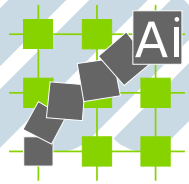
Solvers I — Basics

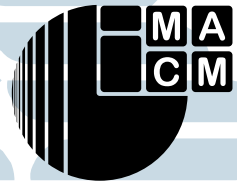
Lattice Practices 2012

Karsten Kahl

Bergische Universität Wuppertal

October 11, 2012





Outline

Linear systems of equations in Lattice QCD

In Lattice QCD codes roughly 85% of time is spent solving linear systems of equations of the type

$$\mathbf{D}\psi = \varphi \qquad (\star)$$

Hence it is of utmost importance to find efficient solvers!

Solving $(\star)$ is required in many situations, e.g.,

- in the calculation of Propagators
- in the hybrid Monte-Carlo process

Depending on the discretization and situation

- $\mathbf{D}$ is sparse (e.g., Wilson) or dense (e.g., Overlap)
- One has to solve only for one rhs or for many rhs'

All this information should influence the choice of solver!

Discretizations of the Dirac operator

Discretizations by covariant finite-differences

$$\mathbf{d}_\mu \psi_x = a^{-1}(U^\mu_{x-a\mu}\psi_{x-a\mu} - (U^\mu_x)^\dagger \psi_{x+a\mu})$$

- Wilson discretization (stabilizing 2^{nd} order term)

$$\mathbf{D}_W = \sum_{\mu=1}^{4}(\gamma_\mu \otimes \mathbf{d}_\mu + a^{-1}\mathbf{d}_\mu^2) \quad \in \mathbb{C}^{12L_s^3L_t \times 12L_s^3L_t}$$

  Non-hermitian, sparse (next-neighbor), $(\gamma_5 D)^\dagger = \gamma_5 D$

- Overlap discretization (Ginsparg-Wilson)

$$\mathbf{D}_O = I + \gamma_5 \operatorname{sign}(\gamma_5(D_W - m)) \quad \in \mathbb{C}^{12L_s^3L_t \times 12L_s^3L_t}$$

Properties of linear systems in Lattice QCD

Typical discretizations yield linear systems $D\psi = \varphi$ where

- D is non-hermitian, yet $(\gamma_5 D)^\dagger = \gamma_5 D$
- $\mathrm{spec}(D)$ lies in the right half-plane
- D is very large (on a $32^3 \times 64$ lattice $\approx$ 25M unknowns)
- D is sparse, i.e., contains only next-neighbor couplings

$$\approx 100 \quad \text{non-zeroes per row}$$

Matrix-Vector operations are cheap $\mathcal{O}(L_s^3 L_t) = \mathcal{O}(V)$

In implementations $D \cdot x$ is often highly optimized

$\rightarrow$ use this in solvers for $D\psi = \varphi$

Notations

- Linear system of equation $\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, n$

$$Ax = b, \quad A \in \mathbb{C}^{n \times n}, x \in \mathbb{C}^n, b \in \mathbb{C}^n$$

- Euclidean inner product

$$\langle x, y \rangle_2 = y^\dagger x = \sum_{i=1}^{n} \bar{y}_i x_i$$

- Adjoint $A^\dagger$ of A w.r.t. $\langle ., . \rangle_2$

$$\langle Ax, y \rangle_2 = \langle x, A^\dagger y \rangle_2$$

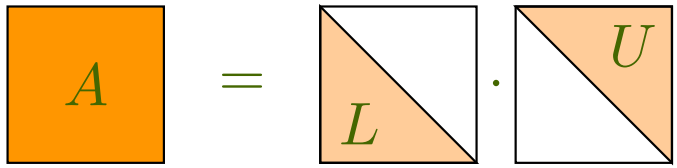
- A hermitian $\iff \quad A^\dagger = A$
- A hermitian positive definite

$$A^\dagger = A \quad \text{and} \quad x^\dagger A x > 0, \quad x \neq 0$$

Direct methods

Idea: Solve $Ax = b$ by row-/column-manipulations

- Usually based on factorizing the system matrix A
- Methods based on Gaussian elimination
  - $A = LU$: LU factorization



  - $A = LDL^*$: Cholesky factorization (A hermitian)

⊕ No restrictions on applications

⊖ Expensive methods ($\mathcal{O}(n^3)$ for dense matrices)

- Methods exploiting sparsity exist, reducing complexity

Other "direct" methods (limited to certain applications)

- Fast-Fourier-Transform ($\mathcal{O}(n\log(n))$)

Iterative solvers

Given: $Ax = b$ with solution $\hat{x}$, A sparse

Find: Approximations $x^{(k)}, k = 1, 2, \ldots$ s.t. $x^{(k)} \to \hat{x}$

1. How do we measure convergence $x^{(k)} \to \hat{x}$?
   - "Computable" measures ($\to$ stopping criteria)?
   - Monotone convergence in suitable norm possible?

2. How do we find suitable $x^{(k)}$ such that
   - the iterative process converges, i.e., $x^{(k)} \to \hat{x}$?
   - there is a "simple" update formula for $x^{(k+1)}$?
   - each iteration only requires the action of A on vector?

How do we measure convergence?

Given: Iterate $x^{(k)}$ in the k^{th} iteration

- Using the error $e^{(k)} = \hat{x} - x^{(k)}$

$$x^{(k)} \to \hat{x} \Longrightarrow ||e^{(k)}|| \to 0$$

In most cases the error is not readily computable!

- Using the residual $r^{(k)} = b - Ax^{(k)}$

$$x^{(k)} \to \hat{x} \Longrightarrow ||r^{(k)}|| \to 0$$

The residual is a computable quantity! Note that

$$r^{(k)} = b - Ax^{(k)} = A\hat{x} - Ax^{(k)} = Ae^{(k)}$$

In what follows we assume that $x^{(0)} = 0$

How do we find suitable $x^{(k)}$?

Task: Given b find x s.t. $Ax = b$ or

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, n \qquad (*)$$

Idea: Solve for x_i in $(*)$ for each i

- Jacobi iteration for $i = 1, \ldots, n$

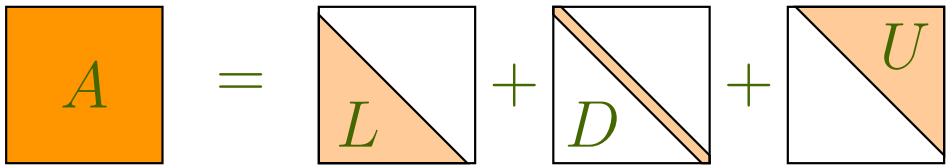
$$x_i^{(k+1)} = x_i^{(k)} \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right)$$

- Gauss-Seidel iteration for $i = 1, \ldots, n$

$$x_i^{(k+1)} = x_i^{(k)} \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right)$$

Splitting methods

Splitting methods use the additive decomposition of A



- Jacobi iteration $x^{(k+1)} = x^{(k)} + D^{-1}r^{(k)}$
- Gauss-Seidel iteration $x^{(k+1)} = x^{(k)} + (D+L)^{-1}r^{(k)}$
- Successive over-relaxation

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega(D+L)^{-1}r^{(k)}$$

- In general if $A = M + N$

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

Convergent iff $\|I - M^{-1}A\| < 1$

Oftentimes used as preconditioners ($\rightarrow$ Solvers II)

Linear Algebra (Minimal polynomial)

Let $p_m^\star$ be the polynomial of smallest degree m s.t.

$$p_m^\star(A) = 0 \quad \text{with} \quad p_m^\star(t) = 1 - t q_{m-1}^\star(t).$$

Then $A^{-1} = q_{m-1}^\star(A)$ is a polynomial in A!

$\Rightarrow$ Solution $\hat{x}$ of $Ax = b$ given by $q_{m-1}^\star(A)b$

Idea: Polynomial approximations $x^{(k)}$ of $\hat{x}$ by

$$x^{(k)} = q_k(A)b, \quad q_k(t) \in \Pi_k = \{p(t) = \sum_{\ell=0}^{k} \alpha_\ell t^\ell\}$$

- Computation of $x^{(k+1)}$ requires
  - multiplication by A
  - update of coefficients $\alpha_1, \ldots, \alpha_{k+1}$

Krylov subspace methods

Krylov subspace methods

Approximation $x^{(k)}$ of the solution $\hat{x}$ in Krylov subspace

$$\mathcal{K}_k(A,b) = \{p(A)b : p \in \Pi_{m-1}\} = \mathsf{span}\{b, Ab, \ldots, A^{m-1}b\}$$

- For $y \in \mathcal{K}_k(A,b)$ we have $Ay \in \mathcal{K}_{k+1}(A,b)$
- Orthonormal bases of $\mathcal{K}_k(A,b) \rightarrow$ Arnoldi iteration

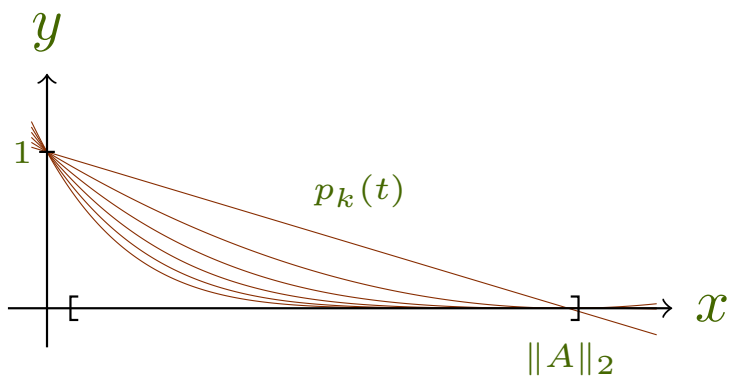
- Many ways to choose $x^{(k)}$ from $\mathcal{K}_k(A,b)$
  - stationary iterations (e.g. Richardson, Chebyshev)
  - non-stationary iterations (e.g. CG, MINRES, GMRES)

Richardson iteration

For A symmetric positive definite, i.e., $\mathrm{spec}(A) \subseteq \mathbb{R}^+$

Richardson Iteration

b given, $x^{(0)} = 0$, $\alpha > \|A\|_2$
for $k = 0, 1, 2, \ldots$ **do**
$x^{(k+1)} = (I - \frac{1}{\alpha}A)x^{(k)} + \frac{1}{\alpha}b$
end for



Hence we have

$$\begin{aligned} r^{(k)} &= (I - \alpha A)^k r^{(0)} = p_k(A)b \\ p_k(t) &= 1 - tq_{k-1}(t) \in \bar{\Pi}_k = \{p \in \Pi_k \mid p(0) = 1\} \\ x^{(k)} &= q_{k-1}(A)b \in \mathcal{K}_k(A, b) \end{aligned}$$

a stationary Krylov subspace method!

Krylov subspace methods — Theory

Does Richardson iteration converge? We have

$$\|r^{(k)}\| = \|(I - \alpha A)^k b\| \leq \underbrace{\|(I - \alpha A)\|}_{<1}^k \|b\| \longrightarrow 0$$

By Definition all Krylov subspace methods are described by

$$\begin{aligned} r^{(k)} &= p_k(A)b \\ e^{(k)} &= p_k(A)e^{(0)} \\ p_k(t) &= 1 - tq_{k-1}(t) \\ x^{(k)}(t) &= q_{k-1}(A)b \end{aligned}$$

Hence $\|r^{(k)}\| \leq \|p_k(A)\|\|b\|$ and $\|e^{(k)}\| \leq \|p_k(A)\|\|e^{(0)}\|$

Task: Find optimal $p_k \in \bar{\Pi}_k$ for specific choices of $\|.\|$!

Optimal Krylov subspace methods I — Conjugate Gradients

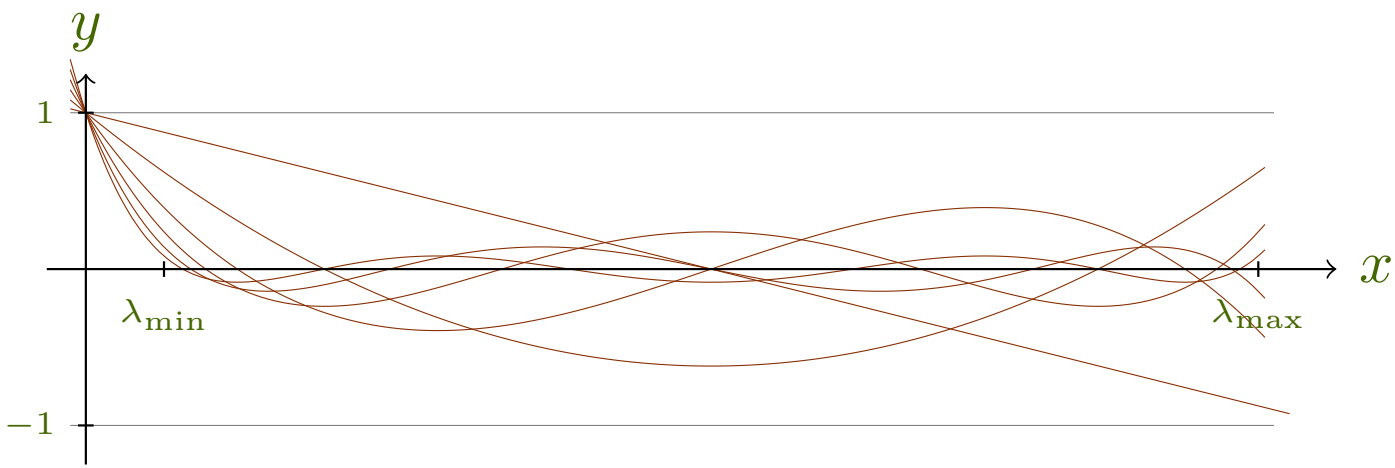
For A hermitian positive definite define the A-norm $\|.\|_A$ by

$$\|x\|_A = \langle x, x\rangle_A^{\frac{1}{2}} \quad \text{with} \quad \langle x, y\rangle_A = \langle Ax, y\rangle$$

What is the optimal Krylov subspace method w.r.t. $\|.\|_A$?

$$\min_{p_k \in \bar{\Pi}_k} \|p_k(A)\|_A = \min_{p_k \in \bar{\Pi}_k} \max_{\lambda \in \mathrm{spec}(A)} |p_k(\lambda)|$$

The well-known Conjugate Gradients (CG) method!



Conjugate Gradients — Algorithm

Conjugate Gradients

$r^{(0)} = b, p^{(0)} = r^{(0)}$

for $k = 1, 2, \ldots$ **do**

$$\alpha_{k-1} = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle_2}{\langle A p^{(k-1)}, p^{(k-1)} \rangle_2}$$

$$x^{(k)} = x^{(k-1)} + \alpha_{k-1} p^{(k-1)}$$

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

$$\beta_{k-1} = \frac{\langle r^{(k)}, r^{(k)} \rangle_2}{\langle r^{(k-1)}, r^{(k-1)} \rangle_2}$$

$$p^{(k)} = r^{(k)} + \beta_{k-1} p^{(k-1)}$$

end for

Minimization of the functional $\mathcal{L}(x) = \frac{1}{2}\langle x, x \rangle_A - \langle x, b \rangle_2$

- $p^{(k)}$ conjugate gradients of $\mathcal{L}$

Conjugate Gradients — Summary

1. Minimal error in $\|.\|_A$ for $x^{(k)} \in \mathcal{K}_k(A,b)$

$$\|e^{(k)}\| = \min_{p_k \in \bar{\Pi}_k} \|p_k(A)\|_A \|e^{(0)}\|_A \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|e^{(0)}\|_A$$

2. Variational property

$$r^{(k)} \perp \mathcal{K}_k(A,b)$$

3. Minimization of the functional

$$\mathcal{L}(x) = \frac{1}{2}\langle x,x\rangle_A - \langle x,b\rangle_2, \quad x \in \mathcal{K}_k(A,b)$$

   - $p^{(k)}$ conjugate gradients (i.e., A-orthogonal) of $\mathcal{L}$

4. Short recurrence, i.e., $x^{(k+1)}$ requires only $x^{(k)}$

Optimal Krylov subspace methods II — MINRES

For A hermitian find the Krylov subspace method with

$$\min_{x^{(k)} \in \mathcal{K}_k(A,b)} \|b - Ax^{(k)}\|_2 = \min_{x^{(k)} \in \mathcal{K}_k(A,b)} \|r^{(k)}\|_2$$

The optimal method w.r.t. $\|.\|_2$ is known as MINRES*

Similar to CG, MINRES introduces search directions $p^{(k)}$

- The residuals are conjugated, i.e., mutually A-orthogonal
- The $Ap^{(k)}$ are mutually orthogonal
- Short recurrence, i.e., $x^{(k+1)}$ requires only $x^{(k)}$

*Conjugate Residuals (CR)

MINRES — Algorithm

Conjugate Residuals (CR)

$r^{(0)} = b, p^{(0)} = r^{(0)}$

for $k = 1, 2, \ldots$ **do**

$\alpha_{k-1} = \frac{\langle r^{(k-1)}, Ar^{(k-1)} \rangle_2}{\langle Ap^{(k-1)}, Ap^{(k-1)} \rangle_2}$

$x^{(k)} = x^{(k-1)} + \alpha_{k-1} p^{(k-1)}$

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

$\beta_{k-1} = \frac{\langle r^{(k)}, Ar^{(k)} \rangle_2}{\langle r^{(k-1)}, Ar^{(k-1)} \rangle_2}$

$p^{(k)} = r^{(k)} + \beta_{k-1} p^{(k-1)}$

Compute $Ap_k = Ar_k + \beta_{k-1} Ap_{k-1}$

end for

MINRES is a stable realization of CR

Optimal Krylov subspace methods II — GMRES

What if A is not hermitian? Optimality w.r.t. $\|.\|_2$ possible

$$\min_{x^{(k)} \in \mathcal{K}_k(A,b)} \|b - Ax^{(k)}\|_2 = \min_{x^{(k)} \in \mathcal{K}_k(A,b)} \|r^{(k)}\|_2$$

Idea: For orthonormal basis $v_1, \ldots, v_k$ of $\mathcal{K}_k(A, b)$

$$x \in \mathcal{K}_k(A,b) \Longrightarrow x = \sum_{\ell=1}^{k} v_\ell y_\ell = [v_1 \mid \ldots \mid v_k] y = V_k y$$

Hence we find

$$\min_{x^{(k)} \in \mathcal{K}_k(A,b)} \|b - Ax^{(k)}\|_2 = \min_{y} \|b - AV_k y\|_2 \qquad (\star)$$

The Arnoldi Iteration

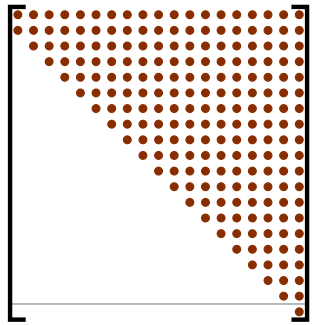
Compute orthonormal basis $\{v_1, \ldots, v_k\}$ of $\mathcal{K}_k(A, b)$

Arnoldi Iteration

$\beta = \|b\|_2, v_1 = \beta^{-1} b$
for $k = 1, 2, \ldots$ **do**
  $q = Av_k$
  for $j = 1, \ldots, k$ **do**
    $h_{j,k} = \langle v_j, q \rangle_2$
    $q = q - h_{j,k} v_j$
  end for
  $h_{k+1,k} = \|q\|_2$
  $v_{k+1} = h_{k+1,k}^{-1} q$
end for

With $V_k = [v_1 \mid \ldots \mid v_k]$ and

$$H_{k+1,k} =$$



we have the Arnoldi relation

$$AV_k = V_{k+1} H_{k+1,k}$$

with $V_k^\dagger V_k = I$

Optimal Krylov subspace methods III — GMRES

Using the Arnoldi relation in $(\star)$ we find

$$\min_y \|b - AV_k y\|_2 = \min_y \|b - V_{k+1} H_{k+1,k} y\|_2$$

Since V_{k+1} has orthonormal columns and $v_1 = \|b\|_2^{-1} b$

$$\min_y \|b - V_{k+1} H_{k+1,k} V_k y\|_2 = \min_y \|\|b\|_2 e_1 - H_{k+1,k} y\|_2$$

Naïve GMRES

for $k = 1, 2, \ldots$ **do**
  Compute $v_k, H_{k+1,k}$ *(Arnoldi)*
  Solve $\operatorname{argmin}_y \|\|b\|_2 e_1 - H_{k+1,k} y\|_2$
  $x^{(k)} = V_k y$
end for

Restarted GMRES

Although an optimal method GMRES has severe drawbacks:

- The computation of $x^{(k)}$ requires $V_k = [v_1 \mid \ldots \mid v_k]$
  $\Rightarrow$ Storage requirements grow with k
- Least-Squares solution requires $\mathcal{O}(k^3)$ operations
  $\Rightarrow$ Computation time per iterate grows with k

Idea: Restart GMRES every m-iterations ($\Rightarrow$ GMRES(m))

GMRES(m)

for $\ell = 0, 1, \ldots$ **do**
$r^{(0)} = b - Ax^{(0)}, \beta = \|r^{(0)}\|_2, v_1 = \beta^{-1} r^{(0)}$
Compute $V_m, H_{m+1,m}$ *(Arnoldi)*
$y_m = \operatorname{argmin}_y \|\beta e_1 - H_{m+1,m} y\|_2$
$x^{(0)} = x^{(0)} + V_m y_m$
end for

Optimal Krylov subspace methods — Summary

	requirements	optimality	recurrence
CG	$A = A^\dagger$ $\langle x, x \rangle_A > 0, x \neq 0$	$\|.\|_A$	short
MINRES[†]	$A = A^\dagger$	$\|.\|_2$	short
GMRES[‡]	none	$\|.\|_2$	long

- What are the requirements for short recurrence?
- Do non-optimal methods exists with short recurrence?

[†] CR
[‡] GCR

Faber-Manteuffel Theorem

The Faber-Manteuffel Theorem

There exists an optimal method with $(s+2)$-term recurrence iff A is s-normal, i.e., $A^\dagger = p(A), \quad p \in \Pi_s$

- A normal $\Longrightarrow A^\dagger = p(A), \quad p \in \Pi_{n-1}$
- A hermitian, $A^\dagger = A \longrightarrow$ (3-term recurrence)
- A anti-hermitian $A^\dagger = -A \longrightarrow$ (3-term recurrence)
  - Chiral operator $\gamma_5 D = -D\gamma_5 \Longrightarrow D^\dagger = -D$

Barth-Manteuffel Theorem — Ginsparg-Wilson relation

The Barth-Manteuffel Theorem

There exists an optimal method with $(s+2,t)$-term recurrence iff A is (s,t)-normal, i.e.,

$$A^\dagger = \frac{p(A)}{q(A)}, \quad p \in \Pi_s, q \in \Pi_t$$

- Methods have multiple recursions
- Occurrence in Lattice QCD: Ginsparg-Wilson relation

$$\begin{aligned} D\gamma_5 + \gamma_5 D = aD\gamma_5 D &\iff \gamma_5(I - aD)D^\dagger = -\gamma_5 D \\ &\iff D^\dagger = (I - aD)^{-1} D \end{aligned}$$

- D fulfills Ginsparg-Wilson $\iff$ D is $(1,1)$-normal

Non-optimal Krylov subspace methods — BCG

What if no optimal short recurrence method exists for A?

Ansatz: Throw optimality over board!

- Instead of building one Krylov subspace build two

$$\mathcal{K}_k(A, r^{(0)}) \quad \text{and} \quad \mathcal{K}_k(A^\dagger, \tilde{r}^{(0)})$$

  - bi-orthogonalization
- Similar to CG, the residuals of BCG fulfill

$$r^{(k)} \perp \mathcal{K}_k(A^\dagger, \tilde{r}^{(0)})$$

- Not optimal in any norm
  - erratic convergence behaviour ($\rightarrow$ excercises)
  - breakdowns can occur, i.e., convergence not guaranteed
- BiCGstab is a stabilized variant of BCG

BiCGstab

$r^{(0)} = b, \beta_0 = 0$

$\hat{r} = r$ shadow residual $\langle p, \hat{p} \rangle_2 \neq 0$

for $k = 0, 1, \ldots$ **do**

$\rho_k = \langle r^{(k)}, \hat{r} \rangle_2$

$\beta_k = \frac{\rho_k}{\rho_{k-1}} \cdot \frac{\alpha_{k-1}}{\omega_{k-1}}$

$p^{(k)} = r^{(k)} + \beta_k (p^{k-1} - \omega_{k-1} v^{(k-1)})$

$\alpha_k = \frac{\rho_k}{\langle Ap^{(k)}, \hat{r} \rangle_2}$

$x^{(k+\frac{1}{2})} = x^{(k)} + \alpha_k p^{(k)}$

$s^{(k)} = r^{(k)} - \alpha_k A p^{(k)}$ $s^{(k)} \equiv r^{(k+\frac{1}{2})}$

$\omega_k = \frac{\langle s^{(k)}, As^{(k)} \rangle_2}{\langle As^{(k)}, As^{(k)} \rangle_2}$

$x^{(k+1)} = x^{(k+\frac{1}{2})} + \omega_k s^{(k)}$

$r^{(k+1)} = s^{(k)} - \omega_k A s^{(k)}$

end for

List of Methods

	requirements	optimal	recurrence	
CG	hpd	$\|.\|_A$	short	
MINRES[§]	hermitian	$\|.\|_2$	short	
GMRES[¶]	none	$\|.\|_2$	long	$\rightarrow$ restarts
CGN	none	$\|.\|_{A^\dagger A}$	short	$A^\dagger A x = A^\dagger b$
BCG	none	no	short	similar to CG unstable
QMR	none	no	short	similar to GMRES
BiCGstab	none	no	short	breakdowns
SUMR	shifted unitary	$\|.\|_2$	short	multiple recursion

For more Krylov subspace methods see [3, 6].

[§] equivalent to Conjugate Residuals

[¶] equivalent to Generalized Conjugate Residuals

Krylov subspace methods are all-duty solvers

- require only multiplication by $A\cdot$ and inner products
  - easy to implement (especially if $A\cdot$ is already done)
  - easy to parallelize ($\log(p)$-scaling due to inner products)
- Whenever short-term recurrence is possible
  - constant cost per iteration
  - constant memory consumption
- If only long-term recurrence is possible
  - restarts limit amount of work
  - deflated-restarts are even more efficient ($\rightarrow$ Solvers II)
- Convergence speed depends on the spectrum of A
  - separation from the origin important (since $p_k \in \bar{\Pi}_k$)
  - preconditioning improves performance ($\rightarrow$ Solvers II)


T. Barth and T. Manteuffel.
Multiple recursion conjugate gradient algorithms. I. sufficient conditions.
SIAM J. Matrix Anal. Appl., 21, 2000.


V. Faber and T. Manteuffel.
Necessary and sufficient conditions for the existence of a conjugate gradient method.
SIAM J. Numer. Anal., 21, 1984.


A. Greenbaum.
Iterative Methods for Solving Linear Systems, volume 17 of *Frontiers in Applied Mathematics*.
Society for Industrial and Applied Mathematics, 1997.


M. Hestenes and E. Stiefels.
Methods of conjugate gradients for solving linear systems.
Journal of Research of the National Bureau of Standards, Section B, 49, 1952.


C. Jagels and L. Reichel.
A fast minimal residual algorithm for shifted unitary matrices.
Numer. Linear Algebra Appl., 1, 1994.


Y. Saad.
Iterative Methods for Sparse Linear Systems.
Society for Industrial and Applied Mathematics, 2nd edition, 2003.