Numerical behavior of matrix splitting iteration methods

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

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

MatTriad’2013 - Conference on Matrix Analysis and its Applications,
September 16–20, 2013, Herceg-Novi, Montenegro
Delay of convergence and maximum attainable accuracy
Stationary iterative methods

- $Ax = b$, $A = M - N$, $G = M^{-1}N$, $F = NN^{-1}$

- $A$: $Mx_{k+1} = Nx_k + b$
  $B$: $x_{k+1} = x_k + M^{-1}(b - Ax_k)$

- Inexact solution of systems with $M$: every computed solution $\bar{y}$ of $My = z$ is interpreted as an exact solution of a system with perturbed data and relative perturbation bounded by parameter $\tau$ such that

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

- Higham, Knight 1993: $M$ triangular, $\tau = O(u)$
Accuracy of the computed approximate solution

A: \[ Mx_{k+1} = Nx_k + b \]
\[
\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \leq \tau \frac{\|M^{-1}\| (\|M\| + \|N\|)}{1 - \|G\|} \max_{i=0, \ldots, k} \{\|\hat{x}_i\|\}
\]
\[
\frac{\|b - A\hat{x}_{k+1}\|}{\|b\| + \|A\|\|\hat{x}_{k+1}\|} \leq \tau \frac{\|M\| \|I - F\|}{\|A\|} \max_{i=0, \ldots, k} \{\|\hat{x}_i\|\}
\]

B: \[ x_{k+1} = x_k + M^{-1}(b - Ax_k) \]
\[
\frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \leq O(u) \frac{\|M^{-1}\| (\|M\| + \|N\|)}{1 - \|G\| - 2\tau\|M^{-1}\||\|M\||} \max_{i=0, \ldots, k} \{\|\hat{x}_i\|\}
\]
\[
\frac{\|b - A\hat{x}_{k+1}\|}{\|b\| + \|A\|\|\hat{x}_{k+1}\|} \leq O(u) \frac{\|M\| + \|N\|}{\|A\|} \frac{\|I - F\|}{1 - \|F\| - 2\tau\|M^{-1}\||\|M\||} \max_{i=0, \ldots, k} \{\|\hat{x}_i\|\}
\]

Higham, Knight 1993, Bai, R, 2012
Numerical experiments: small model example

\[ A = \text{tridiag}(1, 4, 1) \in \mathbb{R}^{100 \times 100}, \quad b = A \cdot \text{ones}(100, 1), \]
\[ \kappa(A) = \|A\| \cdot \|A^{-1}\| = 5.9990 \cdot 0.4998 \approx 2.9983 \]

\[ A = \mathcal{M} - \mathcal{N}, \quad \mathcal{M} = D - L, \quad \mathcal{N} = U \]
iteration number $k$

normwise backward error

$\tau = 10^{-2}$

$\tau = 10^{-6}$

$\tau = 10^{-10}$

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

\[ M_1 x_{k+1/2} = N_1 x_k + b, \quad A = M_1 - N_1 \]
\[ M_2 x_{k+1} = N_2 x_{k+1/2} + b, \quad A = M_2 - N_2 \]

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

\[ \frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \lesssim \left[ \tau_1 \|M_2^{-1}N_2\| \|M_1^{-1}\| (\|M_1\| + \|N_1\|) + \tau_2 \|M_2^{-1}\| (\|M_2\| + \|N_2\|) \right] \]
\[ \max_{i=0,1/2,...,k+1/2} \left\{ \|\hat{x}_i\| \right\} \]
\[ \frac{\|x\|}{\|x\|} \]
Two-step splitting iteration methods

\[ x_{k+1/2} = x_k + M_1^{-1} (b - Ax_k) \]
\[ x_{k+1} = x_{k+1/2} + M_2^{-1} (b - Ax_{k+1/2}) \]
\[ \Leftrightarrow \]
\[ x_{k+1} = x_k + (M_1^{-1} + M_2^{-1} - M_2^{-1} AM_1^{-1}) (b - Ax_k) \]
\[ = x_k + (I + M_2^{-1} N_1) M_1^{-1} (b - Ax_k) \]
\[ = x_k + M_2^{-1} (I + N_2 M_1^{-1}) (b - Ax_k) \]

\[ \frac{\|\hat{x}_{k+1} - x\|}{\|x\|} \lesssim O(u) \|M_2^{-1} (I + N_2 M_1^{-1})\| (\|M\| + \|N\|) \max_{i=0,\ldots,k} \{\|\hat{x}_i\|\} \]
Numerical experiments: small model example

\[ A = \text{tridiag}(2, 4, 1) \in \mathbb{R}^{100 \times 100}, \ b = A \cdot \text{ones}(100, 1), \]
\[ \kappa(A) = \|A\| \cdot \|A^{-1}\| = 5.9990 \cdot 0.4998 \approx 2.9983 \]

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

\[ H = \text{tridiag}\left(\frac{3}{2}, 4, \frac{3}{2}\right), \quad S = \text{tridiag}\left(\frac{1}{2}, 0, -\frac{1}{2}\right) \]
\[ \tau_1 = \mathcal{O}(u), \tau_2 = 10^{-10} \]
\[ \tau_1 = 10^{-6}, \tau_2 = 10^{-10} \]
\[ \tau_1 = 10^{-10}, \tau_2 = 10^{-10} \]
\[ \tau_1 = 10^{-10}, \tau_2 = 10^{-10} \]
relative error norms

iteration number $k$

$\tau_1 = 10^{-10}, \tau_2 = O(u)$

$\tau_1 = 10^{-10}, \tau_2 = 10^{-2}$

$\tau_1 = 10^{-10}, \tau_2 = 10^{-6}$

$\tau_1 = 10^{-10}, \tau_2 = 10^{-10}$
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
  \[
  B^T A^{-1} By = B^T A^{-1} f,
  \]
- 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
  \[
  (A + \Delta A)\hat{u} = b + \Delta b, \quad \|\Delta A\| \leq \tau \|A\|, \quad \|\Delta b\| \leq \tau \|b\|, \quad \tau \kappa(A) \ll 1.
  \]
Iterative solution of the Schur complement system

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

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

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

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

inner iteration

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

outer iteration

$$r_k^{(y)} = r_k^{(y)} - \alpha_k B^T p_k^{(x)}.$$
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

<table>
<thead>
<tr>
<th></th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A:</strong> Generic update</td>
<td>(\tau)</td>
<td>(u)</td>
</tr>
<tr>
<td>(x_{k+1} = x_k + \alpha_k p_k(x))</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B:</strong> Direct substitution</td>
<td>(\tau)</td>
<td>(\tau)</td>
</tr>
<tr>
<td>(x_{k+1} = A^{-1}(f - By_{k+1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C:</strong> Corrected dir. subst.</td>
<td>(u)</td>
<td>(\tau)</td>
</tr>
<tr>
<td>(x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1}))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
-B^T A^{-1} f + B^T A^{-1} By_k = -B^T x_k - B^T A^{-1}(f - Ax_k - By_k)
\]
Numerical experiments: a small model example

\[ \text{\( A = \text{tridiag}(1, 4, 1) \in \mathbb{R}^{100\times100}, \ B = \text{rand}(100, 20), \ f = \text{rand}(100, 1), \)} \]

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

\[ \kappa(B) = \|B\| \cdot \|B^\dagger\| = 7.1695 \cdot 0.4603 \approx 3.3001. \]

[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: \( x_{k+1} = x_k + A^{-1}(f - A x_k - B y_{k+1}) \)
Conclusions

"new_value = old_value + small_correction"

- Fixed-precision iterative refinement for improving the computed solution $x_{\text{old}}$ to a system $Ax = b$: solving update equations $Az_{\text{corr}} = r$ that have residual $r = b - Ay_{\text{old}}$ as a right-hand side to obtain $x_{\text{new}} = x_{\text{old}} + z_{\text{corr}}$, 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 $x_{\text{new}} = x_{\text{old}} + M^{-1}(b - Ax_{\text{old}})$ rather than $x_{\text{new}} = M^{-1}(Nx_{\text{old}} + b)$, [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 $M_1 x_{1/2} = N_1 x_{\text{old}} + b$, $M_2 x_{\text{new}} = N_2 x_{1/2} + b$, 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.cs.cas.cz/~miro


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.

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
  \[(B + \Delta B)\bar{v} \approx c + \Delta c, \quad \|\Delta B\| \leq \tau\|B\|, \quad \|\Delta c\| \leq \tau\|c\|, \quad \tau\kappa(B) \ll 1.\]
Null-space projection method

choose \( x_0 \), solve \( By_0 \approx f - Ax_0 \)

compute \( \alpha_k \) and \( p_k^{(x)} \in N(B^T) \)

\[
x_{k+1} = x_k + \alpha_k p_k^{(x)}
\]

solve \( Bp_k^{(y)} \approx r_k^{(x)} - \alpha_k A p_k^{(x)} \)

back-substitution:

A: \( y_{k+1} = y_k + p_k^{(y)} \),

B: solve \( By_{k+1} \approx f - Ax_{k+1} \),

C: solve \( Bv_k \approx f - Ax_{k+1} - By_k \),

\[
y_{k+1} = y_k + v_k.
\]

\[
r_{k+1}^{(x)} = r_k^{(x)} - \alpha_k A p_k^{(x)} - B p_k^{(y)}
\]
Accuracy in the saddle point system

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

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

\[ X_k \equiv \max \{ \| x_i \| \mid i = 0, 1, \ldots, k \}. \]

<table>
<thead>
<tr>
<th>Back-substitution scheme</th>
<th>( \alpha_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A:</strong> Generic update</td>
<td>( u )</td>
</tr>
<tr>
<td>( y_{k+1} = y_k + p^{(y)}_k )</td>
<td></td>
</tr>
<tr>
<td><strong>B:</strong> Direct substitution</td>
<td>( \tau )</td>
</tr>
<tr>
<td>( y_{k+1} = B^\dagger (f - Ax_{k+1}) )</td>
<td></td>
</tr>
<tr>
<td><strong>C:</strong> Corrected dir. subst.</td>
<td>( u )</td>
</tr>
<tr>
<td>( y_{k+1} = y_k + B^\dagger (f - Ax_{k+1} - By_k) )</td>
<td></td>
</tr>
</tbody>
</table>

\{ \text{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: \( y_{k+1} = y_k + B^\dagger(f - Ax_{k+1} - By_k) \)