Relaxed Mixed Constraint Preconditioners for Ill-conditioned Symmetric Saddle Point Linear Systems
Solution of indefinite systems, TU Eindhoven – April 17, 2012

Luca Bergamaschi
Department of Civil Environmental and Architectural Engineering
Ángeles Martínez
Department of Mathematics
University of Padua, via Trieste 63, 35121 Padova, Italy
e-mail luca.bergamaschi@unipd.it, angeles.martinez@unipd.it

April 17, 2012
Outline

- Mathematical model and discretization of coupled consolidation problems → Saddle point type systems of linear equations
- Inexact Block Preconditioners – Spectral Analysis of the preconditioned matrices
- Implementation: Mixed Constraint Preconditioners
- Numerical results on eigenvalue bounds
- Relaxed Mixed Constraint Preconditioners – Spectral Analysis
- Numerical results on large realistic problems
- Parallel RMCP based on FSAI.
- Another model problem: MFE discretization of Darcy’s law in porous media
- Conclusions and Future Perspectives
Mathematical Problem

- System of PDE for the 3D coupled consolidation process in fully saturated porous media derived from the classical Biot’s formulation

\[(\psi + \mu)\nabla \varepsilon + \mu \nabla^2 u = \alpha \nabla p\]

\[\frac{1}{\gamma} \nabla \cdot (k \nabla p) = [\phi \beta + c_{br}(\alpha - \phi)] \frac{\partial p}{\partial t} + \alpha \frac{\partial \varepsilon}{\partial t}\]

- **unknowns**: \(p\) incremental pore pressure and \(u_x, u_y\) and \(u_z\) the displacements.
- \(c_{br}\) and \(\beta\) volumetric compressibility of solid grains and water, resp.
- \(\phi\) porosity, \(k\) hydraulic conductivity, \(\varepsilon\) volumetric dilatation, \(\alpha\) Biot coefficient
- \(\psi\) the Lamé constant, \(\mu\) the shear modulus of the porous medium
- \(\gamma\) is the specific weight of water
Finite Element discretization

- Linear Galerkin FE in space yields a system of 1st order ODE
- Integration in time by the Crank-Nicolson scheme
- Resulting linear system to be solved at each timestep with coefficient matrix

\[ T = \begin{bmatrix}
\frac{A}{2} & -\frac{Q}{2} \\
\frac{Q^T}{\Delta t} & \frac{H}{2} + \frac{P}{\Delta t}
\end{bmatrix} \]

- \( A \) and \( H \) are elastic and flow stiffness, both SPD, matrices.
- \( P \) and \( Q \) flow capacity and flow-stress coupling matrices.
- Symmetrization of \( T \) yields an equivalent Saddle Point type system

\[ A x = b, \quad \text{where} \quad A = \begin{bmatrix}
A & B^T \\
B & -C(t)
\end{bmatrix} \quad B = -Q^T, \quad C(t) = \Delta t H/2 + P. \]

- \( A \) is \( n \times n \), \( C(t) \) \( m \times m \), \( B \) is a rectangular \( m \times n \) matrix. Here \( n = 3m \).
Solution of saddle point system

- Sparse linear systems of very large size. Direct methods NOT recommended.
- Preconditioned Krylov subspace methods to be preferred.
- Ill conditioning of saddle point system caused by the difference in magnitude among the blocks which may be large depending on $\Delta t$ value.
- In long-term simulations small $\Delta t$ needed in the early stage of the consolidation process produces the most ill-conditioned situation.
- **Consequence:** Standard preconditioners (such as e.g. ILUT) converge very slowly or even fail to converge.
- **Ad hoc** preconditioners exploiting the block structure of the problem must be devised to yield (possibly fast) convergence of iterative methods.
- Among these we analyze the **Inexact Constraint preconditioners**
Exact Constraint Preconditioner (ECP)

Defined as $K^{-1}$ where

$$
K(\approx A) = \begin{bmatrix}
P_A & B^T \\
B & -C
\end{bmatrix}
$$

$$
= \begin{bmatrix}
I & 0 \\
BP_A^{-1} & I
\end{bmatrix}
\begin{bmatrix}
P_A & 0 \\
0 & -S
\end{bmatrix}
\begin{bmatrix}
I & P_A^{-1}B^T \\
0 & I
\end{bmatrix}
$$

$P_A$ is an SPD approximaton for $A$, $S = C + BP_A^{-1}B^T$ is the minus Schur complement of $K$.

- Very popular in a number of applications such as Constrained Optimization (where usually $P_A = \text{diag}(A)$), Mixed finite elements.
- $P_A^{-1}$ must be explicitly known to form $S$ in order to find a preconditioner for $S$ itself.
Exact Constraint Preconditioner (ECP)

- Application of ECP is performed by computing $K^{-1}r$ for some vector $r$.

$$K^{-1} = \begin{bmatrix} I & -P_A^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} P_A^{-1} & 0 \\ 0 & -S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -BP_A^{-1} & I \end{bmatrix}$$

- At each iteration need to solve a linear system with $S$.
- Theory: eigenvalues of the preconditioned matrix $K^{-1}A$ are all real and positive, at least $n - m$ of them equal to 1.
- The clustering of remaining eigenvalues around 1 depends on the quality of $P_A$ approximation to $A$.
- Preconditioned conjugate gradient method is proved to converge if starting from a suitable initial guess.
- Convergence is usually achieved in small to moderate number of yet very costly iterations.
Inexact Constraint Preconditioners

- Save on the cost of $K^{-1}$ application
- Further approximation: $S$ is replaced by the SPD matrix $P_S$
- $P_A^{-1}$ and $P_S^{-1}$ can be regarded as preconditioners for matrices $A$ and $S$, respectively.
- Define two Inexact Constraint Preconditioners $M_1^{-1}$ and $M_2^{-1}$

**Inexact Constraint Preconditioner (ICP)**

$$M_1 = \begin{bmatrix} I & 0 \\ BP_A^{-1} & I \end{bmatrix} \begin{bmatrix} P_A & 0 \\ 0 & -P_S \end{bmatrix} \begin{bmatrix} I & P_A^{-1}B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} P_A & B^T \\ B & S - P_S \end{bmatrix}$$

**Triangular Inexact Constraint Preconditioner (TICP).**

$$M_2 = \begin{bmatrix} P_A & 0 \\ 0 & -P_S \end{bmatrix} \begin{bmatrix} I & P_A^{-1}B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} P_A & B^T \\ 0 & -P_S \end{bmatrix}$$
Theoretical preliminaries

Given $A$ SPD, $n \times n$, $B$ rectangular $m \times n$, with $m < n$, and $C$, SPD $m \times m$, we are interested in the eigenvalues of:

$$M_- u = \lambda u \quad \text{where} \quad M_- = \begin{bmatrix} A & B^T \\ -B & C \end{bmatrix} \quad (1)$$

- **Notation:**
  
  $$0 < \alpha_A = \lambda_{\min}(A), \quad \beta_A = \lambda_{\max}(A),$$
  
  $$0 \leq \alpha_S = \lambda_{\min}(BB^T), \quad \beta_S = \lambda_{\max}(BB^T)$$
  
  $$0 \leq \alpha_C = \lambda_{\min}(C), \quad \beta_C = \lambda_{\max}(C),$$

- **If** $u = (u_1^T, 0)^T$ with $Bu_1 = 0$ then $\lambda$ satisfies $\alpha_A \leq \lambda \leq \beta_A$.

- **Bounds on the real eigenvalues of $M_-$ not lying in $\alpha_A, \beta_A$.** Define, for some $s, u_2 \neq 0$,

  $$\eta_A = \frac{s^T As}{s^T s} \in [\alpha_A, \beta_A], \quad \eta_C = \frac{u_2^T C u_2}{u_2^T u_2} \in [\alpha_C, \beta_C], \quad \eta_S = \frac{u_2^T BB^T u_2}{u_2^T u_2} \in [\alpha_S, \beta_S].$$
Theorem: bounds on real eigenvalues of $\mathcal{M}_-$

**Theorem**

The real eigenvalues of (1) not lying in $[\alpha_A, \beta_A]$ satisfy:

$$\alpha_C + \frac{\alpha_S}{\beta_A} \leq \eta_C + \frac{\eta_S}{\eta_A} \leq \lambda \leq \eta_C \leq \beta_C$$

In order to prove this Theorem a Lemma is needed

**Lemma**

Let $\lambda \not\in [\alpha_A, \beta_A]$. Then, for every $z \neq 0$, there exists a vector $s \neq 0$ such that

$$\frac{z^\top (A - \lambda I)^{-1} z}{z^\top z} = \left( \frac{s^\top As}{s^\top s} - \lambda \right)^{-1} = (\eta_A - \lambda)^{-1}.$$
Proof of the Theorem

Let $\lambda \in \mathbb{R}$ with $\lambda \notin [\alpha_A, \beta_A]$ so that $Bu_1 \neq 0, u_2 \neq 0$.

\[
\begin{align*}
Au_1 + B^\top u_2 &= \lambda u_1 \\
-Bu_1 + Cu_2 &= \lambda u_2
\end{align*}
\]  
(2)

Since $A - \lambda I$ is invertible: $u_1 = -(A - \lambda I)^{-1} B^\top u_2$.

Substituting in the second equation and multiplying by $\frac{u_2^\top}{u_2^\top u_2}$ yields

\[
\frac{u_2^\top B (A - \lambda I)^{-1} B^\top u_2}{u_2^\top u_2} + \eta_C - \lambda = 0
\]  
(3)

where $\eta_C = \frac{u_2^\top C u_2}{u_2^\top u_2} \in [\alpha_C, \beta_C]$. Defining $z = B^\top u_2$, and applying the previous Lemma we obtain

\[
\frac{z^\top (A - \lambda I)^{-1} z}{z^\top z} \frac{u_2^\top BB^\top u_2}{u_2^\top u_2} = \eta_S (\eta_A - \lambda)^{-1}, \quad \text{with} \quad \eta_A \in [\alpha_A, \beta_A], \; \eta_S \in [\alpha_S, \beta_S].
\]
Proof

we now rewrite (3) as

\[(\eta_A - \lambda)^{-1} \eta S + \eta C - \lambda = 0 \implies \lambda^2 - (\eta_C + \eta_A) \lambda + \eta S + \eta_A \eta_C = 0 \quad (4)\]

whose larger solution is

\[
\lambda_2 = \frac{\eta_A + \eta_C + \sqrt{(\eta_A + \eta_C)^2 - 4(\eta_A \eta_C + \eta S)}}{2} \leq \max\{\eta_A, \eta_C\} = \eta_C \leq \beta_C
\]

while the smaller one

\[
\lambda_1 = \frac{\eta_A + \eta_C - \sqrt{(\eta_A + \eta_C)^2 - 4(\eta_A \eta_C + \eta S)}}{2} \geq \frac{2(\eta_A \eta_C + \eta S)}{\eta_A + \eta_C + \sqrt{(\eta_A - \eta_C)^2 - 4\eta S}} \geq \frac{2(\eta_A \eta_C + \eta S)}{2 \max\{\eta_A, \eta_C\}} = \eta_C + \frac{\eta S}{\eta_A} \geq \alpha_C + \frac{\alpha_S}{\beta_A}
\]

hence the thesis for the lower bound. Summarizing the real eigenvalues of \( M_\perp \) satisfy:

\[
\min \left\{ \alpha_A, \alpha_C + \frac{\alpha_S}{\beta_A} \right\} \leq \lambda \leq \max\{\beta_A, \beta_C\}.
\]
Spectral Analysis of the Preconditioned Matrices

- Let $P_A^{-1}$ be a preconditioner for $A$, $P_S^{-1}$ for $S$. We define
  \[ A_P = P_A^{-1/2} A P_A^{-1/2} \quad \text{and} \quad S_P = P_S^{-1/2} S P_S^{-1/2} \]

- Let us assume that
  \begin{align*}
  0 < \alpha_A &= \lambda_{\min}(A_P) < 1 < \lambda_{\max}(A_P) = \beta_A, \\
  0 < \alpha_S &= \lambda_{\min}(S_P) < 1 < \lambda_{\max}(S_P) = \beta_S, \\
  0 \leq \alpha_C &= \lambda_{\min}(\hat{C}) < \lambda_{\max}(\hat{C}) = \beta_C
  \end{align*}

  where $\hat{C} = P_S^{-1/2} C P_S^{-1/2}$.

- The conditions $1 \in [\alpha_A, \beta_A]$ and $1 \in [\alpha_S, \beta_S]$ are very often fulfilled in practice since preconditioners $P_A$ and $P_S$ are expected to cluster eigenvalues around unit.

- Recall the two preconditioners
  \[ M_1 = \begin{bmatrix} P_A & B^\top \\ B & S - P_S \end{bmatrix} \quad M_2 = \begin{bmatrix} P_A & B^\top \\ 0 & -P_S \end{bmatrix} \]
Spectral Analysis of the Preconditioned Matrices

In order to characterize the eigenvalues of preconditioned matrix in the two cases, it is useful to define a matrix $P$ as

$$
P = \begin{bmatrix}
    P_A^{-1/2} & 0 \\
    0 & P_S^{-1/2}
\end{bmatrix}
$$

so that, the problem of finding the eigenvalues of $M_1^{-1}A$ and $M_2^{-1}A$ is equivalent to solve $PAP\nu = \lambda PM_1P\nu$, and $PAP\nu = \lambda PM_2P\nu$. Exploiting the blocks:

**ICP**: $PAP\nu = \lambda PM_1P\nu \rightarrow \begin{bmatrix}
    A_P & R^\top \\
    R & -\hat{C}
\end{bmatrix} \begin{bmatrix}
    \nu_1 \\
    \nu_2
\end{bmatrix} = \lambda \begin{bmatrix}
    I & R^\top \\
    R & S_P - I
\end{bmatrix} \begin{bmatrix}
    \nu_1 \\
    \nu_2
\end{bmatrix}$

where $R = P_S^{-1/2}BP_A^{-1/2}$. Note that $RR^\top = S_P - \hat{C}$.

**TICP**: $PAP\nu = \lambda PM_2P\nu \rightarrow \begin{bmatrix}
    A_P & R^\top \\
    R & -\hat{C}
\end{bmatrix} \begin{bmatrix}
    \nu_1 \\
    \nu_2
\end{bmatrix} = \lambda \begin{bmatrix}
    I & R^\top \\
    0 & -I
\end{bmatrix} \begin{bmatrix}
    \nu_1 \\
    \nu_2
\end{bmatrix}$
Spectral Analysis of ICP with $C = 0$

The inverse of the projected preconditioner, with $C \equiv 0$, can be written as

$$(PM_1P)^{-1} = \begin{bmatrix} I & -R^T \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ R & -I \end{bmatrix} = UL$$

so that the eigenvalues of $M_1^{-1}A$ are the same as those of $LAUw = \lambda w$ which reads:

$$\begin{bmatrix} A_P & (I - A_P)R^T \\ -R(I - A_P) & R(2I - A_P)R^T + \hat{C} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \lambda \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (7)$$

Define also the projected matrix $A_R = (RR^T)^{-1}RA_PR^T$. Let us denote

$$\alpha^R_A = \lambda_{\min}(A_R), \quad \text{and} \quad \beta^R_A = \lambda_{\max}(A_R).$$

It is easy to show that

- $[\alpha^R_A, \beta^R_A] \subset [\alpha_A, \beta_A]$
- the eigenvalues of $A_R$ do not depend on $P_S$.

If $\beta^R_A < 2$ the above matrix satisfy the hypotheses of previous Theorem.
Theorem

We will use the following notation:

\[
\theta_S = \frac{u_2^\top S u_2}{u_2^\top u_2}, \quad \theta_A^R = \frac{u_2^\top R A P R^\top u_2}{u_2^\top R R^\top u_2}, \quad \theta_A = \frac{s^\top A P s}{s^\top s}, \quad \theta_C = \frac{u_2^\top \hat{C} u_2}{u_2^\top u_2},
\]

for some \( s, u_2 \neq 0 \). It follows that \( \theta_A^R \in [\alpha_A^R, \beta_A^R] \) and \( \frac{u_2^\top R R^\top u_2}{u_2^\top u_2} = \theta_S - \theta_C (> 0) \).

Theorem

If \( \beta_A^R < 2 \) and \( C \equiv 0 \) then the real eigenvalues of (7) satisfy:

\[
\min \left\{ \alpha_A, \frac{\alpha_S}{\beta_A} \right\} \leq \lambda \leq \max\{\beta_A, (2 - \alpha_A^R)\beta_S\}.
\]

If \( \lambda_I \neq 0 \) then

\[
\frac{\alpha_A + \alpha_S (2 - \beta_A^R)}{2} \leq \lambda_R \leq \frac{\beta_A + \beta_S (2 - \alpha_A^R)}{2},
\]

\[
|\lambda_I| \leq \sqrt{\beta_S} \max\{1 - \alpha_A^R, \beta_A^R - 1\}.
\]
Proof of Theorem in the simpler case $C \equiv 0$

Bounds on smallest real eigenvalues.

- If the eigenvector is like $(u_1^\top, 0)^\top$ then $\lambda \geq \alpha_A$;
- Off-diagonal blocks of (7) do not have maximum rank since $1 \in \sigma(A_P)$, there are eigenvectors of the form $(0, u_2^\top)^\top$ which satisfy $(I - A_P)R^\top u_2 = 0$ and hence, from the second of (7) $(RR^\top + \hat{C}) u_2 = \lambda u_2$ which implies $\lambda \in [\alpha_S, \beta_S]$.
- To bound the remaining eigenvalues we write $\eta_A, \eta_S, \eta_C$ of (9) as

$$\eta_A = \theta_A, \quad \eta_S = \theta_S(1 - \theta_A^R)^2, \quad \eta_C = \theta_S(2 - \theta_A^R).$$

Then using Theorem 1 and observing that the function $f(t) = 2 - t + \frac{(1 - t)^2}{\theta_A}$, with $t \in [0, \theta_A]$ is decreasing and hence satisfies $f(t) \geq f(\theta_A) = \frac{1}{\theta_A}$,

$$\lambda \geq \eta_A \frac{\eta_C}{\eta_A} = \theta_S \left(2 - \theta_A^R + \frac{(1 - \theta_A^R)^2}{\theta_A}\right) \geq \frac{\theta_S}{\theta_A} \geq \frac{\alpha_S}{\beta_A}.$$
Proof of Theorem in the simpler case $C \equiv 0$

Bounds on largest real eigenvalues.
The largest one satisfy, according to the Theorem regarding $M_-$:

$$\lambda \leq \max \left\{ \lambda_{\max}(A_P), \lambda_{\max}\left(R(2I - A_P)R^\top\right) \right\} = \max\{\beta_A, (2 - \alpha_R^A)\beta_S\}$$

Bounds on complex eigenvalues.
The results regarding the complex eigenvalues follows from Proposition 2.12 in [Benzi Simoncini 2006].

$$\frac{\alpha_A + \alpha_S(2 - \beta_A^R)}{2} \leq \lambda_R \leq \frac{\beta_A + \beta_S(2 - \alpha_A^R)}{2}$$

$$|\lambda_I| \leq \| (A_P - I)R^\top \| \leq \| A_P - I \| \sqrt{\beta_S}$$
Case $C \neq 0$

The eigenvalues of $Au = \lambda M_1 u$ are now the same as those of

$$
\begin{pmatrix}
A_P & (I - A_P)R^T \\
-R(I - A_P) & R(2I - A_P)R^T + \hat{C}
\end{pmatrix}
\begin{pmatrix}
w_1 \\
w_2
\end{pmatrix} = \lambda
\begin{pmatrix}
w_1 \\
w_2
\end{pmatrix}
$$

(8)

Theorem

Let $\beta_A^R < 2$. The real eigenvalues of (8) satisfy:

$$
\min \left\{ \alpha_A, \frac{\alpha_S}{\beta_A^R} + \frac{\alpha_C(\beta_A^R - 1)}{\beta_A^R} \right\} \leq \lambda \leq \max\{\beta_A, (2 - \alpha_A^R)\beta_S - \alpha_C(1 - \alpha_A^R)\}
$$

If $\lambda_I \neq 0$ then

$$
\frac{\alpha_A + \alpha_S(2 - \beta_A^R) + \alpha_C(\beta_A^R - 1)}{2} \leq \lambda_R \leq \frac{\beta_A + \beta_S(2 - \alpha_A^R) - \alpha_C(1 - \alpha_A^R)}{2}
$$

$$
|\lambda_I| \leq \sqrt{\beta_S - \alpha_C} \max\{1 - \alpha_A, \beta_A - 1\}
$$
Alternative bound on complex eigenvalues

Theorem

Let

\[ g(a) = \frac{a - 1}{2 - a} \quad \text{and} \quad \gamma = \max_{a \in [\alpha_A, \beta_A]} |g(a)| \]

If \( \beta_A < 2 \) then the complex eigenvalues of ICP (with \( C = 0 \)) satisfy

\[ |\lambda - 1| < 2\gamma, \quad \lambda_I \leq \gamma, \quad |\lambda_R - 1| < 2\gamma \]

Similar bounds hold for \( C \neq 0 \).
Spectral Analysis of TICP. Case $C \equiv 0$

We analyze the eigenvalues of

$$
\begin{bmatrix}
A_P & R^\top \\
R & 0
\end{bmatrix}
\begin{bmatrix}
v_1 \\
v_2
\end{bmatrix}
= \lambda
\begin{bmatrix}
I & R^\top \\
0 & -I
\end{bmatrix}
\begin{bmatrix}
v_1 \\
v_2
\end{bmatrix}
$$

(9)

Now $RR^\top = S_P$.

Lemma

1. **The real eigenvalues of** $M_2^{-1}A$ **which do not lie in** $[\alpha_A, \beta_A]$ **(if any) satisfy the following equation:**

$$
\lambda^2 - (\theta_S + \theta_A) \lambda + \theta_S = 0
$$

where $\alpha_A^R \leq \theta_A \leq \beta_A^R$ and $\alpha_S \leq \theta_S \leq \beta_S$.

2. **All the eigenvalues satisfy a similar equation**

$$
\lambda^2 - (\theta'_S + \theta'_A) \lambda + \theta'_S = 0
$$

where now $\alpha_A \leq \theta'_A \leq \beta_A$ and $0 < \theta'_S \leq \beta_S$. 
PROOF of 1.

Let $\lambda \in \mathbb{R}$ with $\lambda < \alpha_A$ or $\lambda > \beta_A$. From the first equation in (9)

$$v_1 = (A_P - \lambda I)^{-1} (\lambda - 1) R^\top v_2.$$ 

Substituting in the second one we obtain $R (A_P - \lambda I)^{-1} (\lambda - 1) R^\top v_2 + \lambda v_2 = 0$.

Multiplying by $\frac{v_2^\top}{v_2 v_2}$ yields

$$(\lambda - 1) \frac{v_2^\top R (A_P - \lambda I)^{-1} R^\top v_2}{v_2 v_2} + \lambda = 0.$$ 

Defining $z = R^\top v_2$, we can apply initial Lemma to matrix $A_P$ thus obtaining

$$(\lambda - 1) \frac{v_2^\top R (A_P - \lambda I)^{-1} R^\top v_2}{v_2 v_2} = \frac{z^\top (A_P - \lambda I)^{-1} z}{z^\top z} \frac{v_2^\top R R^\top v_2}{v_2 v_2} = (\theta_A - \lambda)^{-1} \theta_S,$$

so that red equation can be rewritten as

$$(\lambda - 1) (\theta_A - \lambda)^{-1} \theta_S + \lambda = 0.$$ 

which can be written, after some very simple algebra, as

$$\lambda^2 - (\theta_S + \theta_A) \lambda + \theta_S = 0.$$
PROOF of 2.

Write $v_2$ from the second set of equations in (9) as

$$v_2 = -\frac{1}{\lambda} R v_1 \quad (10)$$

and substitute in the first set.

$$A_P v_1 - \lambda v_1 + \frac{\lambda - 1}{\lambda} R^T R v_1 = 0.$$  

Then premultiplying by $\frac{v_1^*}{v_1^* v_1}$ and setting $\theta'_A = \frac{v_1^* A_P v_1}{v_1^* v_1}$ and $\theta'_S = \frac{v_1^* R^T R v_1}{v_1^* v_1}$, we obtain

$$(\theta'_A - \lambda) + \theta'_S \frac{\lambda - 1}{\lambda} = 0,$$

which can be written as

$$\lambda^2 - (\theta'_S + \theta'_A) \lambda + \theta'_S = 0.$$
Theorem

Case with $C \equiv 0$. If $\lambda_I \neq 0$ then the eigenvalues of $M_2^{-1}A$ satisfy:

$$|\lambda - 1| \leq \sqrt{1 - \alpha_A}, \quad \text{and} \quad \frac{\alpha_A}{2} \leq \lambda_R \leq \min \left\{ \frac{1 + \beta_S}{2}, 2 - \frac{\alpha_A}{2} \right\}$$

The real eigenvalues satisfy:

$$\min \left\{ \alpha_A, \frac{\alpha_S}{\beta_A + \alpha_S} \right\} \leq \lambda_R \leq \beta_S + \beta_A$$

The proof of this theorem is based on exploiting the roots of the second order equations of previous Lemma.
Case with $C \neq 0$.
With some more effort the following results can be proved in the general case:

**Theorem**

*If* $\lambda_I \neq 0$ *then*

\[
|\lambda - 1| \leq \sqrt{(1 - \alpha_A)(1 - \alpha_C)} \quad \text{and} \quad \frac{\alpha_A + \alpha_C}{2} \leq \lambda_R \leq 2 - \frac{\alpha_A + \alpha_C}{2}
\]

*The real eigenvalues satisfy:*

\[
\min \left\{ \alpha_A, \frac{\alpha_C (\beta_A - 1) + \alpha_S}{\beta_A^R + \alpha_S} \right\} \leq \lambda \leq \beta_S + \beta_A
\]
Inexact Constraint Preconditioners

- ICP preconditioners can be written as $M^{-1}$ where
  \[
  ICP \quad M = \begin{bmatrix}
  I & 0 \\
  BP_A^{-1} & I
  \end{bmatrix}
  \begin{bmatrix}
  P_A & 0 \\
  0 & -P_S
  \end{bmatrix}
  \begin{bmatrix}
  I & P_A^{-1}B^T \\
  0 & I
  \end{bmatrix}
  \]

- $P_A$ \{SPD approximation to the (1,1) block $A$. $P_A^{-1}$ can be viewed as a preconditioner for $A$.\}

- $P_S$ \{SPD approximation to $S = BP_A^{-1}B^T + C(t)$ (negative Schur complement). $P_S^{-1}$ can be viewed as a preconditioner for $S$.\}

- The inverse of $P_A$ should be known explicitly to form $S$. Symmetric SAINV preconditioner (Benzi, Cullum, Tuma SISC 2000) used in sequential computation, FSAI (Kolotilina, Yeremin, SIMAX, 1993) in parallel computations.

- **Problem**: $P_A^{-1}$ should be chosen very sparse to yield not too dense Schur complement $\longrightarrow$ low quality preconditioner.

---

LB, M. Ferronato, and G. Gambolati,
Novel preconditioners for the iterative solution to FE-discretized coupled consolidation equations, CMAME, 196 (2007),
How to construct a fast preconditioner while keeping the Schur complement sparse

Compute two preconditioner for $A$:
1. $P_A$ which is used in the definition of ICP.
2. $\tilde{P}_A$ which is used only to construct an approximate Schur complement
   
   $$\tilde{S} = B\tilde{P}_A^{-1}B^\top + C(t)$$

Now $P_A$ can be e.g. an incomplete Cholesky factorization with fill-in.
$\tilde{P}_A$, as before, is computed as an AINV preconditioner.
$\tilde{P}_S$, as before, is a preconditioner (e.g. IC0) for $\tilde{S}$.

It is now useful to write the preconditioner explicitly as

$$M^{-1} = \begin{bmatrix}
I & -P_A^{-1}B^\top \\
0 & I
\end{bmatrix}
\begin{bmatrix}
P_A^{-1} & 0 \\
0 & -\tilde{P}_S^{-1}
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
-BP_A^{-1} & I
\end{bmatrix}$$

Call this variant which mixes up explicit (AINV) and implicit (Cholesky) preconditioners for the (1,1) block: Mixed Constraint Preconditioner (MCP).

MCP suited to both ICP and TICP
MCP application

Explicit form for the inverse of $T(MCP)$ preconditioners.

$$M_1^{-1} = \begin{bmatrix} L_A^{-T} & -L_A^{-T}L_A^{-1}B^TL_S^{-T} \\ 0 & L_S^{-T} \end{bmatrix} \begin{bmatrix} L_A^{-1} & 0 \\ L_S^{-1}BL_A^{-T}L_A^{-1} & -L_S^{-1} \end{bmatrix} = UL$$

$$M_2^{-1} = \begin{bmatrix} L_A^{-T} & -L_A^{-T}L_A^{-1}B^TL_S^{-T} \\ 0 & L_S^{-T} \end{bmatrix} \begin{bmatrix} L_A^{-1} & 0 \\ 0 & -L_S^{-1} \end{bmatrix} = UL$$

where

- $P_A = L_A L_A^T$ using incomplete ILLT($\tau_A,1\text{fill}$).
- $\tilde{S}$ is evaluated as $\tilde{S} = BZZ^T B^T + C(t) = S_0 + C(t)$, where $Z$ is the AINV upper triangular factor of the approximate inverse of $A$ depending on dropping tolerance $\tau_Z$.
- $P_S = L_S L_S^T$, using IC(0)

MCP in the split form can be implemented within the BiCGSTAB iterative method.
Implementation

Recent developments in the solution of indefinite systems, TU Eindhoven

Algorithm: preconditioned Bi-CGStab

1. Compute $r_0 = L(b - Ax_0)$, $\tilde{r}_0$ arbitrary
2. $p_0 = r_0$
3. For $j = 0, 1, \ldots$, until convergence, do
   4. $q_j = LAU p_j$
   5. $\alpha_j = (\tilde{r}_0^T r_j) / (\tilde{r}_0^T q_j)$
   6. $s_j = r_j - \alpha_j q_j$
   7. $t_j = LAU s_j$
   8. $\rho_j = \| t_j \|$
   9. $\omega_j = (s_j^T t_j) / \sqrt{\rho_j}$
10. $x_{j+1} = x_j + \alpha_j p_j + \omega_j s_j$
11. $r_{j+1} = s_j - \omega_j t_j$
12. $\beta_j = (\alpha_j \tilde{r}_0^T r_{j+1}) / (\omega_j \tilde{r}_0^T r_j)$
13. $p_{j+1} = r_{j+1} + \beta_j (p_j - \omega_j q_j)$
14. End For
Test problem # 1. M3Dsmall

A vertical cross-section of the cylindrical porous volume.
Medium: sequence of alternating sandy and clayey layers, with the hydraulic conductivity $k_{\text{sand}} = 10^{-5}$ m/s and $k_{\text{clay}} = 10^{-8}$ m/s, porosity $\phi = 0.20$, the Poisson ratio $\nu = 0.25$, and the Young modulus $E = 833.33$ MPa.
Standard Dirichlet bc, with fixed outer and bottom boundaries, and zero pore pressure variation on the top and outer surfaces.
The upper boundary is a traction-free plane. Fully three-dimensional tetrahedral grid.
Simulation of the compaction of a shallow confined aquifer due to groundwater withdrawal in a representative 3D sedimentary basin at a regional scale.

The discretized medium consists of an alternating sequence of sand and clay layers down to 5500 m depth. $k_{\text{sand}} = 10^{-4}$ m/s and $k_{\text{clay}} = 10^{-7}$ m/s, porosity 0.20 and Poisson ratio 0.30.

The mechanical properties of the porous medium vary with depth and are representative of the Northern Adriatic sediments, Italy.

Discretization. The 3D tetrahedral grid is made of 541,190 nodes and 3,122,280 elements. We solve two linear systems corresponding to $\Delta t = 1$ (Large3Da) and $\Delta t = 10^6$ (Large3Db).
Test case # 3. PoRiver.

- Simulation of the consolidation of a real gas reservoir of the Po Valley, Italy, used for underground gas storage purposes.
- The reservoir is a complex multi-layer structure consisting of 5 mineralized pools about 1,200-m deep connected to regional active waterdrives with several interbedded clay lenses.
- The discretized medium has an areal extent of $50 \times 50$ km and goes down to 10,000 m depth.
- The problem is discretized with a 3D tetrahedral grid totaling 299,734 nodes and 1,746,044 elements for 1,198,936 unknowns.
- In Table below we summarize the size and nonzeros of the three tests.

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>nnz($A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3dsmall</td>
<td>1,4212</td>
<td>805,904</td>
</tr>
<tr>
<td>Large3D</td>
<td>2,117,700</td>
<td>124,408,336</td>
</tr>
<tr>
<td>PoRiver</td>
<td>1,198,936</td>
<td>70,812,224</td>
</tr>
</tbody>
</table>
Parameters of the runs

- Parameters of ILLT factorization: drop tolerance $\tau_A$, fill-in level $l\text{fill}$.
- Parameter of AINV: drop tolerance $\tau_Z$.
- BiCGSTAB exit test: $\|r_k\|/\|b\| \leq 10^{-15} \rightarrow$ relative error of $\approx 10^{-8}$.
- $\rho$ measures the density of the preconditioner: $\rho \approx 2 \frac{\text{nnz}(L_A)}{\text{nnz}(A)} + 2 \frac{\text{nnz}(L_S)}{\text{nnz}(A)}$
- Timings (in seconds)
  1. $T_{P1}$ time to construct $L_A$ plus $Z$ plus $S_0$ (preprocessing not depending on time).
  2. $T_{P2}$ time to construct the incomplete Cholesky factorization of $\tilde{S}$,
  3. $T_s$ CPU time required by the iterative solver,
  4. $T = T_{P2} + T_s$ total CPU.
- The Fortran 90 code has been run on an IBM Power6 with 4.7 GHz.
**Computed spectrum vs bounds. Smallest problem**

- 12 instances of the problem obtained by varying the parameters of the ILLT and AINV preconditioners. To yield $P_S$ we always used IC(0). Timestep $\Delta t = 1$.
- Parameters and the extremal eigenvalues of $A_P(A_R), S_P$ and $\hat{C}$.

<table>
<thead>
<tr>
<th># run</th>
<th>$\tau_A$</th>
<th>lfill</th>
<th>$\tau_Z$</th>
<th>$\alpha_A, \beta_A$ ($\alpha_A^R, \beta_A^R$)</th>
<th>$\alpha_S$</th>
<th>$\beta_S$</th>
<th>$\alpha_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10^{-4}$</td>
<td>50</td>
<td>0.30</td>
<td>0.255 1.255</td>
<td>0.110</td>
<td>23.731</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>$10^{-4}$</td>
<td>50</td>
<td>0.10</td>
<td></td>
<td>0.107</td>
<td>7.743</td>
<td>0.002</td>
</tr>
<tr>
<td>3</td>
<td>$10^{-4}$</td>
<td>50</td>
<td>0.05</td>
<td>(0.761 1.217)</td>
<td>0.313</td>
<td>5.111</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>$10^{-4}$</td>
<td>50</td>
<td>0.01</td>
<td></td>
<td>0.708</td>
<td>4.112</td>
<td>0.012</td>
</tr>
<tr>
<td>5</td>
<td>$10^{-2}$</td>
<td>30</td>
<td>0.30</td>
<td>0.059 1.662</td>
<td>0.110</td>
<td>23.323</td>
<td>0.008</td>
</tr>
<tr>
<td>6</td>
<td>$10^{-2}$</td>
<td>30</td>
<td>0.10</td>
<td></td>
<td>0.107</td>
<td>7.577</td>
<td>0.012</td>
</tr>
<tr>
<td>7</td>
<td>$10^{-2}$</td>
<td>30</td>
<td>0.05</td>
<td>(0.411 1.423)</td>
<td>0.312</td>
<td>3.100</td>
<td>0.007</td>
</tr>
<tr>
<td>8</td>
<td>$10^{-2}$</td>
<td>30</td>
<td>0.01</td>
<td></td>
<td>0.680</td>
<td>1.583</td>
<td>0.012</td>
</tr>
<tr>
<td>9</td>
<td>$10^{-1}$</td>
<td>10</td>
<td>0.30</td>
<td>0.027 2.046</td>
<td>0.110</td>
<td>23.019</td>
<td>0.008</td>
</tr>
<tr>
<td>10</td>
<td>$10^{-1}$</td>
<td>10</td>
<td>0.10</td>
<td></td>
<td>0.107</td>
<td>7.500</td>
<td>0.002</td>
</tr>
<tr>
<td>11</td>
<td>$10^{-1}$</td>
<td>10</td>
<td>0.05</td>
<td>(0.384 1.590)</td>
<td>0.312</td>
<td>3.059</td>
<td>0.007</td>
</tr>
<tr>
<td>12</td>
<td>$10^{-1}$</td>
<td>10</td>
<td>0.01</td>
<td></td>
<td>0.419</td>
<td>1.388</td>
<td>0.012</td>
</tr>
</tbody>
</table>
## Results of ICP and TICP

Iteration number and CPU times for BiCGSTAB preconditioned by ICP and TICP.

<table>
<thead>
<tr>
<th># run</th>
<th>( \rho )</th>
<th>iter</th>
<th>( T_p )</th>
<th>( T_s )</th>
<th>( T )</th>
<th>iter</th>
<th>( T_p )</th>
<th>( T_s )</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.93</td>
<td>104</td>
<td>0.08</td>
<td>1.14</td>
<td>1.25</td>
<td>100</td>
<td>0.08</td>
<td>0.96</td>
<td>1.06</td>
</tr>
<tr>
<td>2</td>
<td>1.58</td>
<td>65</td>
<td>0.23</td>
<td>0.77</td>
<td>1.04</td>
<td>59</td>
<td>0.24</td>
<td>0.60</td>
<td>0.88</td>
</tr>
<tr>
<td>3</td>
<td>2.22</td>
<td>30</td>
<td>0.39</td>
<td>0.38</td>
<td>0.81</td>
<td>29</td>
<td>0.39</td>
<td>0.31</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>4.91</td>
<td>22</td>
<td>0.83</td>
<td>0.31</td>
<td>1.21</td>
<td>23</td>
<td>0.84</td>
<td>0.28</td>
<td>1.19</td>
</tr>
<tr>
<td>5</td>
<td>0.60</td>
<td>133</td>
<td>0.08</td>
<td>1.14</td>
<td>1.24</td>
<td>119</td>
<td>0.08</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>1.25</td>
<td>80</td>
<td>0.22</td>
<td>0.74</td>
<td>1.00</td>
<td>76</td>
<td>0.23</td>
<td>0.64</td>
<td>0.89</td>
</tr>
<tr>
<td>7</td>
<td>1.88</td>
<td>51</td>
<td>0.38</td>
<td>0.50</td>
<td>0.91</td>
<td>55</td>
<td>0.38</td>
<td>0.48</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>4.57</td>
<td>35</td>
<td>0.81</td>
<td>0.38</td>
<td>1.27</td>
<td>48</td>
<td>0.82</td>
<td>0.48</td>
<td>1.37</td>
</tr>
<tr>
<td>9</td>
<td>0.45</td>
<td>181</td>
<td>0.07</td>
<td>1.30</td>
<td>1.39</td>
<td>187</td>
<td>0.08</td>
<td>1.19</td>
<td>1.29</td>
</tr>
<tr>
<td>10</td>
<td>1.10</td>
<td>112</td>
<td>0.22</td>
<td>0.89</td>
<td>1.14</td>
<td>129</td>
<td>0.23</td>
<td>0.92</td>
<td>1.17</td>
</tr>
<tr>
<td>11</td>
<td>1.73</td>
<td>78</td>
<td>0.37</td>
<td>0.65</td>
<td>1.07</td>
<td>93</td>
<td>0.37</td>
<td>0.71</td>
<td>1.13</td>
</tr>
<tr>
<td>12</td>
<td>4.42</td>
<td>67</td>
<td>0.81</td>
<td>0.64</td>
<td>1.52</td>
<td>111</td>
<td>0.82</td>
<td>0.97</td>
<td>1.86</td>
</tr>
</tbody>
</table>
Eigenvalue distribution of $M^{-1}A$

ICP (left) and TICP (right). Test # 4 (top), test # 6 (bottom).
**Numerical results**  Recent developments in the solution of indefinite systems, TU Eindhoven

### Computed real eigenvalues vs bounds

$R_{\text{min}}, R_{\text{max}}$: smallest and largest real eigenvalue

<table>
<thead>
<tr>
<th>#</th>
<th>precond.</th>
<th>computed $R_{\text{min}}, R_{\text{max}}$</th>
<th>bounds $R_{\text{min}}, R_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ICP</td>
<td>0.107 8.20</td>
<td>0.088 9.59</td>
</tr>
<tr>
<td>2</td>
<td>TICP</td>
<td>0.107 7.69</td>
<td>0.081 8.96</td>
</tr>
<tr>
<td>3</td>
<td>ICP</td>
<td>0.278 5.68</td>
<td>0.255 6.33</td>
</tr>
<tr>
<td>3</td>
<td>TICP</td>
<td>0.287 4.98</td>
<td>0.205 6.33</td>
</tr>
<tr>
<td>4</td>
<td>ICP</td>
<td>0.279 4.61</td>
<td>0.255 5.09</td>
</tr>
<tr>
<td>4</td>
<td>TICP</td>
<td>0.288 3.96</td>
<td>0.255 5.33</td>
</tr>
<tr>
<td>6</td>
<td>ICP</td>
<td>0.073 7.86</td>
<td>0.059 12.04</td>
</tr>
<tr>
<td>6</td>
<td>TICP</td>
<td>0.074 7.54</td>
<td>0.059 9.00</td>
</tr>
<tr>
<td>11</td>
<td>ICP</td>
<td>0.031 3.12</td>
<td>0.027 4.94</td>
</tr>
<tr>
<td>11</td>
<td>TICP</td>
<td>0.031 3.05</td>
<td>0.027 4.65</td>
</tr>
<tr>
<td>12</td>
<td>ICP</td>
<td>0.031 1.61</td>
<td>0.027 2.23</td>
</tr>
<tr>
<td>12</td>
<td>TICP</td>
<td>0.031 2.49</td>
<td>0.027 2.98</td>
</tr>
</tbody>
</table>
The importance of being real

Plot of scaled square root of condition number $7\sqrt{\kappa}$ and number of iteration for each test case. ICP on the left, TICP on the right.

Even if the preconditioned matrix is not symmetric, again real eigenvalues keep on driving convergence of the iterative method.
To simplify we used $\alpha_A, \beta_A$ instead of $\alpha_A^R, \beta_A^R$.

- **Theorem** Let $\beta_A < 2$. Then the eigenvalues of ICP satisfy

$$\min \left\{ \alpha_A, \frac{\alpha_S}{\beta_A} \right\} \leq \lambda \leq \max\{(2 - \alpha_A)\beta_S, \beta_A\}.$$  

$$\frac{\alpha_A + \alpha_S(2 - \beta_A)}{2} \leq \lambda_R \leq \frac{\beta_A + \beta_S(2 - \alpha_A)}{2}$$  

$$|\lambda_I| \leq \sqrt{\beta_S \max\{1 - \alpha_A, \beta_A - 1\}}.$$  

- Clustering of eigenvalues of the preconditioned matrices relies (apparently) on clustering around one of those of $A_P$ and $S_P \rightarrow$ good preconditioners for $A$ and $S$.

- BUT. With the MCP approach, $S_P = \tilde{P}_S^{-1}S$ i.e. $S_P$ is the exact Schur complement preconditioned with $\tilde{P}_S$ which is in its turn computed from the approximate Schur complement.

- Consequence: the spectrum of $S_P$ is shifted away from unity even using optimal $\tilde{P}_S(\approx \tilde{S}^{-1})$. 

Relaxed MCP

- Introduce a real (relaxation) parameter $\omega$
- Define a family of MCP preconditioners as the inverse of $M(\omega)$, where

$$M(\omega) = \begin{bmatrix} I & 0 \\ BP^{-1}A & I \end{bmatrix} \begin{bmatrix} PA & 0 \\ 0 & \omega \tilde{P}_S \end{bmatrix} \begin{bmatrix} I & P^{-1}A^{-1}B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} PK & B^T \\ B & BP^{-1}B^T - \omega \tilde{P}_S \end{bmatrix}.$$  

- Explicit form of the preconditioner:

$$M^{-1}(\omega) = \begin{bmatrix} I & -P^{-1}A^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} P^{-1}A & 0 \\ 0 & -\frac{1}{\omega} \tilde{P}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -BP^{-1}A & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} P^{-1}A & 0 \\ 0 & -\frac{1}{\omega} \tilde{P}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -BP^{-1}A & I \end{bmatrix}.$$  

- Proper choice of $\omega$ should possibly accelerate Mixed Constraint Preconditioners.
Estimates of extremal eigenvalues of RMCP

Theorem.

Let $\beta_A < 2$, then any real eigenvalue $\lambda$ of $M(\omega)^{-1}A$ satisfy the following bounds:

$$\min \left\{ \alpha_A, \frac{\omega \alpha_S}{2} \right\} \leq \lambda \leq \max\{2\omega \beta_S, \beta_A\}$$

Moreover the complex eigenvalues satisfy

$$\frac{\alpha_A}{2} \leq \lambda_R \leq \frac{\beta_A}{2} + \omega \beta_S \quad |\lambda_I| \leq \sqrt{\omega \beta_S}$$

Proof. (sketch) With RMCP all eigenvalues of $S_P$ are multiplied by $\omega$ so that:

$$\min \left\{ \alpha_A, \frac{\omega \alpha_S}{\beta_A} \right\} \leq \lambda(\omega) \leq \max\{\beta_A, \omega (2 - \alpha_A) \beta_S\}$$

The thesis holds using $\alpha_A > 0$, $\beta_A < 2$. Regarding complex eigenvalues, the thesis holds from

$$\frac{\alpha_A + \omega \alpha_S (2 - \beta_A)}{2} \leq \lambda_R(\omega) \leq \frac{\beta_A + \omega \beta_S (2 - \alpha_A)}{2} \quad |\lambda_I(\omega)| \leq \sqrt{\omega \beta_S} \max\{1 - \alpha_A, \beta_A - 1\}$$
Practical assessment of $\omega$

- Convergence of iterative methods preconditioned by RMCP depends strongly on the ratio between the largest and the smallest real eigenvalues of $M(\omega)^{-1}A$. We want to minimize

$$\kappa_R = \frac{\max_k \{\lambda_k(\omega)\}}{\min_k \{\lambda_k(\omega)\}} \leq \frac{\max \{\beta_A, 2\omega\beta_S\}}{\min \left\{\frac{\alpha_A}{2}, \frac{\omega\alpha_S}{2}\right\}}$$

- Computing smallest eigenvalue(s) is costly. Let us rely only on largest eigenvalues of $A_P$ and $S_P$.
- Define $c_A = \kappa(A_P)$ and $c_S = \kappa(S_P)$.

**Theorem.** Let $\beta_A < 2$. If $\omega = \frac{\beta_A}{\beta_S}$ then

$$\kappa_R \leq \max\{2c_A, 4c_S\}$$

Moreover the complex eigenvalues satisfy

$$\frac{\alpha_A}{2} \leq \lambda_R \leq \frac{3\beta_A}{2} \quad |\lambda_I| \leq \sqrt{\beta_A}(< \sqrt{2})$$

LB, A Martínez,
RMCP: Relaxed Mixed Constraint Preconditioners for Saddle Point Linear Systems arising in Geomechanics
CMAME, 221–222, 2012.
Sequential Computations

- Solution of a Coupled Consolidation system by RMCP-BiCGSTAB.
- Exit test: when relative residual below $tol = 10^{-12}$.
- Parameter to be assessed:
  1. $P_A$: IC depending on $\tau_A$ and $\text{lfil}$.
  2. $\tilde{P}_A$: AINV depending on threshold parameter $\tau_Z$.
  3. $\tilde{P}_S$: IC(0). No parameters to assess.
- Results for a Fortran 90 code on an IBM Power6 with 4.7 GHz RAM.
- CPU times: $T_p =$ CPU time for $\tilde{P}_S$, $T_{sol} =$ CPU time for BiCGSTAB $T_{tot} = T_p + T_{sol}$.
- Other time-consuming tasks: computation of $P_A$, $\tilde{P}_A$ and $B\tilde{P}_A^{-1}B^\top$ independent on $\Delta t$ and hence considered as preprocessing.
- $\rho$ relative density of preconditioner.
**Convergence vs $\omega$.**

- M3Dsmall problem. $\tau_A = \tau_Z = 0.1$ and $1\text{f}1 = 10$. We computed exact eigenvalues finding $\alpha_A = 0.027$, $\beta_A = 1.922$ while $\alpha_S = 0.108$, $\beta_S = 7.390$. Since $\beta_S > \beta_A$ and $\alpha_S > \alpha_A$ $\omega$ must be less than one.

- Plot number of iterations (stars) and “real condition number” $\kappa_R$ (circles) vs $\omega$.

- RMCP with $\omega_{opt} = 0.31$ requires 71 iteration while MCP ($\omega = 1$) 109.

- **Note.** Using previous Theorem we would have obtained $\omega = 0.27$.

- The qualitative plots of $\kappa_R(\omega)$ and $\text{iter}(\omega)$ are much similar. Minimizing $\kappa_R$ yields an $\omega$-value very close to the optimal one.
Numerical Results

Recent developments in the solution of indefinite systems, TU Eindhoven

Eigenvalue distribution

- Eigenvalue distributions in the complex plane for $\omega = 1$ and $\omega = 0.31$.
- Clustering of eigenvalues, both real and complex, around one, using RMCP with $\omega = 0.31$.

Figure: Eigenvalue distribution of preconditioned matrix with $\omega = 1$ (left) and $\omega = 0.31$ (right).
Po878 matrix. MCP results

- Three combinations of parameters.
- The table reports the extremal real eigenvalues of $S_P, A_P$ together with those of $M^{-1}A$ using $\omega = 1$ (MCP).

<table>
<thead>
<tr>
<th># run</th>
<th>lfil</th>
<th>$\tau_A$</th>
<th>$\tau_Z$</th>
<th>$\rho_A$</th>
<th>$\rho_S$</th>
<th>$\beta_A$</th>
<th>$\alpha_A$</th>
<th>$\beta_S$</th>
<th>$\alpha_S$</th>
<th>$\kappa_R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>$10^{-4}$</td>
<td>1</td>
<td>1.21</td>
<td>0.14</td>
<td>1.876</td>
<td>0.0112</td>
<td>37.253</td>
<td>0.194</td>
<td>3105</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>$10^{-4}$</td>
<td>1</td>
<td>0.75</td>
<td>0.14</td>
<td>1.875</td>
<td>0.0067</td>
<td>37.255</td>
<td>0.194</td>
<td>5190</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>$10^{-4}$</td>
<td>1</td>
<td>0.51</td>
<td>0.14</td>
<td>1.885</td>
<td>0.0042</td>
<td>37.243</td>
<td>0.195</td>
<td>8195</td>
</tr>
</tbody>
</table>

- The smallest (largest) eigenvalue of $A_P$ and $S_P$ computed using few iterations of a non preconditioned CG procedure to minimize (maximize) the Rayleigh Quotient (DACG).

<table>
<thead>
<tr>
<th># run</th>
<th>$T(L_A)$</th>
<th>iter</th>
<th>$T_p$</th>
<th>$T_{sol}$</th>
<th>$T_{tot}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151.4</td>
<td>263</td>
<td>3.5</td>
<td>527.8</td>
<td>531.3</td>
</tr>
<tr>
<td>2</td>
<td>77.4</td>
<td>315</td>
<td>3.5</td>
<td>520.2</td>
<td>523.7</td>
</tr>
<tr>
<td>3</td>
<td>44.9</td>
<td>540</td>
<td>3.5</td>
<td>737.3</td>
<td>740.8</td>
</tr>
</tbody>
</table>
Numerical Results

Recent developments in the solution of indefinite systems, TU Eindhoven

Po878 matrix. RMCP results

<table>
<thead>
<tr>
<th># run</th>
<th>$T(L_A)$</th>
<th>iter</th>
<th>$T_p$</th>
<th>$T_{sol}$</th>
<th>$T_{tot}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151.4</td>
<td>263</td>
<td>3.5</td>
<td>527.8</td>
<td>531.3</td>
</tr>
<tr>
<td>2</td>
<td>77.4</td>
<td>315</td>
<td>3.5</td>
<td>520.2</td>
<td>523.7</td>
</tr>
<tr>
<td>3</td>
<td>44.9</td>
<td>540</td>
<td>3.5</td>
<td>737.3</td>
<td>740.8</td>
</tr>
<tr>
<td>MCP ($\omega = 1$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>71</td>
<td>3.5</td>
<td>12.8</td>
<td>145.9</td>
<td>162.2</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>3.5</td>
<td>11.2</td>
<td>165.5</td>
<td>180.2</td>
</tr>
<tr>
<td>3</td>
<td>142</td>
<td>3.5</td>
<td>9.4</td>
<td>192.4</td>
<td>205.7</td>
</tr>
</tbody>
</table>

- RMCP with $\omega = 0.05 \approx \frac{\beta_A}{\beta_S} = \frac{1.875}{37.25}$ as suggested by the Theorem.

- The table reports the results of these RMCP runs. Also shows the CPU time to approximate $\beta_A$ and $\beta_S$ ($T_{eig}$) and the value of $\kappa_R$.

- Three times CPU time and iter number reduction provided by RMCP.
- Note also reduction of $\kappa_R$ by a factor $15-20$.
- Eigenvalue preprocessing negligible.
Convergence of RMCP vs MCP

Convergence profile of RMCP-BiCGSTAB with $\omega = 1$ and $\omega = 0.05$ for run # 2.

![Graph showing convergence profile](image-url)
Parallel implementation is based on FSAI (factorized sparse approximate inverse) to yield $P_A$, $\tilde{P}_A$ and $\tilde{P}_S$.

Parallel FSAI based on prefiltration, postfiltration and sparsity pattern based on powers of $A(S)$ ($d_A = 1 \cdots 4$).

Setting $\hat{W}_S = \omega^{-1/2} W_S$, the FSAI-RMCP can be written directly as:

$$M(\omega)^{-1} = \begin{bmatrix} W_1^T & -W_1^T W_1 B^T \hat{W}_S^T \\ 0 & \hat{W}_S^T \end{bmatrix} \begin{bmatrix} W_1 \\ \hat{W}_S B W_1^T W_1 & 0 \end{bmatrix}$$

$W_1$ is a relatively dense FSAI factor of $A$ ($P_A^{-1} = W_1^T W_1$)

$W_S$ is the FSAI factor of $\tilde{S}$, $P_S^{-1} = W_S^T W_S$.

$\tilde{S} = B W_2^T W_2 B^T + C = S_0 + C$, $W_2$ obtained from $W_1$ by a further dropping.

Preconditioner application strongly based on matrix vector-products.

**References**

- LB, A. Martínez,
  *FSAI-based Parallel Mixed Constraint Preconditioners for Saddle Point Problems Arising in Geomechanics*
  *JCAM, 2011*

- LB, A. Martínez,
  *Parallel Inexact Constraint Preconditioners for Saddle Point Problems*
  *Euro-Par 2011, LNCS 6853, 2011*
Results on the Large3d problem.

- Fortran 90 – MPI code.
- IBM SP6/5376 cluster at the CINECA Centre for HPC
- Define as $S_p^{(\bar{p})}$ the pseudo speedup computed wrt the smallest number of processors $(\bar{p})$ $E_p^{(\bar{p})}$ the corresponding efficiency:

\[
S_p^{(\bar{p})} = \frac{T_{\bar{p}\bar{p}}}{T_p}, \quad E_p^{(\bar{p})} = \frac{S_p^{(\bar{p})}}{p} = \frac{T_{\bar{p}\bar{p}}}{T_p\bar{p}}.
\]

- Timings, iteration numbers and pseudo-efficiencies for the Large3D test case.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$T_P$</th>
<th>$\omega = 1$</th>
<th>$\omega = 0.74$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>iter</td>
<td>$T_{sol}$</td>
</tr>
<tr>
<td>4</td>
<td>10.3</td>
<td>376</td>
<td>610.9</td>
</tr>
<tr>
<td>8</td>
<td>6.7</td>
<td>379</td>
<td>298.7</td>
</tr>
<tr>
<td>16</td>
<td>4.2</td>
<td>327</td>
<td>136.1</td>
</tr>
<tr>
<td>32</td>
<td>3.3</td>
<td>359</td>
<td>77.6</td>
</tr>
<tr>
<td>64</td>
<td>2.5</td>
<td>357</td>
<td>42.0</td>
</tr>
<tr>
<td>128</td>
<td>1.7</td>
<td>387</td>
<td>22.0</td>
</tr>
<tr>
<td>256</td>
<td>1.4</td>
<td>433</td>
<td>10.8</td>
</tr>
<tr>
<td>512</td>
<td>1.1</td>
<td>413</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Scalability of FSAI-RMCP

- Using $\omega = 0.74$, obtained from the Theorem provides a generalized reduction of the number of iterations and CPU time, irrespective of $p$.
- CPU time to approximate extremal eigenvalue has not been reported, being in all cases less than 5 percent the total time.
- Scalability of the two codes are very satisfactory as also accounted by the pseudo speedups in the figure below.
Mixed Finite Element discretizations

- The fluid mass balance is prescribed by the continuity equation:

\[
\text{div} \cdot \vec{v} + \frac{\partial}{\partial t} (\phi \beta p + \alpha) = -\frac{\partial}{\partial t} (\text{div} \cdot \vec{u}) + f
\]  

(11)

where \( \vec{u} \) the (known) medium displacements and \( p \) the pore pressure; \( \phi \) the medium porosity, \( \beta \) the fluid compressibility, \( t \) time, \( f \) a flow source or sink and \( \vec{v} \) the Darcy flux.

- Equation (11) must be coupled with the Darcy law defining \( \vec{v} \):

\[
\rho g K^{-1} \vec{v} + \vec{\nabla} p = 0
\]

(12)

with \( K \) the hydraulic conductivity tensor and \( \rho g \) the fluid specific weight.

- Discretization in space: \( P_0 – RT0 \), (LBB condition satisfied).

- After discretization, a saddle point linear system has to be solved at each timestep.

- Here \( A \) is the matrix of the scalar product between RT0 basis functions, \( B (B^\top) \) discretizes the gradient (divergence) operator and \( C \) is a diagonal matrix depending on \( \Delta t \).

- If the steady state problem has to be solved, (or if \( \beta = 0 \) i.e. the fluid is incompressible) then \( C \equiv 0 \).
Solution of the steady state MFE problem

- The 3D domain is subdivided into 9 zones, each of them characterized by a different value of the hydraulic conductivity tensor whose norm varies by six order of magnitudes from $8.64 \times 10^{-7}$ to $8.64 \times 10^{-1}$ m·s$^{-1}$.
- Very ill-conditioned steady-state problem. The problem has $N = 253216$ and a number of nonzeros $nnz = 1336168$.
- Here (1,1) block $A$ is well-conditioned, $\kappa(A)$ not growing with meshsize $h$.
- By contrast, $\kappa(S) = O(h^{-2})$.
- We chose the following parameters:
  1. IC preconditioner for $A$: $\tau_A = 0.1$, $\text{lfil}_A = 4$;
  2. AINV preconditioner for $A$: $\tau_Z = 0.5$;
  3. IC preconditioner for $S$: $\tau_S = 10^{-4}$, $\text{lfil}_S = 50$.
- Note that for this problem a simple IC(0) preconditioner for $S$ is not sufficient to guarantee convergence of the iterative method. These parameters yields the following density values for the preconditioners: $\rho_A = 0.45$, $\rho_S = 2.79$. 

Table: Iteration number and CPU times for MCP and RMCP with experimentally computed optimal value of $\omega$ for previously defined values of the tolerance $\text{tol}$.

<table>
<thead>
<tr>
<th>$\log_{10}(\text{tol})$</th>
<th>$\omega$</th>
<th>iter</th>
<th>$T_{\text{eig}}$</th>
<th>$T_{\text{prec}}$</th>
<th>$T_{\text{sol}}$</th>
<th>$T_{\text{tot}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-12$</td>
<td>1</td>
<td>746</td>
<td>–</td>
<td>3.91</td>
<td>80.84</td>
<td>84.75</td>
</tr>
<tr>
<td>$-10$</td>
<td>1</td>
<td>589</td>
<td>–</td>
<td>3.88</td>
<td>63.15</td>
<td>67.13</td>
</tr>
<tr>
<td>$-12$</td>
<td>$0.025$</td>
<td>443</td>
<td>1.14</td>
<td>3.91</td>
<td>50.19</td>
<td>55.24</td>
</tr>
<tr>
<td>$-10$</td>
<td>$0.025$</td>
<td>259</td>
<td>1.18</td>
<td>3.88</td>
<td>32.45</td>
<td>37.51</td>
</tr>
</tbody>
</table>

- Table 1 summarizes the timing and iterations results of RMCP with $\omega = 1$ using two different values of the tolerance $\text{tol} = 10^{-12}$ and $\text{tol} = 10^{-10}$.
- RMCP with the experimentally computed value of $\omega = \frac{\beta_A}{\beta_S}$. Leading eigenvalues approximated using ten iterations of the DACG method.
- Improvement in terms of iteration number an CPU time provided by RMCP with optimal $\omega$.
- The elapsed time is reduced by a factor 1.5 ($\text{tol} = 10^{-12}$) or 1.8 ($\text{tol} = 10^{-10}$) with respect to the MCP.
The experimental value of $\omega (= 0.025)$ is very close to the minimum of both graphs.

RMCP improvement regarding iteration number not sensitive to the value of $\omega$. 
Conclusions

- Saddle Point systems arising from discretization of coupled consolidation equations solved by preconditioned Krylov subspace methods.
- Block preconditioners efficiently accelerate such methods.
- Spectral analysis reveals that eigenvalues of preconditioned matrices can be clustered around 1 provided that two effective preconditioners for $A$ and $S$ are available.
- Acceleration of MCP by means of a real parameter $\omega$ effective in both iteration number and CPU time.
- Assessment of “optimal” $\omega$ obtained at the price of few matrix-vector products.
- RMCP also effective when employed in the solution of saddle point systems arising from MFE discretization of fluid flow in porous media.

However

- RMCP very fast but does not scale with the mesh related parameter $h$.
- In our parallel implementation the acceleration is less pronounced. In fact, the two inverse approximations of block $A$ differ only in drop tolerance $\rightarrow$ the optimal $\omega$ is very close to 1. $\rightarrow$ modest acceleration of convergence.
Plans for the Future

- Try the “pressure Schur complement preconditioner”:
  \[(BA^{-1}B^\top)^{-1} \rightarrow (BB^\top)^{-1}\tilde{A},\] with \(\tilde{A}\) the discretized scalar counterpart of \(\nabla^2 u\).
- How to handle matrix \(C\)? Suggestions welcome.
References

V. Simoncini, Block triangular preconditioners for symmetric saddle-point problems, ANM, 49 (2004).

M. Benzi and V. Simoncini, On the eigenvalues of a class of saddle point matrices Numerische Matematik, 103 (2006).


LB On eigenvalue distribution of constraint-preconditioned symmetric saddle point matrices NLAA, published online, OCT 2011.

