Convergence of some iterative methods for symmetric saddle point linear systems

Yvan Notay*

Service de Métrologie Nucléaire
Université Libre de Bruxelles (C.P. 165/84)
50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium.
email : ynotay@ulb.ac.be

Report GANMN 18-02

August 2018

Abstract

We consider the iterative solution of linear systems with symmetric saddle point system matrix. We address a family of solution techniques that exploit the knowledge of a preconditioner (or approximate solution procedure) for both the top left block of the matrix on the one hand, and for the Schur complement resulting from its elimination on the other hand. This includes many “segregated” of “Schur complement” iterations such as the inexact Uzawa method and pressure correction techniques, and also many “block” preconditioners, based on the approximate block factorization of the system matrix. An analysis is developed which proves convergence in norm of stationary iterations. It is more rigorous than eigenvalue analyses which ignore nonnormality effects, while being more general than previous norm analyses. The analysis also clarifies the relations that exist between the many members of this family of methods, and offers practical guidelines to select the scheme most appropriate to a situation at hand.

Key words. Saddle Point, Preconditioning, Uzawa method, Block Triangular, SIMPLE, Convergence Analysis, Linear Systems, Stokes problem, PDE-constrained optimization

AMS subject classification. 65F08, 65F10, 65F50, 65N22

*Yvan Notay is Research Director of the Fonds de la Recherche Scientifique – FNRS
1 Introduction

We consider linear systems of the form

\[
\begin{pmatrix}
A & B^T \\
B & -C
\end{pmatrix}
\begin{pmatrix}
u_A \\
u_C
\end{pmatrix} =
\begin{pmatrix}
b_A \\
b_C
\end{pmatrix},
\]

where the system matrix

\[
K = \begin{pmatrix}
A & B^T \\
B & -C
\end{pmatrix}
\]

is symmetric. We further assume that \(A\) is symmetric and positive definite (SPD) and that \(C\) is nonnegative definite while being positive definite on the on the null space of \(B^T\). These assumptions entail that the system is nonsingular \([3]\). Relaxing these assumptions, we moreover include cases where \(K\) is singular because \(B^T\) and \(C\) have a common null space, while the system (1) is compatible. This often happens with linear systems resulting from the discretization of Stokes problems; see, e.g., \([3]\). We further refer to this work for an overview of the many applications in which such linear systems arise and the discussion of their algebraic properties, as well as for a general introduction to the different solution methods.

Iterative methods for saddle point linear systems often alternate updates of the first block of unknowns \(u_A\) (“velocity solves” in the context of Stokes problems) and updates of the second block \(u_C\) (“pressure corrections” in the same context). That is, each step corresponds to the combination of two or more substeps, in which the different types of unknowns are considered separately—reason for which such approaches are sometimes referred to as “segregated iterations”. They are also called “Schur complement approaches” because the updates of the second block \(u_C\) are based on the approximate solution of the Schur complement system obtained when eliminating the first block of unknowns. Such methods include (inexact) Uzawa \([6, 11, 23]\), SIMPLE \([10, 19]\), pressure correction techniques \([1, 13]\).

These methods are related to the solution process associated with the block factorization of the system matrix

\[
K = \begin{pmatrix}
A & \_ \\
B & -S
\end{pmatrix}
\begin{pmatrix}
\_ & I \\
A^{-1}B^T & I
\end{pmatrix},
\]

where the Schur complement \(S\) writes

\[
S = C + B A^{-1}B^T.
\]

Block preconditioners are also derived from this factorization. For instance, exchanging \(A\) and \(S\) in the right hand side of (3) for proper approximations (or preconditioners) \(M_A\) and \(M_S\) yields the inexact block factorization preconditioner \([2, 24]\)

\[
P = \begin{pmatrix}
M_A & \_ \\
B & -M_S
\end{pmatrix}
\begin{pmatrix}
\_ & I \\
M_A^{-1}B^T & I
\end{pmatrix}.
\]

A cruder approximation consists in discarding either the lower or the upper triangular factor, yielding, respectively, the inexact block triangular preconditioner \([5, 21]\)

\[
P = \begin{pmatrix}
M_A & B^T \\
\_ & -M_S
\end{pmatrix}.
\]
and the inexact Uzawa preconditioner

\[ P = \begin{pmatrix} M_A & -M_S \\ B & -M_S \end{pmatrix}. \]  

(7)

This latter name stems from the fact that using this preconditioner in stationary iterations just implements the inexact Uzawa algorithm [24]. More generally speaking, segregated iterations or Schur complement approaches mentioned above are intimately connected with block preconditioners, as will be further clarified in the next section.

For these methods or preconditioners, there exists several careful eigenvalue analyses and related spectral radius assessments which offer a quite complete picture of the situation; see, in particular, [17, 21, 24]. However, as is well known (e.g., [12]), while the eigenvalue analyses tell something about the asymptotic convergence, they ignore non-normality effects, and they can therefore not guaranty anything for a modest number of iterations (even with Krylov subspace acceleration).

On the other hand, several works prove monotone convergence of stationary iterations by bounding a well chosen norm of the iteration matrix. However, the state of the art seems less satisfactory than for the spectral radius assessment. The published works we are aware of are all restricted to the case where \( C = 0 \); see [4, 24] for the inexact Uzawa algorithm and [21, 22, 24] for the inexact block factorization preconditioner. Moreover, norms considered are nonstandard while restrictive assumptions are introduced on \( M_A \) and \( M_S \).

In this paper, we develop a convergence analysis which avoids unnecessarily restrictive assumptions while using a natural norm for the problem; that is, the energy norm \( \| u \|_E = (\cdot, E \cdot)^{1/2} \) associated with the matrix

\[ E = \begin{pmatrix} A \\ S \end{pmatrix}. \]  

(8)

Thus, for \( u = (u_A^T, u_C^T)^T \),

\[ \| u \|_E^2 = \| u_A \|_A^2 + \| u_C \|_S^2 \]  

combines the standard energy norm associated with the solution of a system with \( A \) (computation of \( u_A \) when \( u_C \) is fixed) and the standard energy norm associated with the Schur complement system.

For many methods belonging to the family just described, we prove that the iteration matrix \( T \) satisfies a relation of the form

\[ \| T^m \|_E \leq \xi \hat{\rho}^m \quad \text{or} \quad \| T^m \|_E \leq \xi \hat{\rho}^{m-1}, \]  

(10)

where the constants \( \xi \) and \( \hat{\rho} \) only depend on the parameters that naturally measure how good are the preconditioners \( M_A \) and \( M_S \); i.e., on

\[ \nu = \min \left( 1, \lambda_{\min} \left( M_A^{-1} A \right) \right), \quad \overline{\nu} = \max \left( 1, \lambda_{\max} \left( M_A^{-1} A \right) \right), \]  

(11)

\[ \nu = \min \left( 1, \lambda_{\min} \left( M_S^{-1} S \right) \right), \quad \overline{\nu} = \max \left( 1, \lambda_{\max} \left( M_S^{-1} S \right) \right), \]  

(12)
and the related spectral radii

\[
\rho_A = \rho(I - M_A^{-1}A) = \max \left( \mu - 1, 1 - \mu \right), \quad (13)
\]

\[
\rho_S = \rho(I - M_S^{-1}S) = \max \left( \nu - 1, 1 - \nu \right). \quad (14)
\]

Moreover, \( \xi \) is moderate and never larger than \( 3 + \sqrt{5} + O(\max(\rho_A, \rho_S)) \), while \( \hat{\rho} \) is close to the best spectral radius estimates. This way, we offer a middle ground between eigenvalue analyses which ignore possible nonnormality effects, and analyses focused on monotone convergence proofs.

Regarding assumptions, for some variant of segregated iterations close to the inexact block factorization preconditioner, nothing is needed besides \( \rho_A, \rho_S < 1 \); i.e., besides that the preconditioners \( M_A \) and \( M_S \) themselves define a convergent iteration for \( A \) and \( S \) (respectively). For other variants including inexact block triangular, inexact Uzawa and inexact block factorization preconditioners, some assumptions are needed on the largest eigenvalue of \( M_A^{-1}A \); i.e., on the scaling of \( M_A \). However, these assumptions are realistic in the sense that, when they are not satisfied, the convergence may deteriorate or cease.

Of course, all analyzed methods can be combined with Krylov subspace acceleration. This is not considered in this work, because, except for symmetric matrices with SPD preconditioners, it is nearly impossible to obtain rigorous convergence proofs in this context. For Krylov subspace methods that minimize the residual norm, one may use the general argument that the convergence should be at least as fast as with stationary iterations, but we note that this argument remains heuristic if the convergence of stationary iterations is proved in another norm than the residual norm. For the class of problems at hand, this sometimes motivates the use of MINRES combined with a block diagonal SPD preconditioner (not considered in this work), for which rigorous convergence analysis can be developed; see [20] and the references therein. Nevertheless, the family of methods considered here may lead to faster solution schemes [17], motivating the development of their convergence analysis.

The remainder of this work is organized as follows. In Section 2 we describe more in detail the iterative schemes under investigation, highlighting their close relationship. Singular systems are discussed in Section 3. The theoretical analysis is developed in Section 4, while Section 5 is devoted to a few illustrative numerical experiments. Concluding remarks are given in Section 6.

## 2 Segregated iterations and block preconditioners

A very general form of segregated iterations can be described as follows, where, considering updates of the first block of unknowns, we refer to the inverse preconditioner \( Y_A^{(i)} \) (\( i = 1, 2 \)) instead of the preconditioner \( M_A^{(i)} \). If this latter is nonsingular, we just have \( Y_A^{(i)} = M_A^{(i)-1} \), but it is not needed to assume this. In particular, this formalism allows us to include cases where either \( Y_A^{(1)} = 0 \) or \( Y_A^{(2)} = 0 \) (but not both), meaning in practice that the first \( (Y_A^{(1)} = 0) \) or the last \( (Y_A^{(2)} = 0) \) substep is skipped. In this way, we do not need to separately describe schemes with only two substeps.
1. Execute 1 stationary iteration for \( u_A \) with inverse preconditioner \( Y_A^{(1)} \):
\[
\tilde{u}_A^{(m+1)} = u_A^{(m)} + Y_A^{(1)} \left( b_A - A u_A^{(m)} - B^T u_C^{(m)} \right)
\]

2. Execute 1 stationary iteration for \( u_C \) with preconditioner \((-M_S)\):
\[
u_C^{(m+1)} = u_C^{(m)} - M_S^{-1} \left( b_C - B \tilde{u}_A^{(m+1)} + C u_C^{(m)} \right)
\]

3. Execute 1 stationary iteration for \( u_A \) with inverse preconditioner \( Y_A^{(2)} \):
\[
u_A^{(m+1)} = \tilde{u}_A^{(m+1)} + Y_A^{(2)} \left( b_A - A \tilde{u}_A^{(m+1)} - B^T u_C^{(m+1)} \right)
\]

For the updates of \( u_A \), one typically uses \( Y_A^{(i)} \) which approximate \( A^{-1} \); i.e., the first and last substeps approximately solve the first block of equations for \( u_A \) assuming \( u_C \) is fixed. Regarding the update of \( u_C \), one typically uses a preconditioner \( M_S \) that approximates the Schur complement \( S \). This is motivated by the observation that, if \( Y_A^{(i)} = A^{-1} \), then the exact \( u_C \) is provided by solving the Schur complement system; that is, by using \( M_S = S \) at the second step. Using then again the exact inverse of \( A \) as \( Y_A^{(i)} \), the last step will provide the exact solution for \( u_A \) as well; that is, with such “exact” preconditioners the above process just implements the solution of the linear system via the block factorization \( \mathcal{D} \).

When using standard (“inexact”) preconditioners, the above process is related with the block preconditioners mentioned in Section \( \mathcal{D} \). This relation is clarified in the following theorem. Before stating it, observe that, since this process combines stationary iterations, the corresponding iteration matrix is the product of the iteration matrices associated with the different substeps; that is, if \( (\tilde{u}_A^T \tilde{u}_C^T)^T \) denotes the exact solution of the linear system \( \mathcal{D} \), the matrix \( T \) such that
\[
\begin{pmatrix}
\tilde{u}_A - u_A^{(m+1)} \\
\tilde{u}_C - u_C^{(m+1)}
\end{pmatrix} = T \begin{pmatrix}
\tilde{u}_A - u_A^{(m)} \\
\tilde{u}_C - u_C^{(m)}
\end{pmatrix}
\]
satisfies
\[
T = T_{(Y_A^{(2)},0)} T_{(0,M_S^{-1})} T_{(Y_A^{(1)},0)}
\]
where
\[
T_{(Y_A^{(i)},0)} = I - \begin{pmatrix} Y_A^{(i)} & 0 \\ 0 & 0 \end{pmatrix} K = \begin{pmatrix} I - Y_A^{(i)} & -Y_A^{(i)} B^T \\ I & I \end{pmatrix}
\]
\[
T_{(0,M_S^{-1})} = I - \begin{pmatrix} 0 & -M_S^{-1} \\ -M_S^{-1} B & I - M_S^{-1} C \end{pmatrix}
\]
\[
\begin{pmatrix}
\overline{Y}_A^{(i)} - A \overline{Y}_A^{(i)} \\
\overline{Y}_A^{(i)} - A Y_A^{(i)}
\end{pmatrix} = M_S^{-1} T_{(0,M_S^{-1})}
\]

Theorem 1. Let \( K \) be a matrix of the form \( \mathcal{D} \) such that \( A \) is an \( n \times n \) matrix and \( C \) is a \( m \times m \) matrix. Let \( Y_A^{(1)} \), \( Y_A^{(2)} \) be \( n \times n \) matrices and \( M_S \) be a nonsingular \( m \times m \) matrix. Let \( T = T_{(Y_A^{(2)},0)} T_{(0,M_S^{-1})} T_{(Y_A^{(1)},0)} \) with \( T_{(Y_A^{(i)},0)} \) \((i = 1, 2)\) and \( T_{(0,M_S^{-1})} \) defined by \( \mathcal{D} \) and \( \mathcal{D} \), respectively. If \( Y_A^{(1)} + Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} \) is nonsingular, there holds
\[
T = I - P^{-1} K
\]
with

\[ P = \begin{pmatrix} I \\ B Y_A^{(1)} \\ I \end{pmatrix} \begin{pmatrix} \left( Y_A^{(1)} + Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} \right)^{-1} \\ -M_S \end{pmatrix} \begin{pmatrix} I \\ Y_A^{(2)} B^T \\ I \end{pmatrix}. \] (18)

**Proof.** The matrix \( Y \) such that \( T_{(0, M_S^{-1})}^T Y_A^{(1), 0} = I - Y K \) satisfies

\[ Y = \begin{pmatrix} Y_A^{(1)} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -M_S^{-1} \end{pmatrix} - \begin{pmatrix} 0 \\ -M_S^{-1} \end{pmatrix} K \begin{pmatrix} Y_A^{(1)} \\ 0 \end{pmatrix} = \begin{pmatrix} Y_A^{(1)} \\ M_S^{-1} BY_A^{(1)} - M_S^{-1} \end{pmatrix}. \]

Then, \( P \) satisfies

\[ P^{-1} \begin{pmatrix} I \\ B Y_A^{(1)} \\ I \end{pmatrix} = \begin{pmatrix} Y + \begin{pmatrix} Y_A^{(2)} \\ 0 \end{pmatrix} - \begin{pmatrix} Y_A^{(2)} \\ 0 \end{pmatrix} K Y \end{pmatrix} \begin{pmatrix} I \\ B Y_A^{(1)} \\ I \end{pmatrix} = \begin{pmatrix} Y_A^{(1)} \\ -M_S^{-1} \end{pmatrix} + \begin{pmatrix} Y_A^{(2)} \\ 0 \end{pmatrix} - \begin{pmatrix} Y_A^{(2)} A \end{pmatrix} \begin{pmatrix} Y_A^{(1)} \\ Y_A^{(2)} B^T M_S^{-1} \end{pmatrix} \begin{pmatrix} Y_A^{(1)} \\ -M_S^{-1} \end{pmatrix} = \begin{pmatrix} Y_A^{(1)} + Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} - Y_A^{(2)} B^T M_S^{-1} \end{pmatrix} = \begin{pmatrix} I \\ -Y_A^{(2)} B^T \\ I \end{pmatrix} \begin{pmatrix} Y_A^{(1)} \\ Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} - M_S^{-1} \end{pmatrix}. \]

Observe that if \( Y_A^{(1)} = M_A^{-1} \) while \( Y_A^{(2)} = 0 \), the preconditioner (18) corresponding to substeps 1 and 2 above (substep 3 being skipped) is the inexact Uzawa preconditioner (7). On the other hand, when \( Y_A^{(1)} = 0 \) and \( Y_A^{(2)} = M_A^{-1} \), the preconditioner (18) corresponding to substeps 2 and 3 above (substep 1 being skipped) is the block triangular preconditioner (6).

However, when \( Y_A^{(1)} = Y_A^{(2)} = M_A^{-1} \), we do not recover the inexact block factorization preconditioner (5), but a slightly different variant\(^1\)

\[ P = \begin{pmatrix} I \\ B M_A^{-1} \\ I \end{pmatrix} \begin{pmatrix} M_A \left( 2 M_A - A \right)^{-1} M_A \\ -M_S \end{pmatrix} \begin{pmatrix} I \\ M_A^{-1} B^T \\ I \end{pmatrix}. \] (19)

which we refer to as *symmetrized inexact block Gauss–Seidel preconditioner*. Observe that, \( A \) and \( M_A \) being SPD, \( 2 M_A^{-1} - M_A^{-1} A M_A^{-1} \) is nonsingular as required if and only if \( \rho_M < 2 \); that is, if and only if \( \rho_A < 1 \). Then, letting

\[ \overline{M_A} = M_A \left( 2 M_A - A \right)^{-1} M_A \] (20)

\(^1\)To implement the inexact block factorization preconditioner, one has to slightly modify the third step of the above procedure into \( u_A^{(m+1)} = u_A^{(m+1)} - Y_A^{(2)} B^T \left( u_C^{(m+1)} - u_C^{(m)} \right) \); see the discussion in [24, Section 2.2] for the rationale behind this.
there holds $\sigma(\overline{M}_A^{-1}A) \subset [1 - \rho_A^2, 1]$.

Conversely, let $Y_A^{(1)}, Y_A^{(2)}$ be any matrices such that $Y_A^{(1)} + Y_A^{(2)} - Y_A^{(1)} A Y_A^{(2)}$ is SPD

$$\overline{M}_A = \left( Y_A^{(1)} + Y_A^{(2)} - Y_A^{(1)} A Y_A^{(2)} \right)^{-1}$$

(21)

satisfying $\sigma(\overline{M}_A^{-1}A) \subset (0, 1]$. Then, there exist a SPD matrix

$$M_A = A^{1/2} \left( I - (I - A^{1/2} \overline{M}_A^{-1} A^{1/2})^{1/2} \right)^{-1} A^{1/2}$$

(22)

such that (20) holds. Moreover, (16) implies then

$$T_{(\overline{M}_A^{-1},0)} = T_{(Y_A^{(1)},0)} T_{(Y_A^{(2)},0)} = \left( T_{(M_A^{-1},0)} \right)^2$$

(23)

This allows us to state the following theorem, which is clearly inspired by the discussion in [9, Section 3]; see also [24, Remark 2.1], [8, Section 2.3] and Theorem 3.1 in [17], where it is proved that inexact Uzawa and block triangular preconditioners lead to the same eigenvalue distribution, which is further related to that associated with the symmetrized inexact block Gauss-Seidel preconditioner.

\textbf{Theorem 2.} Let $K$ be a matrix of the form (2) such that $A$ is an $n \times n$ SPD matrix and $C$ is a $m \times m$ matrix. Let $Y_A^{(1)}, Y_A^{(2)}$ be $n \times n$ matrices and $M_S$ be a nonsingular $m \times m$ matrix. Let $T = T_{(Y_A^{(2)},0)} T_{(0,M_S^{-1})} T_{(Y_A^{(1)},0)}$ with $T_{(Y_A^{(i)},0)} (i = 1, 2)$ and $T_{(0,M_S^{-1})}$ defined by (16) and (17), respectively. Assume that $Y_A^{(1)} + Y_A^{(2)} - Y_A^{(1)} A Y_A^{(2)}$ is SPD and such that, letting $\overline{M}_A$ be defined by (21), on has $\sigma(\overline{M}_A^{-1}A) \subset (0, 1]$.

Then, letting $M_A$ be defined by (22), there holds, for any integer $m \geq 2$,

$$T^m = T_{(Y_A^{(2)},0)} T_{(0,M_S^{-1})} T_{(M_A^{-1},0)} \hat{T}^{m-2} T_{(M_A^{-1},0)} T_{(0,M_S^{-1})} T_{(Y_A^{(1)},0)}$$

(24)

where

$$\hat{T} = T_{(M_A^{-1},0)} T_{(0,M_S^{-1})} T_{(M_A^{-1},0)} = I - P^{-1} K$$

(25)

with $P$ defined by (19).

Moreover, for any SPD matrix $\overline{M}_A$ satisfying $\sigma(\overline{M}_A^{-1}A) \subset (0, 1]$, the inexact Uzawa preconditioner

$$P_{uza} = \begin{pmatrix} \overline{M}_A & B \\ B^T & -M_S \end{pmatrix}$$

(26)

and the inexact block triangular preconditioner

$$P_{bt} = \begin{pmatrix} \overline{M}_A & B^T \\ B & -M_S \end{pmatrix}$$

(27)

\footnote{Compared with the expression appearing in [18], we deliberately revert the order of the terms $Y_A^{(1)}$ and $Y_A^{(2)}$, in view of the needs of Theorem 2 below.}
satisfy, respectively, for any integer \( m \geq 1 \),
\[
( I - P^{-1}_{11} K )^m = T_{(0, M_S^{-1})} T_{(M_A^{-1}, 0)} \hat{T}^{m-1} T_{(M_A^{-1}, 0)}
\]
and
\[
( I - P^{-1}_{11} K )^m = T_{(M_A^{-1}, 0)} \hat{T}^{m-1} T_{(M_A^{-1}, 0)} T_{(0, M_S^{-1})},
\]
where \( M_A \) and \( \hat{T} \) are defined by (22) and (25).

Proof. Straightforward corollary of Theorem 1 and equations (23). □

With this theorem, one sees that the analysis of many methods can be reduced to that of the symmetrized inexact block Gauss-Seidel preconditioner. This will be exploited in Section 4.

### 3 Compatible singular systems

Linear systems (1) arising from Stokes problems are often singular, \( B^T \) and \( C \) having both the constant vector in their null space, while \( b_C \) is orthogonal to this vector, ensuring that the system is compatible. In practice, compatible singular systems can most often be solved trouble free by the same iterative methods as regular ones. However, formulating the convergence theory in a way that comprises the singular case requires many cares for tiny details and a somewhat heavy formalism. To avoid this while not excluding an important class of applications, we show in this section that the convergence of the iterative schemes sketched in the previous section is, in the singular case, identical to that of these schemes applied to a regularized version of the problem.

So, assume that \( B^T \) and \( C \) have a common null space, and let \( V \) be a rectangular matrix whose columns form an orthonormal basis of this null space. Thus, the general assumption on \( C \) is weakened and we only require that it is positive definite on \( \mathcal{N}(B^T) \setminus \mathcal{R}(V) \). We also assume that the system (1) is compatible; i.e., that \( V^T b_C = 0 \).

Consider then sequences \( \bar{u}_A^{(m)}, \bar{u}_C^{(m)} \) generated by the substeps 1–3 sketched in the preceding section, in which one has set \( \bar{u}_A^{(0)} = u_A^{(0)} \) and \( \bar{u}_C^{(0)} = (I - VV^T)u_C^{(0)} \), and in which one has exchanged \( C \) and \( M_S^{-1} \) for, respectively,
\[
\bar{C} = C + VV^T \quad \text{and} \quad \bar{M}_S^{-1} = (I - VV^T)M_S^{-1}(I - VV^T) + VV^T.
\]

First observe that \( \bar{C} \) is positive definite on the null space of \( B^T \), implying that \( \bar{S} = \bar{C} + BA^{-1}B^T (= S + VV^T) \) is SPD while \( \bar{M}_S \) is SPD if \( M_S \) is SPD. Hence the sequences \( \bar{u}_A^{(m)}, \bar{u}_C^{(m)} \) are associated with a regular problem satisfying our general assumptions.

On the other hand, because \( CV = 0, B^T V = 0, V^T C = 0 \) (by symmetry of \( C \)), \( V^T B = 0 \) and \( V^T b_C = 0 \), one can see by induction that \( \bar{u}_A^{(m)} = u_A^{(m)} \) and \( \bar{u}_C^{(m)} = (I - VV^T)u_C^{(m)} \) hold for all \( m \). Therefore, if \( (\hat{u}_A^T, \hat{u}_C^T)^T \) denotes the exact solution of the linear system (1) satisfying \( V^T \hat{u}_C = 0 \), there holds
\[
(\hat{u}_A - u_A^{(m)})^T A(\hat{u}_A - u_A^{(m)}) + (\hat{u}_C - u_C^{(m)})^T S(\hat{u}_C - u_C^{(m)})
\]
\[
= (\hat{u}_A - u_A^{(m)})^T A(\hat{u}_A - u_A^{(m)}) + (\hat{u}_C - u_C^{(m)})^T S(\hat{u}_C - u_C^{(m)})
\]
\[
= \|\hat{u}_A - u_A^{(m)}\|^2_A + \|\hat{u}_C - u_C^{(m)}\|^2_S.
\]
Hence the convergence of the iterative method for the singular system in the seminorm associated with $A$ and $S$ is exactly the same as that of a companion regular problem in the norm $\| \cdot \|_E$ associated with $A$ and $S$.

Moreover, consider the eigenvalue problem
\[
M^{-1}Sz = \nu z \iff ((I - V V^T)M^{-1}S + V V^T)z = \nu z. \tag{28}
\]

Let $(\nu, z)$ be an eigenpair such that $\nu \neq 1$. Hence one has $V^Tz = 0$, implying that $(I - V(V^TM_SV)^{-1}V^TM_S)z$ is a nontrivial vector. Moreover, multiplying to the left both sides of (28) by $(I - V(V^TM_SV)^{-1}V^TM_S)$, one obtains, since $SZ = Sz$ and $V^TS = 0$,
\[
M^{-1}S z = \nu z. \tag{29}
\]

Conversely, if $(\nu, z)$ is an eigenpair of $M^{-1}S$ such that $\nu \neq 0$, there holds $V^TM_Sz = 0$, implying that $z = (I - V V^T)z$ is a nontrivial vector satisfying (28), as seen by multiplying to the left both sides of (29) by $I - V V^T$. In other words, we have just shown that
\[
\nu = \min \left(1, \lambda_{\min} \left( M^{-1}_S \right) \right) = \min \left(1, \min_{\nu \in \sigma(M^{-1}_S) \backslash \{0\}} \nu \right) \tag{30}
\]
and
\[
\nu = \min \left(1, \lambda_{\min} \left( M^{-1}_S \right) \right) = \max \left(1, \lambda_{\max} \left( M^{-1}_S \right) \right). \tag{31}
\]

Therefore, we may restrict ourselves in the next sections to the regular case (assuming that $C$ is positive definite on the null space of $B^T$), and, nevertheless, cover singular cases as considered here. To interpret correctly the stated convergence results in such cases, it suffices to exchange the standard definition of $\nu$ for (30) and to read the given bound on $\| T^m \|_E$ as a bound on the maximal relative error measured in the semi norm associated with $A$ and $S$.

4 Convergence analysis

4.1 Field of values of a class of saddle point matrices

Our analysis is based on a similarity transformation of the iteration matrix: $T = H^{-1}ZH$, allowing us to use
\[
\| T^m \|_E = \| E^{1/2}H^{-1}Z^m H E^{-1/2} \| \leq \| E^{1/2}H^{-1} \| \| Z^m \| \| H E^{-1/2} \|. \tag{32}
\]

Regarding $Z$, we do not analyze directly its norm, but its field of value (or numerical range)
\[
W(Z) = \{ z^* Z z : z \in \mathbb{C}^n, \| z \| = 1 \}
\]
and corresponding numerical radius
\[
w(Z) = \max_{z \in W(Z)} | z |.
\]
This indeed yields a bound on $\|Z^m\|$ using the well known inequalities $\|Z^m\| \leq 2w(Z^m)$ and $w(Z^m) \leq (w(Z))^m$ (for any matrix $Z$).

The matrices $Z$ to be analyzed will have the form

$$Z = I - \hat{K},$$

where $\hat{A}$ and $\hat{C}$ are SPD. Observe that, compared with (2), the second block of rows has opposite sign. Hence symmetry is lost but the symmetric part is positive definite, implying that the field of values is in the right half plane [4]. This field of values is further analyzed in the following theorem, which provides the main tools for the convergence analysis developed thereafter. This theorem can be seen as the counterpart of Theorem 4.1 in [17], where similar inequalities are proved regarding the eigenvalues of $\hat{K}$. In general, the inequalities in [17] are somewhat stricter, but the inequality associated with $\gamma$ is new. (Hence Theorem 4.1 in [17] would be enhanced by adding its counterpart for the eigenvalues.)

**Theorem 3.** Let $\hat{K}$ be a matrix of the form (31), where $\hat{A}$ is an $n \times n$ SPD matrix and $\hat{C}$ is an $m \times m$ SPD matrix with $m \leq n$. Let

$$S_{\hat{A}} = \hat{A} + \hat{B}^T \hat{C}^{-1} \hat{B}, \quad S_{\hat{C}} = \hat{C} + \hat{B} \hat{A}^{-1} \hat{B}^T.$$

Let

$$\begin{align*}
\alpha &= \min \left( \lambda_{\min}(\hat{A}), \lambda_{\min}(\hat{C}) \right), \\
\bar{\alpha} &= \max \left( \lambda_{\max}(\hat{A}), \lambda_{\max}(\hat{C}) \right), \\
\beta &= \max \left( \lambda_{\max}(S_{\hat{A}}), \lambda_{\max}(S_{\hat{C}}) \right), \\
\gamma &= \left\| \hat{C}^{-1/2} \hat{B} \hat{A}^{-1/2} \right\|, \\
\delta &= \|\hat{B}\|.
\end{align*}$$

There holds

$$W(I - \hat{K}) \subset \mathcal{F}(\alpha, \bar{\alpha}, \beta, \gamma, \delta)$$

with

$$\mathcal{F}(\alpha, \bar{\alpha}, \beta, \gamma, \delta) = \left\{ z \in \mathbb{C}^n \left| \begin{array}{l}
\alpha \leq \Re(1 - z) \leq \bar{\alpha} \quad \text{and} \\
|1 - z - \frac{\beta}{2}| \leq \frac{\beta}{2} \\
\text{and} \\
|\Im(z)| \leq \gamma \Re(1 - z) \quad \text{and} \\
|\Im(z)| \leq \delta \end{array} \right. \right\}.$$
Proof. For any $\mathbf{z}$ with $\|\mathbf{z}\| = 1$, $\lambda = \mathbf{z}^* \hat{K} \mathbf{z}$ satisfies $\Re(\lambda) = \mathbf{z}^* \hat{K}_s \mathbf{z}$ and $\Im(\lambda) = -i \mathbf{z}^* \hat{K}_a \mathbf{z}$, where the symmetric and antisymmetric parts of $\hat{K}$ satisfy, respectively,

$$
\hat{K}_s = \begin{pmatrix} \hat{A} \\ \hat{C} \end{pmatrix} \quad \text{and} \quad \hat{K}_a = \begin{pmatrix} -\hat{B} \\ \hat{B}^T \end{pmatrix}.
$$

The first and the last condition in (33) follow immediately, observing that $\|\hat{K}_a\| = \|\hat{B}\|$ because

$$
\hat{K}_a^T \hat{K}_a = \begin{pmatrix} \hat{B}^T \hat{B} \\ \hat{B} \hat{B}^T \end{pmatrix}.
$$

Regarding the third condition, $\gamma \Re(\lambda) \geq \pm \Im(\lambda)$ is guaranteed if

$$
\gamma \hat{K}_s \pm i \hat{K}_a = \begin{pmatrix} \gamma \hat{A} & \pm i \hat{B}^T \\ \mp i \hat{B} & \gamma \hat{C} \end{pmatrix}
$$

is Hermitian nonnegative definite. Because this matrix is Hermitian and $\hat{A}$ is SPD, this holds if and only if the Schur complement $\gamma \hat{C} - \gamma^{-1} \hat{B} \hat{A}^{-1} \hat{B}^T$ is nonnegative definite; that is, if and only if $\gamma \geq \|\hat{C}^{-1/2} \hat{B} \hat{A}^{-1/2}\|$.

On the other hand, the second condition in (33) follows from [18, Theorem 2.2] if there holds $\beta \geq \lambda_{\text{min}}(\frac{1}{2}(\hat{K}^{-1} + \hat{K}^{-T}))$, which we deduce from the relation

$$
\frac{1}{2}(\hat{K}^{-1} + \hat{K}^{-T}) = \begin{pmatrix} S^{-1}_A & S^{-1}_C \\ \overline{S}_A & S^{-1}_C \end{pmatrix}.
$$

To prove (34), we show that its right hand side corresponds to the maximal distance to origin of any boundary point of $\mathcal{F}(\alpha, 0, 1, \gamma, \infty)$ (i.e., $\mathcal{F}(\alpha, 0, \gamma, \infty)$) when skipping the conditions $\Re(1 - z) \leq \hat{\alpha}$ and $|\Im(z)| \leq \hat{\gamma}$. The geometry of the problem is depicted on Figure 1. Note that when $\beta < 2$, the distance to origin along the circle $|1 - z - \frac{\beta}{2}| = \frac{\beta}{2}$ is increasing with the real part of $z$. Hence, if the situation is as on the left picture, the maximal distance is $d(\alpha, p_1)$. If it is as on the right picture, the maximal distance

\[ p_2 = (\alpha = 0.7, \beta = 1, \gamma = 1) \]

\[ p_3 = (\alpha = 0.2, \beta = 1, \gamma = 0.8) \]
correspond to $p_2$ if it is reached along the circle, and to $p_3$ if it is reached along the line $\Re(z) = 1 - \alpha$. As the distance to origin from $p_2$ to $p_3$ along the line $\Im(z) = \gamma \Re(1 - z)$ is either monotonically increasing or monotonically decreasing or first decreasing and next increasing, there are in fact no more possibilities. Thus, when the situation is as on the right picture, the maximal distance to origin of a boundary point of $F(\alpha, \infty, \beta, \gamma, \infty)$ is the maximum of $(d(o, p_2), d(o, p_3))$. Since the situation of the left picture is characterized by $d(o, p_1) \leq d(o, p_2), d(o, p_3)$ and that of the right picture by $d(o, p_1) \geq d(o, p_2), d(o, p_3)$ we have thus proved

$$\max_{z \in F(\alpha, \infty, \beta, \gamma, \infty)} |z| \leq \min \left( d(o, p_1), \max(d(o, p_2), d(o, p_3)) \right).$$

The inequality (34) just translates this result, replacing $d(o, p_i)$ by $\sqrt{x_i^2 + y_i^2}$, where $x_i, y_i$ are obtained by solving the following couples of equations.

For $x_1, y_1$:

$$\begin{cases} (1 - x - \frac{\beta}{2})^2 + y^2 = \frac{\beta^2}{4} \\ x = 1 - \alpha \end{cases}$$

For $x_2, y_2$:

$$\begin{cases} (1 - x - \frac{\beta}{2})^2 + y^2 = \frac{\beta^2}{4} \\ y = \gamma(1 - x) \end{cases}$$

For $x_3, y_3$:

$$\begin{cases} x = 1 - \alpha \\ y = \gamma(1 - x) \end{cases}$$

4.2 Symmetrized inexact block Gauss–Seidel preconditioner

We now give our first convergence proof, which also provides the building blocks for the results presented in the next subsections.

**Theorem 4** (Analysis of the symmetrized inexact block Gauss–Seidel preconditioner). Let $K$ be a matrix of the form (2) such that $A$ is an $n \times n$ SPD matrix and $C$ is a $m \times m$ symmetric nonnegative definite matrix with $m \leq n$. Assume that $B$ has rank $m$ or that $C$ is positive definite on the null space of $B^T$. Let $P$ be the preconditioner (19) where $M_A$ and $M_S$ are SPD matrices. Let $\mu$, $\overline{\mu}$, $\nu$, $\overline{\nu}$, $\rho_A$, $\rho_S$ be defined by (11), (12), (13), (14), and assume $\rho_A, \rho_S < 1$.

Let $F(\alpha, \overline{\alpha}, \beta, \gamma, \delta)$ be the region (33) of the complex plane described by the following parameters:

$$\begin{align*}
\alpha &= (1 - \rho_A^2) \, \nu, \\
\overline{\alpha} &= \beta = \overline{\nu}, \\
\gamma &= \frac{\rho_A}{\sqrt{1 - \rho_A^2}}, \\
\delta &= \begin{cases} \rho_A \sqrt{\overline{\nu}(1 - \rho_A^2)} & \text{if } \rho_A^2 < \frac{1}{2} \\ \frac{1}{2} \sqrt{\overline{\nu}} & \text{otherwise.} \end{cases}
\end{align*}$$
There holds, for any positive integer \( m \),

\[
\| (I - P^{-1} K)^m \|_E \leq \xi \left( \max_{z \in \mathcal{F}(\alpha, \alpha, \beta, \gamma, \delta)} |z| \right)^m \quad (35)
\]

where

\[
\xi = \frac{\sqrt{\nu} \left( \bar{\mu} (3 + \sqrt{5}) + 2 (\bar{\mu} - 1)^2 \right)}{\min \left( \sqrt{1 - \rho_A^2} , \sqrt{\nu} \right)} . \quad (36)
\]

Moreover,

\[
\hat{\rho} = \max_{z \in \mathcal{F}(\alpha, \alpha, \beta, \gamma, \delta)} |z| \leq \sqrt{\rho_A^2 + \rho_S^2 - \rho_A^2 \rho_S^2} = \sqrt{1 - (1 - \rho_A^2)(1 - \rho_S^2)} < 1. \quad (37)
\]

**Proof.** Let \( \tilde{M}_A = M_A (2 M_A - A)^{-1} M_A \) and

\[
L = \begin{pmatrix} I & BM_{A}^{-1} I \\ \end{pmatrix}, \quad D = \begin{pmatrix} \tilde{M}_A & -M_S \\ \end{pmatrix},
\]

so that \( P = LDL^T \). Noting that \( \rho_A < 1 \) implies that \( \tilde{M}_A \) is SPD, let further

\[
F = \begin{pmatrix} (A^{1/2} \tilde{M}_A^{-1} A^{1/2})^{-1/2} \\ M_S^{1/2} S^{-1/2} \end{pmatrix} .
\]

Consider then

\[
Z = I - (F E^{1/2} L^T)^{-1} K (F E^{1/2} L^T)^{-1}.
\]

Because

\[
T^m = (F E^{1/2} L^T)^{-1} Z^m (F E^{1/2} L^T)
\]

there holds

\[
\| T^m \|_E \leq \| E^{1/2} (F E^{1/2} L^T)^{-1} \| \| Z^m \| \| (F E^{1/2} L^T) E^{-1/2} \| \leq 2 \| E^{1/2} L^{-T} E^{-1/2} \| \| E^{1/2} L^T E^{-1/2} \| \| F \| \| F^{-1} \| \| w(Z) \|_E .
\]

The inequality (35) will be obtained by showing that

\[
\| E^{1/2} L^{-T} E^{-1/2} \| \| E^{1/2} L^T E^{-1/2} \| \| F \| \| F^{-1} \| \leq \xi^2 \quad (38)
\]

and that \( W(Z) \subset \mathcal{F}(\alpha, \alpha, \beta, \gamma, \delta) \).

We first note that, since \( \sigma(A^{1/2} \tilde{M}_A^{-1} A^{1/2}) \subset [1 - \rho_A^2, 1] \),

\[
\| F \|^2 = (\lambda_{\min} (F^{-T} F^{-1}))^{-1} = (\min (1 - \rho_A^2 , \nu))^{-1} \quad (39)
\]

and

\[
\| F^{-1} \|^2 = \lambda_{\max} (F^{-T} F^{-1}) = \nu . \quad (40)
\]
We next observe that
\[ E^{1/2}L^{-T}E^{-1/2} = \begin{pmatrix} I & -A^{1/2}M^{-1}_A B^T S^{-1/2} \end{pmatrix} \]
and
\[ E^{1/2}L^T E^{-1/2} = \begin{pmatrix} I & A^{1/2} M^{-1}_A B^T S^{-1/2} \end{pmatrix} \]
have same norm, which is equal to the square root of the norm of
\[
\left( E^{1/2}L^T E^{-1/2} \right)^T E^{1/2}L^T E^{-1/2}
\]
\[
= \left( \begin{array}{cc}
I & \frac{A^{1/2} M^{-1}_A B^T S^{-1/2}}{2} \\
S^{-1/2}B & I + S^{-1/2}B M^{-1}_A A M^{-1}_A B^T S^{-1/2}
\end{array} \right)
\]
\[
= I + \begin{pmatrix} I \\ S^{-1/2}B A^{1/2} \end{pmatrix}
\begin{pmatrix}
0 & A^{1/2} M^{-1}_A A^{1/2} \\
A^{1/2} M^{-1}_A A^{1/2} & (A^{1/2} M^{-1}_A A^{1/2})^2
\end{pmatrix}
\begin{pmatrix}
I \\ A^{1/2} B^T S^{-1/2}
\end{pmatrix}
\]
Because \( \| A^{1/2} B^T S^{-1/2} \| = \lambda_{\max}(S^{-1/2} B A^{-1} B^T) \leq 1 \), we then find
\[
\| E^{1/2}L^T E^{-1/2} \|^2 \leq 1 + \lambda_{\max}(M^{-1}_A A) \| \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \|
+ \max(\lambda_{\max}( (M^{-1}_A A)^2 - M^{-1}_A A ) , 0)
\]
\[
= 1 + \bar{\mu} \left( \frac{1+\sqrt{5}}{2} \right) + \max(\bar{\mu}^2 - \bar{\mu} , 0)
\]
\[
= \bar{\mu} \left( \frac{3+\sqrt{5}}{2} \right) + (\bar{\mu} - 1)^2 .
\]
Remembering that \( \| E^{1/2}L^{-T}E^{-1/2} \| = \| E^{1/2}L^T E^{-1/2} \| \) this inequality, combined with (39), (40), proves (38) with \( \xi \) as in (36).
To prove (35), it remains then only to show that \( W(Z) \subset \mathcal{F}(\alpha, \beta, \gamma, \delta) \). One has
\[
I - Z = FD^{-1}L^{-1}K L^{-T}F^{-1}
\]
\[
= \left( A^{1/2} \bar{M}^{-1}_A A^{1/2} \right)^{-1/2} A^{1/2} \bar{M}^{-1}_A
\]
\[
- M_S^{-1/2}
\begin{pmatrix}
I & -B M^{-1}_A I \\
M^{-1}_A & A^{1/2} \bar{M}^{-1}_A A^{1/2} M_S^{-1/2}
\end{pmatrix}
\]
\[
\begin{pmatrix}
I \\ -M^{-1}_A B^T
\end{pmatrix}
\begin{pmatrix}
A^{1/2} \bar{M}^{-1}_A A^{1/2} \\
A^{1/2} \bar{M}^{-1}_A A^{1/2} M_S^{-1/2}
\end{pmatrix}
\]
\[
\begin{pmatrix}
A \\\n-B(I-M^{-1}_A B^T)
\end{pmatrix}
\begin{pmatrix}
(I - A M^{-1}_A B^T) \\
C + B(2M^{-1}_A - M^{-1}_A A M^{-1}_A) B^T
\end{pmatrix}
\begin{pmatrix}
A^{1/2} \bar{M}^{-1}_A A^{1/2} \\
A^{1/2} \bar{M}^{-1}_A A^{1/2} M_S^{-1/2}
\end{pmatrix}
\]
Let $X$ be an orthogonal matrix that diagonalizes $A^{1/2}M_A^{-1}A^{1/2}$: $X^T A^{1/2}M_A^{-1}A^{1/2}X = \Lambda$, where $\Lambda$ is the diagonal matrix with the eigenvalues of $M_A^{-1}A$ on its diagonal. Note that this matrix also diagonalizes $A^{1/2} \overline{M}_A^{-1}A^{1/2}$: $X^T A^{1/2} \overline{M}_A^{-1}A^{1/2}X = \overline{\Lambda}$ with $\overline{\Lambda} = 2\Lambda - \Lambda^2$. Let further $G = M_S^{-1/2}BA^{-1/2}X$ and $\overline{C} = M_S^{-1/2}CM_S^{-1/2}$. Because the numerical range is invariant under unitary similarity transformations, $W(Z) = W(I - \hat{K})$ with

$$\hat{K} = \left(\begin{array}{cc} X^T & I \\ I & \end{array}\right)(I - Z)\left(\begin{array}{cc} X & I \\ I & \end{array}\right) = \left(\begin{array}{cc} \overline{\Lambda} & \overline{\Lambda}^{1/2}(I - \Lambda)G^T \\ -G(I - \Lambda)\overline{\Lambda}^{1/2} & \end{array}\right) = \left(\begin{array}{cc} \hat{A} & \hat{B}^T \\ -\hat{B} & \hat{C} \end{array}\right).$$

We may apply Theorem 3 to this matrix. Moreover, the same matrix was already analyzed during the proof of Theorem 4.3 in [17], yielding bounds on most quantities involved:

$$\lambda_{\min}(\hat{A}) \geq 1 - \rho_A^2,$$
$$\lambda_{\max}(\hat{A}) \leq 1,$$
$$\lambda_{\max}(S_\hat{A}) \leq 1,$$
$$\lambda_{\min}(\hat{C}) \geq (1 - \rho_A^2)\nu,$$
$$\lambda_{\max}(\hat{C}) \leq \overline{\nu},$$
$$\lambda_{\max}(S_\hat{C}) \leq \overline{\nu},$$

and

$$\|\hat{B}\|^2 \leq \begin{cases} \overline{\nu}\rho_A^2(1 - \rho_A^2) & \text{if } \rho_A^2 < \frac{1}{2} \\ \frac{1}{7}\overline{\nu} & \text{otherwise.} \end{cases}$$

This straightforwardly provides the required result (35) if we can further show that $\|\hat{C}^{-1/2}\hat{B}\hat{A}^{-1/2}\| = \|\hat{C}^{-1/2}G(I - \Lambda)\| \leq \rho_A(1 - \rho_A^2)^{-1/2}$. In this view, observe that, for any $z$,

$$\frac{z^*G(I - \Lambda)^2G^Tz}{z^*\overline{C}z} \leq \frac{z^*G(I - \Lambda)^2G^Tz}{z^*G(2\Lambda - \Lambda^2)G^Tz} \leq \max_{\lambda \in [\mu, \overline{\nu}]} \frac{(1 - \lambda)^2}{2\lambda - \lambda^2} = \frac{\rho_A^2}{1 - \rho_A^2}.$$

It remains thus only to prove (37). Here we use (34) of Theorem 3 yielding

$$\left(\max_{z \in \mathcal{F}(\alpha, \nu, \beta, \gamma, \delta)} |z|^2\right) \leq \max\left(1 - \frac{\beta(2 - \beta)}{1 + \gamma}, \frac{1}{2} - 2\alpha + \alpha^2(1 + \gamma^2)\right) \leq \max\left(1 - (1 + \rho_S)(1 - \rho_S)(1 - \rho_A^2), 1 - 2(1 - \rho_A^2)(1 - \rho_S) + \frac{(1 - \rho_A^2)^2(1 - \rho_S)}{1 - \rho_A^2}\right) = \rho_A^2 + \rho_S^2 - \rho_A^2\rho_S^2.$$

Perhaps the best previous norm estimate for this preconditioner is the one that would be obtained by combining the results in [6, 24] for the inexact Uzawa algorithm with...
Theorem 2 in Section 2; see Section 4.4 for a short discussion of these results. Regarding eigenvalue analyses, the best spectral radius estimate is given in [17, Corollary 4.5]: \( \rho(T) \leq \max(\rho_A, \rho_S) \). On the other hand, Theorem 7 in Appendix shows that this latter bound is sharp in the two limit cases corresponding to either \( M_A = A \) or \( m = n \), \( C = 0 \) and \( M_S = S \). Observe that our bound (37) is thus also optimal in these cases as well. Moreover, it is never much larger than \( \max(\rho_A, \rho_S) \), and near equal if either \( \rho_A \ll \rho_S \) or \( \rho_S \ll \rho_A \).

4.3 Inexact block factorization preconditioner

The analysis of this scheme cannot be reduced to that of the symmetrized inexact block Gauss–Seidel preconditioner, but most of the previous proof can be reused.

**Theorem 5** (Analysis of inexact block factorization preconditioner). Let \( K \) be a matrix of the form (2) such that \( A \) is an \( n \times n \) SPD matrix and \( C \) is a \( m \times m \) symmetric nonnegative definite matrix with \( m \leq n \). Assume that \( B \) has rank \( m \) or that \( C \) is positive definite on the null space of \( B^T \). Let \( P \) be the preconditioner (5) where \( M_A \) and \( M_S \) are SPD matrices.

Let \( \hat{\rho}, \bar{\mu}, \bar{\nu}, \rho_A, \rho_S \) be defined by (11), (12), (13), (14), and assume \( \rho_A, \rho_S < 1 \).

Let \( F(\alpha, \bar{\mu}, \beta, \gamma, \delta) \) be the region (33) of the complex plane described by the following parameters:

\[
\alpha = \min \left( \mu, (1 - \rho_A^2) \nu \right), \\
\bar{\alpha} = \max (\bar{\mu}, \bar{\nu}), \\
\beta = \max \left( (2 - \bar{\mu})^{-1}, \bar{\nu} \right), \\
\gamma = \frac{\rho_A}{\sqrt{1 - \rho_A^2}}, \\
\delta = \begin{cases} \\
\sqrt{\bar{\nu}} \max \left( \frac{2\sqrt{3}}{9}, (\bar{\mu} - 1)\sqrt{\bar{\nu}} \right) & \text{if } \mu < \frac{1}{3} \\
\sqrt{\bar{\nu}} \max \left( (1 - \mu)\sqrt{\bar{\mu}}, (\mu - 1)\sqrt{\bar{\nu}} \right) & \text{otherwise}.
\end{cases}
\]

There holds, for any positive integer \( m \),

\[
\| (I - P^{-1}K)^m \|_E \leq \xi \left( \max_{z \in F(\alpha, \bar{\alpha}, \beta, \gamma, \delta)} |z| \right)^m \tag{42}
\]

where

\[
\xi = \frac{\max(\sqrt{\bar{\mu}}, \sqrt{\bar{\nu}}) (\bar{\mu} (3 + \sqrt{5}) + 2(\bar{\mu} - 1)^2)}{\min(\sqrt{\bar{\mu}}, \sqrt{\bar{\nu}})}. \tag{43}
\]

Moreover, if \( \bar{\mu} \leq 1 + \frac{\rho_S}{1 + \rho_A} \),

\[
\hat{\rho} = \max_{z \in F(\alpha, \bar{\alpha}, \beta, \gamma, \delta)} |z| \leq \max \left( \sqrt{\rho_A^2 + \rho_S^2 - \rho_A^2 \rho_S^2}, \rho_A \sqrt{\frac{3}{1 + \rho_A}} \right) \leq \max \left( \sqrt{1 - (1 - \rho_A^2)(1 - \rho_S^2)}, \sqrt{1 - (1 - \rho_A)^{1+2\rho_A}} \right) < 1. \tag{44}
\]
Avoiding to repeat identical steps, we have then
\[ F = \begin{pmatrix} (A^{1/2}M_{A}^{-1}A^{1/2})^{-1/2} & M_{S}^{1/2}S^{-1/2} \end{pmatrix}. \]

Proof. The proof follows that of Theorem \[ \text{[4]} \] in which we have just to exchange everywhere \( \hat{M}_{A} \) for \( M_{A} \), including in the definition of \( F \):

\[ \| F \|^{2} = (\lambda_{\min}(F^{-T}F^{-1}))^{-1} = (\min(\mu, \eta))^{-1} \]

and

\[ \| F^{-1} \|^{2} = \lambda_{\max}(F^{-T}F^{-1}) = \max(\bar{\mu}, \bar{\eta}) \]

instead of (39) and (40), while the other norm estimates are unchanged. This provides the value of \( \xi \) in (43).

Regarding \( Z \), we obtain here \( W(Z) = W(I - \hat{K}) \) with

\[ \hat{K} = \begin{pmatrix} \Lambda & \Lambda^{1/2}(I - \Lambda)\Lambda^{1/2} \hline -G(I - \Lambda)\Lambda^{1/2} & C + G(2\Lambda - A^{2})G^{T} \end{pmatrix} = \begin{pmatrix} \hat{A} & \hat{B}^{T} \hline -\hat{B} & \hat{C} \end{pmatrix}. \]

Again, the proof of Theorem 4.3 in [17] provides for this matrix bounds on \( \lambda_{\min}(\hat{A}) \), \( \lambda_{\min}(\hat{C}) \), \( \lambda_{\max}(\hat{A}) \), \( \lambda_{\max}(\hat{C}) \), \( \lambda_{\max}(S_{\hat{A}}) \), \( \lambda_{\max}(S_{\hat{C}}) \) and \( \| \hat{B} \| \) which straightforwardly justify the given expressions for \( \alpha, \bar{\alpha}, \beta, \delta \). Finally, \( \hat{C}^{-1/2}\hat{B}\hat{A}^{-1/2} \) takes here exactly the same expression as in the proof of Theorem 4, justifying to set \( \hat{\delta} = \frac{\hat{\rho}_{\hat{A}}}{1-\hat{\rho}_{\hat{A}}} \) here as well.

Regarding (44), first note that if \( \bar{\mu} \leq 1 + \frac{\hat{\rho}_{\hat{S}}}{1+\hat{\rho}_{\hat{S}}} \), then \( \beta \leq 1 + \rho_{S} \) as in Theorem 4. If, in addition, \( \min(1 - \rho_{A}, (1 - \rho_{A}^{2}))(1 - \rho_{S}) = (1 - \rho_{A}^{2})(1 - \rho_{S}) \), the parameter \( \alpha, \beta \) and \( \delta \) have all identical bounds, yielding

\[ \max\left(1 - \frac{\beta(2 - \beta)}{1+\gamma^{2}}, 1 - 2\alpha + \alpha^{2}(1 + \gamma^{2}) \right) \leq \rho_{A}^{2} + \rho_{S}^{2} - \rho_{A}^{2}\rho_{S}^{2} \]

with exactly the same arguments. On the other hand, if \( 1 - \rho_{A} < (1 - \rho_{A}^{2})(1 - \rho_{S}) \), we find

\[ \max\left(1 - \frac{\beta(2 - \beta)}{1+\gamma^{2}}, 1 - 2\alpha + \alpha^{2}(1 + \gamma^{2}) \right) \leq \max\left(1 - (1 + \rho_{S})(1 - \rho_{S})(1 - \rho_{A}^{2}), 1 - 2(1 - \rho_{A}) + \frac{(1 - \rho_{A}^{2})^{2}}{1 - \rho_{A}^{2}} \right) \]

\[ = \max\left(\rho_{A}^{2} + \rho_{S}^{2} - \rho_{A}^{2}\rho_{S}^{2}, \frac{2\rho_{A}^{2}}{1+\rho_{A}} \right). \]

The inequality (44) follows because \( \rho_{A}^{2} + \rho_{S}^{2} - \rho_{A}^{2}\rho_{S}^{2} > \frac{2\rho_{A}^{2}}{1+\rho_{A}} \) if and only if \( 1 - \rho_{A} > (1 - \rho_{A}^{2})(1 - \rho_{S}) \).

Assuming \( C = 0 \) and using the norm (9) with different weights for both terms in the right hand side, the analysis in [22] proves monotone convergence with factor \( \rho = \max(\rho_{A}, 2\rho_{S}/(1 - \rho_{S})) \). It is therefore better than ours if \( \rho_{S} \) is small enough but fails for larger \( \rho_{S} \). This is improved in [22] and further in [21] using a preconditioner dependent norm. However, besides \( C = 0 \), some strict conditions are still needed to get a useful
result; for instance, Theorem 5.3 in [24] fails to prove convergence if $\rho_A = \rho_S = \frac{1}{2}$ or if $\rho_A = \sqrt{2}$ and $\rho_S = 1 + \frac{\rho_S}{1 + \rho_S S} = \sqrt{2}$. (Observe that in the latter case our condition to have (44) holds.)

Nevertheless, this comparison shows that our analysis is suboptimal for small $\rho_S$, at least when $C = 0$. This is further confirmed by Theorem 7 in Appendix, which proves $\rho(T) = \rho_A$ when $m = n$, $C = 0$ and $M_S = S$. Hence it is likely a shortcoming if the bound (44) is less attractive when the maximum corresponds to the second term (i.e., when $\rho_S < \frac{\rho_A}{1 + \rho_A}$). On the other hand, if $\rho_A \to 0$, we have $\hat{\rho} \to \rho_S$; i.e., we recover the optimal result for $M_A = A$ (see Theorem 7).

Regarding the condition $\nu \leq 1 + \frac{\rho_S}{1 + \rho_S}$ to have (44), first note that it is not a requirement on the quality of $M_S$, but a condition on the scaling of $M_S$. Next, this condition is not always necessary to have $\hat{\rho} < 1$. For instance, if $\rho_A = \rho_S = \frac{1}{2}$, using the inequalities associated with $\alpha$, $\bar{\alpha}$, $\gamma$ and $\delta$, we get $\hat{\rho} \leq \frac{\sqrt{3}}{4}$. This is, however, significantly less interesting than $\sqrt{\rho_A^2 + \rho_S^2 - \rho_A^2 \rho_S^2} = \frac{\sqrt{7}}{4}$, but realistic in the sense that, in practice, the convergence may deteriorate or cease when $\nu \leq 1 + \frac{\rho_S}{1 + \rho_S}$ does not hold; see the numerical results in the next section.

4.4 Inexact Uzawa and block triangular preconditioners

When $\nu \leq 1$, it is straightforward to obtain a convergence proof for both preconditioners (6) and (7), by combining Theorem 4 with Theorem 2 and by supplying an analysis of the norm of the additional terms appearing in the expression of $T^m$. This is, basically, what is done in the next theorem, except that the analysis of these additional terms is grouped with that of the similarity transformation used in the proof of Theorem 4 in order to obtain a better estimate for the constant $\xi$ in (45).

**Theorem 6** (Analysis of inexact block triangular and inexact Uzawa preconditioners). Let $K$ be a matrix of the form (2) such that $A$ is an $n \times n$ SPD matrix and $C$ is a $m \times m$ symmetric nonnegative definite matrix with $m \leq n$. Assume that $B$ has rank $m$ or that $C$ is positive definite on the null space of $B^T$. Let $M_A$ and $M_S$ be SPD matrices, let $\nu$, $\rho$, $\mu$, $\alpha$, $\beta$, $\gamma$ and $\delta$ be defined by (11), (12), (13), (14), and assume $\rho_A, \rho_S < 1$ and $\nu \leq 1$.

Let $F(\alpha, \bar{\alpha}, \beta, \gamma, \delta)$ be the region (33) of the complex plane described by the following parameters:

$$\alpha = (1 - \rho_A) \nu, \quad \bar{\alpha} = \beta = 1, \quad \gamma = \frac{\sqrt{\rho_A}}{\sqrt{1 - \rho_A}}, \quad \delta = \begin{cases} \sqrt{\rho_A(1 - \rho_A)} & \text{if } \rho_A < \frac{1}{2} \\ \frac{1}{2} \sqrt{\frac{\rho_A}{\sqrt{1 - \rho_A}}} & \text{otherwise.} \end{cases}$$

There holds, for both preconditioners $P$ defined by (6) and (7)

$$\| (I - P^{-1}K)^m \|_E \leq \xi \max_{z \in F(\alpha, \bar{\alpha}, \beta, \gamma, \delta)} \left| z \right|^{m-1}$$

(45)
where $m$ is any positive integer and

$$\xi = \left( \frac{2\nu (1 + \nu)(2 + \nu(1 + \sqrt{5}))}{\nu} \right)^{1/2} \max \left( 1, \sqrt{\frac{\rho_A}{1 - \rho_A}} \right).$$ \hspace{1cm} (46)

Moreover,

$$\hat{\rho} = \max_{z \in \mathcal{F}(\alpha, \alpha, \beta, \gamma, \delta)} |z| \leq \sqrt{\rho_A + \rho_S^2 - \rho_A \rho_S^2} = \sqrt{1 - (1 - \rho_A)(1 - \rho_S^2)} < 1. \hspace{1cm} (47)$$

Proof. For the clarity of the proof, we adopt the notation of Theorem 2; that is, we exchange $M_A$ for $\overline{M}_A$ in the definition of the preconditioner, and let $M_A$ be defined by (22). This entails that, with $\lambda_{\max}(M_A^{-1}A) = 1$ and $\lambda_{\min}(M_A^{-1}A) = 1 - \rho_A$, one has $\lambda_{\max}(M_A^{-1}A) = 1$ and $\lambda_{\min}(M_A^{-1}A) = 1 - \sqrt{\rho_A}$ (which have to be used here instead of (11) and (13)).

We thus apply Theorem 2 and use (26) to denote the inexact Uzawa preconditioner while defining $\hat{T}$ by (25). We obtain, with $L, F$ and $Z$ as in the proof of Theorem 4

$$(I - P_{\text{aza}}^{-1}K)^m = T_{(0,M_A^{-1})} \ T_{(M_A^{-1},0)} \ \hat{T}^{m-1} \ T_{(M_A^{-1},0)} = T_{(0,M_A^{-1})} \ T_{(M_A^{-1},0)} \ L^{-T} E^{-1/2} F^{-1} \ Z^{m-1} \ F E^{1/2} L^T \ T_{(M_A^{-1},0)}.$$  \hspace{1cm} (48)

The norm of $Z$ can be analyzed exactly as in the proof of Theorem 4, taking only care to exchange $\rho_A$ for $\sqrt{\rho_A}$ in the obtained results. This yields (45) and (47) if we can prove

$$\|E^{1/2} T_{(0,M_A^{-1})} \ T_{(M_A^{-1},0)} \ L^{-T} E^{-1/2} F^{-1}\| \ \|F E^{1/2} L^T \ T_{(M_A^{-1},0)} E^{-1/2}\| \leq \xi.$$  \hspace{1cm} (49)

Setting

$$R = \begin{pmatrix} A^{1/2} \\ M_S^{1/2} \end{pmatrix},$$

this will be done by showing that

$$\|E^{1/2} R^{-1}\| \ \|R T_{(0,M_A^{-1})} \ R^{-1}\| \ \|R T_{(M_A^{-1},0)} \ L^{-T} E^{-1/2} F^{-1}\| \ \|F E^{1/2} L^T \ T_{(M_A^{-1},0)} E^{-1/2}\| \ \|R E^{-1/2}\| \leq \frac{\xi}{2}. \hspace{1cm} (50)$$

We proceed similarly for the inexact block triangular preconditioner: applying Theorem 2 yields

$$(I - P_{\text{bt}}^{-1}K)^m = T_{(M_A^{-1},0)} \ \hat{T}^{m-1} \ T_{(M_A^{-1},0)} \ T_{(0,M_S^{-1})} = T_{(M_A^{-1},0)} \ L^{-T} E^{-1/2} F^{-1} \ Z^{m-1} \ F E^{1/2} L^T \ T_{(M_A^{-1},0)} \ T_{(0,M_S^{-1})},$$

which provides the required results if there holds

$$\|E^{1/2} T_{(M_A^{-1},0)} \ L^{-T} E^{-1/2} F^{-1}\| \ \|F E^{1/2} L^T \ T_{(M_A^{-1},0)} T_{(0,M_S^{-1})} E^{-1/2}\| \leq \frac{\xi}{2}. \hspace{1cm} (51)$$
Using \( E^{1/2}T_{(M_A^{-1},0)} L^{-T}E^{-1/2}F^{-1} = (E^{1/2}R^{-1})(RT_{(M_A^{-1},0)} L^{-T}E^{-1/2}F^{-1}) \) and

\[
F E^{1/2}L^T T_{(M_A^{-1},0)} T_{(0,M_S^{-1})} E^{-1/2} = (F E^{1/2}L^T T_{(M_A^{-1},0)} R^{-1}) (RT_{(0,M_S^{-1})} R^{-1}) (R E^{-1/2})
\]

this is again proved by (48).

We are thus left with the analysis of all terms in the left hand side of (48). We find

\[
\| E^{1/2}R^{-1} \| \| RE^{-1/2} \| \leq \sqrt{\frac{\nu}{2}}.
\]

Further,

\[
RT_{(0,M_S^{-1})} R^{-1} = \begin{pmatrix}
I \\
M_S^{-1/2} B A^{-1/2} & I - M_S^{-1/2} C M_S^{-1/2}
\end{pmatrix}
\]

implying

\[
RT_{(0,M_S^{-1})} R^{-1} (RT_{(0,M_S^{-1})} R^{-1})^T = \begin{pmatrix}
I & (I - M_S^{-1/2} C M_S^{-1/2})^2 \\
(I - M_S^{-1} A M_S^{-1} C M_S^{-1/2}) (0 & I) & (I - A^{-1/2} B^T M_S^{-1/2})
\end{pmatrix}
\]

Since the eigenvalues of \( M_S^{-1/2} C M_S^{-1/2} \) are nonnegative and not larger than \( \nu < 2 \), the norm of the first term of the right hand side is equal to 1, whereas the norm the the second one is bounded by \( \|M_S^{-1/2} B A^{-1/2}\|^2 \) (\( \leq \nu \)) times

\[
\left\| \begin{pmatrix} 0 & I \\ I & I \end{pmatrix} \right\| = \frac{1 + \sqrt{5}}{2}.
\]

Thus we have proved

\[
\| RT_{(0,M_S^{-1})} R^{-1} \| \leq \left( 1 + \nu \frac{1 + \sqrt{5}}{2} \right)^{1/2}.
\]

Setting \( \Delta = A^{1/2} \bar{M}_A^{-1} A^{1/2} \), we pursue with the analysis of

\[
RT_{(M_A^{-1},0)} L^{-T}E^{-1/2}F^{-1} = \begin{pmatrix}
A^{1/2} \\
M_S^{1/2}
\end{pmatrix} \begin{pmatrix}
I - M_A^{-1} A & -M_A^{-1} B^T \\
I & I
\end{pmatrix} \begin{pmatrix}
I - M_A^{-1} B^T \\
I
\end{pmatrix}
\]

\[
= \begin{pmatrix}
A^{1/2} (A^{1/2} \bar{M}_A^{-1} A^{1/2})^{1/2} \\
M_S^{-1/2}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
(I - A^{1/2} M_A^{-1} A^{1/2}) \Delta^{1/2} A^{1/2} (M_A^{-1} A M_A^{-1} - 2 M_A^{-1}) B^T M_S^{-1/2} \\
I
\end{pmatrix}
\]

20
Using \((I - A^{1/2} M_{A}^{-1} A^{1/2})^2 = I - \Delta\) and \(A^{1/2}(M_{A}^{-1} A M_{A}^{-1} - 2 M_{A}^{-1}) A^{1/2} = -\Delta\), we then find

\[
(\nu + 1) I - RT_{(M_{A}^{-1}, 0)} L^{-T} E^{-1/2} F^{-1} \left( RT_{(M_{A}^{-1}, 0)} L^{-T} E^{-1/2} F^{-1} \right)^T
= \begin{pmatrix}
(\nu + 1) I - (I - \Delta) \Delta - \Delta A^{-1/2} B^{T} M_{S}^{-1} B^{T} A^{-1/2} \Delta
& \Delta A^{-1/2} B^{T} M_{S}^{-1/2}\end{pmatrix}
\]

and check that this latter matrix is nonnegative definite by computing the Schur complement associated with the elimination of the bottom right block:

\[
(\nu + 1) I - (I - \Delta) \Delta - (1 + \frac{1}{\nu}) \Delta A^{-1/2} B^{T} M_{S}^{-1} B^{T} A^{-1/2} \Delta
= (\nu + 1) \Delta (I - \frac{1}{\nu} A^{-1/2} B^{T} M_{S}^{-1} B^{T} A^{-1/2}) \Delta + (I - \Delta) + \nu (I - \Delta^2)
\]

is the sum of nonnegative definite terms because \(\Delta\) is SPD with \(\lambda_{\text{max}}(\Delta) \leq 1\) while

\[
\lambda_{\text{max}} \left( A^{-1/2} B^{T} M_{S}^{-1} B^{T} A^{-1/2} \right) = \lambda_{\text{max}} \left( M_{S}^{-1} B A^{-1/2} B^{T} \right) \leq \lambda_{\text{max}} \left( M_{S}^{-1} S \right) \leq \nu .
\]

We have thus proved

\[
\| RT_{(M_{A}^{-1}, 0)} L^{-T} E^{-1/2} F^{-1} \| \leq (1 + \nu)^{1/2}
\]

and it remains to analyze \(FE^{1/2} L^{T} T_{(M_{A}^{-1}, 0)} R^{-1} \). This is more straightforward because

\[
FE^{1/2} L^{T} T_{(M_{A}^{-1}, 0)} R^{-1} = \begin{pmatrix}
(A^{1/2} M_{A}^{-1} A^{1/2})^{-1/2} A^{1/2}
& I
\end{pmatrix}
\begin{pmatrix}
I
& M_{A}^{-1} A
M_{S}^{1/2}
& I
\end{pmatrix}
\begin{pmatrix}
I
& -M_{A}^{-1} A
-I
& I
\end{pmatrix}
\begin{pmatrix}
A^{-1/2}
& M_{A}^{-1} A^{1/2}
M_{S}^{-1/2}
& I
\end{pmatrix}
\begin{pmatrix}
I
& A^{-1/2} M_{A}^{-1} A^{1/2}
\end{pmatrix}
\]

Hence the square norm is equal to the maximum of 1 and

\[
\| (A^{1/2} M_{A}^{-1} A^{1/2})^{-1} (I - A^{1/2} M_{A}^{-1} A^{1/2})^2 \| = \| A^{1/2} M_{A}^{-1} A^{1/2} (I - A^{1/2} M_{A}^{-1} A^{1/2}) \|
= \max_{\xi \in [1 - \rho_{A}, 1]} | \frac{1}{\xi} - 1 | ;
\]

that is, \(\| FE^{1/2} L^{T} T_{(M_{A}^{-1}, 0)} R^{-1} \| \leq \max \left( 1, \sqrt{\frac{\rho_{A}}{1 - \rho_{A}}} \right)\). Combining this and the above inequalities yields \([48]\), concluding the proof.

In \([3]\) Theorem 3.1, monotone convergence of the inexact Uzawa algorithm is proved with a factor that is only slightly larger than the right hand side of \([47]\), using a rather exotic norm and assuming \(C = 0, \nu \leq 1\) and \(\nu \leq 1\). Using another preconditioner dependent norm, the same bound is recovered in \([23]\) (see Remark 4.4 in this work), as a particular case of a more general result compatible with \(\nu, \nu > 1\). However, it is
clear from the discussion in [17] that the convergence estimate rapidly deteriorates when \( \rho \) increases above 1. This is confirmed with Theorem 7 in Appendix, which proves that \( \rho(T) \geq \rho - 1 + \sqrt{\rho(\rho - 1)} \) in the limit case \( m = n, C = 0 \) and \( M_S = S \). One may then see that convergence is lost as soon as \( \rho \geq 4/3 \).

On the other hand, staying with the condition \( \rho \leq 1 \), the best spectral radius estimate for the inexact Uzawa and block triangular preconditioners is given in [17, Corollary 4.5] (combining results from there and Theorem 2 in [21]): \( \rho(T) \leq \max(\sqrt{\rho_A}, \rho_S) \). Here again, see Theorem 7, this result is sharp in the two limit cases corresponding to either \( M_A = A \) or \( m = n, C = 0 \) and \( M_S = S \). Hence our bound (37) is optimal in these case as well, while being never much larger than \( \max(\sqrt{\rho_A}, \rho_S) \), and near equal if either \( \sqrt{\rho_A} \ll \rho_S \) or \( \rho_S \ll \sqrt{\rho_A} \).

5 Numerical illustration

Here we illustrate our theoretical result with a simple example. We consider the linear system resulting from the discretization of a time dependent Stokes problem defined on the unit square with Dirichlet boundary conditions for velocities, using for the spatial discretization the marker and cell (MAC) scheme on a uniform staggered grid with mesh size \( h = 1/40 \), and for the time discretization the backward Euler method with time step \( \tau = 1/100 \). Since this latter scheme is implicit, a system of the form (1) has to be solved at each time step. Here we consider a sample obtained setting \( b_A \) to a random vector of unit norm, while \( b_C = 0 \).

As is well known, \( A \) contains two diagonal blocks, and each of them corresponds to a negative discrete Laplace operator plus a constant stemming from the time discretization. Since this latter term implies strict diagonal dominance, it is consistent to use for \( M_A \) the symmetrized Gauss-Seidel preconditioner \( M_A = \text{low}(A)(\text{diag}(A))^{-1}\text{upp}(A) \). However, multigrid techniques can be more efficient, hence we also tested \( M_A \) as resulting from one application of the preconditioner implemented in the AGMG software package [15] (see [13] [16] for a description of the method). For \( M_S \), we selected the Cahouet–Chabard preconditioner [7] which is state-of-the-art for time dependent Stokes problems. Its inverse amounts to a constant plus another constant times the inverse of a negative discrete Laplacian on the pressure space. Here, we further exchange this inverse for one application of the AGMG preconditioner.

For the sake of brevity, we keep the description of these settings minimal. In the context of the present study, it is likely more important to have a look at the parameters involved in the theoretical analysis. Computed values are reported in Table 5 where \( \omega_A \) and \( \omega_S \) refer to the overrelaxation factor applied when using \( M_A \) and \( M_S \) (respectively); that is, \( \omega_A \) and \( \omega_S \) are the inverse of the scaling factor applied to \( M_A \) and \( M_S \). Indeed, without rescaling or (over)relaxation (\( \omega_A, \omega_S = 1 \)), all largest eigenvalues are equal to 1. Hence there is room to decrease the corresponding spectral radius by rescaling the preconditioner, and the provided values of \( \omega_A \) and \( \omega_S \) correspond to a rough optimization of \( \rho_A \) and \( \rho_S \).

The convergence curves for the three main methods investigated here are reported in Figure 2 and 3. One sees that the symmetrized inexact block Gauss–Seidel preconditioner is the winner or among the winners in all cases. Moreover, as predicted by our bound (37),
Table 1: Parameters for the example as a function of the applied overrelaxation factor $\omega_A$ or $\omega_S$.

<table>
<thead>
<tr>
<th></th>
<th>$\omega_A$</th>
<th>$\mu$</th>
<th>$\mu$</th>
<th>$\rho_A$</th>
<th>$\rho_A^2$</th>
<th>$\omega_A$</th>
<th>$\mu$</th>
<th>$\mu$</th>
<th>$\rho_A$</th>
<th>$\rho_A^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGS for $A$</td>
<td>1.0</td>
<td>0.07</td>
<td>1.0</td>
<td>0.93</td>
<td>0.86</td>
<td>1.7</td>
<td>0.12</td>
<td>1.7</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td>AGMG for $A$</td>
<td>1.0</td>
<td>0.54</td>
<td>1.0</td>
<td>0.46</td>
<td>0.21</td>
<td>1.3</td>
<td>0.70</td>
<td>1.3</td>
<td>0.30</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$\omega_S$</td>
<td>$\nu$</td>
<td>$\nu$</td>
<td>$\rho_S$</td>
<td>$\rho_S^2$</td>
<td>$\omega_S$</td>
<td>$\nu$</td>
<td>$\nu$</td>
<td>$\rho_S$</td>
<td>$\rho_S^2$</td>
</tr>
<tr>
<td>........</td>
<td>1.0</td>
<td>0.28</td>
<td>1.0</td>
<td>0.72</td>
<td>0.51</td>
<td>1.5</td>
<td>0.43</td>
<td>1.5</td>
<td>0.57</td>
<td>0.33</td>
</tr>
</tbody>
</table>

improving $\rho_A$ with rescaling enhances the convergence when $\rho_A$ dominates $\rho_S$ (Figure 2) and has little effect otherwise (Figure 3). Conversely, improving $\rho_S$ with rescaling enhances the convergence when $\rho_S$ dominates $\rho_A$ (Figure 3) and has little effect otherwise (Figure 2).

Regarding the inexact block factorization preconditioner, it is generally as fast as the symmetrized inexact block Gauss–Seidel preconditioner when our condition $\mu \leq 1 + \frac{\rho_S}{1 + \rho_A}$ to have (44) holds (it is slightly slower in the bottom right picture of Figure 3 but the difference is not significant). Otherwise, it may be competitive (Figure 2, bottom left picture), but convergence may also be lost (Figure 2, bottom right picture).

As expected the inexact Uzawa algorithm never converges when rescaling $M_A$. Hence the restriction $\mu \leq 1$ in our theoretical analysis is a sensible one. Next, again as expected from the theory, this method converges roughly as fast as the more costly symmetrized inexact block Gauss–Seidel preconditioner if and only if $\rho_S^2$ dominates $\rho_A$. This is what happens in Figure 3 top left picture. Observe, however, that when rescaling $M_S$ (top right picture, $\rho_S^2$ still dominates $\rho_A^2$, but $\rho_S^2 \approx \rho_A$, entailing that this rescaling little helps the inexact Uzawa method while it significantly improves the symmetrized inexact block Gauss–Seidel preconditioner.

6 Conclusions

We have developed a rigorous convergence analysis which applies to many segregated iteration schemes and related block preconditioners. It is more general and more accurate than previous ones, and therefore helps to clarify which conditions are needed for a given scheme to converge, as well as its respective potentialities.

The symmetrized inexact Gauss-Seidel preconditioner (19) (see Section 2 for its definition as segregated iteration) can be recommended unconditionally, as it converges with near optimal convergence rate as soon as $\rho_A < 1$ and $\rho_S < 1$.

The inexact block factorization preconditioner (5) has about the same cost per iteration and is often as fast, but it needs to satisfy a scaling condition on $\lambda_{\text{max}}(M_A^{-1} A)$ to be on the safe side. This, in turn, may prevent the improvement of the convergence via a scaling of $M_A$ aiming at minimizing $\rho_A$.

The inexact Uzawa algorithm and the closely related block triangular preconditioners (6), (7) need also to satisfy a scaling condition to define a convergent iteration. Although our assumption $\lambda_{\text{max}}(M_A^{-1} A) \leq 1$ is likely somewhat stricter than necessary, it is clear that
Figure 2: Residual norm as a function of the number of iterations with SGS preconditioner for $A$ ($\rho_A > \rho_S$ in all cases, whereas $\overline{\mu} = 1$ for both top pictures and $\overline{\mu} > 1 + \frac{\rho_S}{\Gamma + \rho_S} > 1$ for both bottom pictures).

\begin{align*}
\omega_A = 1, \quad \omega_S = 1 & \quad (\rho_S^2 > \rho_A) \\
\omega_A = 1.7, \quad \omega_S = 1 & \quad (\rho_S^2 \approx \rho_A \gg \rho_A^2)
\end{align*}

Figure 3: Residual norm as a function of the number of iterations with AGMG preconditioner for $A$ ($\rho_S > \rho_A$ in all cases, whereas $\overline{\mu} = 1$ for both top pictures and $1 < \overline{\mu} < 1 + \frac{\rho_S}{\Gamma + \rho_S}$ with $\rho_S^2 \gg \rho_A^2$ for both bottom pictures).

$$\omega_A = 1, \quad \omega_S = 1$$

$$\omega_A = 1.3, \quad \omega_S = 1$$

$$\omega_A = 1, \quad \omega_S = 1.5$$

$$\omega_A = 1.3, \quad \omega_S = 1.5$$
convergence ceases when it is strongly violated. These approaches are further penalized by a less interesting convergence rate, but are cheaper per iteration, as they require only one approximate solve with $A$ where the symmetrized inexact Gauss-Seidel preconditioner requires two. They can thus be cost effective, and are in fact recommended if $\rho_S^2 \gg \rho_A$. Opposite to this, if $\rho_A \gg \rho_S$, the symmetrized inexact Gauss-Seidel preconditioner is likely better despite its higher cost, because its faster convergence can further be enhanced by optimizing $\rho_A$ with the rescaling of $M_A$. In intermediate situations, both approach are viable, and which one is most cost effective is likely application dependent.

**Appendix: Exact spectral radius in some limit cases**

**Theorem 7.** Let $K$ be a matrix of the form (2) such that $A$ is an $n \times n$ SPD matrix and $C$ is a $m \times m$ symmetric nonnegative definite matrix with $m \leq n$. Assume that $B$ has rank $m$ or that $C$ is positive definite on the null space of $B^T$. Let $M_A$ and $M_S$ be SPD matrices.

1. If $m = n$, $C = 0$ and $M_S = BA^{-1}B^T = S$, then, for the inexact block factorization preconditioner (5) and the symmetrized inexact block Gauss–Seidel preconditioner (19) there holds

$$\rho \left( I - P^{-1}K \right) = \rho_A,$$

(49) whereas, for the inexact block triangular preconditioner (6) and the inexact Uzawa preconditioner (7) one has

$$\rho \left( I - P^{-1}K \right) = \max \left( 1 - \mu, \bar{\mu} - 1 + \sqrt{\mu(\bar{\mu} - 1)} \right).$$

(50)

2. If $M_A = A$, for all four preconditioners defined by (5), (6), (7), (19), one has

$$\rho \left( I - P^{-1}K \right) = \rho_S.$$  

(51)

**Proof.** To prove (49), we can consider that $I - T$ is similar to the matrix $\tilde{K}$ seen in the proof of Theorem 4 (eq. (41)), where $\Lambda$ is diagonal with diagonal entries equal to the eigenvalues of $M_A^{-1}A$, where $\Lambda = \Lambda$ for the inexact block factorization preconditioner (5) and $\tilde{\Lambda} = 2\Lambda - \Lambda^2$ for the symmetrized inexact block Gauss–Seidel preconditioner (19), where $\tilde{C} = M_S^{-1/2}C M_S^{-1/2}$ is here the zero matrix, and where $G = M_S^{1/2}BA^{-1/2}X$ with $X$ orthogonal is here orthogonal because $M_S = S = BA^{-1}B^T$; that is, $T$ is similar to $I - \begin{pmatrix} \tilde{\Lambda} & \tilde{\Lambda}^{1/2}(I - \Lambda) \\ -(I - \Lambda)\tilde{\Lambda}^{1/2} & 2\Lambda - \Lambda^2 \end{pmatrix}$

$$= \begin{pmatrix} I - \Lambda & -\tilde{\Lambda}^{1/2}(I - \Lambda) \\ (I - \Lambda)\tilde{\Lambda}^{1/2} & (I - \Lambda)^2 \end{pmatrix}.$$  

Thus, for each eigenvalue $\lambda \in \sigma(M_A^{-1}A)$ there are two eigenvalues of $T$, which are equal to the eigenvalues of:

Prec. (5): $(1 - \lambda) \begin{pmatrix} 1 & -\sqrt{\lambda} \\ \sqrt{\lambda} & 1 - \lambda \end{pmatrix}$; Prec. (19): $(1 - \lambda) \begin{pmatrix} 1 - \lambda & -\sqrt{2\lambda - \lambda^2} \\ \sqrt{2\lambda - \lambda^2} & 1 - \lambda \end{pmatrix}$.  

25
In both cases, the second term has two complex conjugate eigenvalues of unit modulus; that is, the modulus of both eigenvalue is equal to $|1 - \lambda|$, proving (49).

For (50), we first note that by Theorem 2, the spectral radius of the iteration matrix is necessarily the same for the inexact block triangular and the inexact Uzawa preconditioners (see also [17, Theorem 3.1]). Hence we may analyze this latter only, for which we find, using $I = S^{-1}BA^{-1}B$,

$$
T = I - \begin{pmatrix}
    M_A^{-1} & A \\
    S^{-1}B M_A^{-1} & B
\end{pmatrix}
= \begin{pmatrix}
    I - M_A^{-1}A & -M_A^{-1}B^T \\
    S^{-1}B(I - M_A^{-1}A) & I - S^{-1}B M_A^{-1}B^T
\end{pmatrix}
= \begin{pmatrix}
    A^{-1/2} & (I - A^{1/2}M_A^{-1}A^{1/2})
\end{pmatrix}
\begin{pmatrix}
    I - A^{1/2}M_A^{-1}A^{1/2} & -A^{1/2}M_A^{-1}A^{1/2} \\
    S^{-1}B A^{-1/2} & I - A^{1/2}M_A^{-1}A^{1/2}
\end{pmatrix}
= \begin{pmatrix}
    A^{1/2} & A^{-1/2}B^T
\end{pmatrix}
$$

Each eigenvalue $\lambda \in \sigma(M_A^{-1}A)$ leads thus to two eigenvalues of $T$, which are here equal to the eigenvalues of

$$
\begin{pmatrix}
    1 - \lambda & -\lambda \\
    1 - \lambda & 1 - \lambda
\end{pmatrix}.
$$

If $\lambda \leq 1$, there are 2 complex conjugate eigenvalues of modulus $\sqrt{1 - \lambda}$, whereas, if $\lambda > 1$, the eigenvalues are $1 - \lambda \pm \sqrt{\lambda^2 - \lambda}$, proving (50).

Regarding (51), we first observe that, when $M_A = A$, there is no more difference between the inexact block factorization preconditioner ([5]) and the symmetrized inexact block Gauss–Seidel preconditioner ([19]). Moreover, $T_{(A^{-1},0)} = (T_{(A^{-1},0)})^2$, hence one sees with Theorem 2 that the symmetrized inexact block Gauss–Seidel preconditioner and both triangular preconditioners lead to the same spectral radius for the iteration matrix. Therefore, we may consider only one of the four preconditioners; e.g., the inexact Uzawa preconditioner, for which $T = T_{(0,M_S^{-1})}T_{(A^{-1},0)}$ satisfies

$$
T = \begin{pmatrix}
    I & 0 \\
    M_S^{-1}B & I - M_S^{-1}C
\end{pmatrix}
= \begin{pmatrix}
    0 & -A^{-1}B^T \\
    0 & I - M_S^{-1}S
\end{pmatrix},
$$

yielding straightforwardly the required result.

References


