# Numerical behavior of iterative methods

Miro Rozložník
joint results with Zhong-zhi Bai and Pavel Jiránek

Institute of Mathematics, Czech Academy of Sciences,
Prague, Czech Republic

Seminarium Pierścienie, macierze i algorytmy numeryczne,
Politechnika Warszawska, Warszawa, April 3, 2017

# Iterative methods in exact arithmetic

generate a sequence of approximate solutions $x_0, x_1, \dots, x_n \to x$ to the solution of $Ax = b$ with residual vectors $r_0 = b - Ax_0, \dots, r_n = b - Ax_n \to 0$

**Iterative methods in finite precision arithmetic**

compute approximations $x_0, \hat{x}_1, \dots, \hat{x}_n$ and updated residual vectors $\hat{r}_0, \hat{r}_1, \dots, \hat{r}_n$ which are usually close to (but different from) the true residuals $b - A\hat{x}_n$

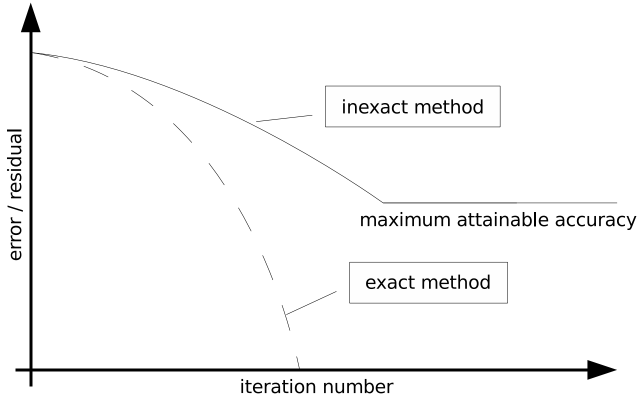
## Two main questions and two main effects

- How good is the computed approximate solution $\hat{x}_n$? How many (extra) steps do we need to reach the same accuracy as one can get in the exact method?
- How well the computed vector $\hat{r}_n$ approximates the (true) residual $b - A\hat{x}_n$? Is there a limitation on the accuracy of the computed approximate solution?

**Two effects of rounding errors:**

- **Delay of convergence**
- **Maximum attainable accuracy**

# Delay of convergence and maximum attainable accuracy



# Stationary iterative methods

- $\mathcal{A}x = b$, $\mathcal{A} = \mathcal{M} - \mathcal{N}$, $\mathcal{G} = \mathcal{M}^{-1}\mathcal{N}$, $\mathcal{F} = \mathcal{N}\mathcal{M}^{-1}$
- A: $\mathcal{M}x_{k+1} = \mathcal{N}x_k + b$

  B: $x_{k+1} = x_k + \mathcal{M}^{-1}(b - \mathcal{A}x_k)$
- Inexact solution of systems with $\mathcal{M}$: **every computed solution $\hat{y}$ of $\mathcal{M}y = z$ is interpreted as an exact solution of a system with perturbed data and relative perturbation bounded by parameter** $\tau$ such that

$$(\mathcal{M} + \Delta\mathcal{M})\hat{y} = z, \quad \|\Delta\mathcal{M}\| \leq \tau\|\mathcal{M}\|, \quad \tau k(\mathcal{M}) \ll 1$$

- Higham, Knight 1993: $\mathcal{M}$ triangular, $\tau = O(u)$

# Accuracy of the computed approximate solution

A: $\mathcal{M}x_{k+1} = \mathcal{N}x_k + b$

$$\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \leq \tau \frac{\|\mathcal{M}^{-1}\|\|(\|\mathcal{M}\| + \|\mathcal{N}\|)}{1 - \|\mathcal{G}\|} \frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}$$

$$\frac{\|b - \mathcal{A}\hat{x}_{k+1}\|}{\|b\| + \|\mathcal{A}\|\|\hat{x}_{k+1}\|} \leq \tau \frac{\|\mathcal{M}\|}{\|\mathcal{A}\|} \frac{\|I - \mathcal{F}\|}{1 - \|\mathcal{F}\|} \frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}$$

B: $x_{k+1} = x_k + \mathcal{M}^{-1}(b - \mathcal{A}x_k)$

$$\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \leq O(u) \frac{\|\mathcal{M}^{-1}\| \, (\|\mathcal{M}\| + \|\mathcal{N}\|)}{1 - \|\mathcal{G}\| - 2\tau\|\mathcal{M}^{-1}\|\|\mathcal{M}\|} \frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}$$

$$\frac{\|b - \mathcal{A}\hat{x}_{k+1}\|}{\|b\| + \|\mathcal{A}\|\|\hat{x}_{k+1}\|} \leq O(u) \frac{\|\mathcal{M}\| + \|\mathcal{N}\|}{\|\mathcal{A}\|} \frac{\|I - \mathcal{F}\|}{1 - \|\mathcal{F}\| - 2\tau\|\mathcal{M}^{-1}\|\|\mathcal{M}\|} \frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}$$

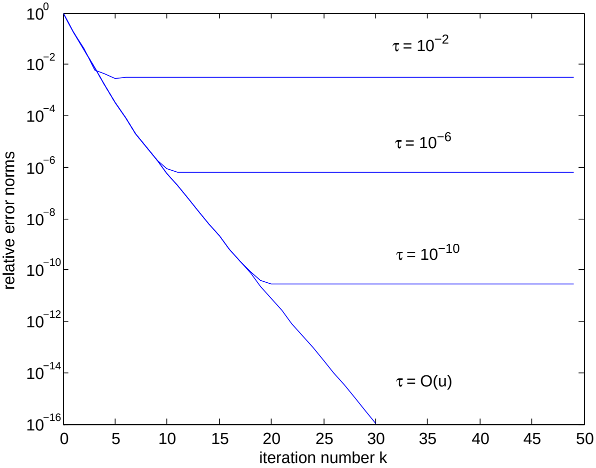
Higham, Knight 1993, Bai, R, 2012

# Numerical experiments: small model example

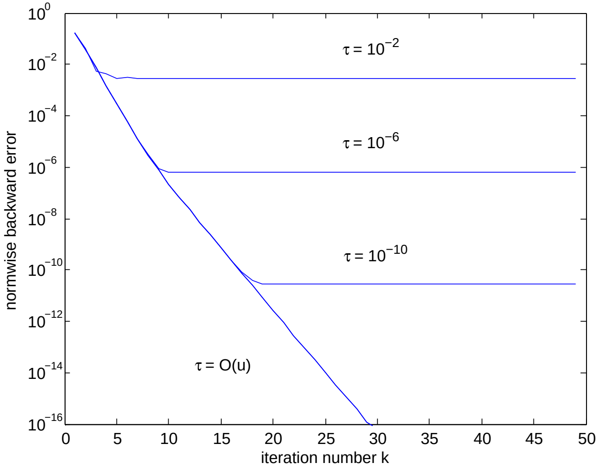
$$\mathcal{A} = \mathrm{tridiag}(1, 4, 1) \in \mathbb{R}^{100\times 100},\ b = \mathcal{A}\cdot \mathrm{ones}(100, 1),$$

$$\kappa(A) = \|A\|\cdot\|A^{-1}\| = 5.9990\cdot 0.4998 \approx 2.9983$$

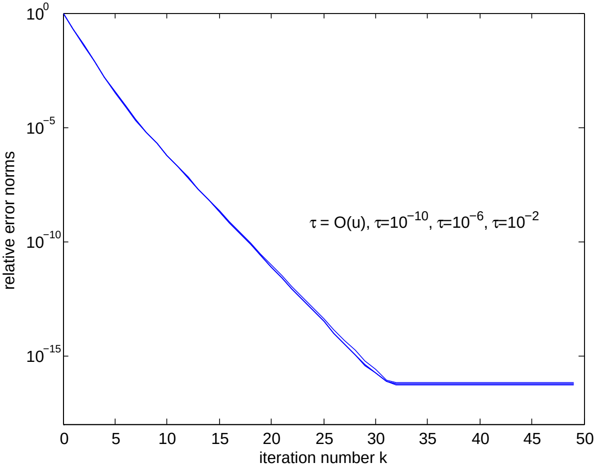
$$\mathcal{A} = \mathcal{M} - \mathcal{N},\ \mathcal{M} = D - L,\ \mathcal{N} = U$$



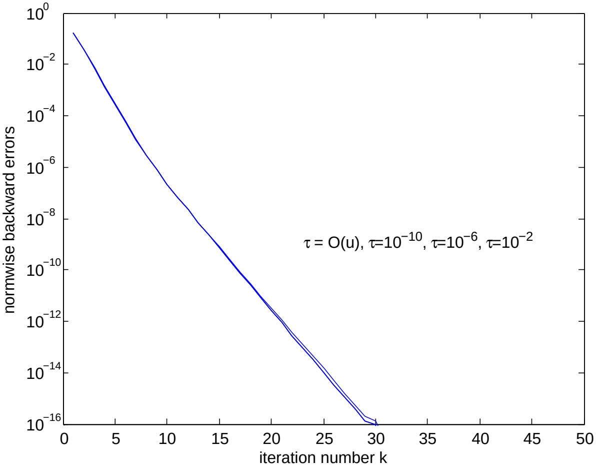
$\tau = 10^{-2}$
$\tau = 10^{-6}$
$\tau = 10^{-10}$
$\tau = O(u)$
relative error norms
iteration number k
$10^{0}$
$10^{-2}$
$10^{-4}$
$10^{-6}$
$10^{-8}$
$10^{-10}$
$10^{-12}$
$10^{-14}$
$10^{-16}$
0
5
10
15
20
25
30
35
40
45
50



normwise backward error
iteration number k
τ = 10^−2
τ = 10^−6
τ = 10^−10
τ = O(u)



$10^0$
$10^{-5}$
$10^{-10}$
$10^{-15}$
relative error norms
$\tau = O(u)$, $\tau=10^{-10}$, $\tau=10^{-6}$, $\tau=10^{-2}$
0
5
10
15
20
25
30
35
40
45
50
iteration number k


normwise backward errors
iteration number k
τ = O(u), τ=10^−10, τ=10^−6, τ=10^−2
10^0
10^−2
10^−4
10^−6
10^−8
10^−10
10^−12
10^−14
10^−16
0
5
10
15
20
25
30
35
40
45
50

# Two-step splitting iteration methods

$$\mathcal{M}_1 x_{k+1/2} = \mathcal{N}_1 x_k + b, \qquad \mathcal{A} = \mathcal{M}_1 - \mathcal{N}_1$$
$$\mathcal{M}_2 x_{k+1} = \mathcal{N}_2 x_{k+1/2} + b, \quad \mathcal{A} = \mathcal{M}_2 - \mathcal{N}_2$$

Numerous solution schemes: Hermitian/skew-Hermitian (HSS) splitting, modified Hermitian/skew-Hermitian (MHSS) splitting, normal Hermitian/skew-Hermitian (NSS) splitting, preconditioned variant of modified Hermitian/skew-Hermitian (PMHSS) splitting and other splittings, ...

Bai, Golub, Ng 2003, 2007, 2008; Bai 2009
Bai, Benzi, Chen 2010, 2011; Bai, Benzi, Chen, Wang 2012

$$\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \lessapprox \left[\tau_1 \|\mathcal{M}_2^{-1}\mathcal{N}_2\| \|\mathcal{M}_1^{-1}\| (\|\mathcal{M}_1\| + \|\mathcal{N}_1\|) + \tau_2 \|\mathcal{M}_2^{-1}\| (\|\mathcal{M}_2\| + \|\mathcal{N}_2\|)\right] \frac{\max_{i=0,1/2,\dots,k+1/2}\{\|\hat{x}_i\|\}}{\|x\|}$$

# Two-step splitting iteration methods

$$\begin{aligned} x_{k+1/2} &= x_k + \mathcal{M}_1^{-1}(b - \mathcal{A}x_k) \\ x_{k+1} &= x_{k+1/2} + \mathcal{M}_2^{-1}(b - \mathcal{A}x_{k+1/2}) \end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned} x_{k+1} &= x_k + (\mathcal{M}_1^{-1} + \mathcal{M}_2^{-1} - \mathcal{M}_2^{-1}\mathcal{A}\mathcal{M}_1^{-1})(b - \mathcal{A}x_k) \\ &= x_k + (\mathcal{I} + \mathcal{M}_2^{-1}\mathcal{N}_1)\mathcal{M}_1^{-1}(b - \mathcal{A}x_k) \\ &= x_k + \mathcal{M}_2^{-1}(\mathcal{I} + \mathcal{N}_2\mathcal{M}_1^{-1})(b - \mathcal{A}x_k) \end{aligned}$$

$$\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \lessapprox O(u)\|\mathcal{M}_2^{-1}(\mathcal{I} + \mathcal{N}_2\mathcal{M}_1^{-1})\|(\|\mathcal{M}\| + \|\mathcal{N}\|)\frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}$$

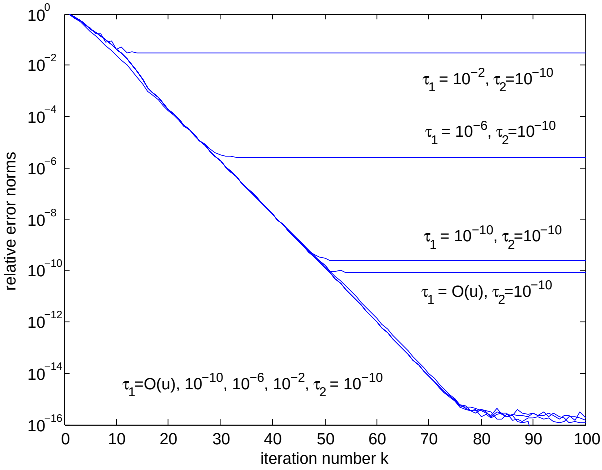
# Numerical experiments: small model example

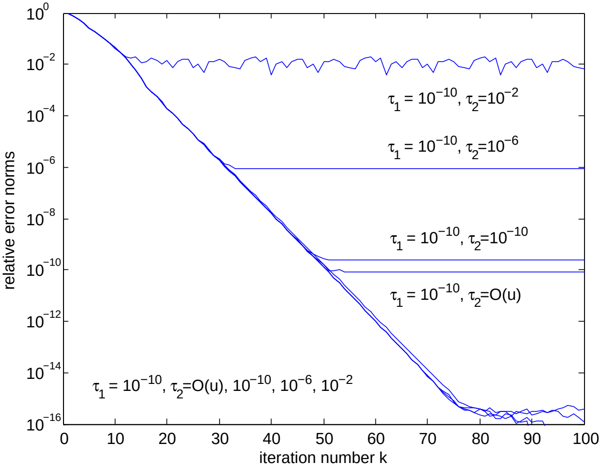
$$\mathcal{A} = \text{tridiag}(2, 4, 1) \in \mathbb{R}^{100\times 100},\ b = \mathcal{A} \cdot \text{ones}(100, 1),$$

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| = 5.9990 \cdot 0.4998 \approx 2.9983$$

$$\mathcal{A} = \mathcal{H} + \mathcal{S}, \quad \mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T), \quad \mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$$

$$\mathcal{H} = \text{tridiag}(\frac{3}{2}, 4, \frac{3}{2}),\ \mathcal{S} = \text{tridiag}(\frac{1}{2}, 0, -\frac{1}{2})$$


$10^{0}$
$10^{-2}$
$10^{-4}$
$10^{-6}$
$10^{-8}$
$10^{-10}$
$10^{-12}$
$10^{-14}$
$10^{-16}$
relative error norms
0
10
20
30
40
50
60
70
80
90
100
iteration number k
$\tau_1 = 10^{-2}$, $\tau_2=10^{-10}$
$\tau_1 = 10^{-6}$, $\tau_2=10^{-10}$
$\tau_1 = 10^{-10}$, $\tau_2=10^{-10}$
$\tau_1 = O(u)$, $\tau_2=10^{-10}$
$\tau_1=O(u)$, $10^{-10}$, $10^{-6}$, $10^{-2}$, $\tau_2 = 10^{-10}$



$10^{0}$
$10^{-2}$
$10^{-4}$
$10^{-6}$
$10^{-8}$
$10^{-10}$
$10^{-12}$
$10^{-14}$
$10^{-16}$
relative error norms
0 10 20 30 40 50 60 70 80 90 100
iteration number k
$\tau_1 = 10^{-10}$, $\tau_2=10^{-2}$
$\tau_1 = 10^{-10}$, $\tau_2=10^{-6}$
$\tau_1 = 10^{-10}$, $\tau_2=10^{-10}$
$\tau_1 = 10^{-10}$, $\tau_2=O(u)$
$\tau_1 = 10^{-10}$, $\tau_2=O(u)$, $10^{-10}$, $10^{-6}$, $10^{-2}$

# Saddle point problems

We consider a saddle point problem with the symmetric $2 \times 2$ block form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

- $A$ is a square $n \times n$ nonsingular (symmetric positive definite) matrix,
- $B$ is a rectangular $n \times m$ matrix of (full column) rank $m$.

# Schur complement reduction method

- Compute $y$ as a solution of the Schur complement system

$$B^T A^{-1} B y = B^T A^{-1} f,$$

- compute $x$ as a solution of

$$Ax = f - By.$$

- Segregated vs. coupled approach: $x_k$ and $y_k$ approximate solutions to $x$ and $y$, respectively.
- Inexact solution of systems with $A$: **every computed solution $\hat{u}$ of $Au = b$ is interpreted as an exact solution of a perturbed system**

$$(A + \Delta A)\hat{u} = b + \Delta b, \; \|\Delta A\| \leq \tau \|A\|, \; \|\Delta b\| \leq \tau \|b\|, \; \tau \kappa(A) \ll 1.$$

# Iterative solution of the Schur complement system

choose $y_0$, solve $Ax_0 = f - By_0$

compute $\alpha_k$ and $p_k^{(y)}$

$y_{k+1} = y_k + \alpha_k p_k^{(y)}$

$$\left.\begin{array}{|l}
\text{solve } Ap_k^{(x)} = -Bp_k^{(y)} \\
\text{back-substitution:} \\
\textbf{A: } x_{k+1} = x_k + \alpha_k p_k^{(x)}, \\
\textbf{B: } \text{solve } Ax_{k+1} = f - By_{k+1}, \\
\textbf{C: } \text{solve } Au_k = f - Ax_k - By_{k+1}, \\
\qquad x_{k+1} = x_k + u_k.
\end{array}\right\} \text{inner iteration}$$

$r_{k+1}^{(y)} = r_k^{(y)} - \alpha_k B^T p_k^{(x)}$

outer iteration

# Accuracy in the saddle point system

$$\|f - A\hat{x}_k - B\hat{y}_k\| \leq \frac{O(\alpha_1)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\hat{Y}_k),$$

$$\|-B^T\hat{x}_k - \hat{r}_k^{(y)}\| \leq \frac{O(\alpha_2)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\hat{Y}_k),$$

$$\hat{Y}_k \equiv \max\{\|\hat{y}_i\| \mid i = 0, 1, \ldots, k\}.$$

| Back-substitution scheme | $\alpha_1$ | $\alpha_2$ |
|---|---|---|
| **A**: Generic update $x_{k+1} = x_k + \alpha_k p_k^{(x)}$ | $\tau$ | $u$ |
| **B**: Direct substitution $x_{k+1} = A^{-1}(f - By_{k+1})$ | $\tau$ | $\tau$ |
| **C**: Corrected dir. subst. $x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1})$ | $u$ | $\tau$ |

additional system with A

$$-B^T A^{-1} f + B^T A^{-1} B\hat{y}_k = -B^T\hat{x}_k - B^T A^{-1}(f - A\hat{x}_k - B\hat{y}_k)$$

# Numerical experiments: a small model example

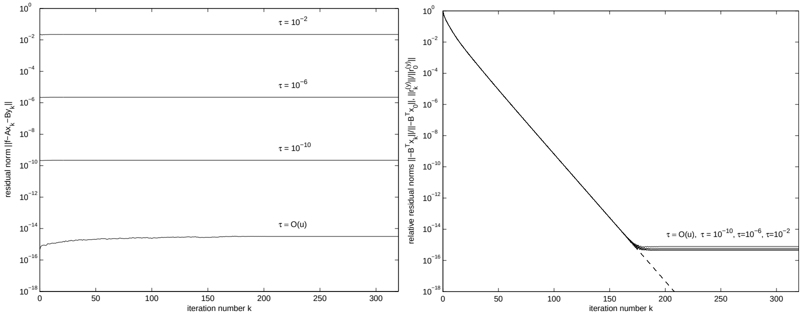
$$A = \mathrm{tridiag}(1, 4, 1) \in \mathbb{R}^{100\times 100},\ B = \mathrm{rand}(100, 20),\ f = \mathrm{rand}(100, 1),$$

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| = 5.9990 \cdot 0.4998 \approx 2.9983,$$

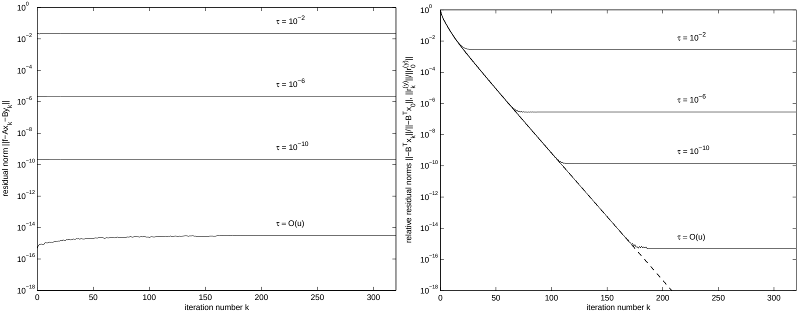
$$\kappa(B) = \|B\| \cdot \|B^{\dagger}\| = 7.1695 \cdot 0.4603 \approx 3.3001.$$

[R, Simoncini, 2002]

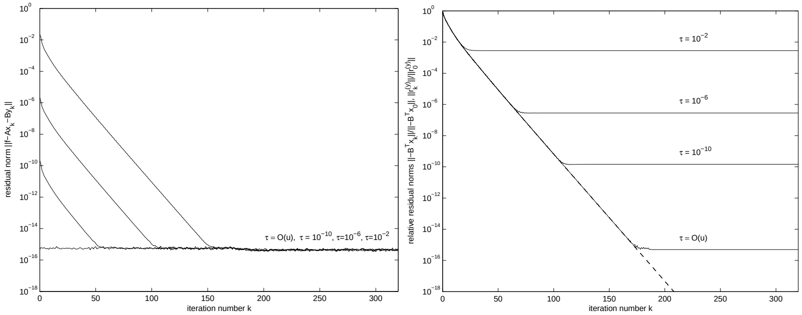
## Generic update: $x_{k+1} = x_k + \alpha_k p_k^{(x)}$



# Direct substitution: $x_{k+1} = A^{-1}(f - By_{k+1})$



# Corrected direct substitution: $x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1})$



# Conclusions

**"new_value = old_value + small_correction"**

- Fixed-precision iterative refinement for improving the computed solution $x_{\rm old}$ to a system $Ax = b$: solving update equations $Az_{\rm corr} = r$ that have residual $r = b - Ay_{\rm old}$ as a right-hand side to obtain $x_{\rm new} = x_{\rm old} + z_{\rm corr}$, see [Wilkinson, 1963], [Higham, 2002].
- Stationary iterative methods for $Ax = b$ and their maximum attainable accuracy [Higham and Knight, 1993]: assuming splitting $A = M - N$ and inexact solution of systems with $M$, use $x_{\rm new} = x_{\rm old} + M^{-1}(b - Ax_{\rm old})$ rather than $x_{\rm new} = M^{-1}(Nx_{\rm old} + b)$, [Higham, 2002; Bai, R].
- Two-step splitting iteration framework: $A = M_1 - N_1 = M_2 - N_2$ assuming inexact solution of systems with $M_1$ and $M_2$, reformulation of $M_1 x_{1/2} = N_1 x_{\rm old} + b$, $M_2 x_{\rm new} = N_2 x_{1/2} + b$, Hermitian/skew-Hermitian splitting (HSS) iteration [Bai, Golub and Ng 2003; Bai, R].
- Saddle point problems and inexact linear solvers: Schur complement and null-space approach [Jiránek, R 2008]

# Thank you for your attention.

`http://www.math.cas.cz/rozloznik`


Zhong-Zhi Bai and M. Rozložník, On the behavior of two-step splitting iteration methods, *SIAM J. Numer. Analysis, 53(4) (2015), pp. 1716–1737.*

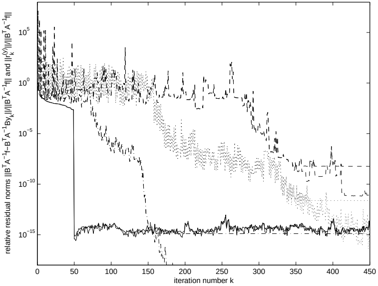
P. Jiránek and M. Rozložník. Maximum attainable accuracy of inexact saddle point solvers. *SIAM J. Matrix Anal. Appl.*, 29(4):1297–1321, 2008.

P. Jiránek and M. Rozložník. Limiting accuracy of segregated solution methods for nonsymmetric saddle point problems. *J. Comput. Appl. Math.* 215 (2008), pp. 28-37.

M. Rozložník and V. Simoncini, Krylov subspace methods for saddle point problems with indefinite preconditioning, *SIAM J. Matrix Anal. Appl., 24 (2002), pp. 368–391.*

# The maximum attainable accuracy of saddle point solvers

- The accuracy measured by the residuals of the saddle point problem depends on the choice of the back-substitution scheme [Jiránek, R, 2008]. The schemes with (generic or corrected substitution) updates deliver approximate solutions which satisfy either the first or second block equation to working accuracy.
- Care must be taken when solving nonsymmetric systems [Jiránek, R, 2008], all bounds of the limiting accuracy depend on the maximum norm of computed iterates, cf. [Greenbaum 1994,1997], [Sleijpen, et al. 1994].



## Null-space projection method

- compute $x \in N(B^T)$ as a solution of the projected system

$$(I - \Pi)A(I - \Pi)x = (I - \Pi)f,$$

- compute $y$ as a solution of the least squares problem

$$By \approx f - Ax,$$

  $\Pi = B(B^TB)^{-1}B^T$ is the orthogonal projector onto $R(B)$.

- Schemes with the inexact solution of least squares with $B$. Every computed approximate solution $\hat{v}$ of a least squares problem $Bv \approx c$ is interpreted as an exact solution of a perturbed least squares

$$(B + \Delta B)\hat{v} \approx c + \Delta c, \ \|\Delta B\| \leq \tau\|B\|, \ \|\Delta c\| \leq \tau\|c\|, \ \tau\kappa(B) \ll 1.$$

# Null-space projection method

$$
\left.\begin{array}{l}
\text{choose } x_0,\ \text{solve } By_0 \approx f - Ax_0\\[4pt]
\text{compute } \alpha_k \text{ and } p_k^{(x)} \in N(B^T)\\[4pt]
x_{k+1} = x_k + \alpha_k p_k^{(x)}\\[4pt]
\quad \left|\begin{array}{l}
\text{solve } Bp_k^{(y)} \approx r_k^{(x)} - \alpha_k A p_k^{(x)}\\[4pt]
\text{back-substitution:}\\[4pt]
\textbf{A: } y_{k+1} = y_k + p_k^{(y)},\\[4pt]
\textbf{B: } \text{solve } By_{k+1} \approx f - Ax_{k+1},\\[4pt]
\textbf{C: } \text{solve } Bv_k \approx f - Ax_{k+1} - By_k,\\[4pt]
\qquad y_{k+1} = y_k + v_k.
\end{array}\right\} \begin{array}{l}\text{inner}\\ \text{iteration}\end{array}\\[4pt]
r_{k+1}^{(x)} = r_k^{(x)} - \alpha_k A p_k^{(x)} - Bp_k^{(y)}
\end{array}\right\} \begin{array}{l}\text{outer}\\ \text{iteration}\end{array}
$$

# Accuracy in the saddle point system

$$\|f - A\hat{x}_k - B\hat{y}_k - \hat{r}_k^{(x)}\| \leq \frac{O(\alpha_3)\kappa(B)}{1-\tau\kappa(B)}(\|f\| + \|A\|\hat{X}_k),$$

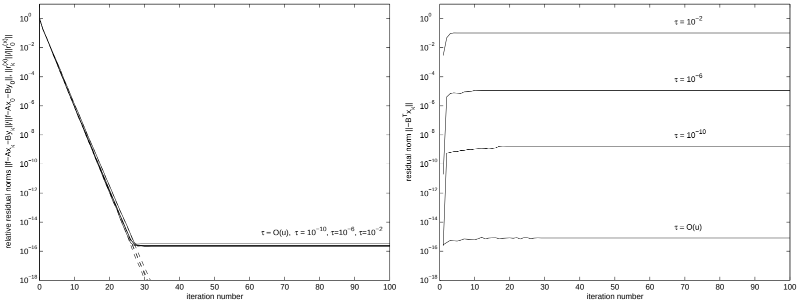
$$\|-B^T\hat{x}_k\| \leq \frac{O(\tau)\kappa(B)}{1-\tau\kappa(B)}\|B\|\hat{X}_k,$$

$$\hat{X}_k \equiv \max\{\|\hat{x}_i\| \mid i = 0, 1, \ldots, k\}.$$

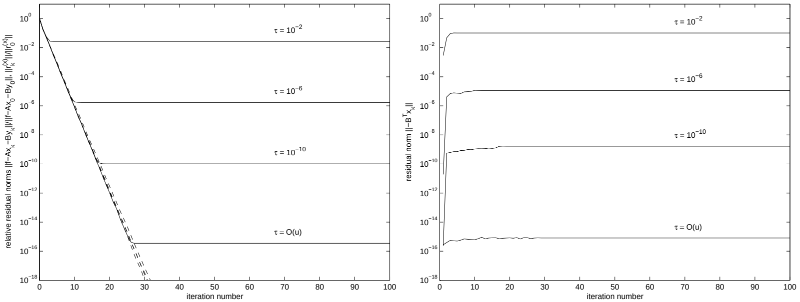
| Back-substitution scheme | $\alpha_3$ |
|---|---|
| **A**: Generic update $y_{k+1} = y_k + p_k^{(y)}$ | $u$ |
| **B**: Direct substitution $y_{k+1} = B^\dagger(f - Ax_{k+1})$ | $\tau$ |
| **C**: Corrected dir. subst. $y_{k+1} = y_k + B^\dagger(f - Ax_{k+1} - By_k)$ | $u$ |

additional least square with B

# Generic update: $y_{k+1} = y_k + p_k^{(y)}$



# Direct substitution: $y_{k+1} = B^{\dagger}(f - Ax_{k+1})$



# Corrected direct substitution: $y_{k+1} = y_k + B^{\dagger}(f - Ax_{k+1} - By_k)$