Tracking the trajectory in finite precision CG computations

Tomáš Gergelits [1] Zdeněk Strakoš [1,2]

[1] Faculty of Mathematics and Physics, Charles University in Prague

[2] Institute of Computer Science AS CR

Seminar on Numerical Analysis 2014

Nymburk, 28th January 2014

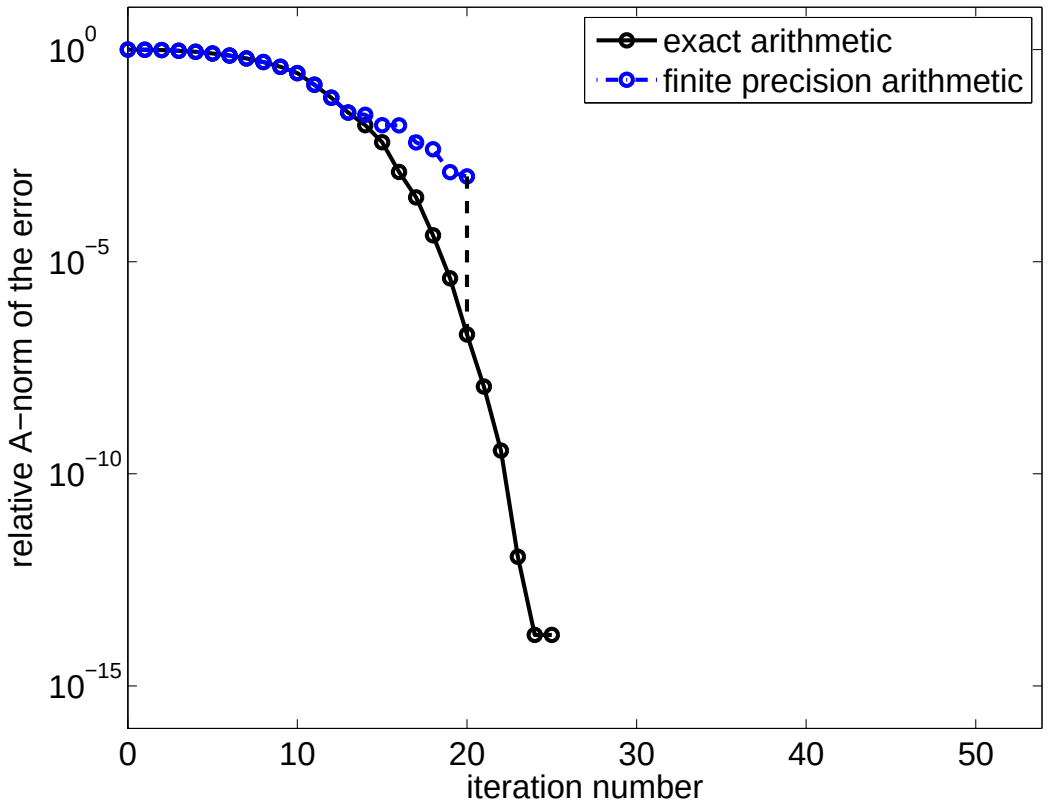
Contents

Contents

Simple numerical experiment

Solve $Ax = b$ by the CG method

A ... 25×25 diagonal and positive definite matrix, $b = \text{ones}(1, \ldots, 1)$



- 20 iterations of CG in exact arithmetic

$$\|x - x_{20}\|_A = 2.0611 \times 10^{-6}$$

- 20 iterations of CG in finite precision arithmetic (FP CG)

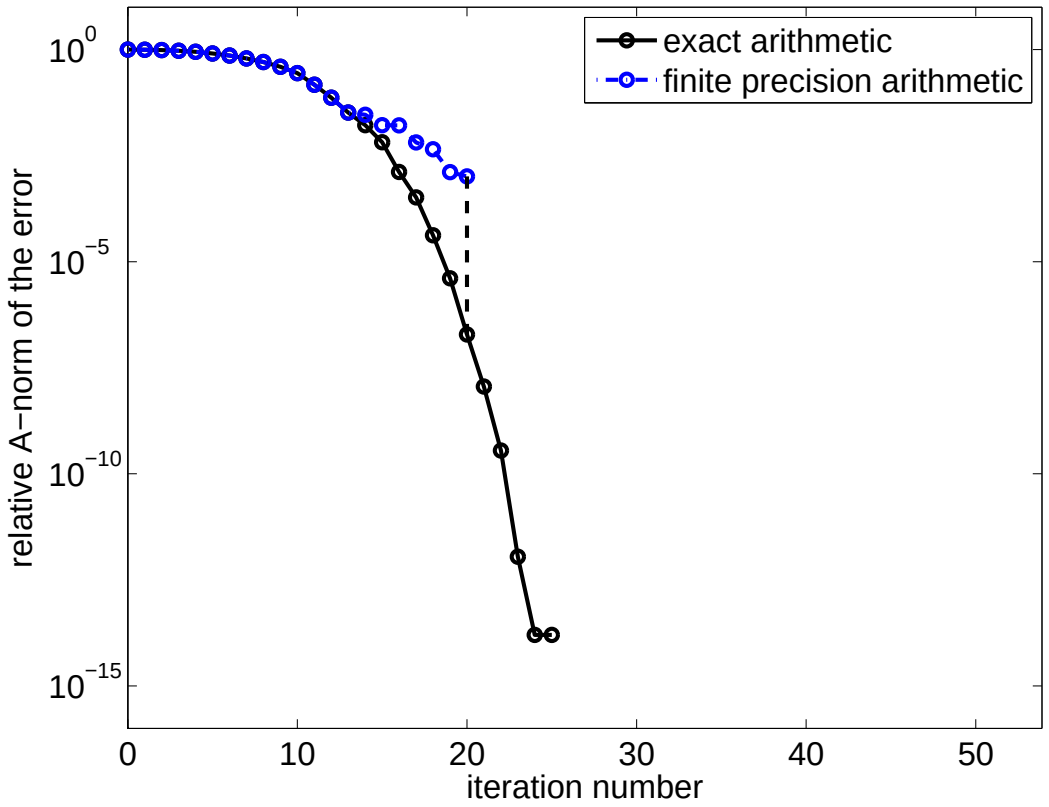
$$\|x - \bar{x}_{20}\|_A = 1.0953 \times 10^{-2}$$

- $\|x_{20} - \bar{x}_{20}\|_\infty = 1.424 \times 10^{-2}$

Simple numerical experiment

Solve $Ax = b$ by the CG method

A ... 25×25 diagonal and positive definite matrix, $b = \text{ones}(1, \ldots, 1)$



- 20 iterations of CG in exact arithmetic

$$\|x - x_{20}\|_A = 2.0611 \times 10^{-6}$$

- 20 iterations of CG in finite precision arithmetic (FP CG)

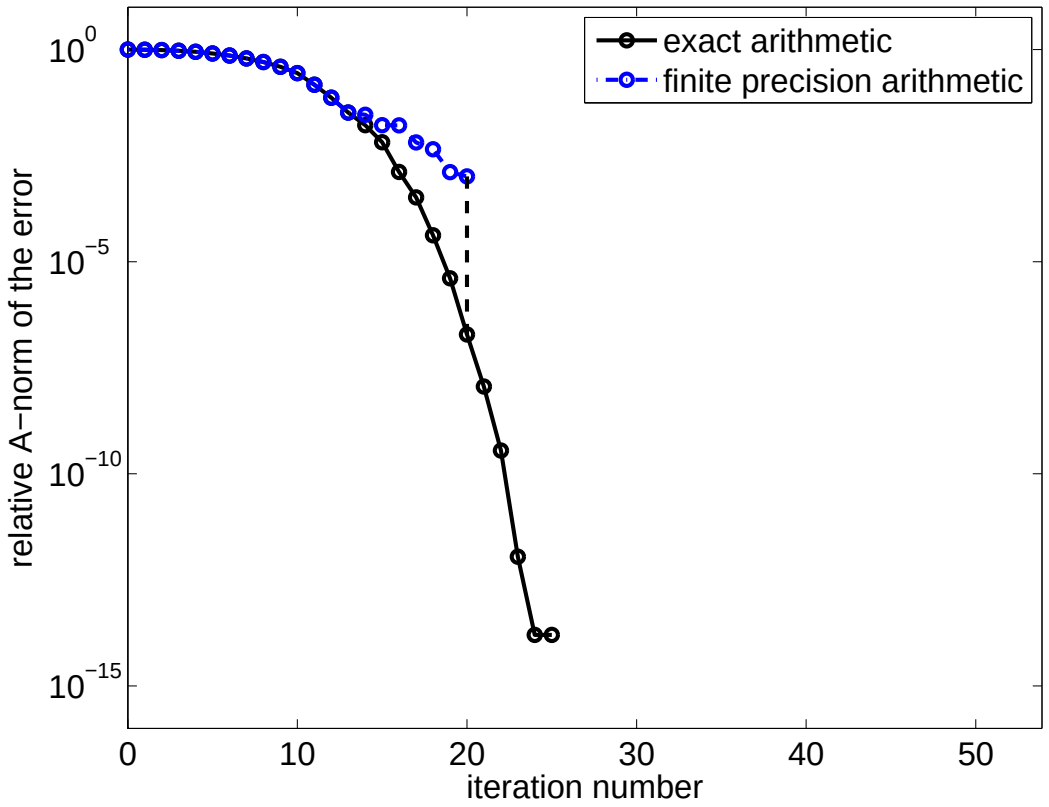
$$\|x - \bar{x}_{20}\|_A = 1.0953 \times 10^{-2}$$

- $\|x_{20} - \bar{x}_{20}\|_\infty = 1.424 \times 10^{-2}$

Simple numerical experiment

Solve $Ax = b$ by the CG method

A ... 25×25 diagonal and positive definite matrix, $b = \text{ones}(1, \ldots, 1)$



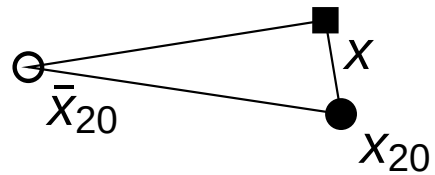
- 20 iterations of CG in exact arithmetic

$$\|x - x_{20}\|_A = 2.0611 \times 10^{-6}$$

- 20 iterations of CG in finite precision arithmetic (FP CG)

$$\|x - \bar{x}_{20}\|_A = 1.0953 \times 10^{-2}$$

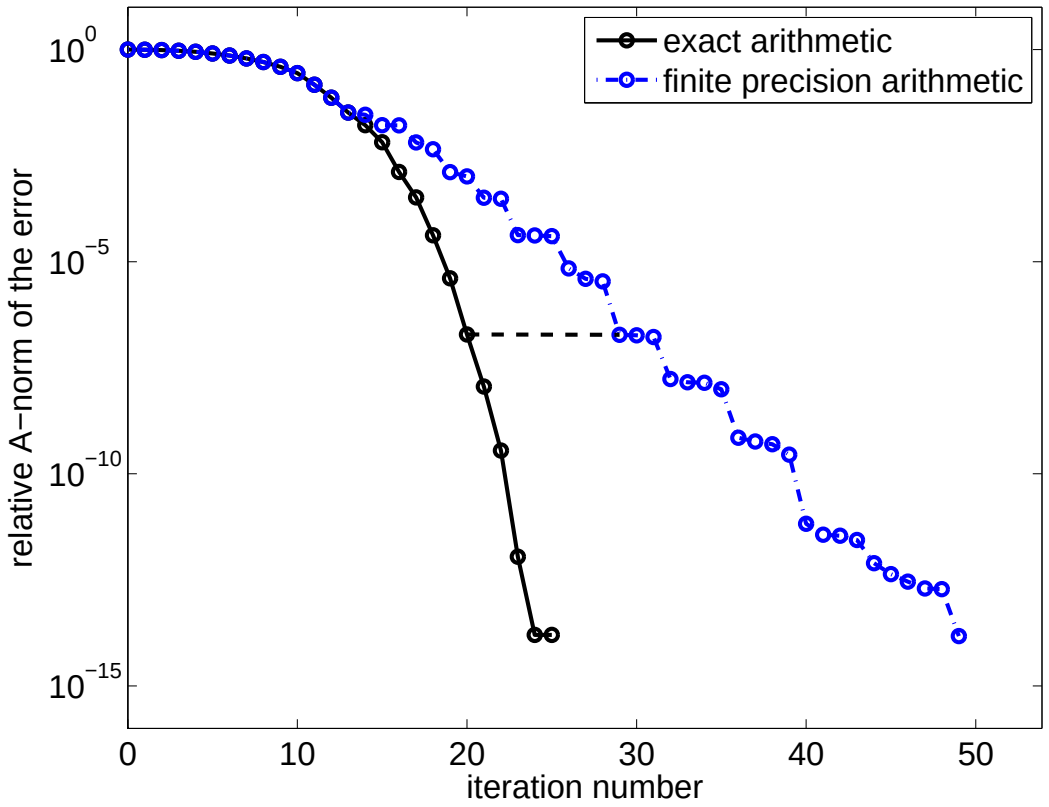
- $\|x_{20} - \bar{x}_{20}\|_\infty = 1.424 \times 10^{-2}$



Simple numerical experiment

Solve $Ax = b$ by the CG method

A ... 25×25 diagonal and positive definite matrix, $b = \text{ones}(1, \ldots, 1)$



- 20 iterations of CG in exact arithmetic

$$\|x - x_{20}\|_A = 2.0611 \times 10^{-6}$$

- 29 iterations of FP CG for the same level of accuracy

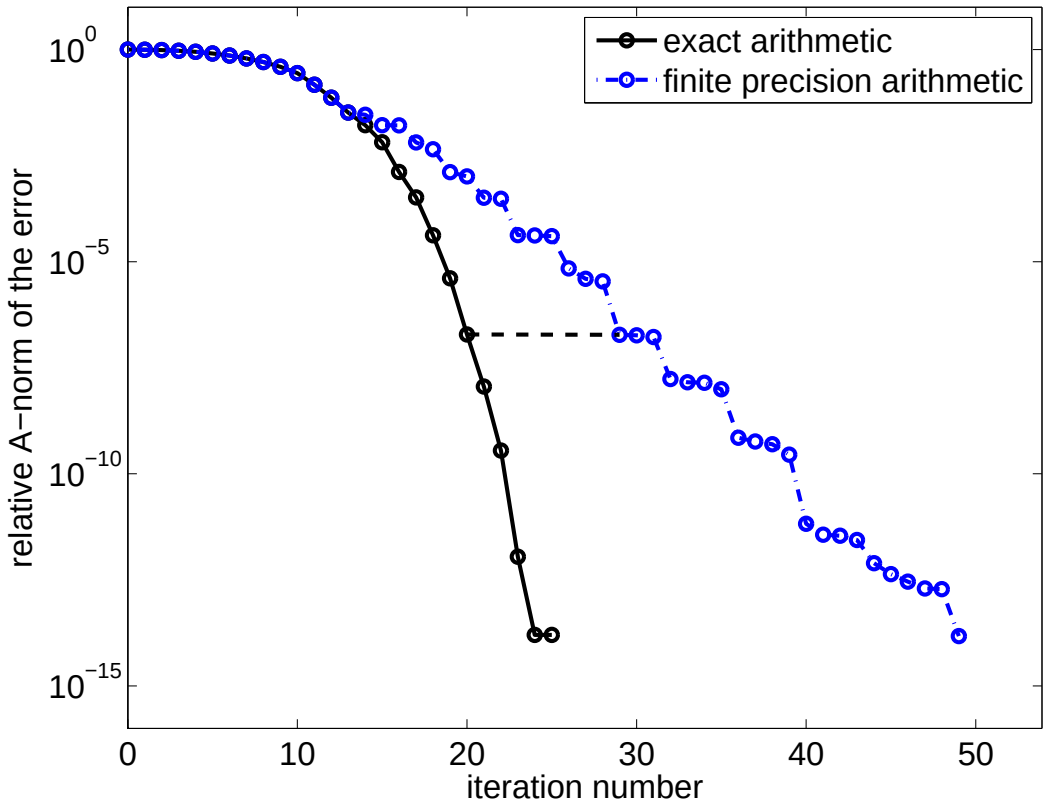
$$\|x - \bar{x}_{29}\|_A = 2.0057 \times 10^{-6}$$

- $\|x_{20} - \bar{x}_{29}\|_\infty = 1.304 \times 10^{-7}$

Simple numerical experiment

Solve $Ax = b$ by the CG method

A ... 25×25 diagonal and positive definite matrix, $b = \text{ones}(1, \ldots, 1)$



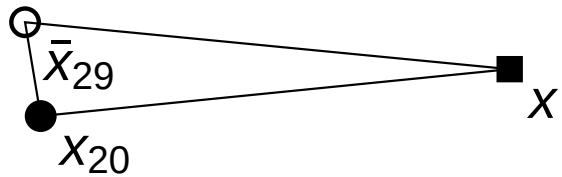
- 20 iterations of CG in exact arithmetic

$$\|x - x_{20}\|_A = 2.0611 \times 10^{-6}$$

- 29 iterations of FP CG for the same level of accuracy

$$\|x - \bar{x}_{29}\|_A = 2.0057 \times 10^{-6}$$

- $\|x_{20} - \bar{x}_{29}\|_\infty = 1.304 \times 10^{-7}$



Contents

The essence of the CG method I.

$$Ax = b, \quad A \in \mathbb{F}^{N\times N} \text{ HPD matrix}, \quad b \in \mathbb{F}^N \quad \text{where } \mathbb{F} \text{ is } \mathbb{C} \text{ or } \mathbb{R}$$

Hestenes, Stiefel (1952), Lanczos (1952)

- CG is a projection method which minimizes the energy norm of the error

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots$$

$$\text{where} \quad \mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k-1} r_0\}$$

$$\|x - x_k\|_A = \min\{\|x - y\|_A : \ y \in x_0 + \mathcal{K}_k(A, r_0)\}$$

- CG is conforming with the Galerkin approximation

$$\|\nabla(\mathbf{u} - \mathbf{u}_h^{(k)})\|^2 = \|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 + \|x - x_k\|_A^2$$

The essence of the CG method I.

$$Ax = b, \quad A \in \mathbb{F}^{N\times N} \text{ HPD matrix}, \quad b \in \mathbb{F}^N \quad \text{where } \mathbb{F} \text{ is } \mathbb{C} \text{ or } \mathbb{R}$$

Hestenes, Stiefel (1952), Lanczos (1952)

- CG is a projection method which minimizes the energy norm of the error

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots$$

$$\text{where} \quad \mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k-1} r_0\}$$

$$\|x - x_k\|_A = \min\{\|x - y\|_A : \ y \in x_0 + \mathcal{K}_k(A, r_0)\}$$

- CG is conforming with the Galerkin approximation

$$\|\nabla(\mathbf{u} - \mathbf{u}_h^{(k)})\|^2 = \|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 + \|x - x_k\|_A^2$$

The essence of the CG method I.

$$Ax = b, \quad A \in \mathbb{F}^{N\times N} \text{ HPD matrix}, \quad b \in \mathbb{F}^N \quad \text{where } \mathbb{F} \text{ is } \mathbb{C} \text{ or } \mathbb{R}$$

Hestenes, Stiefel (1952), Lanczos (1952)

- CG is a projection method which minimizes the energy norm of the error

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots$$

$$\text{where} \quad \mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k-1} r_0\}$$

$$\|x - x_k\|_A = \min\{\|x - y\|_A : \ y \in x_0 + \mathcal{K}_k(A, r_0)\}$$

- CG is conforming with the Galerkin approximation

$$\|\nabla(\mathbf{u} - \mathbf{u}_h^{(k)})\|^2 = \|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 + \|x - x_k\|_A^2$$

The essence of the CG method II.

- CG method is mathematically equivalent to the Lanczos process:

$$AV_k = V_k T_k + \delta_{k+1} v_{k+1} e_k^T \quad \text{with given} \quad v_1 = r_0/\|r_0\|$$

  - The CG approximations x_k are given by

$$T_k y_k = \|r_0\| e_1, \quad x_k = x_0 + V_k y_k .$$

  - Jacobi matrix T_k is the $k \times k$ matrix representation of the OG projected operator

$$A_k \equiv V_k V_k^* A V_k V_k^* : \mathcal{K}_k(A, r_0) \to \mathcal{K}_k(A, r_0).$$

- CG is a matrix formulation of the Gauss-Christoffel quadrature

$$\begin{array}{ccc} A, r_0/\|r_0\| & \longleftrightarrow & \omega(\lambda), \int_0^\infty f(\lambda)\, d\omega(\lambda) \\ \uparrow & & \uparrow \\ & \text{first 2k moments matched} & \\ \uparrow & & \uparrow \\ T_k, e_1 & \longleftrightarrow & \omega^{(k)}(\lambda), \sum_{i=1}^k \omega_i^{(k)} f(\lambda_i^{(k)}) \end{array}$$

The essence of the CG method II.

- CG method is mathematically equivalent to the Lanczos process:

$$AV_k = V_k T_k + \delta_{k+1} v_{k+1} e_k^T \quad \text{with given} \quad v_1 = r_0/\|r_0\|$$

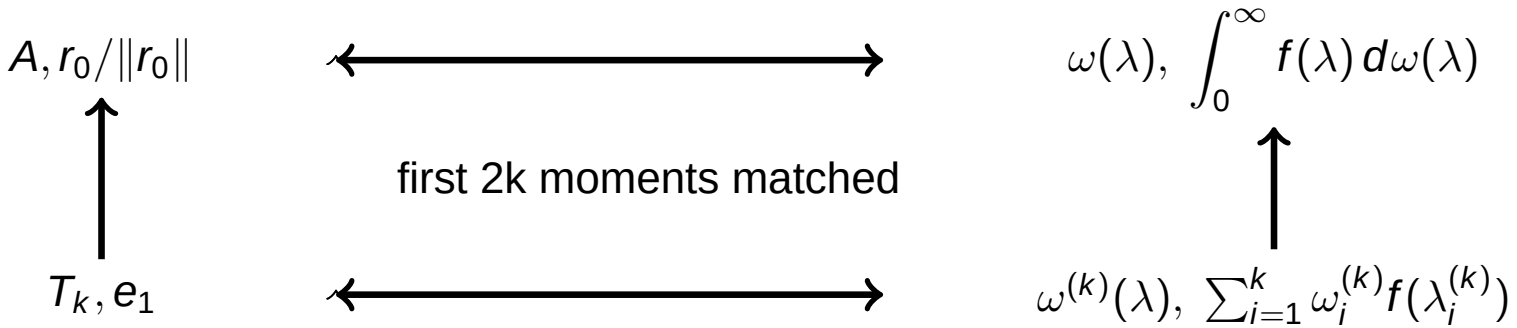
  - The CG approximations x_k are given by

$$T_k y_k = \|r_0\| e_1, \quad x_k = x_0 + V_k y_k .$$

  - Jacobi matrix T_k is the $k \times k$ matrix representation of the OG projected operator

$$A_k \equiv V_k V_k^* A V_k V_k^* : \mathcal{K}_k(A, r_0) \to \mathcal{K}_k(A, r_0).$$

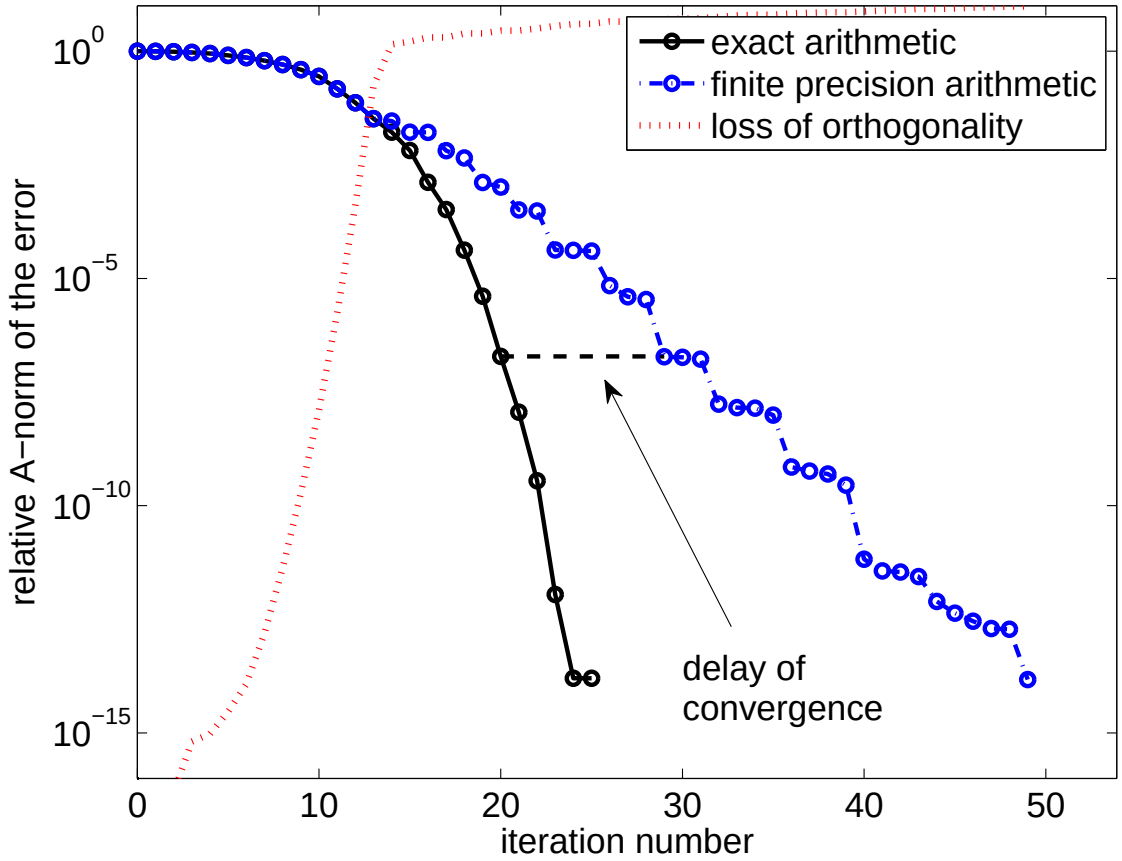
- CG is a matrix formulation of the Gauss-Christoffel quadrature



Contents

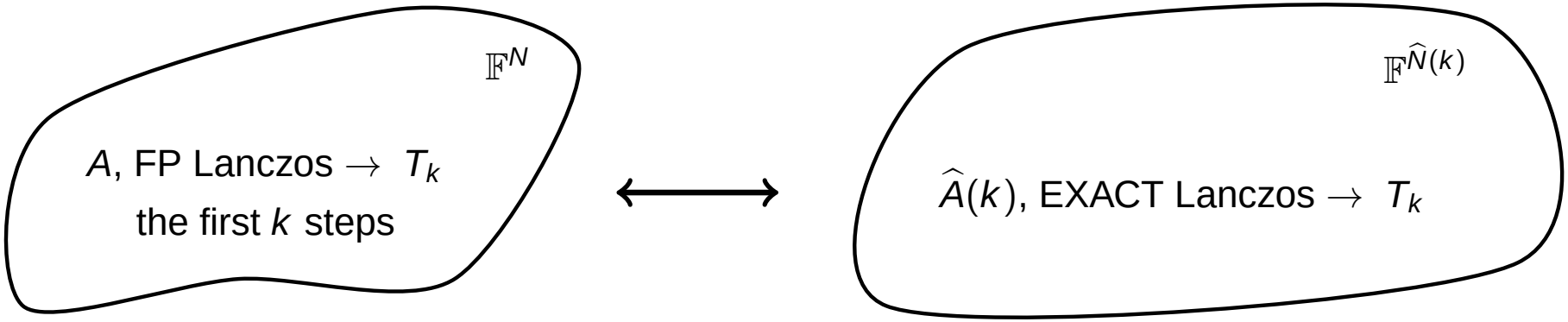
Delay of convergence

Short recurrences $\Longrightarrow$ loss of orthogonality $\Longrightarrow$ delay of convergence



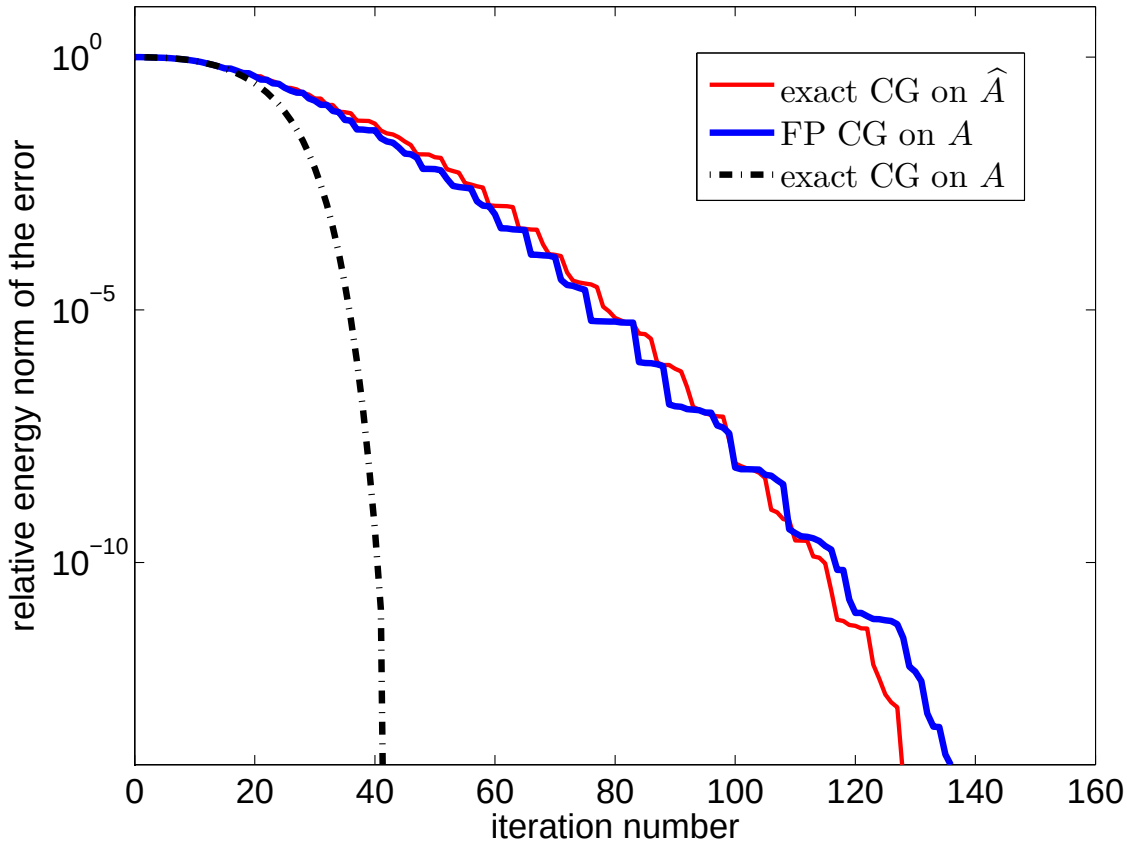
Scheme of backward-like analysis

Paige (1971, 1980), Greenbaum (1989), S (1991), Greenbaum and S (1992)



- Eigenvalues of $\widehat{A}(k)$ are tightly clustered around the eigenvalues of A.
- In numerical experiments, $\widehat{A}(k)$ can be replaced (with small inaccuracy) by an "universal" $\widehat{A}$ with sufficiently many eigenvalues in the tight clusters around the eigenvalues of A.

Illustration of backward-like analysis



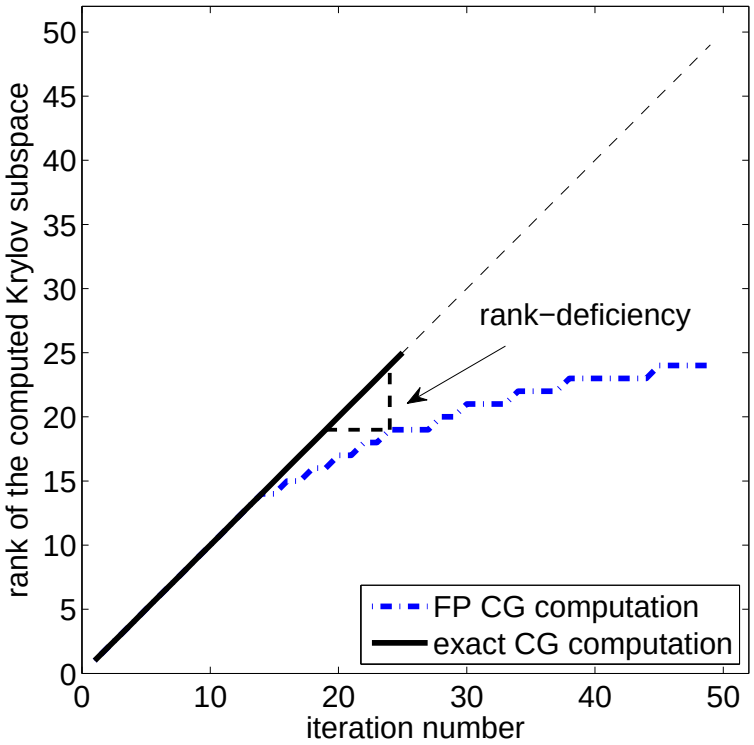
$$\|x - \bar{x}_k\|_A \approx \|\widehat{x} - \widehat{x}_k\|_{\widehat{A}}$$

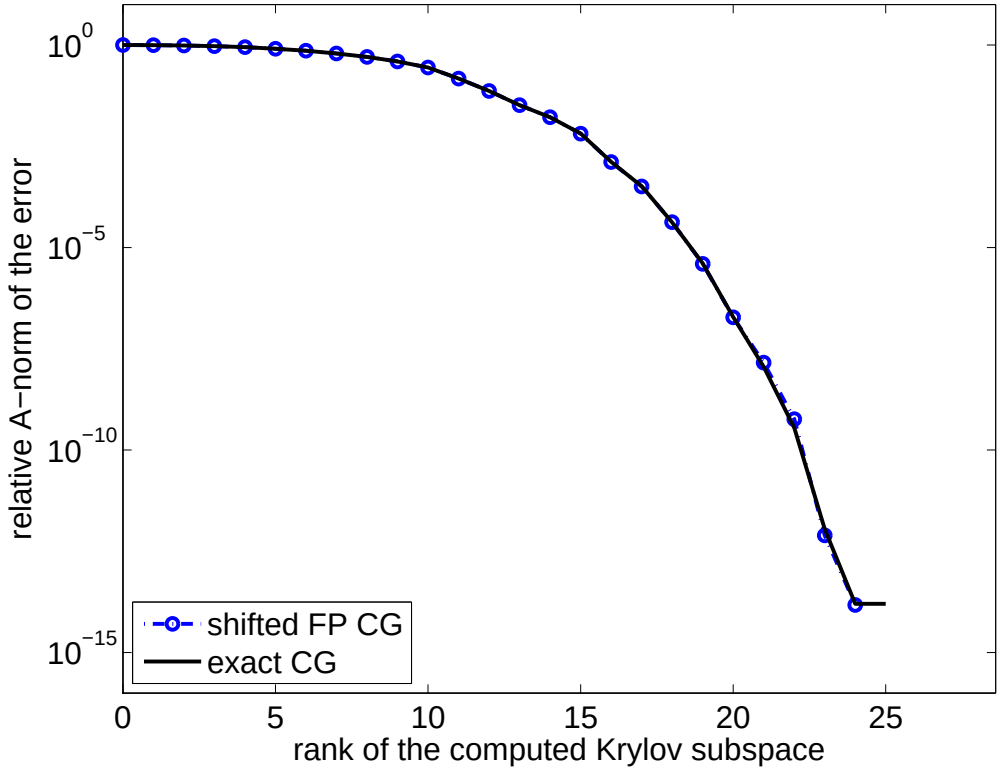
We can relate k-th iteration of finite precision CG applied on A with k-th iteration of exact CG applied on blurred problem $\widehat{A}$.

Contents

Rank-deficiency and delay of convergence

FP CG convergence curve shifted correspondingly to the rank-deficiency of the computed Krylov subspace gives the exact CG convergence curve.

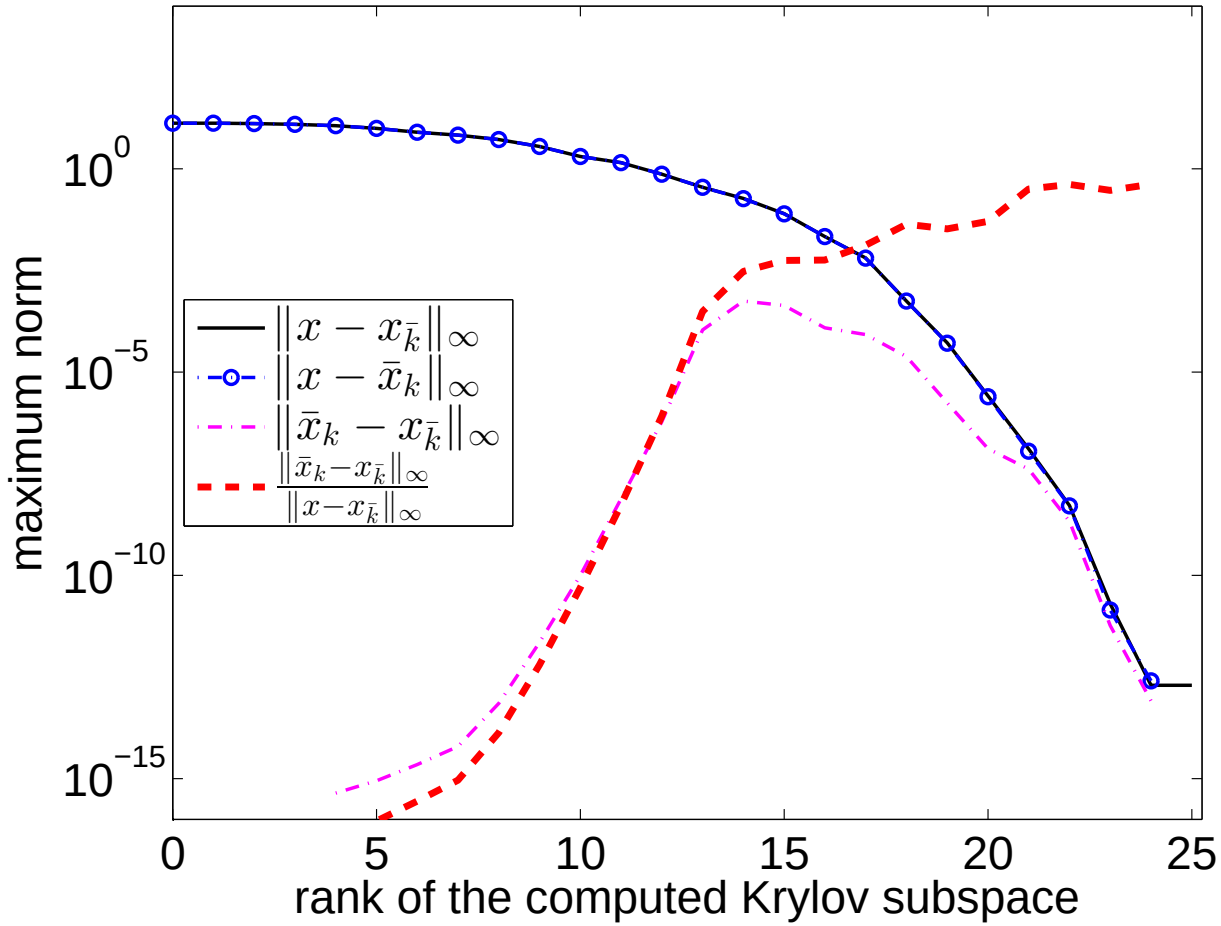




- Rank deficiency: $k - \bar{k}$, where $\bar{k} = \mathrm{rank}(\mathcal{K}_k(A, r_0))$ is the numerical rank of the computed Krylov subspace.
- Threshold: 10^{-1} (correspondence to a significant loss of orthogonality)

Comparison of approximations

We can compare FP and exact CG approximations $\bar{x}_k$ and $x_{\bar{k}}$ themselves, both $\bar{x}_k, x_{\bar{k}} \in \mathbb{F}^N$.

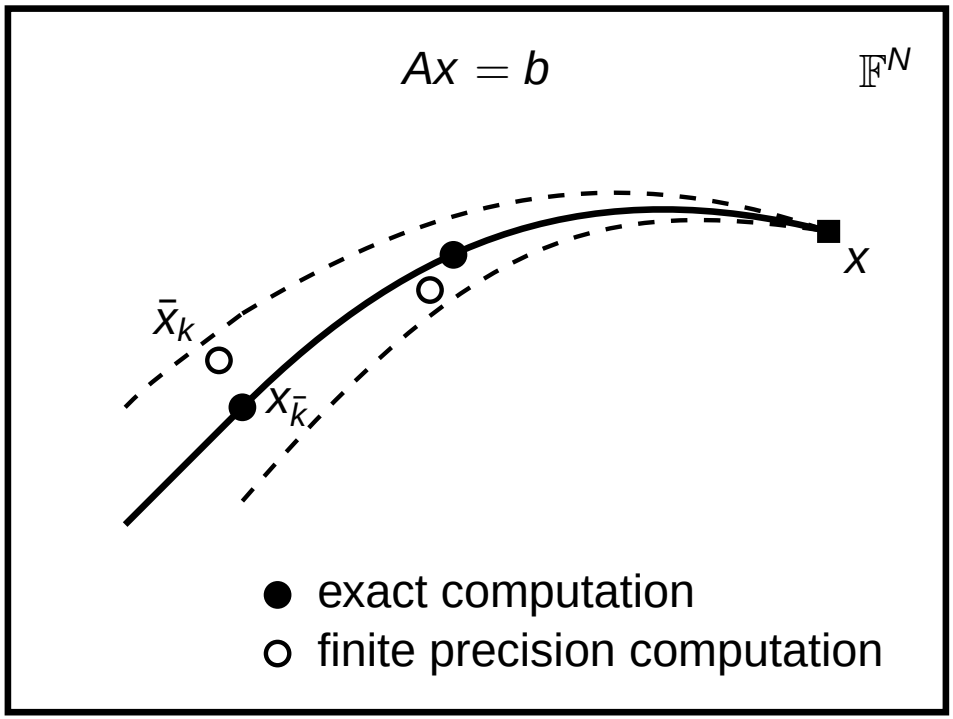


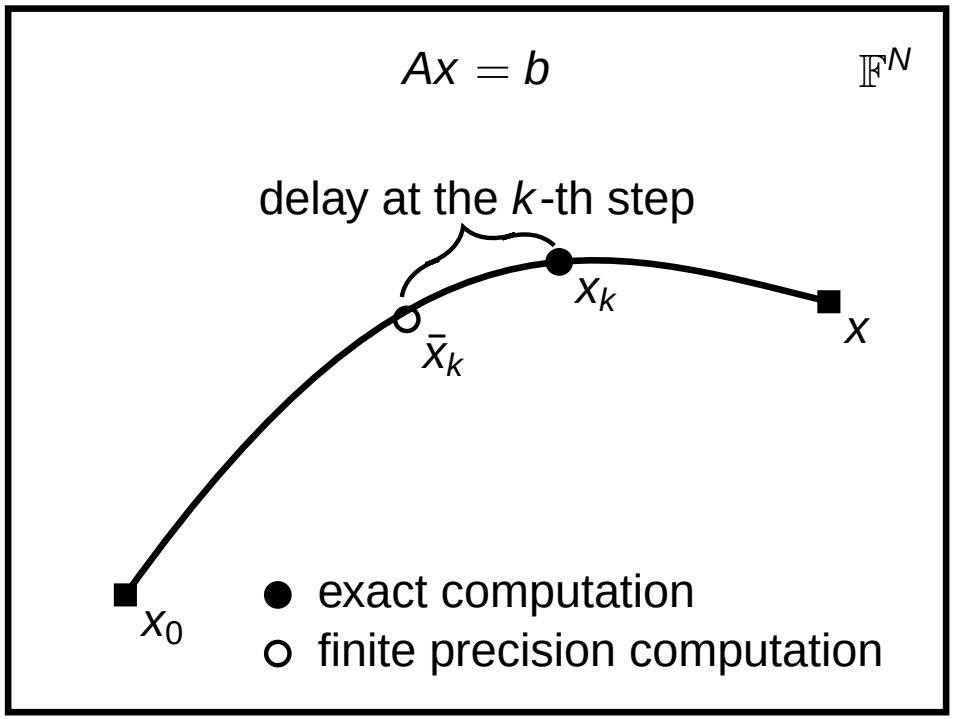
Observation

$$\frac{\|\bar{x}_k - x_{\bar{k}}\|_\infty}{\|x - x_{\bar{k}}\|_\infty} \ll 1,$$

i.e., distance between approximations is small in comparison with the actual level of error.

Comparison of approximations





Finite precision CG computation tightly follows the trajectory of exact CG computations with the delay which is given by the rank-deficiency of the computed Krylov subspace.

Contents

Comparison of numerical ranks

- $\overline{\mathcal{K}}_k(A, r_0)$: k-th Krylov subspace generated by FP CG
- $\mathcal{K}_{\bar{k}}(A, r_0)$: $\bar{k}$-th Krylov subspace generated by exact CG
  ($\mathrm{rank}(\mathcal{K}_{\bar{k}}(A, r_0)) = \bar{k}$)

k	56	80	100	196	362	528	611	664
$\bar{k} = \mathrm{rank}(\overline{\mathcal{K}}_k(A, r_0))$	56	73	80	93	112	126	131	132
$\mathrm{rank}(\overline{\mathcal{K}}_k(A, r_0) \cup \mathcal{K}_{\bar{k}}(A, r_0))$	56	74	80	93	113	127	131	132

Bcsstk04 (from `MatrixMarket`)

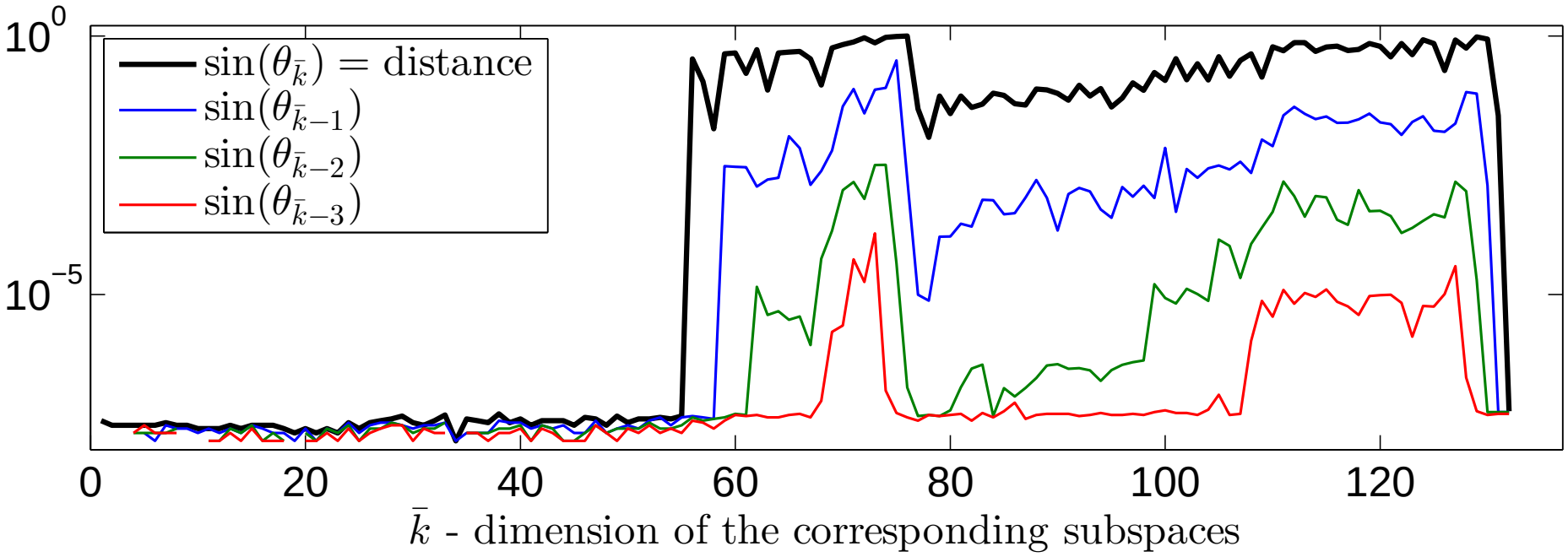
Distance between subspaces

- Principal angles between $\bar{k}$-dimensional subspace $\mathcal{K}_{\bar{k}}(A, r_0)$ and the $\bar{k}$-dimensional restriction of the subspace $\overline{\mathcal{K}}_k(A, r_0)$ which corresponds to its numerical rank:

$$0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_{\bar{k}} \leq \pi/2$$

- Distance:

$$\text{distance}\left(\mathcal{K}_{\bar{k}}(A, r_0),\ \text{restricted}\left(\overline{\mathcal{K}}_k(A, r_0)\right)\right) = \sin(\theta_{\bar{k}})$$



Concluding remarks and future work

- The rate of convergence typically substantially differs for FP and exact CG. However, the trajectories of FP and exact CG approximations seem to be very close to each other.
- Inertia of computed Krylov subspaces represents phenomenon which needs to be studied.

Thank you for your kind attention

Acknowledgement

This work has been supported by the ERC-CZ project LL1202, by the GACR grant 201/13-06684S and by the GAUK grant 695612.