van der Vorst '92]

Bi-CGSTAB: Bi-CG STABilized

Put $\quad r_k := \mathbf{r}_k^{\text{Bi-CG}}, \quad u_k := \mathbf{u}_k^{\text{Bi-CG}}$

$$q_{k+1}(t) = (1 - \omega_k t) q_k$$

$$\begin{aligned} \mathbf{Q}_{k+1}\, r_{k+1} &= (\mathbf{I} - \omega_k \mathbf{A})\, \mathbf{Q}_k\, r_{k+1}, \\ & \quad \mathbf{x}_{k+1} = \hat{\mathbf{x}} + \omega_k\, \mathbf{Q}_k\, r_{k+1} \\ \mathbf{Q}_{k+1}\, u_k &= (\mathbf{I} - \omega_k \mathbf{A})\, \mathbf{Q}_k\, u_k \end{aligned}$$

where ω_k minimizes $\|(\mathbf{I} - \omega\mathbf{A})\, \underbrace{\mathbf{Q}_k\, r_{k+1}}_{= \hat{\mathbf{r}}}\|$

$$\omega_k = \frac{(\hat{\mathbf{r}}, \mathbf{A}\hat{\mathbf{r}})}{(\mathbf{A}\hat{\mathbf{r}}, \mathbf{A}\hat{\mathbf{r}})} \quad \text{and} \quad \theta_k = -\omega_{k-1}$$

[van der Vorst '92]

Bi-CGSTAB

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$
$k = 0$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u}_{-1} = 0, \quad \omega_{-1} = \sigma_{-1} = 1$
while $\|\mathbf{r}_k\| > tol$ do
$\quad \rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_0),\ \beta_k = -\rho_k/(\omega_{k-1}\sigma_{k-1})$
$\quad \hat{\mathbf{u}} \leftarrow \mathbf{r}_k - \beta_k\, \mathbf{u}_{k-1}, \qquad \hat{\mathbf{c}} \leftarrow \mathbf{A}\hat{\mathbf{u}}$
$\quad \sigma_k = (\hat{\mathbf{c}}, \tilde{\mathbf{r}}_0), \quad \alpha_k = \rho_k/\sigma_k$
$\quad \hat{\mathbf{x}} \leftarrow \mathbf{x}_k + \alpha_k\, \hat{\mathbf{u}}$
$\quad \hat{\mathbf{r}} \leftarrow \mathbf{r}_k - \alpha_k\, \hat{\mathbf{c}}, \qquad \hat{\mathbf{s}} \leftarrow \mathbf{A}\hat{\mathbf{r}}$
$\quad \omega_k = (\hat{\mathbf{r}}, \hat{\mathbf{s}})/(\hat{\mathbf{s}}, \hat{\mathbf{s}})$
$\quad \mathbf{u}_k = \hat{\mathbf{u}} - \omega_k\, \hat{\mathbf{c}}$
$\quad \mathbf{x}_{k+1} = \hat{\mathbf{x}} + \omega_k\, \hat{\mathbf{r}}$
$\quad \mathbf{r}_{k+1} = \hat{\mathbf{r}} - \omega_k\, \hat{\mathbf{s}}$
$\quad k \leftarrow k + 1$
end while

[van der Vorst '92]

Bi-CGSTAB

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$
$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{u} = 0, \quad \omega = \sigma = 1$
while $\|\mathbf{r}\| > tol$ do
$\quad \sigma \leftarrow -\omega\sigma$
$\quad \rho \leftarrow (\mathbf{r}, \tilde{\mathbf{r}}_0), \quad \beta \leftarrow \rho/\sigma$
$\quad \mathbf{u} \leftarrow \mathbf{r} - \beta\,\mathbf{u}, \quad \mathbf{c} \leftarrow \mathbf{A}\mathbf{u}$
$\quad \sigma \leftarrow (\mathbf{c}, \tilde{\mathbf{r}}_0), \quad \alpha \leftarrow \rho/\sigma$
$\quad \mathbf{x} \leftarrow \mathbf{x} + \alpha\,\mathbf{u}$
$\quad \mathbf{r} \leftarrow \mathbf{r} - \alpha\,\mathbf{c}, \qquad \mathbf{s} \leftarrow \mathbf{A}\mathbf{r}$
$\quad \omega \leftarrow (\mathbf{r}, \mathbf{s})/(\mathbf{s}, \mathbf{s})$
$\quad \mathbf{u} \leftarrow \mathbf{u} - \omega\,\mathbf{c}$
$\quad \mathbf{x} \leftarrow \mathbf{x} + \omega\,\mathbf{r}$
$\quad \mathbf{r} \leftarrow \mathbf{r} - \omega\,\mathbf{s}$
end while

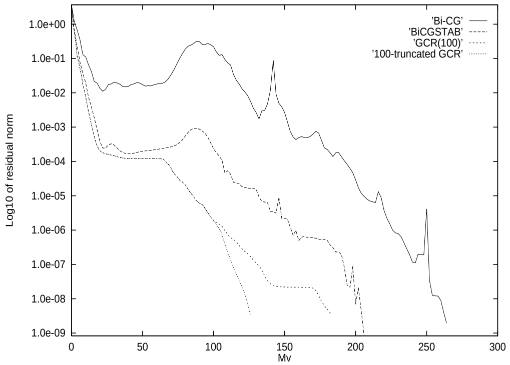
Properties Bi-CGSTAB

Advantages

- Hybrid Bi-CG.
- Converges faster (& smoother) than CGS.
- More accurate than CGS.
- 2 DOT/MV, 3 AXPY/MV.
- Storage: 6 large vectors.

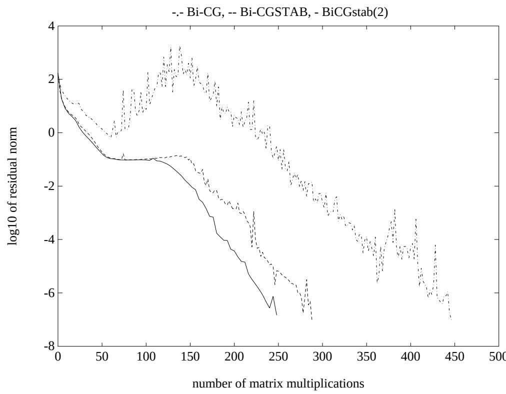
Drawbacks Danger of

(A) Lanczos breakdown $(\rho_k = 0)$,
(B) pivot breakdown $(\sigma_k = 0)$,
(C) breakdown minimization $(\omega_k = 0)$.



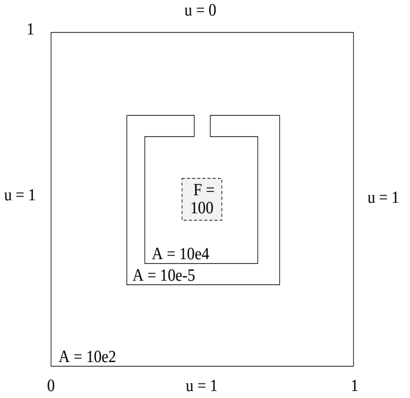
$-(a\,u_x)_x - (a\,u_y)_y = 1$ on $[0, 1] \times [0, 1]$.
$a = 1000$ for $0.1 \leq x, y \leq 0.9$
and $a = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
82×82 volumes. ILU Decomp.

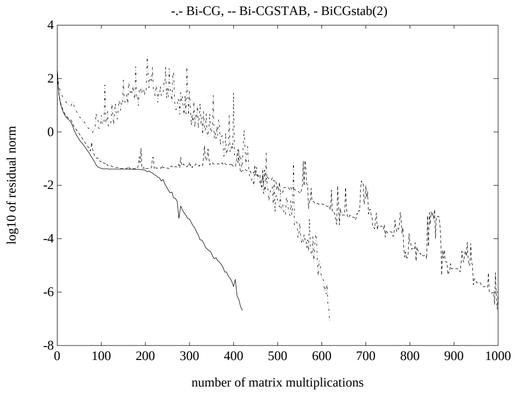


$-(a\,u_x)_x - (a\,u_y)_y + bu_x = f$ on $[0, 1] \times [0, 1]$.
$b(x, y) = 2\exp(2(x^2 + y^2))$, a changes strongly

Dirichlet BC
129×129 volumes. ILU Decomp.



$-(a\,u_x)_x - (a\,u_y)_y + b\,u_x = f$ on $[0,1] \times [0,1]$.
definition of $a = A$ and of $f = F$; $F = 0$ except ...



$-(a\,u_x)_x - (a\,u_y)_y + bu_x = f$ on $[0,1] \times [0,1]$.
$b(x,y) = 2\exp(2(x^2+y^2))$, a changes strongly

Dirichlet BC

201×201 volumes. ILU Decomp.

Breakdown of the minimization

Exact arithmetic, $\omega_k = 0$:

– No reduction of residual by

$$\mathbf{Q}_{k+1}\, r_{k+1} = (\mathbf{I} - \omega_k \mathbf{A}\,)\, \mathbf{Q}_k\, r_{k+1}. \qquad (\star)$$

– q_{k+1} is of degree k:
Bi-CG scalars can not be computed;
breakdown of incorporated Bi-CG.

Finite precision arithmetic, $\omega_k \approx 0$:

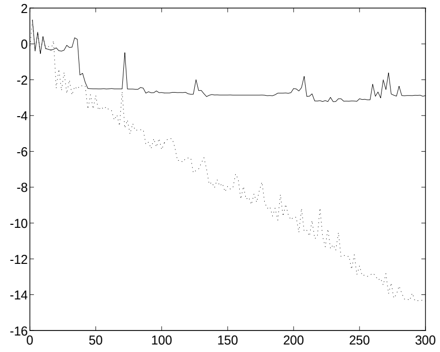
– Poor reduction of residual by $(\star)$

– Bi-CG scalars are seriously affected by evaluation errors: drop of speed of convergence.

$\omega_k \approx 0$ to be expected if

$\mathbf{A}$ has eigenvalues with rel. large imaginary part

ω_k is real!



$$u_{xx} + u_{yy} + u_{zz} + 1000\, u_x = f.$$

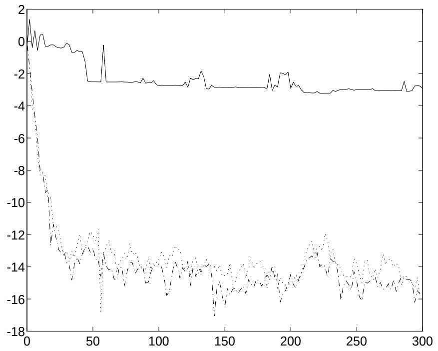
f is defined by the solution
$u(x,y,z) = \exp(xyz)\sin(\pi x)\sin(\pi y)\sin(\pi z)$.

Discretized using $(10 \times 10 \times 10)$ volumes.
No preconditioning was used.

$$\rho_k = (\mathbf{r}_k, \tilde{\mathbf{r}}_0), \qquad \rho_k^\star = \rho_k(1+\epsilon)$$

Accurate Bi-CG coefficients

$$|\epsilon| \le n\,\bar{\xi}\,\frac{\|\mathbf{r}_k\|\,\|\tilde{\mathbf{r}}_0\|}{|(\mathbf{r}_k, \tilde{\mathbf{r}}_0)|} \quad = \frac{n\,\bar{\xi}}{\widehat{\rho}_k} \quad \text{where } \widehat{\rho}_k := \frac{|(\mathbf{r}_k, \tilde{\mathbf{r}}_0)|}{\|\mathbf{r}_k\|\,\|\tilde{\mathbf{r}}_0\|}$$



For optimal accurate Bi-CG coefficient:

minimize

$$\frac{\|\, q_k(\mathbf{A})\, r_k\|\, \|\tilde{\mathbf{r}}_0\|}{|(q_k(\mathbf{A})\, r_k, \tilde{\mathbf{r}}_0)|} = \frac{\|\, q_k(\mathbf{A})\, r_k\|}{|\theta_k|} \cdot \frac{\|\tilde{\mathbf{r}}_0\|}{|(r_k, (\mathbf{A}^*)^k\, \tilde{\mathbf{r}}_0)|}$$

$$\begin{aligned} (q_k(\mathbf{A})\, r_k, \tilde{\mathbf{r}}_0) &= (r_k, q_k(\mathbf{A}^*)\, \tilde{\mathbf{r}}_0) \\ &= (r_k, \theta_k \cdot (\mathbf{A}^*)^k\, \tilde{\mathbf{r}}_0) \end{aligned}$$

$$r_k \perp \mathcal{K}_k(\mathbf{A}^*; \tilde{\mathbf{r}}_0)$$

Solution

$$\left.\begin{array}{l} q_k(\mathbf{A})\, r_k \perp \mathcal{K}_k(\mathbf{A}; r_k) \\ q_k(0) = 1 \end{array}\right\} \quad \begin{array}{l} q_k \text{ is the} \\ \text{FOM polynomial} \end{array}$$

For minimal residuals:

minimize

$$\|\, q_k(\mathbf{A})\, r_k\|$$

Solution

$$\left.\begin{array}{l} q_k(\mathbf{A})\, r_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}; r_k) \\ q_k(0) = 1 \end{array}\right\} \quad \begin{array}{l} q_k \text{ is the} \\ \text{GMRES polynomial} \end{array}$$

[Sleijpen, van der Vorst '95]

$$q_{k+1}(t) = (1 - \omega_k t)\, q_k(t)$$

$$\hat{\mathbf{r}} = q_k(\mathbf{A}) r_{k+1}, \qquad \widehat{\omega} := \frac{(\,\hat{\mathbf{r}}\,, \mathbf{A}\,\hat{\mathbf{r}}\,)}{\|\,\hat{\mathbf{r}}\,\|\,\|\,\mathbf{A}\,\hat{\mathbf{r}}\,\|}$$

Bi-CGSTAB, I

For minimal residuals:

$$\|\,\mathbf{r}^{\mathrm{MR}}\| = \|(\mathbf{I} - \omega\mathbf{A})\,\hat{\mathbf{r}}\,\| \quad \text{minimal for} \quad \omega = \omega^{\mathrm{MR}} = \widehat{\omega}\,\frac{\|\,\hat{\mathbf{r}}\,\|}{\|\,\mathbf{A}\,\hat{\mathbf{r}}\,\|}$$

For optimal accurate Bi-CG coefficient:

$$\frac{\|\,\mathbf{r}^{\mathrm{OR}}\|}{|\,\omega\,|} = \frac{\|(\mathbf{I} - \omega\mathbf{A})\,\hat{\mathbf{r}}\,\|}{|\,\omega\,|} \quad \text{minimal for} \quad \omega = \omega^{\mathrm{OR}} = \frac{1}{\widehat{\omega}}\,\frac{\|\,\hat{\mathbf{r}}\,\|}{\|\,\mathbf{A}\,\hat{\mathbf{r}}\,\|}$$

$$q_{k+1}(t) = (1-\omega_k t)q_k(t)$$

$$\widehat{\mathbf{r}} = q_k(\mathbf{A})r_{k+1}, \qquad \widehat{\omega} := \frac{(\widehat{\mathbf{r}}, \mathbf{A}\,\widehat{\mathbf{r}})}{\|\widehat{\mathbf{r}}\|\,\|\mathbf{A}\,\widehat{\mathbf{r}}\|}$$

Bi-CGSTAB, II

$$\|\mathbf{r}^{\mathrm{OR}}\| = \frac{1}{\widehat{\omega}}\|\mathbf{r}^{\mathrm{MR}}\|, \qquad \frac{\|\mathbf{r}^{\mathrm{MR}}\|}{|\omega^{\mathrm{MR}}|} = \frac{1}{\widehat{\omega}}\frac{\|\mathbf{r}^{\mathrm{OR}}\|}{|\omega^{\mathrm{OR}}|}$$

Compromise:

take

$$\mathbf{r} = \mathbf{r}_{k+1} = (\mathbf{I} - \omega\mathbf{A})\,\widehat{\mathbf{r}}$$

with

$$\omega = \mathrm{sign}(\widehat{\omega})\ \max(0.7, |\widehat{\omega}|)\ \frac{\|\widehat{\mathbf{r}}\|}{\|\mathbf{A}\,\widehat{\mathbf{r}}\|}$$

Updating by pols. of dg. $\ell \geq 2$

Consider $k = m\ell$.

$$q_{m\ell+\ell} = p_m q_{m\ell} = p_m \cdot p_{m-1} \cdot \ldots \cdot p_0,$$

where $\qquad \gamma_\ell \neq 0$

- p_m of exact degree ℓ, $\qquad p_m(\zeta) = \gamma_\ell\,\zeta^\ell + \ldots$
- $p_m(0) = 1$ and, with $\quad \widehat{\mathbf{r}} := q_{m\ell}(\mathbf{A})r_{m\ell+\ell}$,

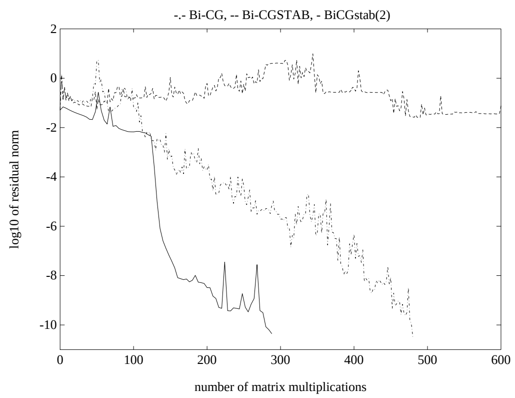
for minimal residuals: [S., Fokkema '93]

- p_m minimizes $\|p_m(\mathbf{A})\,\widehat{\mathbf{r}}\|$ BiCGstab(ℓ)

for accurate ρ's: [S., van der Vorst '95]

- p_m minimizes $\dfrac{\|p_m(\mathbf{A})\,\widehat{\mathbf{r}}\|}{|\gamma_\ell|}$

Compromise can efficiently be computed enhanced ...



$$u_{xx} + u_{yy} + u_{zz} + 1000u_x = f.$$

f s.t. $u(x,y,z) = \exp(xyz)\sin(\pi x)\sin(\pi y)\sin(\pi z)$.
$(52 \times 52 \times 52)$ volumes. No preconditioning.

Updating by pols. dg. $\ell \geq 2$

Consider $k = m\ell$.

$$q_{m\ell+\ell} = p_m q_{m\ell} = p_m \cdot p_{m-1} \cdot \ldots \cdot p_0,$$

where p_m of exact degree ℓ, $p_m(0) = 1$ and

p_m minimizes $\|p_m(\mathbf{A})\,\underbrace{q_{m\ell}(\mathbf{A})\,r_{m\ell+\ell}}_{\widehat{\mathbf{r}}}\|$.

BiCGStab2, $\ell = 2$.

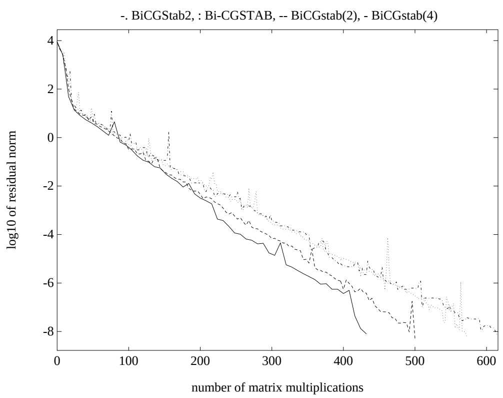
q_{m2}

$q_{m2+1} = (1-\omega_m t)q_{m2} \qquad$ min. w.r.t. ω_m

$q_{m2+2} = p_m q_{m2} = (1-\xi)q_{m2} + (\xi + \eta t)q_{m2+1}$
min. w.r.t. ξ, η

$p_m = 1 - \omega_m' t - \omega_m'' t^2 = (1-\xi) + (\xi + \eta t)(1-\omega_m t)$

Still breakdown dangers in the odd steps!



$-(a\,u_x)_x - (a\,u_y)_y = 1$ on $[0,1]\times[0,1]$.
$a = 1000$ for $0.1 \leq x, y \leq 0.9$
and $a = 1$ elsewhere.

Dirichlet BC on $y = 0$,
Neumann BC on other parts of Boundary.
200×200 volumes. ILU Decomp.

Updating by pol. dg. $\ell \geq 2$

Consider $k = m\ell$.

$$q_{m\ell+\ell} = p_m q_{m\ell} = p_m \cdot p_{m-1} \cdot \ldots \cdot p_0,$$

where p_m of exact degree ℓ, $p_m(0) = 1$ and

p_m minimizes $\|p_m(\mathbf{A})\,\underbrace{q_{m\ell}(\mathbf{A})\,r_{m\ell+\ell}}_{\widehat{\mathbf{r}}}\|$.

BiCGstab(ℓ) $\quad \ell \geq 2$.

With respect to the γ_i, minimize

$$\|\widehat{\mathbf{r}} - \sum_{j=1}^{\ell}\gamma_j\,\mathbf{A}^j\,\widehat{\mathbf{r}}\|, \qquad (\,p_m(t) = 1 - \sum_{j=1}^{\ell}\gamma_j\,t^j\,)$$

Compute $\quad \mathbf{A}\widehat{\mathbf{r}},\ \mathbf{A}^2\widehat{\mathbf{r}},\ \ldots,\ \mathbf{A}^\ell\widehat{\mathbf{r}} \quad$ explicitly.

With $\mathbf{V} := \left[\mathbf{A}\widehat{\mathbf{r}}\,|\,\ldots\,|\,\mathbf{A}^\ell\widehat{\mathbf{r}}\right], \quad \vec{\gamma} := (\gamma_1, \ldots, \gamma_\ell)^T$

we have [Normal Equations, use Choleski]

$$(\mathbf{V}^*\mathbf{V})\vec{\gamma} = \mathbf{V}^*\widehat{\mathbf{r}}$$

$\widehat{\mathbf{r}} := \mathbf{Q}_{m\ell}\,r_{m\ell+\ell} \quad$ and $\quad \widehat{\mathbf{u}} := \mathbf{Q}_{m\ell}\,u_{m\ell+\ell-1}$

Suppose
$\widehat{\mathbf{r}}, \mathbf{A}\widehat{\mathbf{r}}, \ldots, \mathbf{A}^\ell\widehat{\mathbf{r}} \quad$ and $\quad \widehat{\mathbf{u}}, \mathbf{A}\widehat{\mathbf{u}}, \ldots, \mathbf{A}^\ell\widehat{\mathbf{u}}$
have explicitly been computed.

The γ_i minimize

$$\|\widehat{\mathbf{r}} - \sum_{j=1}^{\ell}\gamma_j\,\mathbf{A}^j\,\widehat{\mathbf{r}}\|$$

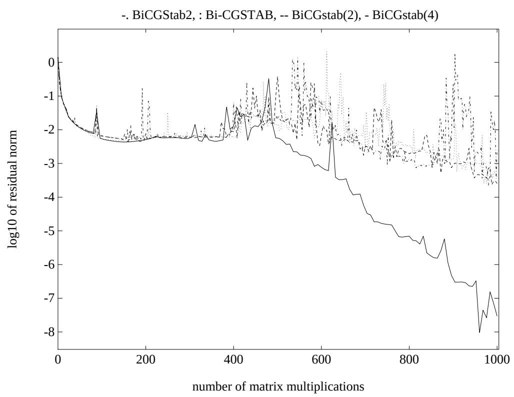
BiCGstab(ℓ) $\quad \ell \geq 2$

If $\qquad p_m(t) := 1 - \sum_{j=1}^{\ell}\gamma_j\,t^j, \qquad$ then

$$\mathbf{r}_{m\ell+\ell} = p_m(\mathbf{A})\widehat{\mathbf{r}} \ = \ \widehat{\mathbf{r}} - \sum_{j=1}^{\ell}\gamma_j\,\mathbf{A}^j\widehat{\mathbf{r}}$$

$$\mathbf{x}_{m\ell+\ell} = \widehat{\mathbf{x}} \ = \ \widehat{\mathbf{x}} + \sum_{j=1}^{\ell}\gamma_j\,\mathbf{A}^{j-1}\widehat{\mathbf{r}}$$

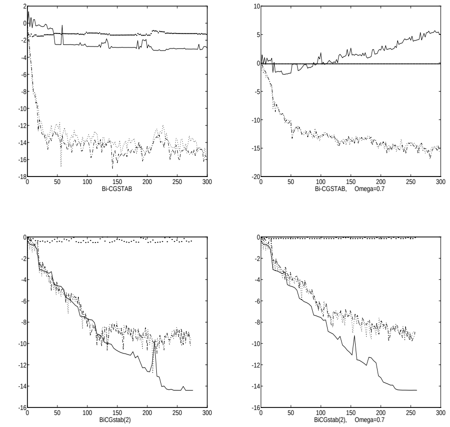
$$\mathbf{u}_{m\ell+\ell-1} = p_m(\mathbf{A})\widehat{\mathbf{u}} \ = \ \widehat{\mathbf{u}} - \sum_{j=1}^{\ell}\gamma_j\,\mathbf{A}^j\widehat{\mathbf{u}}$$



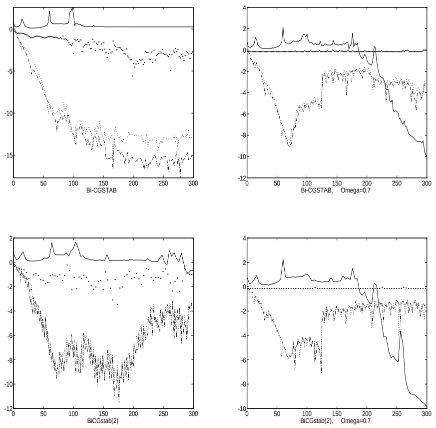
$$-\epsilon(u_{xx} + u_{yy}) + a(x,y)u_x + b(x,y)u_y = 0,$$

on $\quad [0,1]\times[0,1]$, Dirichlet BC $\quad \epsilon = 10^{-1}$

$a(x,y) = 4x(x-1)(1-2y)$,
$b(x,y) = 4y(1-y)(1-2x)$,
$u(x,y) = \sin(\pi x) + \sin(13\pi x) + \sin(\pi y) + \sin(13\pi y)$
(201×201) volumes, no preconditioning.



$u_{xx}+u_{yy}+u_{zz}+1000\,u_x=f,\qquad(10\times10\times10)$ vol.



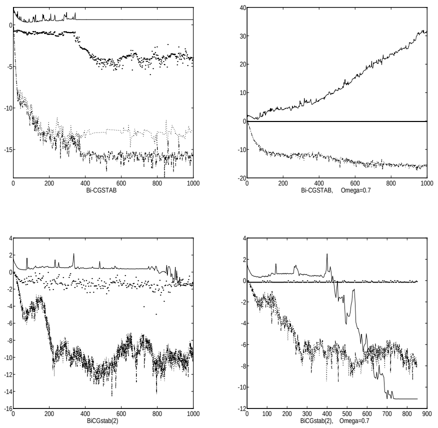
$$-u_{xx}-u_{yy}+\gamma(xu_x+yu_y)+\beta u=f$$

$\gamma=100\quad\&\quad\beta=-200,\quad(63\times63)$ vol.

$$-u_{xx}-u_{yy}+\gamma(xu_x+yu_y)+\beta u=f$$

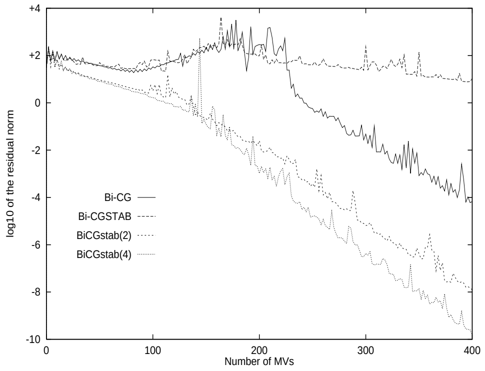
on unit-square, Dirichlet BC,
with discretized solution $x=(1,1,\ldots,1)^T$,
$\gamma=1000$ and $\beta=10$.

(66×66) volumes. No preconditioning.



$$-u_{xx}-u_{yy}+\gamma(xu_x+yu_y)+\beta u=f$$

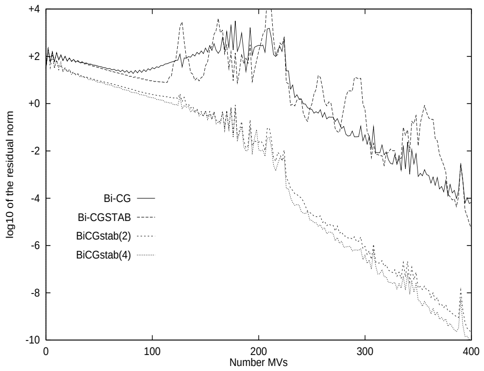
$\gamma=1000\quad\&\quad\beta=10,\quad(66\times66)$ vol.



BiCGstab(ℓ)

$$-u_{xx}-u_{yy}+1000(xu_x+yu_y)+10u=f$$

on the unit cube, Dirichlet BC.
(33×33) volumes. No preconditioning.
f such that $u(x,y)=\exp(xy)\sin(\pi x)\sin(\pi y)$



BiCGstab(ℓ), *using Bi-CG coefficients*

$$-u_{xx}-u_{yy}+1000(xu_x+yu_y)+10u=f$$

on the unit cube, Dirichlet BC
(33×33) volumes. No preconditioning.
f such that $u(x,y)=\exp(xy)\sin(\pi x)\sin(\pi y)$

Properties BiCGstab(ℓ), $\ell=2,4,8$

Advantages

- Hybrid Bi-CG.
- Converges faster than Bi-CGSTAB.
- Stagnation can be avoided
- $1.75+0.25\cdot\ell$ DOT/MV, $2.5+0.5\cdot\ell$ AXPY/MV.
- Storage: $5+2\cdot\ell$ large vectors.

Drawbacks Danger of

(A) Lanczos breakdown $(\rho_k=0)$,

(B) pivot breakdown $(\sigma_k=0)$,

(C) breakdown minimization ($\omega_k=0$) minimized.

Why using pol. fact. of dg.≥ 2?

1 sweep BiCGstab(ℓ) versus ℓ steps Bi-CGSTAB:

- GMRES(ℓ) will reduce residuals better than $\ell\times$ GMRES(1)
- GMRES(ℓ) contributes only once to an increase of $\widehat{\rho}_k$

Why not?

◦ Efficiency:
$1.75+0.25\cdot\ell$ DOT/MV, $2.5+0.5\cdot\ell$ AXPY/MV
Storage: $2\ell+5$ large vector.

◦ Loss of accuracy:

$$\Big|\|\mathbf{r}_k\|-\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|\Big|\leq\ldots+c\,\overline{\xi}\,\max\left(|\gamma_i|\,\Big\|\,|\mathbf{A}|\,|\mathbf{A}^{i-1}\widehat{\mathbf{r}}|\,\Big\|\right)$$

The stability of BiCGstab(ℓ)

Global stability

How is the speed of convergence affected by perturbations on, for instance, the Bi-CG coefficients α_k, β_k?

Local stability

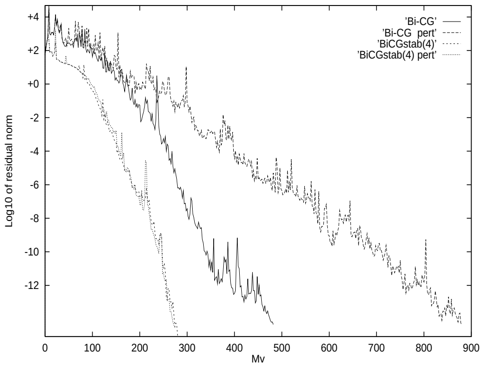
The scaled normal equations are remarkably well-conditioned:

the polynomial coefficients are accurate.

$$\mathbf{V} = [\,\mathbf{A}\hat{\mathbf{r}}\,|\,\mathbf{A}^2\hat{\mathbf{r}}\,|\,\dots\,|\,\mathbf{A}^\ell\hat{\mathbf{r}}\,]$$

$\mathbf{V}\Delta^{-1}$ is well-conditioned,

where $\Delta = \mathrm{diag}(\dots, \|\mathbf{A}^j\hat{\mathbf{r}}\|, \dots)$.



Bi-CG and BiCGstab(4), perturbed and unpert.

Sherman4

Do we need polynomials that are 1 in 0 in the intermediate steps?

For ρ_k, σ_k: $\mathrm{degr}(q_k) = k$. For $\mathbf{r}_k$: $q_k(0) = 1$.

Cure. Use q'_k for ρ_k, σ_k & q''_k for $\mathbf{r}$.

(Bi-CG)

For $i = 0, \dots, \ell-1$, with $k = m\ell + i$,

$$\left\{\begin{array}{ll}
\rho_k = (\mathbf{A}^i\,\mathbf{Q}_{m\ell}\,r_k, \tilde{\mathbf{r}}_0) & q'_k(t) = t^i\,q_{m\ell}(t) \\
\text{for } j = 0, \dots, i & \\
\quad \mathbf{A}^j\,\mathbf{Q}_{m\ell}\,u_k = \mathbf{A}^j\,\mathbf{Q}_{m\ell}\,r_k - \beta_k\,\mathbf{A}^j\,\mathbf{Q}_{m\ell}\,u_{k-1}, & \\
\text{Compute} \quad \mathbf{A}\,\mathbf{A}^i\,\mathbf{Q}_{m\ell}\,u_k & \\
\sigma_k = (\mathbf{A}\,\mathbf{A}^i\,\mathbf{Q}_{m\ell}\,u_k, \tilde{\mathbf{r}}_0) & \\
\text{for } j = 0, \dots, i & \\
\quad \mathbf{A}^j\,\mathbf{Q}_{m\ell}\,r_{k+1} = \mathbf{A}^j\,\mathbf{Q}_{m\ell}\,r_k - \alpha_k\,\mathbf{A}^j\,\mathbf{A}\mathbf{Q}_{m\ell}\,u_k, & \\
\text{Compute} \quad \mathbf{A}\,\mathbf{A}^i\,\mathbf{Q}_{m\ell}\,r_{k+1} & \\
\hat{\mathbf{x}} = \hat{\mathbf{x}} + \alpha_k\,\mathbf{Q}_{m\ell}\,u_k & q''_k(t) = q_{m\ell}(t)
\end{array}\right.$$

$k = m2$

	$\mathbf{u}_{k-1}$ $\mathbf{Q}_k u_{k-1}$	$\mathbf{r}_k$ $\underline{\mathbf{Q}_k r_k}$					B
ρ_k	$\downarrow$ ↙						
	$\mathbf{Q}_k u_k$	$\mathbf{Q}_k r_k$	$\boxed{\mathbf{A}\mathbf{Q}_k u_k}$				i
σ_k		$\downarrow$ ↙					
	$\mathbf{Q}_k u_k$	$\mathbf{Q}_k r_{k+1}$	$\mathbf{A}\mathbf{Q}_k u_k$	$\boxed{\mathbf{A}\mathbf{Q}_k r_{k+1}}$			—
ρ_{k+1}	$\downarrow$ ↙		$\downarrow$ ↙				
	$\mathbf{Q}_k u_{k+1}$	$\mathbf{Q}_k r_{k+1}$	$\mathbf{A}\mathbf{Q}_k u_{k+1}$	$\mathbf{A}\mathbf{Q}_k r_{k+1}$	$\boxed{\mathbf{A}^2\mathbf{Q}_k u_{k+1}}$		C G
σ_{k+1}		$\downarrow$ ↙		$\downarrow$ ↙			
	$\mathbf{Q}_k u_{k+1}$	$\mathbf{Q}_k r_{k+2}$	$\mathbf{A}\mathbf{Q}_k u_{k+1}$	$\mathbf{A}\mathbf{Q}_k r_{k+2}$	$\mathbf{A}^2\mathbf{Q}_k u_{k+1}$	$\boxed{\mathbf{A}^2\mathbf{Q}_k r_{k+2}}$	
	$\mathbf{Q}_k u_{k+1}$	$\mathbf{Q}_k r_{k+2}$	$\mathbf{A}\mathbf{Q}_k u_{k+1}$	$\mathbf{A}\mathbf{Q}_k r_{k+2}$	$\mathbf{A}^2\mathbf{Q}_k u_{k+1}$	$\mathbf{A}^2\mathbf{Q}_k r_{k+2}$	P o l
γ_1, γ_2	$\mathbf{Q}_{k+2} u_{k+1}$ $\mathbf{u}_{k+1}$	$\mathbf{Q}_{k+2} r_{k+2}$ $\mathbf{r}_{k+2}$					

The computational scheme for BiCGstab(2): Power basis variant


```
problem
% Accuracy (residual reduction), choice "ell".
epsilon=10^(-16); ell=4

% Choice initial approximation x
% and shadow residual rt
x=zeros(b); rt=rand(b);

% Initiation scalars and search direction
sigma=1; theta=1; u=zeros(b);
% Computation of initial residual
y=x; MV; y=b-y; r=y;

norm=r'*r;  epsilon=norm*epsilon^2; L=2:ell+1;

while norm > epsilon

   sigma=-theta*sigma; y=r;
   for j=1:ell
        rho=y'*rt;   beta=rho/sigma;
        u=r-beta*u;
        y=u(:,j); MV; u(:,j+1)=y;
        sigma=y'*rt; alpha=rho/sigma;
        x=x+alpha*u(:,1);
        r=r-alpha*u(:,2:j+1);
        y=r(:,j); MV; r(:,j+1)=y;
   end

   G=r'*r; gamma=G(L,L)\G(L,1); theta=gamma(ell);
   x=x+r*[gamma;0]; r=r*[1;-gamma]; u=u*[1;-gamma];
   norm=r'*r;

end
```

[Sleijpen, Fokkema '93]

BiCGstab(ℓ) Power basis variant

Choose $\mathbf{x}_0$ and $\tilde{\mathbf{r}}_0$.
$\mathbf{x} = \mathbf{x}_0$, $\mathbf{r} = [\mathbf{b} - \mathbf{A}\mathbf{x}]$.
$\mathbf{u} = [0]$, $\gamma_\ell = \sigma = 1$.
While $\|\mathbf{r}\| > tol$ do
 $\sigma \leftarrow -\gamma_\ell\sigma$
 for $j = 1$ to ℓ do
  $\rho \leftarrow (\mathbf{r}_j, \tilde{\mathbf{r}}_0)$, $\beta \leftarrow \rho/\sigma$
  $\mathbf{u} \leftarrow \mathbf{r} - \beta\,\mathbf{u}$, $\mathbf{u} \leftarrow [\mathbf{u}, \mathbf{A}\mathbf{u}_j]$
  $\sigma \leftarrow (\mathbf{u}_{j+1}, \tilde{\mathbf{r}}_0)$, $\alpha \leftarrow \rho/\sigma$
  $\mathbf{x} \leftarrow \mathbf{x} + \alpha\,\mathbf{u}_1$
  $\mathbf{r} \leftarrow \mathbf{r} - \alpha\,[\mathbf{u}_2, \dots, \mathbf{u}_{j+1}]$, $\mathbf{r} \leftarrow [\mathbf{r}, \mathbf{A}\mathbf{r}_j]$
 end do
 $\mathbf{R} := [\mathbf{r}_2, \dots, \mathbf{r}_{\ell+1}]$
 solve $\vec{\gamma}$ from $(\mathbf{R}^T\mathbf{R})\vec{\gamma} = \mathbf{R}^T\mathbf{r}_1$
 $\mathbf{u} \leftarrow [\mathbf{u}_1 - (\gamma_1\,\mathbf{u}_2 + \dots + \gamma_\ell\,\mathbf{u}_{\ell+1})]$
 $\mathbf{x} \leftarrow \mathbf{x} + (\gamma_1\,\mathbf{r}_1 + \dots + \gamma_\ell\,\mathbf{r}_\ell)$
 $\mathbf{r} \leftarrow [\mathbf{r}_1 - (\gamma_1\,\mathbf{r}_2 + \dots + \gamma_\ell\,\mathbf{r}_{\ell+1})]$
end while

$\mathbf{r}$ and $\mathbf{u}$ are matrices with n rows, # columns varies:
$\mathbf{r} := [\mathbf{r}_1, \dots, \mathbf{r}_j]$, $\mathbf{u} := [\mathbf{u}_1, \dots, \mathbf{u}_j]$. $\vec{\gamma} := (\gamma_1, \dots, \gamma_\ell)^T$

Conclusions

Hybrid Bi-CG meths are efficient linear solvers.

The auxiliary polynomial can be used

- to gain efficiency and
- to improve residual reduction

but also for improving other properties

- to obtain more accurate Bi-CG coefficients
- to improve inexact Newton schemes

Effects of erratic convergence can be reduced:

- Inaccurate results due to large intermediate residuals can inexpensively be avoided.
- Large residual peaks may deteriorate convergence, but can be avoided (BiCGstab(ℓ), GCGS).

Program

- Galerkin approach
- Arnoldi type of methods
  Arnoldi
  Lanczos
  RKS (rational Krylov Sequence)
  ⋮
- ◦ IRA (Implicitly Restarted Arnoldi)
- Davidson type of methods
- Jacobi-Davidson
- Deflation
- Generalized eigenproblems

$$\boxed{\mathbf{A}\,\mathbf{x}_\lambda = \lambda\,\mathbf{x}_\lambda}$$

Galerkin approach

- Generate a subspace with basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$.
- With $\mathbf{V} := [\mathbf{v}_1|\ldots|\mathbf{v}_k]$ find y and θ such that

$$\mathbf{r} := \mathbf{AV}y - \theta\mathbf{V}y \perp \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$$

$(\theta, \mathbf{V}y)$ **Ritz pair** w.r.t. **search space** span$(\mathbf{V})$.

$(\theta, \mathbf{V}y)$ with **residual** $\mathbf{r}$ approximates $(\lambda, \mathbf{x}_\lambda)$.

In practice. $k \ll n$

- Generate an orthonormal system $\mathbf{v}_1, \ldots, \mathbf{v}_k$:

$\mathbf{z}$ expansion vector $\mathbf{V} \Rightarrow \mathbf{v}_{k+1} := \text{modGS}_{\mathbf{V}}(\mathbf{z})$.

- solve the small problem

$$\mathbf{V}^*\mathbf{AV}y = \theta\, y$$

Subspace iteration

1. **Start:**

   Choose an initial subspace $\mathbf{V}$.

2. **Repeat:**
   - Compute $\mathbf{W} \leftarrow \mathbf{AV}$, $H \leftarrow \mathbf{V}^*\mathbf{W}$.
   - Compute a desired eigenpair (θ, y) of H (with $\|y\| = 1$).
   - Compute $\mathbf{u} \leftarrow \mathbf{V}y$,

     $$\mathbf{r} \leftarrow \mathbf{A}\,\mathbf{u} - \theta\,\mathbf{u}.$$
   - **Stop if** satisfied.
   - Compute an expansion vector $\mathbf{z}$
   - $\mathbf{V} \leftarrow \text{ModGS}([\mathbf{V}\,|\,\mathbf{z}])$

Choices for the expansion vector

with initial search subspace span$(\mathbf{v}_1)$.

Arnoldi's method ['51]

$$\mathbf{z} = \mathbf{A}\mathbf{v}_k.$$

Then $H := \mathbf{V}^*\mathbf{AV}$ is upper Hessenberg.

Lanczos' method if $\mathbf{A}^* = \mathbf{A}$ ['50]

$$\mathbf{z} = \mathbf{A}\mathbf{v}_k.$$

Then $H := \mathbf{V}^*\mathbf{AV}$ is tridiagonal.

Search space span$(\mathbf{V})$ = Krylov subspace

$$\text{span}\big(\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}^{k-1}\mathbf{v}_1\big).$$

Arnoldi (Lanczos) favors extremal eigenvalues

Exponential Arnoldi [Meerbergen & Sadkane '96]

$$\mathbf{z} = \exp(\mathbf{A})\mathbf{v}_k.$$

Approach favors eigenvalues λ with Re(λ) large

(amplifies $\mathbf{x}_\lambda$-comp. with $\exp(\lambda)$)

In practice.

With $\mathbf{U}$ orthornormal matrix of basisvectors of Krylov subspace span$(\mathbf{v}_k, \mathbf{A}\mathbf{v}_k, \ldots, \mathbf{A}^m\mathbf{v}_k)$

and $H' := \mathbf{U}^*\mathbf{AU}$ take $\mathbf{z} = \mathbf{U}[\exp(H')e_1]$

Arnoldi (Lanczos) favors extremal eigenvalues

Shift-and-Invert-Arnoldi

Select target $\tau \in \mathbb{C}$.

$$\mathbf{z} = (\mathbf{A} - \tau\,\mathbf{I})^{-1}\mathbf{v}_k.$$

Then $H = \mathbf{V}^*\mathbf{AV}$ is upper Hessenberg.

Approach favors eigenvalues close to τ

(amplifies $\mathbf{x}_\lambda$-comp. with $\frac{1}{\lambda-\tau}$)

Rational Krylov Sequence [Ruhe '84, '93]

$$\mathbf{z} = (\mathbf{A} - \tau_k\,\mathbf{I})^{-1}(\mathbf{A} - \sigma_k\mathbf{I})\mathbf{v}_k.$$

(amplifies $\mathbf{x}_\lambda$-comp. with $\frac{\lambda-\sigma_k}{\lambda-\tau_k}$)

However, for large problems,

"solve $(\mathbf{A} - \tau\,\mathbf{I})\mathbf{z} = \mathbf{v}_k$"

is expensive and has to be done accurately.

$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ orthogonal basis search space.

Put $\mathbf{V}_k = [\mathbf{v}_1|\ldots|\mathbf{v}_k]$.

Arnoldi ['51]

Expansion vector $\mathbf{v}_{k+1}$ computed as

$$\mathbf{v}_{k+1} = \text{modGS}_{\mathbf{V}_k}(\mathbf{A}\mathbf{v}_k).$$

Then $\boxed{\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k}$

with $\underline{H}_k$ $(k+1) \times k$ upper Hessenberg.

Solve $H_k\,y - \theta\,y = 0$. Take $\mathbf{u} = \mathbf{V}_k\,y$.

$$\boxed{\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k}$$

$$(H_k - \theta\,I)\,y = 0, \qquad \mathbf{u} := \mathbf{V}_k\,y$$

The Arnoldi residual

$$\begin{aligned}\mathbf{r} &:= (\mathbf{A} - \theta\,\mathbf{I})\mathbf{u} \\ &= (\mathbf{A} - \theta\,\mathbf{I})\mathbf{V}_k\,y \\ &= \mathbf{V}_{k+1}\,(\underline{H}_k - \theta\,I)\,y\end{aligned}$$

$$\mathbf{r} = h_{k+1\,k}\; e_k^* y\;\mathbf{v}_{k+1}$$

$$\|\mathbf{r}\| = |h_{k+1\,k}|\;|e_k^* y|$$

$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ orthogonal basis search space.

Put $\mathbf{V}_k = [\mathbf{v}_1|\ldots|\mathbf{v}_k]$.

Arnoldi ['51]

Expansion vector $\mathbf{v}_{k+1}$ computed as

$$\mathbf{v}_{k+1} = \text{modGS}_{\mathbf{V}_k}(\mathbf{A}\mathbf{v}_k).$$

Then $\boxed{\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k}$

with $\underline{H}_k$ $(k+1) \times k$ upper Hessenberg.

Implicitly Restarted Arnoldi [Sorensen 92]

Select polynomial $\Psi(\lambda) = (\lambda - \mu_\ell)\cdots(\lambda - \mu_1)$.

Then, after **implicit restart**,

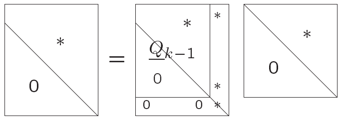
$$\mathbf{A}\widetilde{\mathbf{V}}_{k-\ell} = \widetilde{\mathbf{V}}_{k+1-\ell}\,\underline{\widetilde{H}}_{k-\ell}$$

$$\Psi(\mathbf{A})\mathbf{v}_1 = \gamma\,\tilde{\mathbf{v}}_1.$$

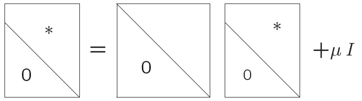
Can be computed without MVs and DOTs!

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

$\underline{H}_k - \mu I \quad = \quad \underline{Q}_k \quad R_k$



$\widetilde{\underline{H}}_{k-1} \quad = \quad R_k \quad \underline{Q}_{k-1} \quad +\mu I$



With $\widetilde{\mathbf{V}}_i := \mathbf{V}_{i+1}\,\underline{Q}_i, \quad \widetilde{\underline{H}}_{k-1} = \underline{Q}_k^*\,\underline{H}_k\,\underline{Q}_{k-1}$

we have $\quad \underline{Q}_k\,\widetilde{\underline{H}}_{k-1} = \underline{H}_k\,\underline{Q}_{k-1}$

$$\mathbf{A}\widetilde{\mathbf{V}}_{k-1} = \widetilde{\mathbf{V}}_k\,\widetilde{\underline{H}}_{k-1}$$

and $\quad (\mathbf{A} - \mu\,\mathbf{I})\mathbf{V}_k = \mathbf{V}_{k+1}(\underline{H}_k - \mu\,I) = \widetilde{\mathbf{V}}_k\,R_k$

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

Lemma. $\underline{H}_k - \mu\,I = \underline{Q}_k\,R_k$.

With $\widetilde{\mathbf{V}}_i := \mathbf{V}_{i+1}\,\underline{Q}_i, \quad \widetilde{\underline{H}}_{k-1} = \underline{Q}_k^*\,\underline{H}_k\,\underline{Q}_{k-1}$

we have $\quad \mathbf{A}\widetilde{\mathbf{V}}_{k-1} = \widetilde{\mathbf{V}}_k\,\widetilde{\underline{H}}_{k-1}$

and, $\quad (\mathbf{A} - \mu\,\mathbf{I})\mathbf{v}_1 = \gamma\,\tilde{\mathbf{v}}_1, \qquad$ *for some* γ.

Theorem. $\Psi(\lambda) := (\lambda - \mu_\ell)\cdots(\lambda - \mu_1)$.

Then $\quad \mathbf{A}\widetilde{\widetilde{\mathbf{V}}}_{k-\ell} = \widetilde{\widetilde{\mathbf{V}}}_{k+1-\ell}\,\widetilde{\widetilde{\underline{H}}}_{k-\ell}$

and $\quad \Psi(\mathbf{A})\mathbf{v}_1 = \gamma\,\tilde{\tilde{\mathbf{v}}}_1, \qquad$ *for some* γ.

where $\quad \widetilde{\widetilde{\mathbf{V}}}_{k-\ell} := \mathbf{V}_k\,Q', \quad \widetilde{\widetilde{\underline{H}}}_{k-\ell} = Q^*\,\underline{H}_k\,Q'$.

An implicit restart step

$$\mathbf{A}\,\mathbf{V}_k = \mathbf{V}_{k+1}\,\underline{H}_k$$

Q' is $\underline{Q}$ without last row and last column

```
Select  μ_1,...,μ_ℓ ∈ ℂ
Set  H = H_k,  Q̃ = I_k
For j = 1,...,ℓ do
   Factorise  H − μ_j I_{k+1−j} = Q R
   Compute
     H ← R Q' + μ_j I_{k−j}
     Q̃ ← Q̃ Q
end for
Set  H̃̃_{k−ℓ} = H
Compute  Ṽ̃_{k+1−ℓ} = V_{k+1} Q̃,
```

Select $\mu_1, \ldots, \mu_\ell \in \mathbb{C}$

Set $\underline{H} = \underline{H}_k, \quad \tilde{Q} = \underline{I}_k$

For $j = 1, \ldots, \ell$ do

Factorise $\underline{H} - \mu_j\,\underline{I}_{k+1-j} = \underline{Q}\,R$

Compute

$\underline{H} \leftarrow R\,Q' + \mu_j\,\underline{I}_{k-j}$

$\tilde{Q} \leftarrow \tilde{Q}\,\underline{Q}$

end for

Set $\widetilde{\widetilde{\underline{H}}}_{k-\ell} = \underline{H}$

Compute $\widetilde{\widetilde{\mathbf{V}}}_{k+1-\ell} = \mathbf{V}_{k+1}\,\tilde{Q}$,

---

$$\mathbf{A}\,\widetilde{\widetilde{\mathbf{V}}}_{k-\ell} = \widetilde{\widetilde{\mathbf{V}}}_{k+1-\ell}\,\widetilde{\widetilde{\underline{H}}}_{k-\ell}$$

$$\underline{H}_k - \mu\,I = \underline{Q}_k\,R_k, \qquad \widetilde{\underline{H}}_{k-1} = R_k\underline{Q}_{k-1} + \mu\,I$$

Proposition. $\underline{H}_k$ irreducible

$$\mu \in \sigma(H_k) \ \Rightarrow\ \sigma(\widetilde{H}_{k-1}) = \sigma(H_k)\backslash\{\mu\}$$

Proof. $\mu \in \sigma(H_k)$ then $H_k - \mu$ singular and $\underline{Q}_k e_k = e_{k+1}$.

Lemma. $\underline{Q}_k e_k = e_{k+1}$.

Then, with $\tilde{Q}_k := [\underline{Q}_{k-1}\,|\,q]$ orthonormal,

$$\tilde{Q}_k^*\,H_k\,\tilde{Q}_k = \begin{bmatrix} \widetilde{H}_{k-1} & * \\ 0 & \mu \end{bmatrix}$$

Proof. With $\widetilde{R}_k := R_k - \rho\,e_k e_k^*$ such that $e_k^*\widetilde{R}_k = 0$,

we have $H_k = \tilde{Q}_k\,\widetilde{R}_k + \mu$.

$$e_k^*\widetilde{R}_k\tilde{Q}_k = 0, \ \ \widetilde{R}_k\tilde{Q}_k[e_1, \ldots, e_{k-1}] = R_k\underline{Q}_{k-1} - \rho\,e_k e_k^*\,\underline{Q}_{k-1}$$
$$= R_k\underline{Q}_{k-1} - \eta\,e_k e_{k-1}^*.$$

Hence $\quad \tilde{Q}_k^*\,H_k\,\tilde{Q}_k = \widetilde{R}_k\,\tilde{Q}_k + \mu = \begin{bmatrix} \widetilde{H}_{k-1} & * \\ 0 & \mu \end{bmatrix}$

$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ orthogonal basis search space.

Put $\quad \mathbf{V}_k = [\mathbf{v}_1|\ldots|\mathbf{v}_k]$.

RKS (Rational Krylov Sequence)

Select $\mu_k \in \mathbb{C}, \quad \vec{t} \in \mathbb{C}^k$.

Compute expansion vector $\mathbf{v}_{k+1}$ as

$$\mathbf{v}_{k+1} = \mathrm{modGS}_{\mathbf{V}_k}\left((\mathbf{A} - \mu_k\,\mathbf{B})^{-1}\mathbf{B}\mathbf{V}_k\,\vec{t}\right).$$

Then $\quad \mathbf{A}\,\mathbf{V}_{k+1}\,\vec{h} = \mathbf{B}\,\mathbf{V}_{k+1}\left(\mu_k\,\vec{h} + \vec{t}\right)$.

Hence $\quad \boxed{\mathbf{A}\,\mathbf{V}_{k+1}\,\underline{H}_k = \mathbf{B}\,\mathbf{V}_{k+1}\,\underline{K}_k}$

with $(\underline{H}_k, \underline{K}_k)$ strongly irreducible Hessenberg

Solve $(K_k - \theta\,H_k)\,\vec{y} = 0$. Take $\mathbf{u} := \mathbf{V}_{k+1}\,\underline{H}_k\,\vec{y}$.

Suppose $\underline{H}_k$ and $\underline{K}_k$ are Hessenberg.

Definition.

- $\underline{H}_k$ is **irreducible** if $h_{j+1,j} \neq 0 \ \ \forall j$
- $(\underline{H}_k, \underline{K}_k)$ **strongly irreducible** if
  $\underline{H}_k$ irreducible and
  $\underline{K}_k - \mu\,\underline{H}_k$ full rank $\forall\mu$.

Property. If $\underline{H}_k$ is irreducible then

$$\underline{K}_k = \underline{H}_k\,D + T$$

for some diagonal $D = D(\underline{H}_k, \underline{K}_k)$
and upper triangular $T = T(\underline{H}_k, \underline{K}_k)$.

Proof. $\quad D = \mathrm{diag}(k_{j+1,j}/h_{j+1,j})$.

In RKS: $\quad D = \mathrm{diag}(\mu_1, \ldots, \mu_k)$

$$T = [\vec{t}_1, \ldots, \vec{t}_k].$$

A RKS expansion step

$$\mathbf{A}\,\mathbf{V}_k\,\underline{H}_{k-1} = \mathbf{B}\,\mathbf{V}_k\,\underline{K}_{k-1}$$

Select a $\mu \in \mathbb{C}$ and a $\vec{t} \in \mathbb{C}^k$

Solve $\tilde{\mathbf{v}}$ from $(\mathbf{A} - \mu\,\mathbf{B})\tilde{\mathbf{v}} = \mathbf{V}_k\,\vec{t}$

$\vec{h} = \mathbf{V}_k^*\,\tilde{\mathbf{v}}, \quad \tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{V}_k\,\vec{h}$

$h = \|\tilde{\mathbf{v}}\|, \quad \mathbf{v}_{k+1} = \tilde{\mathbf{v}}/h$

$\mathbf{V}_{k+1} = [\mathbf{V}_k\,|\,\mathbf{v}_{k+1}]$

$$\underline{H}_k = \begin{bmatrix} \underline{H}_{k-1} & \vec{h} \\ 0\ldots0 & h \end{bmatrix}, \quad \underline{K}_k = \begin{bmatrix} \underline{K}_{k-1} & \mu\,\vec{h} + \vec{t} \\ 0\ldots0 & \mu\,h \end{bmatrix}$$

$$\mathbf{A}\,\mathbf{V}_{k+1}\,\underline{H}_k = \mathbf{B}\,\mathbf{V}_{k+1}\,\underline{K}_k$$

Selecting the anti-shift μ_k

Examples

- $\mu_k = \theta, \quad \vec{t} = \vec{y}$ with $(K_k - \theta\,H_k)\,\vec{y} = 0$.

However:
LU-factorisation $\mathbf{A} - \mu_k\,\mathbf{B}$ may be expensive.

- $\mu_k = \mu_\ell, \quad \vec{t} = e_k$

- $\mu_k = \mu_\ell, \quad \vec{t} = \vec{y}$

$$\boxed{\mathbf{A}\,\mathbf{V}_{k+1}\,\underline{H}_k = \mathbf{B}\,\mathbf{V}_{k+1}\,\underline{K}_k}$$

$(K_k - \theta\, H_k)\,\vec{y} = 0, \qquad \mathbf{u} := \mathbf{V}_{k+1}\,\underline{H}_k\,\vec{y}$

The RKS residual

$$\begin{aligned}\mathbf{r} &:= (\mathbf{A} - \theta\,\mathbf{B})\mathbf{u}\\ &= (\mathbf{A} - \theta\,\mathbf{B})\mathbf{V}_{k+1}\,\underline{H}_k\,\vec{y}\\ &= \mathbf{V}_{k+1}\,(\underline{K}_k - \theta\,\underline{H}_k)\,\vec{y}\end{aligned}$$

$$\mathbf{r} = (k_{k+1\,k} - \theta\,h_{k+1\,k})\;e_k^*\vec{y}\;\mathbf{v}_{k+1}$$

$$\|\mathbf{r}\| = |k_{k+1\,k} - \theta\,h_{k+1\,k}|\;|e_k^*\vec{y}|$$

$\underline{H}_k$, $\underline{K}_k$ Hessenberg. $\quad \underline{H}_k$ with QR-decomp.

$$\underline{H}_k = \underline{Q}_k\,R_k.$$

Let $q \in \mathbb{C}^{k+1}$, $\|q\| = 1$ such that $q^*\underline{Q}_k = 0$.
$\underline{Z}_{k-1}$ orthogonal $k \times (k-1)$ Hessenberg s.t.

$$q^*\underline{K}_k\,\underline{Z}_{k-1} = 0$$

Put

$$\underline{\widetilde{H}}_{k-1} := \underline{Q}^*\,\underline{H}_k\,\underline{Z}_{k-1} \;\&\; \underline{\widetilde{K}}_{k-1} := \underline{Q}^*\,\underline{K}_k\,\underline{Z}_{k-1}$$

Lemma.
$(\underline{H}_{k-1}, \underline{K}_{k-1})$ is strongly irreducible
if $(\underline{H}_k, \underline{K}_k)$ is strongly irreducible.

$\mathbf{C}\,\mathbf{V}_{k+1}\,\underline{K}_k = \mathbf{V}_{k+1}\,\underline{H}_k$ with
$(\underline{H}_k, \underline{K}_k)$ strongly irr. Hessenberg,

---

Lemma. Factorize $\quad \underline{K}_k - \sigma\,\underline{H}_k = \underline{Q}_k\,R_k$.
With q and $\underline{Z}_{k-1}$ such that
$[\underline{Q}_k, q]$ unitary and
$[\underline{Z}_{k-1}, (\bar{\sigma}\underline{K}_k + \underline{H}_k)^*\,q]$ unitary Hessenberg

With $\quad \widetilde{\mathbf{V}}_k := \mathbf{V}_{k+1}\,\underline{Q}_k$

$\quad \underline{\widetilde{H}}_{k-1} := \underline{Q}_k^*\,\underline{H}_k\,\underline{Z}_{k-1}$

$\quad \underline{\widetilde{K}}_{k-1} := \underline{Q}_k^*\,\underline{K}_k\,\underline{Z}_{k-1}$

we have $\quad \mathbf{C}\,\widetilde{\mathbf{V}}_k\,\underline{\widetilde{K}}_{k-1} = \widetilde{\mathbf{V}}_k\,\underline{\widetilde{H}}_{k-1}$,
$(\underline{\widetilde{H}}_{k-1}, \underline{\widetilde{K}}_{k-1})$ strongly irr. Hessenberg,
$D(\underline{\widetilde{H}}_{k-1}, \underline{\widetilde{K}}_{k-1}) := \text{diag}(\mu_2, \ldots, \mu_k)$.

$\mu_k = 0 \;\Rightarrow\; \text{span}(\mathbf{I} - \sigma\,\mathbf{C})\mathbf{V}_k = \text{span}\big(\widetilde{\mathbf{V}}_k\big)$.

$\mathbf{A}\,\mathbf{V}_{k+1}\,\underline{H}_k = \mathbf{B}\,\mathbf{V}_{k+1}\,\underline{K}_k$ with
$(\underline{H}_k, \underline{K}_k)$ strongly irr. Hessenberg,

---

Theorem. For $\tau_1, \ldots, \tau_\ell$, consider

$$\Phi := \prod_{j=1}^{\ell}(\mathbf{A} - \mu_k\mathbf{B})^{-1}(\mathbf{A} - \tau_j\,\mathbf{B})$$

Then $\quad \mathbf{A}\,\widetilde{\mathbf{V}}_{k+1-\ell}\,\underline{\widetilde{H}}_{k-\ell} = \mathbf{B}\,\widetilde{\mathbf{V}}_{k+1-\ell}\,\underline{\widetilde{K}}_{k-\ell}$
$(\underline{\widetilde{H}}_{k-\ell}, \underline{\widetilde{K}}_{k-\ell})$ strongly irr. Hessenberg,
$D(\underline{\widetilde{H}}_{k-\ell}, \underline{\widetilde{K}}_{k-\ell}) := \text{diag}(\mu_{\ell+1}, \ldots, \mu_k)$.
$\text{span}\big(\Phi\,(\mathbf{V}_{k+1-\ell})\big) = \text{span}\big(\widetilde{\mathbf{V}}_{k+1-\ell}\big)$

Proof. With $\mathbf{C} := (\mathbf{A} - \mu_k\mathbf{B})^{-1}\mathbf{B}$, $\sigma_j := \tau_j - \mu_k$

$$(\mathbf{A} - \mu_k\mathbf{B})^{-1}(\mathbf{A} - \tau_j\,\mathbf{B}) = \mathbf{I} - \sigma_j\,\mathbf{C}$$

With $\Psi(\lambda) := (1 - \lambda\,\sigma_\ell)\cdots(1 - \lambda\,\sigma_1)$ we have $\Phi = \Psi(\mathbf{C})$

[De Samblanx, Meerbergen, Bultheel 96]

An implicitly restarted RKS step

$\sigma_j = \tau_j - \mu_k$,

Set $\underline{H} = \underline{H}_k$, $\underline{K} = \underline{K}_k$, $\underline{\tilde{Q}} = \underline{I}_k$

For $j = 1, \ldots, \ell$ do

  Factorise $\underline{K} - \sigma_j\,\underline{H} = \underline{Q}\,R$

  Compute q such that $[\underline{Q}, q]$ unitary

  Compute $\underline{Z}$ such that
    $[\underline{Z}, (\bar{\sigma}_j\,\underline{K} + \underline{H})^*q]$ unitary Hess.

  Compute $\quad \underline{H} \leftarrow \underline{Q}^*\,\underline{H}\,\underline{Z}$

  $\quad \underline{K} \leftarrow \underline{Q}^*\,\underline{K}\,\underline{Z}$

  $\quad \underline{\tilde{Q}} \leftarrow \underline{\tilde{Q}}\,\underline{Q}$

end for

Set $\quad \underline{\widetilde{H}}_{k-\ell} = \underline{H}$, $\underline{\widetilde{K}}_{k-\ell} = \underline{K}$

Compute $\quad \widetilde{\mathbf{V}}_{k+1-\ell} = \mathbf{V}_{k+1}\,\underline{\tilde{Q}}$,

---

Alternatives

Davidson's method [Davidson '75]

$$\mathbf{z} = (\text{diag}(\mathbf{A}) - \theta\,\mathbf{I})^{-1}\mathbf{r}$$

with $\quad \mathbf{r} := \mathbf{A}\mathbf{u} - \theta\,\mathbf{u} \quad (\mathbf{u} := \mathbf{V}y,\; \mathbf{V}^*\mathbf{A}\mathbf{V}y = \theta y)$.

Generalized Davidson [Morgan '90]

$$\mathbf{M} \approx \mathbf{A} - \theta\,\mathbf{I} \quad \text{and} \quad \mathbf{z} = \mathbf{M}^{-1}\mathbf{r}$$

If $\quad \mathbf{z} = (\mathbf{A} - \theta\,\mathbf{I})^{-1}\mathbf{r} \quad$ then **stagnation**.

If $\quad \mathbf{z} = \mathbf{r} = (\mathbf{A} - \theta\,\mathbf{I})\mathbf{u} \quad$ then **Arnoldi**:
$\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{A}\mathbf{v}_k) =$
$\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k, (\mathbf{A} - \theta\,\mathbf{I})\mathbf{u})$.

Expansion with Ritz vectors

$$\mathbf{M} \approx \mathbf{A} - \theta\,\mathbf{I}$$

Inexact Shift-and-Invert,

$$\mathbf{z} = \mathbf{M}^{-1}\mathbf{v}_k,$$

aims for eigensolution $\quad \mathbf{M}$.

Generalized Davidson,

$$\mathbf{z} = \mathbf{M}^{-1}(\mathbf{A} - \theta\,\mathbf{I})\mathbf{u},$$

aims for eigensolution $\quad \mathbf{M}^{-1}(\mathbf{A} - \lambda\,\mathbf{I})\mathbf{x}_\lambda = 0$.

A Gen. Davidson $\approx$ Cayley Transform Arnoldi:

$$\mathbf{z} = (\mathbf{A} - \tau\,\mathbf{I})^{-1}(\mathbf{A} - \sigma\,\mathbf{I})\mathbf{u}.$$

Exponential Arnoldi (for efficient expansion):

$$\mathbf{z} = \exp(\mathbf{A} - \theta\,\mathbf{I})\mathbf{u} \quad (\approx \mathbf{U}[\exp(H')e_1]\,).$$

$$\boxed{\mathbf{A}\,\mathbf{x}_\lambda = \lambda\,\mathbf{x}_\lambda}$$

Towards optimal expansion

Consider the Ritz approximation $(\theta, \mathbf{u} := \mathbf{V}y)$.

$$\theta := \frac{\mathbf{u}^*\mathbf{A}\mathbf{u}}{\mathbf{u}^*\mathbf{u}}, \qquad \mathbf{r} := \mathbf{A}\,\mathbf{u} - \theta\,\mathbf{u} \perp \mathbf{V}$$

Find $\quad \mathbf{z} \perp \mathbf{u} \quad$ such that $\quad \mathbf{x}_\lambda = \mathbf{u} + \mathbf{z}$.

$$(\mathbf{A} - \lambda\,\mathbf{I})\,\mathbf{z} = -\mathbf{r} + (\lambda - \theta)\,\mathbf{u}$$

or, equivalently,

$$\begin{cases}\mathbf{u}^*\mathbf{A}\mathbf{x}_\lambda = \lambda\,\mathbf{u}^*\mathbf{x}_\lambda\\[4pt] \left(\mathbf{I} - \dfrac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)(\mathbf{A} - \lambda\,\mathbf{I})\left(\mathbf{I} - \dfrac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}\end{cases}$$

[S. & van der Vorst '96]

Jacobi-Davidson

Since λ is unknown:

solve $\quad \mathbf{z} \perp \mathbf{u} \quad$ (approximately) from

the correction equation

$$\boxed{\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}}$$

[S., van der Vorst '96]

Jacobi-Davidson

1. **Start:**
   Choose initial orth. matrix $\mathbf{V}$.
2. **Repeat:**
   - Compute $\mathbf{W} \leftarrow \mathbf{A}\mathbf{V}$, $H \leftarrow \mathbf{V}^*\mathbf{W}$.
   - Compute a desired eigenpair (θ, y) of H (with $\|y\| = 1$).
   - Compute $\mathbf{u} \leftarrow \mathbf{V}y$,
     $\mathbf{r} \leftarrow \mathbf{A}\,\mathbf{u} - \theta\,\mathbf{u}$.
   - **Stop if** satisfied.
   - Solve (approximately)
     $$(\mathbf{I} - \mathbf{u}\,\mathbf{u}^*)\,(\mathbf{A} - \theta\,\mathbf{I})\,(\mathbf{I} - \mathbf{u}\,\mathbf{u}^*)\,\mathbf{z} = -\mathbf{r}.$$
   - $\mathbf{V} \leftarrow$ ModGS($[\mathbf{V}\,|\,\mathbf{z}]$)

Solve approximately

$$\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}.$$

Any preconditioned it. method can be used.

Preconditioning: $\quad \mathbf{M} \approx \mathbf{A} - \theta\,\mathbf{I}$

Preconditioning equation:

$$\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{M}\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}$$

Solution:

$$\mathbf{M}\,\mathbf{z} = -\mathbf{r} + \alpha\,\mathbf{u} \quad \text{such that} \quad \mathbf{z} \perp \mathbf{u},$$

or

$$\mathbf{z} = -\mathbf{M}^{-1}\mathbf{r} + \alpha\,\mathbf{M}^{-1}\mathbf{u} \quad \& \quad \alpha = \frac{\mathbf{u}^*\mathbf{M}^{-1}\mathbf{z}}{\mathbf{u}^*\mathbf{M}^{-1}\mathbf{u}}.$$

Solve approximately

$$\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}$$

$$\neq$$

Solve approximately

$$\left.\begin{array}{l}(\mathbf{A} - \theta\,\mathbf{I})\,\mathbf{z}_r = \mathbf{r} \\ (\mathbf{A} - \theta\,\mathbf{I})\,\mathbf{z}_u = \mathbf{u}\end{array}\right\} \quad \& \quad \begin{array}{l}\mathbf{z} = -\mathbf{z}_r + \alpha\,\mathbf{z}_u \\ \text{s.t.} \quad \mathbf{z} \perp \mathbf{u}\end{array}$$

except if
"solve appr." = 1 step with fixed precond. $\mathbf{M}$.

Example. With static polynomial method p:

$$p\left(\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)\right) \neq$$

$$\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)p\,(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right)$$

Related methods

with $\mathbf{M}$ a fixed preconditioner

Expand the search space by the sol. $\mathbf{z}$ of

- $\mathbf{M}\,\mathbf{z} = -\mathbf{r} + \alpha\,\mathbf{u} \quad \& \quad \alpha = \mathbf{u}^*\mathbf{M}^{-1}\mathbf{z}/\mathbf{u}^*\mathbf{M}^{-1}\mathbf{u}$
  Olsen's method ['90 for symm. $\mathbf{A}$]
- $\mathbf{M}\,\mathbf{z} = -\mathbf{r}$
  Generalized Davidson [Morgan '90]
- $\mathbf{M}\,\mathbf{z} = \mathbf{u}$
  Inexact Shift-and-Invert.

$$\mathbf{z} = -\mathbf{M}^{-1}\mathbf{r} + \alpha\,\mathbf{M}^{-1}\mathbf{u}$$

Extreme choices

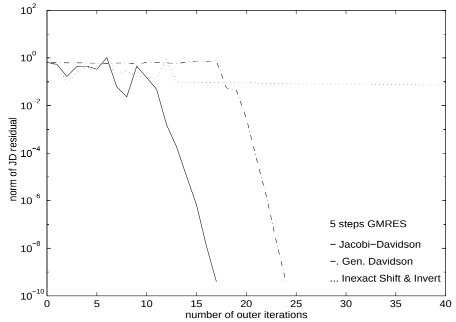
- $\mathbf{M} = \mathbf{I}$:
  $\mathbf{z} = -\mathbf{r} + \alpha\,\mathbf{u} = -\mathbf{A}\,\mathbf{u} + (\theta + \alpha)\mathbf{u} \sim \mathbf{A}\,\mathbf{u}$.
  J-D $\equiv$ Gen. Dav. $\equiv$ Arnoldi.
  Inexact Shift-and-Invert stagnates.
- $\mathbf{M} = \mathbf{A} - \theta\,\mathbf{I}$:
  $\mathbf{z} = -\mathbf{u} + \alpha\,(\mathbf{A} - \theta\,\mathbf{I})^{-1}\mathbf{u} \sim (\mathbf{A} - \theta\,\mathbf{I})^{-1}\mathbf{u}$.
  J-D $\equiv$ Shift-and-Invert with optimal shift.
  Generalized Davidson stagnates.

$\lambda \approx 0, \quad \mathbf{A} = \text{SHERMAN4} - .5\,\mathbf{I}, \quad n = 1104$

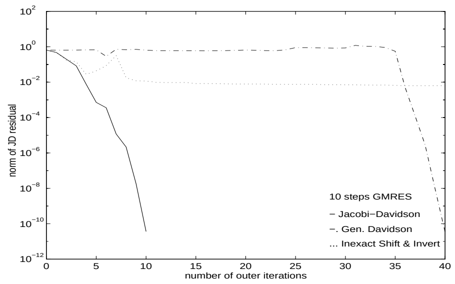
CorEq:

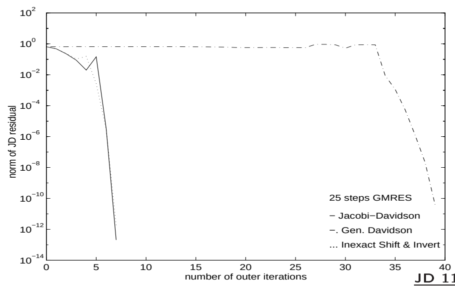
- $(\mathbf{I} - \mathbf{u}\,\mathbf{u}^*)(\mathbf{A} - \theta\,\mathbf{I})(\mathbf{I} - \mathbf{u}\,\mathbf{u}^*)\,\mathbf{z} = -\mathbf{r}$ [J-D]
- $(\mathbf{A} - \theta\,\mathbf{I})\,\mathbf{z} = -\mathbf{r}$ ["Davidson"]
- $(\mathbf{A} - \theta\,\mathbf{I})\,\mathbf{z} = \mathbf{u}$ ["Shift-and-Invert"]

m steps precond. GMRES with $\mathbf{M} = $ ILU(2).



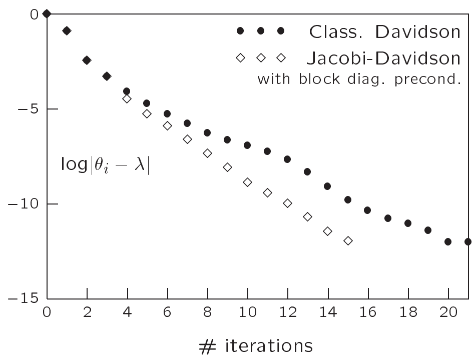
$m = 5$





[van Dam, van Lenthe '94]

Example [Quantum Chemistry]



$\mathbf{A}$ symmetric, $n \approx 50{,}000$

block diag. precond. with max. block size 66

J-D as a Newton scheme

Choose scaling vectors $\tilde{\mathbf{u}}$ and $\mathbf{w}$.

$$\mathbf{F}(\mathbf{u}) \equiv \mathbf{A}\,\mathbf{u} - \theta\,\mathbf{u}, \qquad \theta \equiv \theta(\mathbf{u}) \equiv \frac{\mathbf{w}^*\mathbf{A}\mathbf{u}}{\mathbf{w}^*\mathbf{u}}$$

$$\mathbf{F} : \{\mathbf{u}, \tilde{\mathbf{u}}^*\mathbf{u} = 1\} \to \{\mathbf{r} = \mathbf{F}(\mathbf{u}) \perp \mathbf{w}\}$$

$$\boxed{\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x} \quad \& \quad \tilde{\mathbf{u}}^*\mathbf{x} = 1} \quad \Leftrightarrow \quad \boxed{\mathbf{F}(\mathbf{x}) = 0}$$

Newton:
- solve $\quad \mathbf{z} \perp \tilde{\mathbf{u}}$

  $$\frac{\partial \mathbf{F}}{\partial \mathbf{u}}\mathbf{z} = -\mathbf{r} \equiv -\mathbf{F}(\mathbf{u})$$
- $\mathbf{u} \leftarrow \mathbf{u} + \mathbf{z}$

For $\mathbf{z} \perp \tilde{\mathbf{u}}$, $\quad \dfrac{\partial \mathbf{F}}{\partial \mathbf{u}}\mathbf{z} = \left(\mathbf{I} - \dfrac{\mathbf{u}\,\mathbf{w}^*}{\mathbf{w}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\,\mathbf{z}$

Accelerated Newton:
- solve $\quad \mathbf{z} \perp \tilde{\mathbf{u}}$

  $$\left(\mathbf{I} - \frac{\mathbf{u}\,\mathbf{w}^*}{\mathbf{w}^*\mathbf{u}}\right)(\mathbf{A} - \theta\,\mathbf{I})\left(\mathbf{I} - \frac{\mathbf{u}\,\tilde{\mathbf{u}}^*}{\tilde{\mathbf{u}}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}$$
- accelerated update

Inexact Newton: $\quad \mathbf{M} \approx \mathbf{A} - \theta\,\mathbf{I}$.

Solve $\quad \left(\mathbf{I} - \dfrac{\mathbf{u}\,\mathbf{w}^*}{\mathbf{w}^*\mathbf{u}}\right)\mathbf{M}\left(\mathbf{I} - \dfrac{\mathbf{u}\,\tilde{\mathbf{u}}^*}{\tilde{\mathbf{u}}^*\mathbf{u}}\right)\mathbf{z} = -\mathbf{r}$:

$$\mathbf{z} = -\mathbf{M}^{-1}\mathbf{r} + \alpha\,\mathbf{M}^{-1}\mathbf{u} \quad \& \quad \alpha = \frac{\tilde{\mathbf{u}}^*\mathbf{M}^{-1}\mathbf{r}}{\tilde{\mathbf{u}}^*\mathbf{M}^{-1}\tilde{\mathbf{u}}}$$

$\alpha = 0 \quad$ if $\mathbf{M}$ maps $\tilde{\mathbf{u}}^\perp$ onto $\mathbf{w}^\perp$

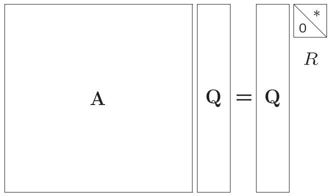
JDQR

[Fokkema, S, van der Vorst '96]

Compute m eigenvalues close to some $\tau \in \mathbb{C}$.

- Deflation
- Schur vectors
- Restart

Partial Schur form



$\mathbf{A}\,\mathbf{Q} = \mathbf{Q}\,R \quad$ with $\quad \mathbf{Q}^*\mathbf{Q} = \mathbf{I}$

R upper Δ

$\mathrm{diag}(R) = (\lambda_1, \ldots, \lambda_m)$

$|\lambda_i - \tau| \quad$ increases

If $\quad R\,y_i = \lambda_i\,y_i \quad$ then $\quad \mathbf{A}\,\mathbf{Q}\,y_i = \lambda_i\,\mathbf{Q}\,y_i$.

Finding other eigenpairs

Expansion of the partial Schur form:

$$\mathbf{A}\,[\mathbf{Q}\,|\,\mathbf{q}] = [\mathbf{Q}\,|\,\mathbf{q}]\begin{bmatrix} R & s \\ & \lambda \end{bmatrix}$$

with $\quad \mathbf{Q}^*\mathbf{q} = 0$.

Hence

$$(\mathbf{I} - \mathbf{Q}\,\mathbf{Q}^*)(\mathbf{A} - \lambda\,\mathbf{I})(\mathbf{I} - \mathbf{Q}\,\mathbf{Q}^*)\mathbf{q} = 0$$

which means that $(\lambda, \mathbf{q})$ is an eigenpair of $\widetilde{\mathbf{A}}$:

$$\widetilde{\mathbf{A}} := (\mathbf{I} - \mathbf{Q}\,\mathbf{Q}^*)(\mathbf{A} - \lambda\,\mathbf{I})(\mathbf{I} - \mathbf{Q}\,\mathbf{Q}^*)$$

JDQR: apply JD for $\widetilde{\mathbf{A}}$.

1. **Start:** Choose an initial space $\mathbf{V}$.
   - Set $\Lambda = [\,]$, $\mathbf{Q} = [\,]$
2. **Repeat:**
   a. Compute $\mathbf{W} \leftarrow \mathbf{A}\mathbf{V}$, $\quad H \leftarrow \mathbf{V}^*\mathbf{W}$

   b. Compute $\quad H = USU^*$
   with $\quad U^*U = I$, $\quad S$ upper Δ,
   $|S_{11} - \tau| \le |S_{22} - \tau| \le \ldots$

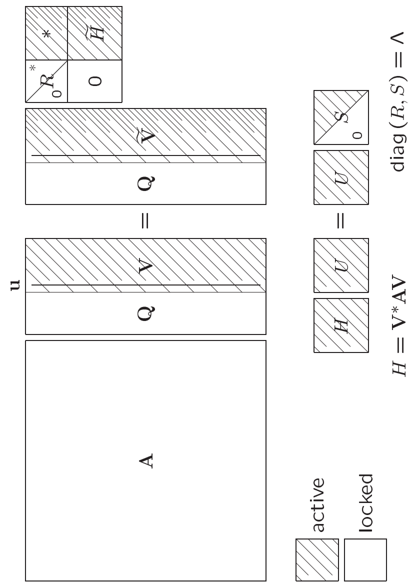
   c. $\theta \leftarrow S_{11}$, $y \leftarrow Ue_1$, $\mathbf{u} \leftarrow \mathbf{V}y$,
   $\mathbf{r} \leftarrow \mathrm{ModGS}_{\mathbf{Q}}(\mathbf{A}\,\mathbf{u} - \theta\,\mathbf{u})$.

   d. **If** $\|\mathbf{r}\| \le$ *tol*
   - $\Lambda \leftarrow [\Lambda\,|\,\theta]$, $\qquad \mathbf{Q} \leftarrow [\mathbf{Q}\,|\,\mathbf{u}]$.
   - $J = [e_2\,|\,\ldots\,|\,e_{\dim(\mathbf{V})}]$,
     $\mathbf{V} \leftarrow \mathbf{V}(UJ)$, $U \leftarrow I$, $S \leftarrow J^*SJ$.
   - **Goto** 2.c.

   e. Solve (app.) $\mathbf{z} \perp \widetilde{\mathbf{Q}} \equiv [\mathbf{Q}\,|\,\mathbf{u}]$

   $(\mathbf{I} - \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^*)\,(\mathbf{A} - \theta\,\mathbf{I})\,(\mathbf{I} - \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^*)\,\mathbf{z} = -\mathbf{r}$.

   f. $\mathbf{V} \leftarrow \mathrm{ModGS}([\mathbf{V}\,|\,\mathbf{z}])$



Solving the preconditioning eq.

$$(\mathbf{I}-\mathbf{Q}\,\mathbf{Q}^*)\,\mathbf{M}\,(\mathbf{I}-\mathbf{Q}\,\mathbf{Q}^*)\,\mathbf{z}=-\mathbf{r}.$$

$\mathbf{M}\,\mathbf{z}=-\mathbf{r}+\mathbf{Q}\,y$ such that $\mathbf{z}\perp\mathbf{Q}$

$\Rightarrow$

$$\mathbf{z}=-\mathbf{M}^{-1}\mathbf{r}+\mathbf{M}^{-1}\mathbf{Q}\,y$$

& $$y=(\underbrace{\mathbf{Q}^*\,\mathbf{M}^{-1}\,\mathbf{Q}}_{\text{small matrix}})^{-1}\,\mathbf{Q}^*\,\mathbf{M}^{-1}\mathbf{r}.$$

Restart

For λ close to a given **target** $\tau\in\mathbb{C}$

- For $\mathbf{V}$, compute $H:=\mathbf{V}^*\mathbf{A}\mathbf{V}$

  $$\left.\begin{array}{c} H=USU^* \text{ with } U^*U=I, \\ S \text{ upper triangular}\end{array}\right\}\text{QR}$$

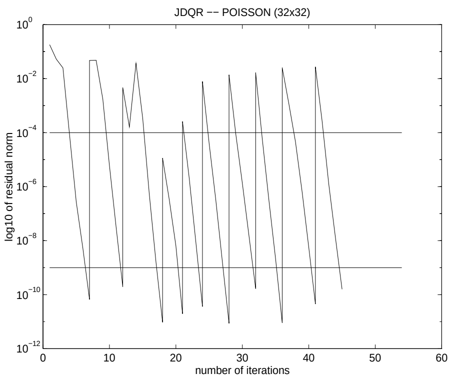
- Order S such that $|s_{j,j}-\tau|$ increases.
- Restart with $\mathbf{V}\leftarrow\mathbf{V}[u_1,\ldots,u_\ell]$ if necessary.

Properties JDQR

- Leads to partial Schur form.
- Employs orthonormal sets of vectors.
- Finds also multiple eigenvalues.
- Easy selection of wanted eigenvalues.
- No matrix inversion involved.
- Preconditioner can be included.
- Ingredients:
  vector updates, Dot-products, MVs,
  preconditioners
  QR for small problems.

Some practical notes

- After convergence of some eigenvectors, solution of correction equation easier.
- Number of pojections can be limited.
- Correction equation can be preconditioned.
  Preconditioner can be used more often: pays to make better preconditioning.
- For generalized eigenproblems:
  JD + QZ-reduction



2D Poisson Problem (32×32)
Target = 0d0
Max Dim = 15, Min Dim = 10
GMRES(10) plus ILU(0)
eps = 1d-9


```
lambda( 1): ( 9.0562E-03, 0.0000E+00 )
lambda( 2): ( 2.2599E-02, 0.0000E+00 )
lambda( 3): ( 2.2599E-02, 0.0000E+00 )
lambda( 4): ( 3.6143E-02, 0.0000E+00 )
lambda( 5): ( 4.5035E-02, 0.0000E+00 )
lambda( 6): ( 4.5035E-02, 0.0000E+00 )
lambda( 8): ( 5.8578E-02, 0.0000E+00 )
lambda( 7): ( 5.8578E-02, 0.0000E+00 )
lambda( 9): ( 7.6160E-02, 0.0000E+00 )
lambda(10): ( 7.6160E-02, 0.0000E+00 )
```

2D Poisson Problem (32×32)
Target = 0d0
Max Dim = 15, Min Dim = 10
GMRES(10) with ILU(0)
eps = 1d-9

Examples

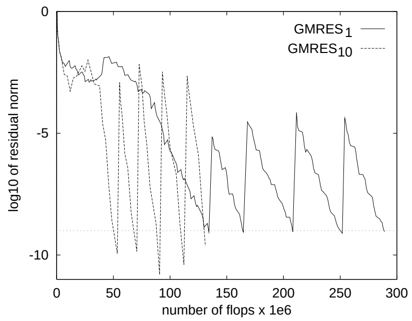
Maximum dimension $\mathbf{V}$ is 15,
after restart dimension $\mathbf{V}$ is 10.

Accuracy correction equation: $\|\tilde{\mathbf{r}}_i\|\leq 2^{-j}\|\tilde{\mathbf{r}}_0\|$

DW4096: Square Dielectric Waveguide standard eigenproblem

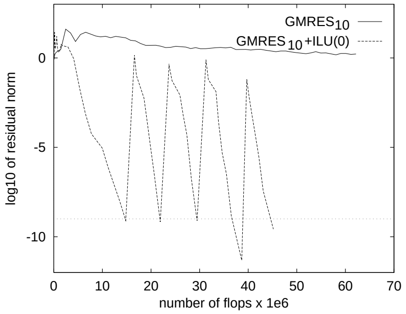
Searching for the five rightmost eigenvalues:

$$8.6479e-01,\quad 8.6528e-01,\quad 8.6564e-01,$$
$$8.6594e-01,\quad 8.6657e-01$$



BWM2000: Brusselator wave model
standard eigenproblem

Searching for the five rightmost eigenvalues.



Generalized eigenproblems

$$\mathrm{A}\,\mathrm{x}_\lambda = \lambda\,\mathrm{B}\,\mathrm{x}_\lambda$$

V <u>search space</u>, $\quad$ W <u>test space</u>

- Petrov-Galerkin approach:

$$\mathrm{W}^*\,\mathrm{A}\,\mathrm{V}\,y - \theta\,\mathrm{W}^*\,\mathrm{B}\,\mathrm{V}\,y = 0.$$

$$\mathrm{u} = \mathrm{V}\,y, \qquad \mathrm{r} := \mathrm{A}\,\mathrm{u} - \theta\,\mathrm{B}\,\mathrm{u}$$

- The correction equation:

$$\left(\mathrm{I} - \frac{\tilde{\mathrm{w}}\,\mathrm{w}^*}{\mathrm{w}^*\tilde{\mathrm{w}}}\right)(\mathrm{A} - \theta\,\mathrm{B})\left(\mathrm{I} - \frac{\mathrm{u}\,\tilde{\mathrm{u}}^*}{\tilde{\mathrm{u}}^*\mathrm{u}}\right)\mathrm{z} = -\mathrm{r}$$

<u>**Th.**</u> $\mathrm{w} \in \mathrm{W}$ $\quad\&\quad$ $\tilde{\mathrm{w}} = (\alpha\mathrm{A} + \beta\mathrm{B})\,\mathrm{u}$

& exact solve $\Rightarrow$ quadratic convergence.

$$\boxed{(\mathrm{A} - \lambda\,\mathrm{B})\,\mathrm{x} = 0}$$

Split $\quad \mathrm{x} = \mathrm{u} + \mathrm{z}_e \quad$ with $\quad \mathrm{z}_e \perp \tilde{\mathrm{u}}$.

z is the computed correction, $\quad \mathrm{z} \perp \tilde{\mathrm{u}}$.

Note $\quad \mathrm{x} - \mathrm{u} = \mathrm{z}_e, \quad \mathrm{x} - (\mathrm{u} + \mathrm{z}) = \mathrm{z}_e - \mathrm{z}$.

$$(\mathrm{A} - \theta\,\mathrm{B})\,\mathrm{z}_e = -\mathrm{r} + (\lambda - \theta)\mathrm{B}\mathrm{x}, \qquad \mathrm{z}_e \perp \tilde{\mathrm{u}}.$$

$$\mathrm{P}(\mathrm{A} - \theta\,\mathrm{B})\,\mathrm{z} = -\mathrm{r}, \qquad \mathrm{z} \perp \tilde{\mathrm{u}}$$

where $\quad \mathrm{P} \equiv \left(\mathrm{I} - \dfrac{\tilde{\mathrm{w}}\,\mathrm{w}^*}{\mathrm{w}^*\tilde{\mathrm{w}}}\right).$

Hence $\quad \mathrm{P}(\mathrm{A} - \theta\,\mathrm{B})_{|\tilde{\mathrm{u}}^\perp}(\mathrm{z}_e - \mathrm{z}) = \mathrm{P}(\lambda - \theta)\,\mathrm{B}\mathrm{x}.$

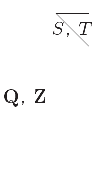
$$\lambda - \theta = \frac{\mathrm{w}^*(\mathrm{A} - \theta\,\mathrm{B})\mathrm{z}_e}{\mathrm{w}^*\mathrm{B}\mathrm{x}} = \mathcal{O}(\|\,\mathrm{z}_e\|).$$

$\mathrm{P}\mathrm{B}\mathrm{x} = \mathcal{O}(\|\mathrm{x} - \mathrm{u}\|) = \mathcal{O}(\|\,\mathrm{z}_e\|) \quad$ iff $\quad \mathrm{P}\mathrm{B}\mathrm{u} = 0.$

Partial generalized Schur form

$\mathrm{A}\,\mathrm{Q} = \mathrm{Z}\,S \qquad \mathrm{Q}, \mathrm{Z}$ orth.,

$\mathrm{B}\,\mathrm{Q} = \mathrm{Z}\,T \qquad S, T$ upper $\triangle$.



$\mathrm{A}\,\mathrm{Q} = \mathrm{B}\,\mathrm{Q}\,\underbrace{T^{-1}S}_{\text{upper } \triangle} \qquad \mathrm{span}(\mathrm{Z}) = \mathrm{span}(\mathrm{B}\,\mathrm{Q})$

$= \mathrm{span}(\mathrm{C}\,\mathrm{Q})$

$\mathrm{C} := \alpha\,\mathrm{A} + \beta\,\mathrm{B}$

$\mathrm{Z}^*\mathrm{A}\,\mathrm{Q} = S \qquad \mathrm{V} \ni \mathrm{u} \;\rightarrow\;$ first column Q

$\mathrm{Z}^*\mathrm{B}\,\mathrm{Q} = T \qquad \mathrm{W} \ni \mathrm{w} \rightarrow\;$ first column Z

Natural choices in connection with QZ:

$\mathrm{span}(\mathrm{W}) = \mathrm{span}(\mathrm{C}\,\mathrm{V}), \quad \mathrm{w} = \mathrm{C}\,\mathrm{u}.$

Choosing spaces & projections

Select $\alpha, \beta \in \mathbb{C}$. $\quad$ Consider $\mathrm{C} := \alpha\,\mathrm{A} + \beta\,\mathrm{B}$.

$$\mathrm{u} = \mathrm{V}\,y, \quad \mathrm{r} = \mathrm{A}\,\mathrm{u} - \theta\,\mathrm{B}\,\mathrm{u}, \quad \mathrm{w} = \mathrm{C}\,\mathrm{u}$$

- The correction equation:

$$\left(\mathrm{I} - \frac{\mathrm{w}\,\mathrm{w}^*}{\mathrm{w}^*\mathrm{w}}\right)(\mathrm{A} - \theta\,\mathrm{B})\left(\mathrm{I} - \frac{\mathrm{u}\,\mathrm{u}^*}{\mathrm{u}^*\mathrm{u}}\right)\mathrm{z} = -\mathrm{r}.$$

- The search space and test space:

$\mathrm{V} = \mathrm{modGS}([\mathrm{V}\,|\,\mathrm{z}]), \quad \mathrm{W} = \mathrm{modGS}([\mathrm{W}\,|\,\mathrm{C}\,\mathrm{z}])$

Preconditioning

$$\mathrm{M} \approx \mathrm{A} - \theta\,\mathrm{B}$$

$$\left(\mathrm{I} - \frac{\mathrm{w}\,\mathrm{w}^*}{\mathrm{w}^*\mathrm{w}}\right)\mathrm{M}\left(\mathrm{I} - \frac{\mathrm{u}\,\mathrm{u}^*}{\mathrm{u}^*\mathrm{u}}\right)\mathrm{z} = -\mathrm{r}.$$

<u>Explicit left preconditioning</u>:

- The correction equation:

$$\left(\mathrm{I} - \frac{\mathrm{v}\,\mathrm{u}^*}{\mathrm{u}^*\mathrm{v}}\right)\mathrm{M}^{-1}(\mathrm{A} - \theta\,\mathrm{B})\left(\mathrm{I} - \frac{\mathrm{v}\,\mathrm{u}^*}{\mathrm{u}^*\mathrm{v}}\right)\mathrm{z} = -\mathrm{r}'.$$

where

$$\mathrm{M}\,\mathrm{v} = \mathrm{w} = \mathrm{B}\,\mathrm{u} \quad \& \quad \mathrm{r}' = \left(\mathrm{I} - \frac{\mathrm{v}\,\mathrm{u}^*}{\mathrm{u}^*\mathrm{v}}\right)\mathrm{r}.$$

Choosing the mixture of A and B

Aim for optimal appr. first "left Schur vector"

$$\mathrm{w} = \alpha\,\mathrm{A}\,\mathrm{u} + \beta\,\mathrm{B}\,\mathrm{u}$$

<u>Inspiration:</u>

$\mathrm{B} = \mathrm{I}$: choose $\quad \mathrm{w} = \mathrm{u} = 0\,\mathrm{A}\,\mathrm{u} + 1\,\mathrm{B}\,\mathrm{u}.$

$\mathrm{A} = \mathrm{I}$: choose $\quad \mathrm{w} = \mathrm{u} = 1\,\mathrm{A}\,\mathrm{u} + 0\,\mathrm{B}\,\mathrm{u}.$

$\mathrm{w} = 1\,\mathrm{A}\,\mathrm{u} + 0\,\mathrm{B}\,\mathrm{u} \quad$ is "instable" choice

if $\quad \mathrm{B} = \mathrm{I} \quad$ and $\quad \mathrm{A}\mathrm{x} = \lambda\mathrm{x} \approx 0$

Select α, β such that

— $\quad |\alpha|^2 + \beta^2 = 1, \quad \beta \geq 0, \quad$ and

— $\quad \|\alpha\,\mathrm{A}\,\mathrm{u} + \beta\,\mathrm{B}\,\mathrm{u}\| \quad$ approximately maximal:

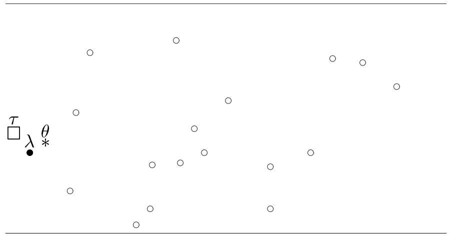
$$\mathrm{C} = \frac{\bar{\tau}}{\sqrt{1 + |\tau|^2}}\mathrm{A} + \frac{1}{\sqrt{1 + |\tau|^2}}\mathrm{B}$$

Selecting Ritz pairs

Suppose A normal, $\mathrm{B} = \mathrm{I}$.

$$\theta = \sum_i |\gamma_i|^2 \lambda_i \quad \text{with} \quad \mathrm{u} = \sum_i \gamma_i \mathrm{x}_i$$

Suppose τ outside hull spectrum:



If $\theta \approx \tau$ then wanted component large.

What if τ is in the interior of the hull spectrum?

Harmonic Ritz values

$(\mathbf{A}-\lambda\,\mathbf{I})\mathbf{x}=\mathbf{0} \quad \Leftrightarrow \quad (\mathbf{A}-\tau\,\mathbf{I})^{-1}\mathbf{x}=\frac{1}{\lambda-\tau}\,\mathbf{x}:$

if $\lambda \approx \tau$ then $1/(\lambda-\tau)$ extremal for $(\mathbf{A}-\tau\,\mathbf{I})^{-1}$.

<u>How to avoid inversion?</u>

Galerkin for $(\mathbf{A}-\tau\,\mathbf{I})^{-1}$, with
search space = test space = span$((\mathbf{A}-\tau\,\mathbf{I})\mathbf{V})$,
is equivalently to Petrov for $\mathbf{A}$, with
search space: $\mathbf{V}$; test space: $\mathbf{W}=(\mathbf{A}-\tau\,\mathbf{I})\mathbf{V}$.

Harmonic Ritz value := this Petrov value

For generalized eigenproblems, take

$$\text{span}(\mathbf{W})=\text{span}((\mathbf{A}-\tau\,\mathbf{B})\mathbf{V}),$$

$$\mathbf{C}=\frac{1}{\sqrt{1+|\tau|^2}}\,\mathbf{A}-\frac{\tau}{\sqrt{1+|\tau|^2}}\,\mathbf{B}.$$

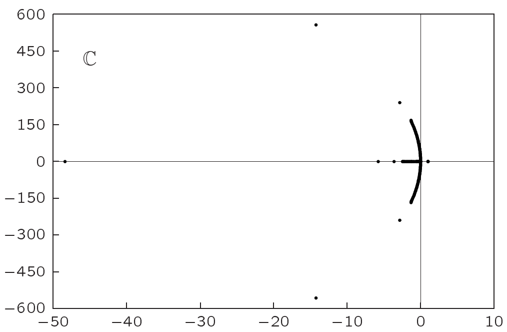
The problem

$$\boxed{\mathbf{A}\,\mathbf{x}_\lambda=\lambda\,\mathbf{B}\,\mathbf{x}_\lambda}$$

For MHD-problems:

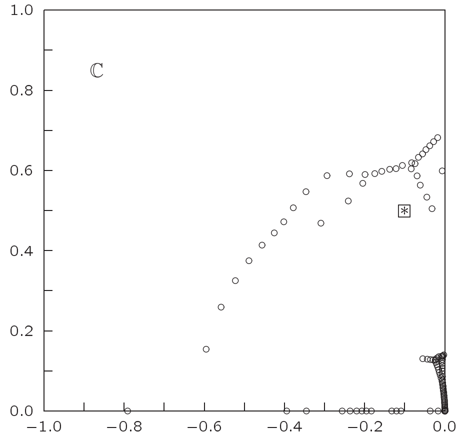
- $\mathbf{A}$ is non Hermitian
- $\mathbf{B}$ Hermitian, posidtive definite
- $\mathbf{A}$, $\mathbf{B}$ block tridiagonal
- interior eigenvalues
- very large order
- inversion unpractical

Spectrum MHD test problem

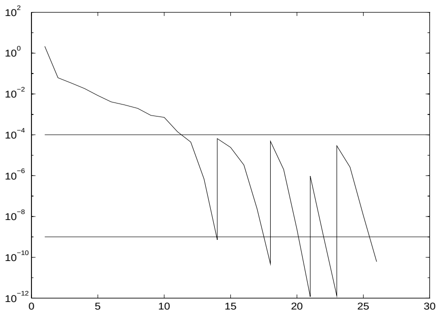


Entire spectrum MHD test problem ($n = 416$).
Correct eigenvalues (.) are shown.

Alfvén spectrum



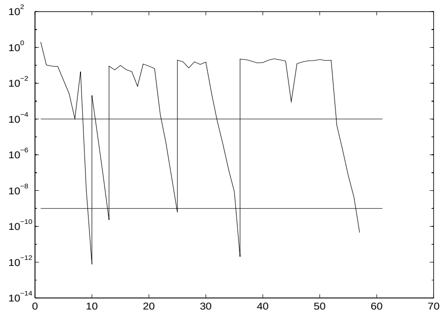
MHD test problem ($n = 416$).
Correct eigenvalues ($\circ$), target ($\boxed{*}$).



MHD Problem (416)
Target = (-0.3, 0.5)
Max Dim = 15,   Min Dim = 10
GMRES(10) with ILU(2)
eps = 1d-9


```
lambda( 1): ( -3.0929E-01, 4.6875E-01 )
lambda( 2): ( -3.4656E-01, 5.4689E-01 )
lambda( 3): ( -2.4103E-01, 5.2381E-01 )
lambda( 4): ( -3.7810E-01, 5.0716E-01 )
lambda( 5): ( -2.9401E-01, 5.8715E-01 )
```

MHD Problem (416)
Target = (-0.3, 0.5)
Max Dim = 15,   Min Dim = 10
GMRES(10) with ILU(2)
eps = 1d-9



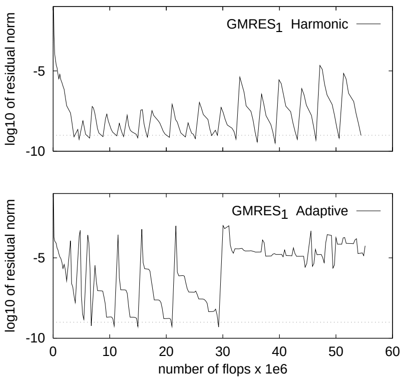
MHD Problem (1600)
Target = (-0.3, 0.5)
Max Dim = 15,   Min Dim = 10
GMRES(10) with ILU(2)
eps = 1d-9


```
lambda( 1): ( -3.3746E-01, 4.8958E-01 )
lambda( 2): ( -3.5776E-01, 4.6050E-01 )
lambda( 3): ( -2.3839E-01, 4.4657E-01 )
lambda( 4): ( -3.8189E-01, 5.0369E-01 )
lambda( 5): ( -2.2157E-01, 5.3364E-01 )
```

MHD Problem (1600)
Target = (-0.3, 0.5)
Max Dim = 15,   Min Dim = 10
GMRES(10) with ILU(2)
eps = 1d-9

MHD416

Searching for the Alfvén branch



ILU(2) for $A - \tau B$ is used as preconditioner.

Generalized eigenproblems

$$A\,x = \lambda\,B\,x$$

[Booten, Fokkema **MHD**]

$$P(\lambda)x := A_0\,x + \lambda\,A_1\,x + \ldots + \lambda^\ell\,A_\ell\,x = 0$$

[van Gijzen, Raeven **acoustics, con. mech.**]

- $$W^* P(\theta) V y = 0$$

  $$u = V y, \quad w = P'(\theta)u, \quad r = P(\theta)u$$

- The correction equation:

  $$\left(I - \frac{w\,w^*}{w^* w}\right) P(\theta) \left(I - \frac{u\,u^*}{u^* u}\right) z = -r$$

- The spaces: $V = \mathrm{modGS}([\,V \,|\, z\,])$, ...

J-D

- is an accelerated inexact Newton scheme
  — quadratic convergence —
- relies on orthogonal projections
- allows any precond. it. method
  to solve the linear (Jacobian) system
- robust
- no inversion
- also for interior eigenvalues

J-D is <u>not</u>

- a Krylov subspace method,
- Davidson with Jacobi-preconditioning,
- a method that produces eigenvectors for
  a preconditioned eigenproblem

Current research

- Fine tuning inner-outer loop and problem.
- Assembly strategies.
- Stopping criterion.
- **Preconditioner.**
- Tracking strategies.

What to use?

- For small problems: QR-algorithm

- If matrices can be inverted
(systems can be solved exactly):
Shift-&-Invert Arnoldi with implicit restart
[Sorensen '92]
Rational Krylov Sequence with implicit restart
[Ruhe '84, '93]

- If matrix inversion is not feasible: JD