Numerical behavior of inexact linear solvers

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

Institute of Computer Science, Czech Academy of Sciences,
Prague, Czech Republic

The fourth international conference on numerical algebra and scientific computing (NASC 2012), Dalian University of Technology, Dalian, China, October 20-24, 2012

Iterative solution of linear systems

Iterative methods in exact arithmetic

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

Iterative methods in finite precision arithmetic

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

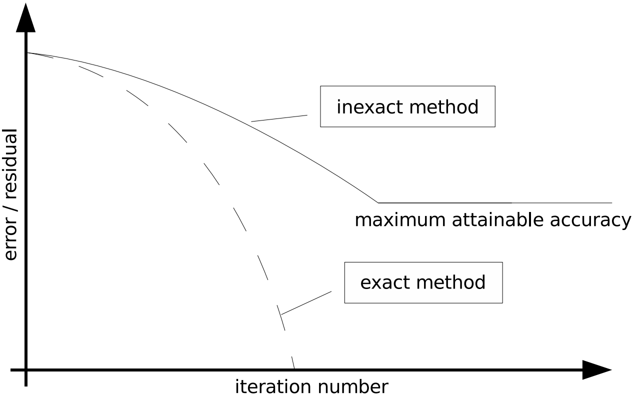
Two main questions and two main effects

- How good is the computed approximate solution $\bar{x}_n$? 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 $\bar{r}_n$ approximates the (true) residual $b - A\bar{x}_n$? Is there a limitation on the accuracy of the computed approximate solution?

Two main effects of rounding errors:

DELAY OF CONVERGENCE AND LIMIT ON THE FINAL (MAXIMUM ATTAINABLE) ACCURACY

Delay of convergence and maximum attainable accuracy



The concept of backward stability

A backward stable algorithm eventually computes the exact answer to a nearby problem, i.e. the approximate solution $\bar{x}_n$ satisfying

$$(A + \Delta A_n)\bar{x}_n = b + \Delta b_n$$
$$\|\Delta A_n\|/\|A\| \leq O(u), \ \|\Delta b_n\|/\|b\| \leq O(u)$$

$\iff$ The normwise backward error associated with the approximate solution $\bar{x}_n$ satisfies

$$\frac{\|b - A\bar{x}_n\|}{\|b\| + \|A\|\|\bar{x}_n\|} \leq O(u)$$

A forward stable algorithm eventually computes the approximate solution $\bar{x}_n$ with the error that satisfies

$$\|\bar{x}_n - x\| \leq O(u)\|A^{-1}\|\|A\|\|x\|$$

Prager, Oettli, 1964; Rigal, Gaches, 1967

see also Higham, 2nd ed. 2002; Stewart, Sun, 1990; Meurant 1999

The level of maximum attainable accuracy

We are looking for the difference between the updated $\bar{r}_n$ and true residual $b - A\bar{x}_n$ (divided by $\|b\| + \|A\|\|\bar{x}_n\|$)

$$\frac{\|b - A\bar{x}_n - \bar{r}_n\|}{\|b\| + \|A\|\|\bar{x}_n\|} \leq ?$$

$$\|\bar{r}_n\| \longrightarrow 0 \Longrightarrow \lim_{n\to\infty} \frac{\|b - A\bar{x}_n\|}{\|b\| + \|A\|\|\bar{x}_n\|} \leq ?$$

In the optimal case the bound is of $O(u)$; then we have a backward stable solution. The backward stability implies the forward stability.

Higham 2002, Higham, Knight 1993, Greenbaum, R, Strakoš, 1997

Stationary iterative methods

- $\mathcal{A}x = b$, $\mathcal{A} = \mathcal{M} - \mathcal{N}$
- 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 $\overline{y}$ of $\mathcal{M}y = z$ is interpreted as an exact solution of a system with perturbed data and relative perturbation bounded by parameter** τ such that

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

Accuracy of the computed approximate solution

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

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

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

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

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

$$\frac{\|b - \mathcal{A}\hat{x}_{k+1}\|}{\|b\| + \|\mathcal{A}\|\|\hat{x}_{k+1}\|} \leq O(u) \left(1 + \frac{\max_{i=0,\dots,k}\{\|\hat{x}_i\|\}}{\|x\|}\right)$$

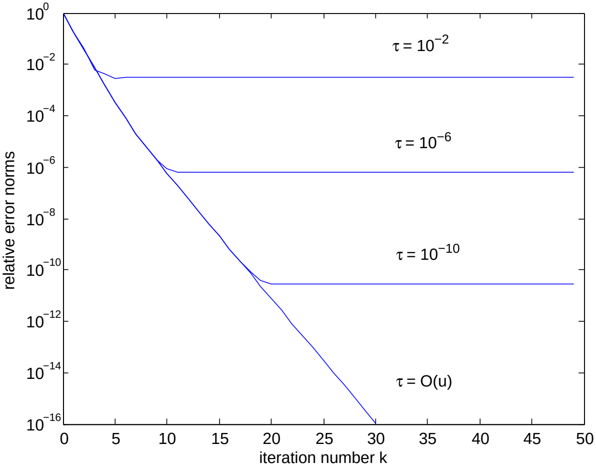
Higham, Knight 1993, Bai, R, 2012

Numerical experiments: small model example

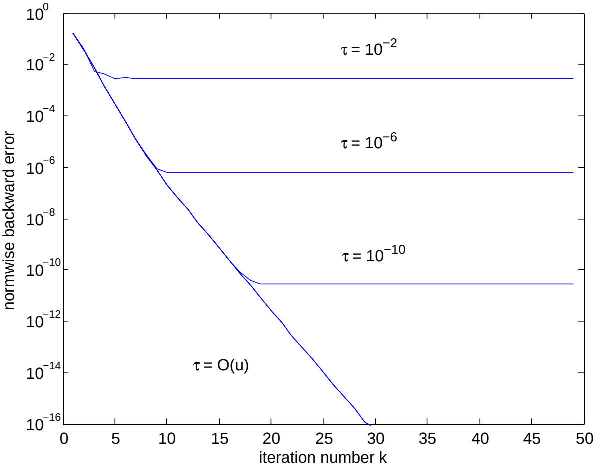
$$\mathcal{A} = \mathrm{tridiag}(1, 4, 1) \in \mathbb{R}^{100\times 100},\ b = \mathcal{A}\cdot \mathrm{ones}(100, 1),$$

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

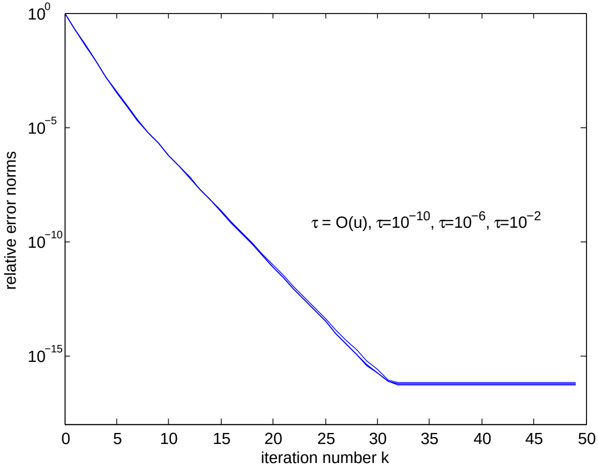
$$\mathcal{A} = \mathcal{M} - \mathcal{N},\ \mathcal{M} = D - L,\ \mathcal{N} = U$$



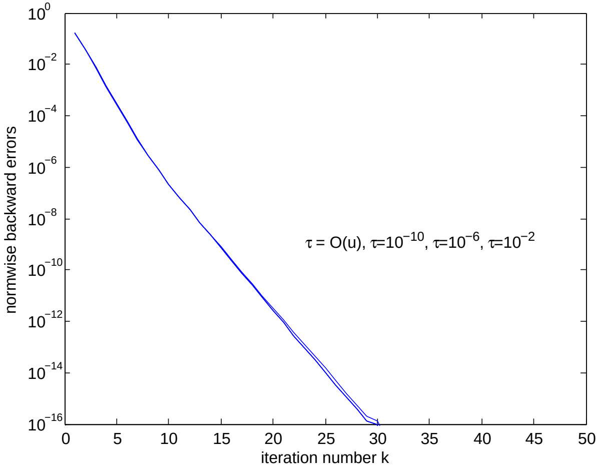
relative error norms
iteration number k
$\tau = 10^{-2}$
$\tau = 10^{-6}$
$\tau = 10^{-10}$
$\tau = O(u)$



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



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

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

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

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

Two-step splitting iteration methods

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

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

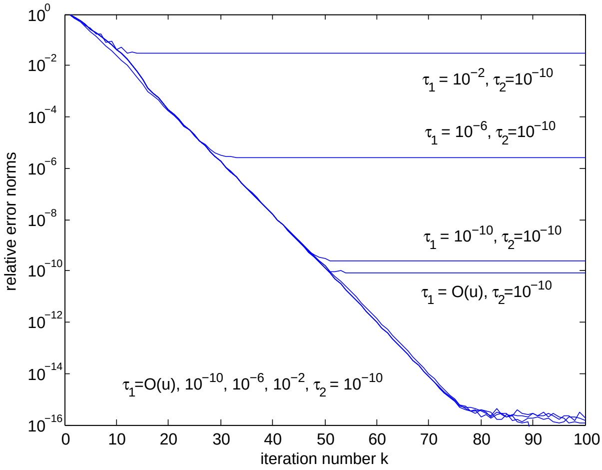
Numerical experiments: small model example

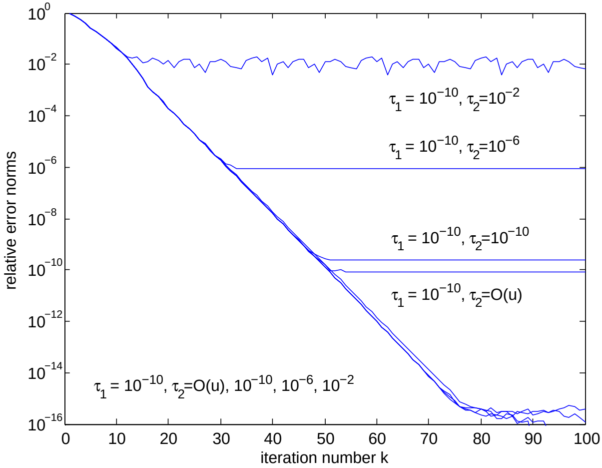
$$\mathcal{A} = \mathrm{tridiag}(2,4,1) \in \mathbb{R}^{100\times 100},\ b = \mathcal{A}\cdot \mathrm{ones}(100,1),$$

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

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

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


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 2×2 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.

Applications: mixed finite element approximations, weighted least squares, constrained optimization, computational fluid dynamics, electromagnetism etc. [Benzi, Golub and Liesen, 2005], [Elman, Silvester, Wathen, 2005]. For the updated list of applications leading to saddle point problems contact [Benzi].

Iterative solution of saddle point problems

Numerous solution schemes: inexact Uzawa algorithms, inexact null-space methods, inner-outer iteration methods, two-stage iteration processes, multilevel or multigrid methods, domain decomposition methods

Numerous preconditioning techniques and schemes: block diagonal preconditioners, block triangular preconditioners, constraint preconditioning, Hermitian/skew-Hermitian preconditioning and other splittings, combination preconditioning

Numerous iterative solvers: conjugate gradient (CG) method, MINRES, GMRES, flexible GMRES, GCR, BiCG, BiCGSTAB, ...

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 α_k and $p_k^{(y)}$

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

solve $Ap_k^{(x)} = -Bp_k^{(y)}$

back-substitution:

A: $x_{k+1} = x_k + \alpha_k p_k^{(x)}$,

B: solve $Ax_{k+1} = f - By_{k+1}$,

C: solve $Au_k = f - Ax_k - By_{k+1}$,

$x_{k+1} = x_k + u_k$.

inner iteration

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

outer iteration

Accuracy in the saddle point system

$$\|f - Ax_k - By_k\| \leq \frac{O(\alpha_1)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|Y_k),$$

$$\| - B^T x_k - r_k^{(y)}\| \leq \frac{O(\alpha_2)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|Y_k),$$

$$Y_k \equiv \max\{\|y_i\| \mid i = 0, 1, \ldots, k\}.$$

Back-substitution scheme	α_1	α_2
A: Generic update $x_{k+1} = x_k + \alpha_k p_k^{(x)}$	τ	u
B: Direct substitution $x_{k+1} = A^{-1}(f - By_{k+1})$	τ	τ
C: Corrected dir. subst. $x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1})$	u	τ

additional system with A

$$-B^T A^{-1} f + B^T A^{-1} B y_k = -B^T x_k - B^T A^{-1}(f - Ax_k - By_k)$$

Numerical experiments: a small model example

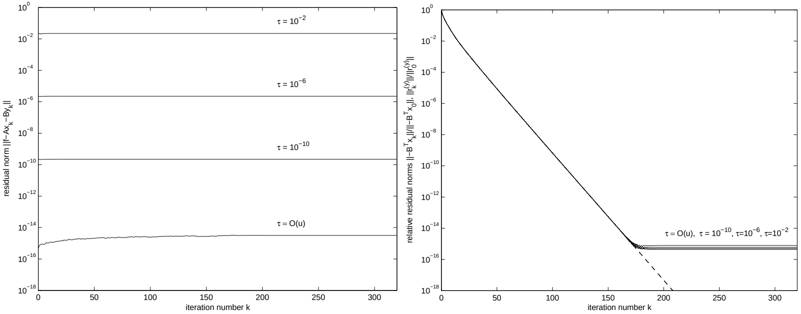
$$A = \mathrm{tridiag}(1,4,1) \in \mathbb{R}^{100\times 100},\ B = \mathrm{rand}(100,20),\ f = \mathrm{rand}(100,1),$$

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

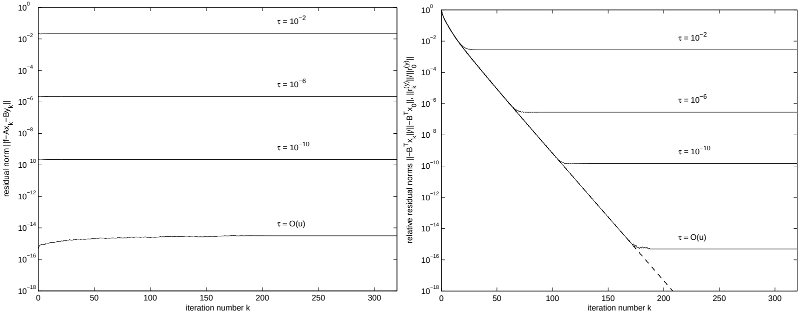
$$\kappa(B) = \|B\| \cdot \|B^{\dagger}\| = 7.1695 \cdot 0.4603 \approx 3.3001.$$

[R, Simoncini, 2002]

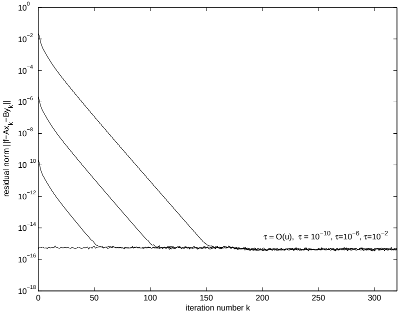
Generic update: $x_{k+1} = x_k + \alpha_k p_k^{(x)}$

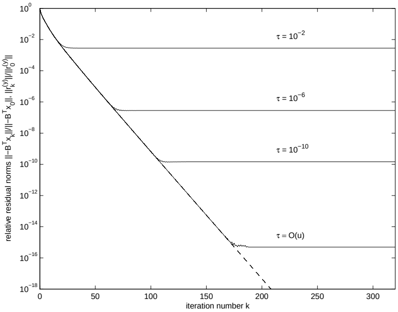


Direct substitution: $x_{k+1} = A^{-1}(f - By_{k+1})$



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





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

$$By \approx f - Ax,$$

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

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

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

Null-space projection method

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

Accuracy in the saddle point system

$$\|f - Ax_k - By_k - r_k^{(x)}\| \leq \frac{O(\alpha_3)\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|X_k),$$

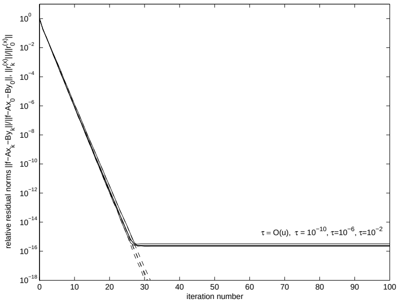
$$\|-B^T x_k\| \leq \frac{O(\tau)\kappa(B)}{1 - \tau\kappa(B)}\|B\|X_k,$$

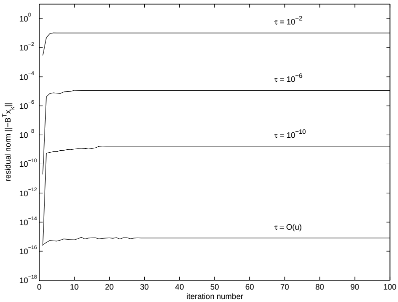
$$X_k \equiv \max\{\|x_i\| \,|\, i = 0, 1, \dots, k\}.$$

Back-substitution scheme		α_3
A:	Generic update $y_{k+1} = y_k + p_k^{(y)}$	u
B:	Direct substitution $y_{k+1} = B^\dagger(f - Ax_{k+1})$	τ
C:	Corrected dir. subst. $y_{k+1} = y_k + B^\dagger(f - Ax_{k+1} - By_k)$	u

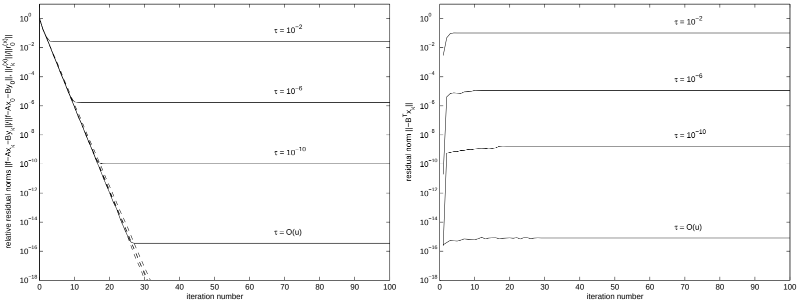
additional least square with B

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

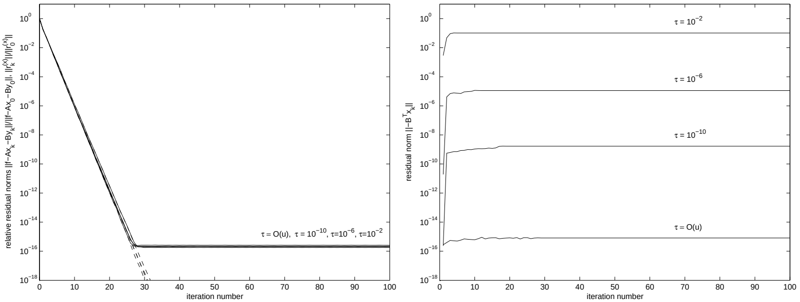




Direct substitution: $y_{k+1} = B^{\dagger}(f - Ax_{k+1})$

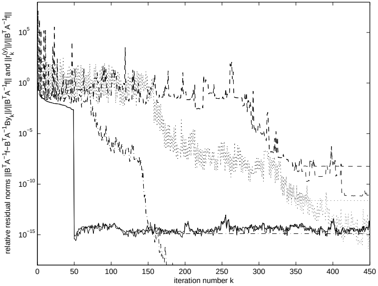


Corrected direct substitution: $y_{k+1} = y_k + B^\dagger(f - Ax_{k+1} - By_k)$



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].



Conclusions

"new_value = old_value + small_correction"

- 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 $x_{\mathrm{new}} = x_{\mathrm{old}} + M^{-1}(b - Ax_{\mathrm{old}})$ rather than $x_{\mathrm{new}} = M^{-1}(Nx_{\mathrm{old}} + b)$, [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 $M_1 x_{1/2} = N_1 x_{\mathrm{old}} + b$, $M_2 x_{\mathrm{new}} = N_2 x_{1/2} + b$, 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] Preconditioners for saddle point problems: SIMPLE and SIMPLE(R) type algorithms [Vuik and Saghir, 2002] and constraint preconditioners [R, Simoncini, 2002].
- Fixed-precision iterative refinement for improving the computed solution x_{old} to a system $Ax = b$: solving update equations $Az_{\mathrm{corr}} = r$ that have residual $r = b - Ay_{\mathrm{old}}$ as a right-hand side to obtain $x_{\mathrm{new}} = x_{\mathrm{old}} + z_{\mathrm{corr}}$, see [Wilkinson, 1963], [Higham, 2002].

Thank you for your attention.

`http://www.cs.cas.cz/~miro`

Zhong-Zhi Bai and M. Rozložník, On the behavior of two-step splitting iteration methods, *in preparation*.

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.*