Numerical behavior of saddle point solvers

Miro Rozložník
joint results with Pavel Jiránek and Valeria Simoncini

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

The International Workshop Numerical Analysis and Scientific Computing (NASCom'08), Rostov-on-Don, Russian federation, October 13-17, 2008

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 etc. [?].

Inexact saddle point solvers

Solution algorithms for saddle point problems:

1. **the segregated or coupled approach**: outer iteration for solving the reduced system;
2. **the inexact solution of inner systems:** inner iteration loop with appropriate stopping criterion;
3. **the rounding errors**: finite precision arithmetic.

Numerous schemes: inexact Uzawa algorithms, inexact null-space methods, inner-outer iteration methods, two-stage iteration processes, multilevel or multigrid methods, domain decomposition methods
[**?**], [**?**], [**?**], [**?**],...

Delay of convergence and limit on the final accuracy

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

$$Ax = f - By.$$

- inexact solution of systems with A: **every computed solution $\hat{u}$ of $Au = b$ is interpreted 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

$$\left.\begin{array}{l}
\text{choose } y_0,\ \text{solve } Ax_0 = f - By_0 \\
\text{compute } \alpha_k \text{ and } p_k^{(y)} \\
y_{k+1} = y_k + \alpha_k p_k^{(y)} \\
\quad \left|\begin{array}{l}
\text{solve } Ap_k^{(x)} = -Bp_k^{(y)} \\
\text{back-substitution:} \\
\textbf{A: } x_{k+1} = x_k + \alpha_k p_k^{(x)}, \\
\textbf{B: } \text{solve } Ax_{k+1} = f - By_{k+1}, \\
\textbf{C: } \text{solve } Au_k = f - Ax_k - By_{k+1}, \\
\qquad x_{k+1} = x_k + u_k.
\end{array}\right\} \begin{array}{l}\text{inner}\\ \text{iteration}\end{array} \\
r_{k+1}^{(y)} = r_k^{(y)} - \alpha_k B^T p_k^{(x)}
\end{array}\right\} \begin{array}{l}\text{outer}\\ \text{iteration}\end{array}$$

Maximum attainable accuracy of inexact Schur complement schemes

The limiting (maximum attainable) accuracy is measured by the ultimate (asymptotic) values of:

1. **the Schur complement residual**: $B^T A^{-1} f - B^T A^{-1} B y_k$;
2. **the residuals in the saddle point system**: $f - Ax_k - By_k$ and $-B^T x_k$;
3. **the forward errors**: $x - x_k$ and $y - y_k$.

Numerical experiments: a small model example

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

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| = 7.1695 \cdot 0.4603 \approx 3.3001,$$

$$\kappa(B) = \|B\| \cdot \|B^{\dagger}\| = 5.9990 \cdot 0.4998 \approx 2.9983.$$

Accuracy in the outer iteration process

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

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

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

$$\|B^T A^{-1} f - B^T A^{-1} B \hat{y}\| \leq \frac{\tau\kappa(A)}{1 - \tau\kappa(A)} \|A^{-1}\| \|B\|^2 \|\hat{y}\|.$$

Accuracy in the saddle point system

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

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

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

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

$\Big\}$ additional system with A

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

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

$$By \approx f - Ax,$$

Π is the orthogonal projector onto $R(B)$.

Analogous results for schemes, where the least squares with B are solved inexactly. 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

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

Accuracy in the saddle point system

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

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

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

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

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

Null-space projection method: saddle point problem and indefinite constraint preconditioner

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

$$\mathcal{P} = \begin{pmatrix} I & B \\ B^T & 0 \end{pmatrix}$$

PCG method starting with particular right-hand side or initial guess:
$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, r_0 = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$

R, Simoncini, 2002, Gould, Keller, Wathen 2000, Lukšan, Vlček, 1998

Saddle point problems: symmetric indefinite system and symmetric indefinite preconditioner

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \quad \mathcal{P} = \begin{pmatrix} I & B \\ B^T & 0 \end{pmatrix}$$

$$\mathcal{A}\mathcal{P}^{-1} = \begin{pmatrix} A(I-\Pi)+\Pi & (A-I)B(B^TB)^{-1} \\ 0 & I \end{pmatrix}$$

$\Pi = B(B^TB)^{-1}B^T$ - orth. projector onto $span(B)$

Indefinite preconditioner: spectral properties of preconditioned system

$\mathcal{A}\mathcal{P}^{-1}$ **nonsymmetric** and **non-diagonalizable**!

but it has a 'nice' spectrum:

$$\begin{aligned}\sigma(\mathcal{A}\mathcal{P}^{-1}) &\subset \{1\} \cup \sigma(A(I-\Pi)+\Pi) \\ &\subset \{1\} \cup \sigma((I-\Pi)A(I-\Pi)) - \{0\}\end{aligned}$$

Perugia, Simoncini 1999, Lukšan, Vlček 1998, Gould, Wathen, Keller, 1999

and only 2 by 2 Jordan blocks!

Preconditioned conjugate gradients (PCG)

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, r_0 = b - \mathcal{A} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix}, k = 0, 1, \ldots$$

$$e_{k+1} = \begin{pmatrix} x - x_{k+1} \\ y - y_{k+1} \end{pmatrix}$$

$$r_{k+1} = \begin{pmatrix} f \\ 0 \end{pmatrix} - \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix}$$

Basic properties of the constraint preconditioner

$$r_0 = \begin{pmatrix} s_0 \\ 0 \end{pmatrix} \Rightarrow r_{k+1} = \begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix}$$

$$\Rightarrow B^T(x - x_{k+1}) = 0$$

$$\Rightarrow x_{k+1} \in Null(B^T)!$$

The (energy)-norm of the error in PCG

$$r_{k+1}^T \mathcal{P}^{-1} r_j = 0,\ j = 0, \dots, k$$

x_{k+1} is an iterate from CG applied to

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

$$\|x - x_{k+1}\|_A = \min_{u \in x_0 + span\{(I-\Pi) s_j\}} \|x - u\|_A$$

Lukšan, Vlček 1998, Gould, Wathen, Keller, 1999

The residual norm in PCG - continued

$$x - x_{k+1} = \phi_{k+1}((I-\Pi)A(I-\Pi))(x-x_0)$$

$$x_{k+1} \to x$$

$$\|r_{k+1}\| = \|\phi_{k+1}(A(I-\Pi)+\Pi)s_0\|$$

$$\sigma((I-\Pi)A(I-\Pi)) \subset \sigma(A(I-\Pi)+\Pi)$$

$$\{1\} \in \sigma((I-\Pi)\alpha A(I-\Pi)) - \{0\}$$

$$\Rightarrow \|r_{k+1}\| = \left\| \begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix} \right\| \to 0!$$

Scaling by a constant $\alpha > 0$ such that

$$\{1\} \in conv(\sigma((I - \Pi)\alpha A(I - \Pi)) - \{0\})$$

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

$$v : \quad \|(I - \Pi)v\| \neq 0$$

$$\alpha = \frac{1}{((I - \Pi)v, A(I - \Pi)v)}!$$

Scaling by a diagonal?

$$A \rightarrow (diag(A))^{-1/2} A (diag(A))^{-1/2}$$

often gives what we want!

Different direction vector?

$$y_{k+1} = y_k + (B^T B)^{-1} B^T s_k$$

so that $\|r_{k+1}\| = \|s_{k+1}\|$ is locally minimized!

Braess, Deuflhard,Lipikov 1999, Hribar, Gould, Nocedal, 1999

Numerical example

$$A = tridiag(1,4,1) \in \mathsf{R}^{25,25},\ B = rand(25,5) \in \mathsf{R}^{25,5}$$
$$f = rand(25,1) \in \mathsf{R}^{25}$$

$$\sigma(A) \subset [2.0146, 5.9854]$$

α	$\sigma(\begin{pmatrix} \alpha A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} I & B \\ B^T & 0 \end{pmatrix}^{-1})$
$1/100$	$[0.0207, 0.0586]\ \{1\}$
$1/10$	$[0.2067, 0.5856]\ \{1\}$
$\mathbf{1/4}$	$[\mathbf{0.5170}, \mathbf{1.4641}]$
1	$\{1\}\ [2.0678, 5.8563]$
4	$\{1\}\ [8.2712, 23.4252]$

Behavior of PCG in finite precision arithmetic

$$\bar{t}_{k+1} = \begin{pmatrix} \bar{x}_{k+1} \\ \bar{y}_{k+1} \end{pmatrix}, \quad \bar{r}_{k+1} = \begin{pmatrix} \bar{s}^{(1)}_{k+1} \\ \bar{s}^{(2)}_{k+1} \end{pmatrix}$$

$$\|x - \bar{x}_{k+1}\|_A \leq \gamma_1 \|\Pi(x - \bar{x}_{k+1})\| + \gamma_2 \|(I - \Pi)A(I - \Pi)(x - \bar{x}_{k+1})\| \qquad (1)$$

Exact arithmetic:

$$\|\Pi(x - \bar{x}_{k+1})\| = 0$$
$$\|(I - \Pi)A(I - \Pi)(x - \bar{x}_{k+1})\| \to 0$$

Forward error of computed approximate solution: finite precision arithmetic

$$\|x-\bar{x}_{k+1}\|_A \leq \gamma_3\|B^T(x-\bar{x}_{k+1})\| + \gamma_2\|(I-\Pi)(f-A\bar{x}_{k+1}-B\bar{y}_{k+1})\|$$

Departure from the null-space of B^T
+ projection of the residual onto it

$$B^T(x-\bar{x}_{k+1}) \sim \bar{s}_{k+1}^{(2)}$$

$$(I-\Pi)(f-A\bar{x}_{k+1}-B\bar{y}_{k+1}) \sim (I-\Pi)\bar{s}_{k+1}^{(1)}$$

Behavior of PCG in finite precision arithmetic

$$
\begin{gathered}
\|(f - A\bar{x}_{k+1} - B\bar{y}_{k+1}) - \bar{s}^{(1)}_{k+1}\|, \\
\|B^T(x - \bar{x}_{k+1}) - \bar{s}^{(2)}_{k+1}\| \leq \\
\leq \| \begin{pmatrix} f \\ 0 \end{pmatrix} - \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \bar{x}_{k+1} \\ \bar{y}_{k+1} \end{pmatrix} - \begin{pmatrix} \bar{s}^{(1)}_{k+1} \\ \bar{s}^{(2)}_{k+1} \end{pmatrix} \| \\
\leq c_1 \varepsilon \kappa(\mathcal{A}) \kappa(\mathcal{P}) \max_{j=0,\ldots,k+1} \|\bar{r}_j\|
\end{gathered}
$$

Greenbaum 1994,1997, Sleijpen, et al. 1994

good scaling: $\|\bar{r}_j\| \rightarrow 0$ nearly monotonically

$$\|\bar{r}_0\| \sim \max_{j=0,\ldots,k+1} \|\bar{r}_j\|$$

Conclusions

- The accuracy measured by the residuals of the saddle point problem depends on the choice of the back-substitution scheme [?]. 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 [?], all bounds of the limiting accuracy depend on the maximum norm of computed iterates, the choice of the method and scaling is very important cf. [?], [?].

Related results in the context of saddle-point problems and Krylov subspace methods

- General framework of inexact Krylov subspace methods: in exact arithmetic the effects of relaxation in matrix-vector multiplication on the ultimate accuracy of several solvers [?], [?].
- The effects of rounding errors in the Schur complement reduction (block LU decomposition) method and the null-space method [?], [?], the maximum attainable accuracy studied in terms of the user tolerance specified in the outer iteration [?], [?].
- Error analysis in computing the projections into the null-space and constraint preconditioning, limiting accuracy of the preconditioned CG , residual update strategy when solving constrained quadratic programming problems [?], or in cascadic multigrid method for elliptic problems [?].
- Theory for a general class of iterative methods based on coupled two-term recursions, all bounds of the limiting accuracy depend on the maximum norm of computed iterates, fixed matrix-vector multiplication, cf. [?].

General comments and considerations, future work

"new_value = old_value + small_correction"

- Fixed-precision iterative refinement for improving the computed solution x_{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 [?], [?].
- Stationary iterative methods for $Ax = b$ and their maximum attainable accuracy [?]: 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 $x_{\text{new}} = M^{-1}(Nx_{\text{old}} + b)$, [?].
- 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 [?].
- Inexact preconditioners for saddle point problems: SIMPLE and SIMPLE(R) type algorithms [?] and constraint preconditioners [?].

Thank you for your attention.

`http://www.cs.cas.cz/∼miro`

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.*

References I