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

$$\mathbf{x}_0 \to \mathbf{x}_1 \to \ldots \to \mathbf{x}_{k-1} \xrightarrow{\text{step } k} \mathbf{x}_k$$

initial guess ($\mathbf{x}_0$) — k-th approximate ($\mathbf{x}_k$)

- 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

$$\begin{array}{c}\mathbf{x}_0\\ \mathbf{r}_0\end{array} \to \begin{array}{c}\mathbf{x}_1\\ \mathbf{r}_1\end{array} \to \ldots \to \begin{array}{c}\mathbf{x}_{k-1}\\ \mathbf{r}_{k-1}\end{array} \to \begin{array}{c}\mathbf{x}_k\\ \mathbf{r}_k\end{array}$$

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 $\mathbf{u}_k$ explicitly 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 $\quad \dfrac{\|\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}_k\| + \|\mathbf{b}\|} \leq \epsilon \quad$ then

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

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

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

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

Stopping criteria II

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

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

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

$$|\Delta_A| \leq n_A\,\overline{\xi}\,|\mathbf{A}|,$$

where $\overline{\xi}$ is the relative machine precision, and n_A max non-zeros of $\mathbb{A}$ per row. Hence,

$$\|\mathrm{fl}(\mathbf{b} - \mathbf{A}\mathbf{x})\| \leq n_A\,\overline{\xi}\,\|\,|\mathbf{A}|\,\|\ \|\mathbf{x}\| \leq n_A\,\overline{\xi}\,\|\,|\mathbf{A}|\,\|\ \|\mathbf{A}^{-1}\|\ \|\mathbf{b}\|$$

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

Proposition. 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 $\mathbf{V} = [\mathbf{v}_0 \mid \ldots \mid \mathbf{v}_{k-1}]$.

Proposition. 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 Krylov subspace method 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 $\mathbf{x}_{k-1}$: $\qquad \mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_{k-1}$);
- 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 $\mathbf{r}_0$.

Proposition. $\mathbf{x} - \mathbf{x}_0 \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$

if $\quad k = \dim\mathcal{K}_k(\mathbf{A};\mathbf{r}_0) = \dim\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$

(*Proof.* $\quad \alpha_0\mathbf{r}_0 + \alpha_1\mathbf{A}\mathbf{r}_0 + \ldots + \alpha_k\mathbf{A}^k\mathbf{r}_0 = 0$,
for some $\quad \alpha_i \in \mathbb{C}, \quad \alpha_0 \neq 0,\ \alpha_k \neq 0. \qquad$ 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): $(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbf{x} = \sum y_i x_i$

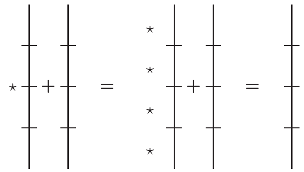
- Portable

- 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.

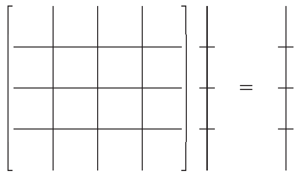
AXPY: $\alpha \mathbf{x} + \mathbf{y}$



Easy to
vectorize
parallelize.

Group expressions
for chaining.

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

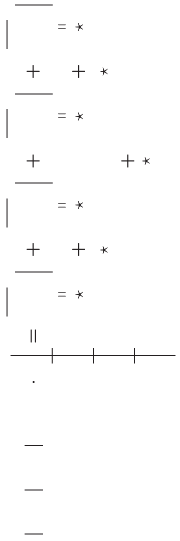


Store compressed
row-wise

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:

$\left.\begin{array}{l}\text{1 MV} \\ k \text{ AXPYs} \\ k \text{ DOTs}\end{array}\right\}$ 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, $\begin{cases}\text{GCR} \\ \text{ORTHODIR} \\ \text{ORTHOMIN} \\ \vdots\end{cases}$

GCRO(ℓ, m)

GMRESR(ℓ, m)

GMRES(m), ℓ–truncated GCR

⋮

some hybrid Bi-CG method??

⋮

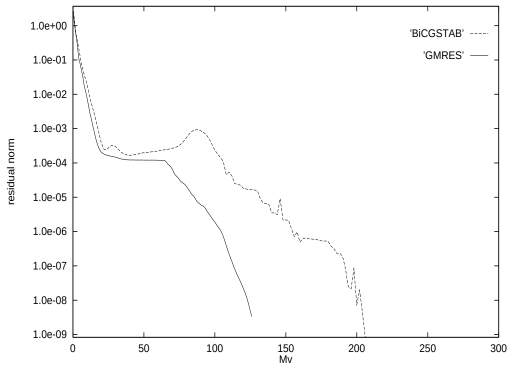
BiCGstab(ℓ),

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 $0.1 \le x, y \le 0.9$

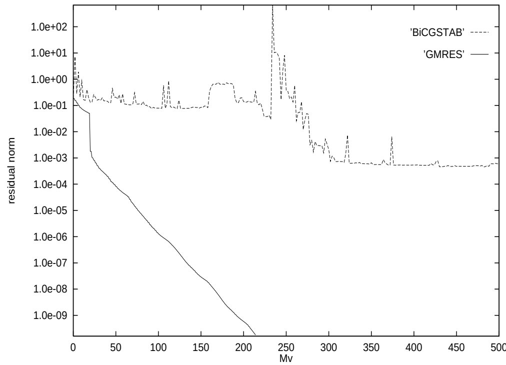
and $D = 1$ elsewhere.

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?

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

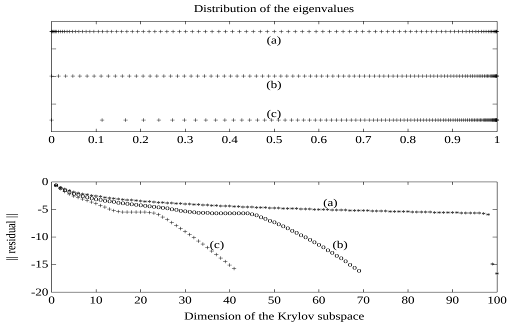
$$\min_{p \in \mathcal{P}_k^1} \max_{\lambda \text{ eigenv. } \mathbf{A}} |p_k(\lambda)|$$

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

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 $q_k(t) = \gamma_0 + \gamma_1 t + \ldots + \gamma_{k-1}t^{k-1}$

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

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

No preconditioning?

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

$\mathbf{A}$ is an 99×99 matrix.

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

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

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

$$\boxed{\begin{array}{l} \text{solve } (\mathbf{U} + \mathbf{I})\mathbf{w}_1 = \mathbf{u} \\ \text{solve } (\mathbf{L} + \mathbf{I})\mathbf{w}_2 = \mathbf{D}\mathbf{w}_1 + \mathbf{u} \\ \text{Compute } \mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2 \end{array}}$$

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

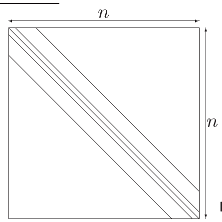
Bi-CGSTAB with D-ILU:
$\mathbf{K} = (\mathbf{L}_A + \Delta)\Delta^{-1}(\Delta + \mathbf{U}_A)$ with
Δ diagonal such that $\mathrm{diag}(\mathbf{K}) = \mathbf{D}_A$
or Δ diagonal such that $\mathbf{K}\,\mathbf{1} = \mathbf{A}\,\mathbf{1}$

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

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

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

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



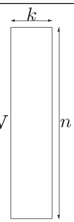
$\mathbb{R}^n = \mathbb{E} \underset{\mathbf{K}^{-1}}{\overset{\mathbf{A},\mathbf{K}}{\rightleftarrows}} \mathbb{F} = \mathbb{R}^n$

$\mathbf{A}^k$, $\mathbf{A}^{-1}$
$\mathbf{K}^{-1}$, $\mathbf{A}$, $\mathbf{A}^T$, $\mathbb{V}$, Λ
only
for theory

not stored as matrices!

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

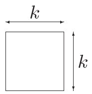
$\mathbf{U}$, $\mathbf{R}$, $\mathbf{V}$



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

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

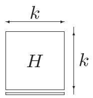
T, H, L, U



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

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

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



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 $j = 0, 1, \ldots, k-1$

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

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

$\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{k-1}$ 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 $\quad j = 0, 1, \dots, k-1$

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

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 $\quad i = 0, \dots, k-1 \quad$ do
    $\tau_{ik} = (\mathbf{c}, \mathbf{c}_i), \quad \beta_{ik} = \tau_{ik}/\sigma_i$
  end for
  for $\quad i = 0, \dots, k-1 \quad$ 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 $i \le k-2$.
  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 $\quad \beta_k = -\rho_k/\rho_{k-1} \quad \& \quad \alpha_k = \rho_k/\sigma_k$

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 $\sigma_k = \mathbf{c}_k^T\mathbf{r}_k$).

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

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

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

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

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

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

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

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

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

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

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

CGNE versus ORTHOMIN, II

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

If $\mathbf{A}\mathbf{x} = \mathbf{r}_0 = e_1$ then

$\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) = \text{span}(e_1, \dots, e_k)$ 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: $<\mathbf{x},\mathbf{y}> = \sum_j x_j \overline{y_j}$,

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

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

[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

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

Convergence CG

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

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

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

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

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

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

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.

$\mathbf{r}_0, \dots, \mathbf{r}_k$ 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

$\text{span}(\mathbf{c}_0, \dots, \mathbf{c}_{k-1}) = \mathbb{A}\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0)$ 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 $\mathbf{r}_0, \dots, \mathbf{r}_{k-1}$.

Hence $\mathcal{K}_k(\mathbb{A}; \mathbf{r}_0) = \{\mathbf{R}_k\vec{\gamma} \mid \vec{\gamma} \in \mathbb{R}^k\}$

where $\mathbf{R}_k := [\mathbf{r}_0 \,|\, \mathbf{r}_1 \,|\, \dots \,|\, \mathbf{r}_{k-1}]$

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:

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

$$\widetilde{\gamma}_k + \widetilde{\alpha}_k + \widetilde{\beta}_k = 0$$

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

$$\underbrace{\mathbf{R}_k = [\,\mathbf{r}_0\,|\,\mathbf{r}_1|\,\ldots\,|\,\mathbf{r}_{k-1}]}_{\text{orthogonal columns}}, \quad \underline{\widetilde{T}}_k = \underbrace{\begin{bmatrix} \widetilde{\alpha}_0 & \widetilde{\beta}_0 & & \\ \widetilde{\gamma}_0 & \widetilde{\alpha}_1 & \ddots & \\ & \widetilde{\gamma}_1 & \ddots & \widetilde{\beta}_k \\ & & \ddots & \widetilde{\alpha}_k \\ & & & \widetilde{\gamma}_k \end{bmatrix}}_{(k+1)\times k \text{ tri-diagonal}}$$

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

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

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

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