CONVERGENCE OF SOME ITERATIVE METHODS FOR SYMMETRIC SADDLE POINT LINEAR SYSTEMS[*]

YVAN NOTAY[†]

Abstract. We consider the iterative solution of linear systems with a symmetric saddle point system matrix. We address a family of solution techniques that exploit the knowledge of a preconditioner (or approximate solution procedure) both for 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" or "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 classifications. 65F08, 65F10, 65F50, 65N22

DOI. 10.1137/18M1208836

1. Introduction. We consider linear systems of the form

$$(1.1) \qquad \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} \mathbf{u}_A \\ \mathbf{u}_C \end{pmatrix} = \begin{pmatrix} \mathbf{b}_A \\ \mathbf{b}_C \end{pmatrix},$$

where the system matrix

$$(1.2) \qquad K = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

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 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 (1.1) 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 $\mathbf{u}_A$ ("velocity solves" in the context of Stokes problems) and updates of the second block $\mathbf{u}_C$ ("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—reasons for which such

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

approaches are sometimes referred to as "segregated iterations." They are also called "Schur complement approaches" because the updates of the second block $\mathbf{u}_C$ 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, 12, 25], SIMPLE [10, 20], and pressure correction techniques [1, 14].

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

(1.3)
$$K = \begin{pmatrix} A & \\ B & -S \end{pmatrix} \begin{pmatrix} I & A^{-1}B^T \\ & I \end{pmatrix},$$

where the (negative) Schur complement S is given by

(1.4)
$$S = C + B A^{-1} B^T.$$

Block preconditioners are also derived from this factorization. For instance, exchanging A and S in the right-hand side of (1.3) for proper approximations (or preconditioners) M_A and M_S yields the *inexact block factorization preconditioner* [2, 26]

(1.5)
$$P = \begin{pmatrix} M_A & \\ B & -M_S \end{pmatrix} \begin{pmatrix} I & M_A^{-1}B^T \\ & I \end{pmatrix}.$$

A cruder approximation consists in discarding either the lower or the upper triangular factor, yielding, respectively, the *inexact block triangular preconditioner* [5, 23]

(1.6)
$$P = \begin{pmatrix} M_A & B^T \\ & -M_S \end{pmatrix}$$

and the *inexact Uzawa* preconditioner

(1.7)
$$P = \begin{pmatrix} M_A & \\ B & -M_S \end{pmatrix}.$$

This latter name stems from the fact that using this preconditioner in stationary iterations just implements the inexact Uzawa algorithm [26]. 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 exist several careful eigenvalue analyses and related spectral radius assessments which offer a quite complete picture of the situation; see, in particular, [18, 23, 26]. However, as is well known (e.g., [13]), while the eigenvalue analyses tell something about the asymptotic convergence, they ignore nonnormality effects, and they can therefore not guarantee 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 [6, 26] for the inexact Uzawa algorithm and [2, 24, 26] 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 $\| \cdot \|_E = (\cdot, E \cdot)^{1/2}$ associated with the matrix

(1.8)
$$E = \begin{pmatrix} A & \\ & S \end{pmatrix}.$$

Thus, for $\mathbf{u} = \left(\mathbf{u}_A^T \ \mathbf{u}_C^T \right)^T$,

$$(1.9) \qquad \|\mathbf{u}\|_E^2 = \|\mathbf{u}_A\|_A^2 + \|\mathbf{u}_C\|_S^2$$

combines the standard energy norm associated with the solution of a system with A (computation of $\mathbf{u}_A$ when $\mathbf{u}_C$ 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

$$(1.10) \qquad \|T^m\|_E \ \leq \ \xi \, \hat{\rho}^m \quad \text{or} \quad \|T^m\|_E \ \leq \ \xi \, \hat{\rho}^{m-1},$$

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

$$(1.11) \qquad \underline{\mu} = \min\left(1, \lambda_{\min}\left(M_A^{-1}A\right)\right), \qquad \overline{\mu} = \max\left(1, \lambda_{\max}\left(M_A^{-1}A\right)\right),$$

$$(1.12) \qquad \underline{\nu} = \min\left(1, \lambda_{\min}\left(M_S^{-1}S\right)\right), \qquad \overline{\nu} = \max\left(1, \lambda_{\max}\left(M_S^{-1}S\right)\right),$$

and the related spectral radii

$$(1.13) \qquad \rho_A = \rho\left(I - M_A^{-1}A\right) = \max\left(\overline{\mu}-1, \, 1-\underline{\mu}\right),$$

$$(1.14) \qquad \rho_S = \rho\left(I - M_S^{-1}S\right) = \max\left(\overline{\nu}-1, \, 1-\underline{\nu}\right).$$

Moreover ξ is moderate and never larger than $3 + \sqrt{5} + \mathcal{O}\left(\max(\rho_A, \rho_S)\right)$, 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 a norm other 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 a rigorous convergence analysis can be developed; see [21] and the references therein. Nevertheless, the family of methods considered here may lead to faster solution schemes [18], motivating the development of their convergence analysis.

The remainder of this work is organized as follows. In section 2 we describe in more 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* $\mathbf{u}_A$ *with inverse preconditioner* $Y_A^{(1)}$:
$$\widetilde{\mathbf{u}}_A^{(m+1)} = \mathbf{u}_A^{(m)} + Y_A^{(1)} \left(\mathbf{b}_A - A\,\mathbf{u}_A^{(m)} - B^T \mathbf{u}_C^{(m)} \right).$$

2. *Execute* 1 *stationary iteration for* $\mathbf{u}_C$ *with preconditioner* $(-M_S)$:
$$\mathbf{u}_C^{(m+1)} = \mathbf{u}_C^{(m)} - M_S^{-1} \left(\mathbf{b}_C - B\,\widetilde{\mathbf{u}}_A^{(m+1)} + C\,\mathbf{u}_C^{(m)} \right).$$

3. *Execute* 1 *stationary iteration for* $\mathbf{u}_A$ *with inverse preconditioner* $Y_A^{(2)}$:
$$\mathbf{u}_A^{(m+1)} = \widetilde{\mathbf{u}}_A^{(m+1)} + Y_A^{(2)} \left(\mathbf{b}_A - A\,\widetilde{\mathbf{u}}_A^{(m+1)} - B^T \mathbf{u}_C^{(m+1)} \right).$$

For the updates of $\mathbf{u}_A$, one typically uses $Y_A^{(i)}$ to approximate A^{-1}; i.e., the first and last substeps approximately solve the first block of equations for $\mathbf{u}_A$ assuming $\mathbf{u}_C$ is fixed. Regarding the update of $\mathbf{u}_C$, one typically uses a preconditioner M_S that approximates the Schur complement (1.4). This is motivated by the observation that if $Y_A^{(1)} = A^{-1}$, then the exact $\mathbf{u}_C$ 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 $Y_A^{(1)}$, the last step will provide the exact solution for $\mathbf{u}_A$ as well; that is, with such "exact" preconditioners the above process just implements the solution of the linear system via the block factorization (1.3).

When using standard ("inexact") preconditioners, the above process is related with the block preconditioners mentioned in section 1. 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 $\left(\hat{\mathbf{u}}_A^T \ \hat{\mathbf{u}}_C^T \right)^T$ denotes the exact solution of the linear system (1.1), the matrix T such that

$$\begin{pmatrix} \hat{\mathbf{u}}_A - \mathbf{u}_A^{(m+1)} \\ \hat{\mathbf{u}}_C - \mathbf{u}_C^{(m+1)} \end{pmatrix} = T \begin{pmatrix} \hat{\mathbf{u}}_A - \mathbf{u}_A^{(m)} \\ \hat{\mathbf{u}}_C - \mathbf{u}_C^{(m)} \end{pmatrix}$$

satisfies

$$(2.1) \qquad T = T_{(Y_A^{(2)},0)} \ T_{(0,M_S^{-1})} \ T_{(Y_A^{(1)},0)},$$

where

$$(2.2) \qquad T_{(Y_A^{(i)},0)} = I - \begin{pmatrix} Y_A^{(i)} \\ & 0 \end{pmatrix} K = \begin{pmatrix} I - Y_A^{(i)}A & -Y_A^{(i)}B^T \\ & I \end{pmatrix}$$

and

$$(2.3) \qquad T_{(0,M_S^{-1})} = I - \begin{pmatrix} 0 \\ & -M_S^{-1} \end{pmatrix} K = \begin{pmatrix} I \\ M_S^{-1}B & I - M_S^{-1}C \end{pmatrix}.$$

THEOREM 2.1. *Let K be a matrix of the form* (1.2) *such that A is an $n \times n$ matrix and C is an $m \times m$ 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* (2.2) *and* (2.3)*, 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

(2.4)

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

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

$$Y = \begin{pmatrix} Y_A^{(1)} & \\ & 0 \end{pmatrix} + \begin{pmatrix} 0 & \\ & -M_S^{-1} \end{pmatrix} - \begin{pmatrix} 0 & \\ & -M_S^{-1} \end{pmatrix} K \begin{pmatrix} Y_A^{(1)} & \\ & 0 \end{pmatrix}$$

$$= \begin{pmatrix} Y_A^{(1)} & \\ M_S^{-1} B Y_A^{(1)} & -M_S^{-1} \end{pmatrix}.$$

Then, P satisfies

$$P^{-1} \begin{pmatrix} I & \\ B Y_A^{(1)} & I \end{pmatrix} = \left(Y + \begin{pmatrix} Y_A^{(2)} & \\ & 0 \end{pmatrix} - \begin{pmatrix} Y_A^{(2)} & \\ & 0 \end{pmatrix} K Y \right) \begin{pmatrix} I & \\ B Y_A^{(1)} & I \end{pmatrix}$$

$$= \begin{pmatrix} Y_A^{(1)} & \\ & -M_S^{-1} \end{pmatrix} + \begin{pmatrix} Y_A^{(2)} & \\ & 0 \end{pmatrix}$$

$$\qquad - \begin{pmatrix} Y_A^{(2)} A & Y_A^{(2)} B^T \\ 0 & \end{pmatrix} \begin{pmatrix} Y_A^{(1)} & \\ & -M_S^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} Y_A^{(1)} + Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} & Y_A^{(2)} B^T M_S^{-1} \\ & -M_S^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} I & -Y_A^{(2)} B^T \\ & I \end{pmatrix} \begin{pmatrix} Y_A^{(1)} + Y_A^{(2)} - Y_A^{(2)} A Y_A^{(1)} & \\ & -M_S^{-1} \end{pmatrix}. \qquad \square$$

Observe that if $Y_A^{(1)} = M_A^{-1}$ while $Y_A^{(2)} = 0$, the preconditioner (2.4) corresponding to substeps 1 and 2 above (substep 3 being skipped) is the inexact Uzawa preconditioner (1.7). On the other hand, when $Y_A^{(1)} = 0$ and $Y_A^{(2)} = M_A^{-1}$, the preconditioner (2.4) corresponding to substeps 2 and 3 above (substep 1 being skipped) is the block triangular preconditioner (1.6).

However, when $Y_A^{(1)} = Y_A^{(2)} = M_A^{-1}$, we do not recover the inexact block factorization preconditioner (1.5), but a slightly different variant[1]

$$(2.5) \qquad P = \begin{pmatrix} I & \\ B M_A^{-1} & I \end{pmatrix} \begin{pmatrix} M_A (2 M_A - A)^{-1} M_A & \\ & -M_S \end{pmatrix} \begin{pmatrix} I & M_A^{-1} B^T \\ & I \end{pmatrix},$$

---

[1]To implement the inexact block factorization preconditioner, one has to slightly modify the third step of the above procedure into $\mathbf{u}_A^{(m+1)} = \tilde{\mathbf{u}}_A^{(m+1)} - Y_A^{(2)} B^T (\mathbf{u}_C^{(m+1)} - \mathbf{u}_C^{(m)})$; see the discussion in [26, section 2.2] for the rationale behind this.

which we refer to as *symmetrized inexact Uzawa preconditioner* (termed the "block symmetric Gauss–Seidel (SGS) preconditioner" in [18]). 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 $\bar{\mu} < 2$, that is, if and only if $\rho_A < 1$. Then, letting

$$(2.6) \qquad \overline{M}_A = M_A\,(2\,M_A - A)^{-1}\,M_A$$

there holds $\sigma\big(\,\overline{M}_A^{-1}\,A\big) \subset [1 - \rho_A^2, 1]$.

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[2] with

$$(2.7) \qquad \overline{M}_A = \left(Y_A^{(1)} + Y_A^{(2)} - Y_A^{(1)} A\,Y_A^{(2)}\right)^{-1}$$

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

$$(2.8) \qquad M_A = A^{1/2}\left(I - \left(I - A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{1/2}\right)^{-1} A^{1/2}$$

such that (2.6) holds. Moreover, (2.2) implies then

$$(2.9) \qquad T_{(\overline{M}_A^{-1},0)} = T_{(Y_A^{(1)},0)}\,T_{(Y_A^{(2)},0)} = \left(T_{(M_A^{-1},0)}\right)^2.$$

This allows us to state the following theorem, which is clearly inspired by the discussion in [9, section 3]; see also [26, Remark 2.1], [8, section 2.3], and Theorem 3.1 in [18], 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 Uzawa preconditioner.

THEOREM 2.2. *Let K be a matrix of the form* (1.2) *such that A is an $n \times n$ SPD matrix and C is an $m \times m$ 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* (2.2) *and* (2.3), *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* (2.7), *one has $\sigma\big(\,\overline{M}_A^{-1}\,A\big) \subset (0, 1]$.*

Then, letting M_A be defined by (2.8), *there holds, for any integer $m \geq 2$,*

$$(2.10) \qquad T^m = T_{(Y_A^{(2)},0)}\,T_{(0,M_S^{-1})}\,T_{(M_A^{-1},0)}\,\,\widehat{T}^{\,m-2}\,\,T_{(M_A^{-1},0)}\,T_{(0,M_S^{-1})}\,T_{(Y_A^{(1)},0)},$$

where

$$(2.11) \qquad \widehat{T} = T_{(M_A^{-1},0)}\,T_{(0,M_S^{-1})}\,T_{(M_A^{-1},0)} = I - P^{-1}K$$

with P defined by (2.5).

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

$$(2.12) \qquad P_{uza} = \begin{pmatrix} \overline{M}_A & \\ B & -M_S \end{pmatrix}$$

---

[2]Compared with the expression appearing in (2.4), we deliberately reverse the order of the terms $Y_A^{(1)}$ and $Y_A^{(2)}$, in view of the needs of Theorem 2.2 below.

and the inexact block triangular preconditioner

$$(2.13) \qquad P_{bt} = \begin{pmatrix} \overline{M}_A & B^T \\ & -M_S \end{pmatrix}$$

satisfy, respectively, for any integer $m \geq 1$,

$$\left(I - P_{uza}^{-1} K\right)^m = T_{(0,M_S^{-1})} \; T_{(M_A^{-1},0)} \quad \widehat{T}^{\,m-1} \quad T_{(M_A^{-1},0)}$$

and

$$\left(I - P_{bt}^{-1} K\right)^m = T_{(M_A^{-1},0)} \quad \widehat{T}^{\,m-1} \quad T_{(M_A^{-1},0)} \; T_{(0,M_S^{-1})},$$

where M_A *and* $\widehat{T}$ *are defined by* (2.8) *and* (2.11).

Proof. This is a straightforward corollary of Theorem 2.1 and (2.9). □

With this theorem, one sees that the analysis of many methods can be reduced to that of the symmetrized inexact Uzawa preconditioner. This will be exploited in section 4.

3. Compatible singular systems. Linear systems (1.1) arising from Stokes problems are often singular, B^T and C both having the constant vector in their null space, while $\mathbf{b}_C$ 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 a somewhat heavy formalism, with much attention to tiny details. 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 a orthonormal basis of this null space. Thus, the general assumption on C is weakened and we only require that it is positive definite on $\mathcal{N}(B^T) \backslash \mathcal{R}(V)$. We also assume that the system (1.1) is compatible, i.e., that $V^T \mathbf{b}_C = 0$.

Consider then sequences $\overline{\mathbf{u}}_A^{(m)}$, $\overline{\mathbf{u}}_C^{(m)}$ generated by substeps 1–3 sketched in the preceding section, in which one has set $\overline{\mathbf{u}}_A^{(0)} = \mathbf{u}_A^{(0)}$ and $\overline{\mathbf{u}}_C^{(0)} = (I - V V^T)\mathbf{u}_C^{(0)}$, and in which one has exchanged C and M_S^{-1} for, respectively,

$$\overline{C} = C + V V^T \quad \text{and} \quad \overline{M}_S^{-1} = (I - V V^T)M_S^{-1}(I - V V^T) + V V^T.$$

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

On the other hand, because $C V = 0$, $B^T V = 0$, $V^T C = 0$ (by symmetry of C), $V^T B = 0$, and $V^T \mathbf{b}_C = 0$, one can see by induction that $\overline{\mathbf{u}}_A^{(m)} = \mathbf{u}_A^{(m)}$ and $\overline{\mathbf{u}}_C^{(m)} = (I - V V^T)\mathbf{u}_C^{(m)}$ hold for all m. Therefore, if $\left(\hat{\mathbf{u}}_A^T \; \hat{\mathbf{u}}_C^T\right)^T$ denotes the exact solution of the linear system (1.1) satisfying $V^T \hat{\mathbf{u}}_C = 0$, there holds

$$(\hat{\mathbf{u}}_A - \mathbf{u}_A^{(m)})^T A(\hat{\mathbf{u}}_A - \mathbf{u}_A^{(m)}) \; + \; (\hat{\mathbf{u}}_C - \mathbf{u}_C^{(m)})^T S(\hat{\mathbf{u}}_C - \mathbf{u}_C^{(m)})$$
$$= (\hat{\mathbf{u}}_A - \overline{\mathbf{u}}_A^{(m)})^T A(\hat{\mathbf{u}}_A - \overline{\mathbf{u}}_A^{(m)}) + (\hat{\mathbf{u}}_C - \overline{\mathbf{u}}_C^{(m)})^T S(\hat{\mathbf{u}}_C - \overline{\mathbf{u}}_C^{(m)})$$
$$= \|\hat{\mathbf{u}}_A - \overline{\mathbf{u}}_A^{(m)}\|_A^2 + \|\hat{\mathbf{u}}_C - \overline{\mathbf{u}}_C^{(m)}\|_{\overline{S}}^2.$$

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 (1.9) associated with A and $\overline{S}$.

Moreover, consider the eigenvalue problem

$$(3.1) \qquad \overline{M}_S^{-1}\,\overline{S}\,\overline{\mathbf{z}} = \nu\,\overline{\mathbf{z}} \iff \left((I - V\,V^T)M_S^{-1}S + V\,V^T\right)\overline{\mathbf{z}} = \nu\,\overline{\mathbf{z}}.$$

Let $(\nu, \overline{\mathbf{z}})$ be an eigenpair such that $\nu \neq 1$. Hence one has $V^T\overline{\mathbf{z}} = 0$, implying that $(I - V(V^TM_SV)^{-1}V^TM_S)\overline{\mathbf{z}}$ is a nontrivial vector. Moreover, multiplying to the left both sides of (3.1) by $(I - V(V^TM_SV)^{-1}V^TM_S)$, one obtains, since $S\,\overline{\mathbf{z}} = S\,\mathbf{z}$ and $V^TS = 0$,

$$(3.2) \qquad M_S^{-1}S\,\mathbf{z} = \nu\,\mathbf{z}.$$

Conversely, if $(\nu, \mathbf{z})$ is an eigenpair of $M_S^{-1}S$ such that $\nu \neq 0$, there holds $V^TM_S\mathbf{z}=0$, implying that $\overline{\mathbf{z}} = (I - V\,V^T)\mathbf{z}$ is a nontrivial vector satisfying (3.1), as seen by multiplying to the left both sides of (3.2) by $I - V\,V^T$. In other words, we have just shown that

$$(3.3) \qquad \underline{\nu} = \min\left(1, \lambda_{\min}\left(\overline{M}_S^{-1}\,\overline{S}\right)\right) = \min\left(1, \min_{\nu \in \sigma(M_S^{-1}S)\setminus\{0\}} \nu\right)$$

and

$$\overline{\nu} = \min\left(1, \lambda_{\min}\left(\overline{M}_S^{-1}\,\overline{S}\right)\right) = \max\left(1, \lambda_{\max}\left(M_S^{-1}S\right)\right).$$

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 $\underline{\nu}$ for (3.3) and to read the given bound on $\|T^m\|_E$ as a bound on the maximal relative error measured in the seminorm 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, $T = H^{-1}Z\,H$, allowing us to use

$$\|T^m\|_E = \|E^{1/2}H^{-1}Z^m\,H\,E^{-1/2}\| \leq \|E^{1/2}H^{-1}\|\,\|Z^m\|\,\|H\,E^{-1/2}\|.$$

Regarding Z, we do not analyze directly its norm but its field of value (or numerical range)

$$W(Z) = \{\mathbf{z}^*\,Z\,\mathbf{z} \;:\; \mathbf{z} \in \mathbb{C}^n,\; \|\mathbf{z}\| = 1\}$$

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 $\|Z^m\| \leq 2\,w(Z^m)$ and $w(Z^m) \leq (w(Z))^m$ (for any matrix Z).

The matrices Z to be analyzed will have the form $Z = I - \widehat{K}$, where

$$(4.1) \qquad \widehat{K} = \begin{pmatrix} \widehat{A} & \widehat{B}^T \\ -\widehat{B} & \widehat{C} \end{pmatrix},$$

with $\widehat{A}$, $\widehat{C}$ SPD. Observe that, compared with (1.2), 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 [18], where similar inequalities are proved regarding the eigenvalues of $\widehat{K}$. In general, the inequalities in [18] are somewhat stricter, but the inequality associated with γ is new. (Hence Theorem 4.1 in [18] would be enhanced by adding its counterpart for the eigenvalues.)

THEOREM 4.1. *Let $\widehat{K}$ be a matrix of the form* (4.1), *where $\widehat{A}$ is an $n \times n$ SPD matrix and $\widehat{C}$ is an $m \times m$ SPD matrix with $m \le n$. Let*

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

Let

$$\underline{\alpha} = \min\left(\lambda_{\min}(\widehat{A}), \lambda_{\min}(\widehat{C})\right),$$
$$\overline{\alpha} = \max\left(\lambda_{\max}(\widehat{A}), \lambda_{\max}(\widehat{C})\right),$$
$$\beta = \max\left(\lambda_{\max}(S_{\widehat{A}}), \lambda_{\max}(S_{\widehat{C}})\right),$$
$$\gamma = \left\|\widehat{C}^{-1/2} \widehat{B} \widehat{A}^{-1/2}\right\|,$$
$$\delta = \|\widehat{B}\|.$$

There holds

(4.2) $$W\left(I - \widehat{K}\right) \subset \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$$

with

(4.3) $$\mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta) = \left\{ z \in \mathbb{C}^n \mid \underline{\alpha} \le \Re e(1 - z) \le \overline{\alpha} \quad and \quad |1 - z - \tfrac{\beta}{2}| \le \tfrac{\beta}{2} \right.$$
$$\left. and \quad |\Im_m(z)| \le \gamma \, \Re e(1 - z) \quad and \quad |\Im_m(z)| \le \delta \right\}.$$

Moreover, if $\beta < 2$,

(4.4) $$\max_{z \in \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)} |z|$$
$$\le \min\left(\sqrt{1 - \underline{\alpha}(2 - \beta)}, \ \max\left(\sqrt{1 - \tfrac{\beta(2 - \beta)}{1 + \gamma^2}}, \ \sqrt{1 - 2\underline{\alpha} + \underline{\alpha}^2(1 + \gamma^2)} \right) \right).$$

Proof. For any $\mathbf{z}$ with $\|\mathbf{z}\| = 1$, $\lambda = \mathbf{z}^* \widehat{K} \mathbf{z}$ satisfies $\Re e(\lambda) = \mathbf{z}^* \widehat{K}_s \mathbf{z}$ and $\Im_m(\lambda) = -\imath \, \mathbf{z}^* \widehat{K}_a \mathbf{z}$, where the symmetric and antisymmetric parts of $\widehat{K}$ satisfy, respectively,

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

The first and the last condition in (4.3) follow immediately, observing that $\|\widehat{K}_a\| = \|B\|$ because

$$\widehat{K}_a^T \widehat{K}_a = \begin{pmatrix} \widehat{B}^T \widehat{B} & \\ & \widehat{B}\,\widehat{B}^T \end{pmatrix}.$$

$$(\underline{\alpha} = 0.7, \ \beta = 1.5, \ \gamma = 1.33) \qquad\qquad (\underline{\alpha} = 0.2, \ \beta = 1.5, \ \gamma = 0.8)$$
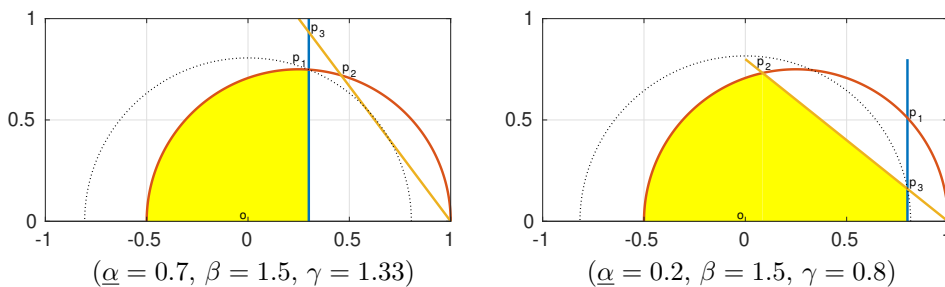
FIG. 4.1. $\mathcal{F}(\underline{\alpha}, \infty, \beta, \gamma, \infty)$ *is depicted in yellow (light gray) together with the lines and the circle that define its boundary (only the upper half plane is represented); the dotted line is the half circle centered at origin and with radius equal to the right hand side of* (4.4).

Regarding the third condition, $\gamma \, \Re\mathrm{e}(\lambda) \geq \pm \, \Im\mathrm{m}(\lambda)$ is guaranteed if

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

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

On the other hand, the second condition in (4.3) follows from [19, Theorem 2.2] if there holds $\beta \geq \lambda_{\min}\big(\frac{1}{2}(\widehat{K}^{-1} + \widehat{K}^{-T})\big)$, which we deduce from the relation

$$\tfrac{1}{2}(\widehat{K}^{-1} + \widehat{K}^{-T}) = \begin{pmatrix} S_{\widehat{A}}^{-1} & \\ & S_{\widehat{C}}^{-1} \end{pmatrix}.$$

To prove (4.4), we show that its right-hand side corresponds to the maximal distance to origin of any boundary point of $\mathcal{F}(\underline{\alpha}, \infty, \beta, \gamma, \infty)$ (i.e., $\mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$ when skipping the conditions $\Re\mathrm{e}(1 - z) \leq \overline{\alpha}$ and $|\Im\mathrm{m}(z)| \leq \delta$). The geometry of the problem is depicted in Figure 4.1. 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 in the left picture, the maximal distance is $d(o, p_1)$. If it is as in the right picture, the maximal distance corresponds to p_2 if it is reached along the circle and to p_3 if it is reached along the line $\Re\mathrm{e}(z) = 1 - \underline{\alpha}$. As the distance to origin from p_2 to p_3 along the line $\Im\mathrm{m}(z) = \gamma \, \Re\mathrm{e}(1 - z)$ 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 in the right picture, the maximal distance to origin of a boundary point of $\mathcal{F}(\underline{\alpha}, \infty, \beta, \gamma, \infty)$ is the maximum of $\big(d(o, p_2), d(o, p_3)\big)$. 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

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

The inequality (4.4) 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 pairs of equations:

$$\text{For } x_1, y_1\text{:} \quad \begin{cases} (1 - x - \frac{\beta}{2})^2 + y^2 = \frac{\beta^2}{4}, \\ x = 1 - \underline{\alpha}; \end{cases}$$

$$\text{For } x_2, y_2: \quad \begin{cases} (1 - x - \frac{\beta}{2})^2 + y^2 = \frac{\beta^2}{4}, \\ y = \gamma(1-x); \end{cases}$$

$$\text{For } x_3, y_3: \quad \begin{cases} x = 1 - \underline{\alpha}, \\ y = \gamma(1-x). \end{cases}$$

$\Box$

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

THEOREM 4.2 (analysis of the symmetrized inexact Uzawa preconditioner). *Let K be a matrix of the form* (1.2) *such that A is an $n \times n$ SPD matrix and C is an $m \times m$ 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* (2.5) *where M_A and M_S are SPD matrices. Let $\underline{\mu}, \overline{\mu}, \underline{\nu}, \overline{\nu}, \rho_A, \rho_S$ be defined by* (1.11), (1.12), (1.13), (1.14), *and assume $\rho_A, \rho_S < 1$.*

Let $\mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$ be the region (4.3) *of the complex plane described by the following parameters:*

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

There holds, for any positive integer m,

$$(4.5) \qquad \|(I - P^{-1}K)^m\|_E \;\leq\; \xi \left(\max_{z \,\in\, \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)} |z| \right)^m,$$

where

$$(4.6) \qquad \xi = \frac{\sqrt{\overline{\nu}}\,\left(\overline{\mu}\,\left(3 + \sqrt{5}\right) + 2\,(\overline{\mu} - 1)^2\right)}{\min\left(\sqrt{1 - \rho_A^2},\,\sqrt{\underline{\nu}}\right)}.$$

Moreover,

(4.7)

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

Proof. Let $\overline{M}_A = M_A\,(2\,M_A - A)^{-1}\,M_A$ and

$$L = \begin{pmatrix} I & \\ B\,M_A^{-1} & I \end{pmatrix}, \qquad D = \begin{pmatrix} \overline{M}_A & \\ & -M_S \end{pmatrix},$$

so that $P = L\,D\,L^T$. Noting that $\rho_A < 1$ implies that $\overline{M}_A$ is SPD, let further

$$F = \begin{pmatrix} \left(A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{-1/2} & \\ & M_S^{1/2} S^{-1/2} \end{pmatrix}.$$

Consider then

$$Z = I - \left(F \, E^{1/2} \, L^T \right) P^{-1} K \left(F \, E^{1/2} L^T \right)^{-1}.$$

Because

$$T^m = \left(F \, E^{1/2} \, L^T \right)^{-1} Z^m \, \left(F \, E^{1/2} L^T \right)$$

there holds

$$\|T^m\|_E \leq \left\| E^{1/2} \left(F \, E^{1/2} \, L^T \right)^{-1} \right\| \, \|Z^m\| \, \left\| \left(F \, E^{1/2} \, L^T \right) E^{-1/2} \right\|$$

$$\leq 2 \, \left\| E^{1/2} L^{-T} E^{-1/2} \right\| \, \left\| E^{1/2} L^T E^{-1/2} \right\| \, \|F\| \, \|F^{-1}\| \, \left(w(Z) \right)^m.$$

The inequality (4.5) will be obtained by showing that

$$(4.8) \qquad \left\| E^{1/2} L^{-T} E^{-1/2} \right\| \, \left\| E^{1/2} L^T E^{-1/2} \right\| \, \|F\| \, \|F^{-1}\| \; \leq \; \tfrac{\xi}{2}$$

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

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

$$(4.9) \qquad \left\| F \right\|^2 = \left(\lambda_{\min} \left(F^{-T} F^{-1} \right) \right)^{-1} = \left(\min \left(1 - \rho_A^2, \, \underline{\nu} \right) \right)^{-1}$$

and

$$(4.10) \qquad \left\| F^{-1} \right\|^2 = \lambda_{\max} \left(F^{-T} F^{-1} \right) = \overline{\nu}.$$

We next observe that

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

and

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

have the same norm, which is equal to the square root of the norm of

$$\left(E^{1/2} L^T E^{-1/2} \right)^T E^{1/2} L^T E^{-1/2}$$

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

$$= I + \begin{pmatrix} I & \\ & S^{-1/2} B \, A^{1/2} \end{pmatrix} \begin{pmatrix} 0 & A^{1/2} M_A^{-1} A^{1/2} \\ A^{1/2} M_A^{-1} A^{1/2} & \left(A^{1/2} M_A^{-1} A^{1/2} \right)^2 \end{pmatrix}$$

$$\begin{pmatrix} I & \\ & A^{1/2} B^T S^{-1/2} \end{pmatrix}.$$

Because $\|A^{1/2} B^T S^{-1/2}\| = \lambda_{\max}(S^{-1} B \, A^{-1} B^T) \leq 1$, we then find

$$\left\| E^{1/2} \, L^T E^{-1/2} \right\|^2 \leq 1 \; + \; \lambda_{\max}(M_A^{-1} A) \, \left\| \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \right\|$$

$$+ \; \max \left(\lambda_{\max} \left((M_A^{-1} A)^2 - M_A^{-1} A \right), \, 0 \right)$$

$$= 1 + \overline{\mu} \, \tfrac{1 + \sqrt{5}}{2} + \max \left(\overline{\mu}^2 - \overline{\mu}, \, 0 \right)$$

$$= \overline{\mu} \, \tfrac{3 + \sqrt{5}}{2} + \left(\overline{\mu} - 1 \right)^2.$$

Remembering that $\left\| E^{1/2} L^{-T} E^{-1/2} \right\| = \left\| E^{1/2} L^T E^{-1/2} \right\|$ this inequality, combined with (4.9), (4.10), proves (4.8) with ξ as in (4.6).

To prove (4.5), it remains then only to show that $W(Z) \subset \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$. One has

$$
\begin{aligned}
I - Z &= F\, D^{-1} L^{-1} K\, L^{-T} F^{-1} \\
&= \begin{pmatrix} (A^{1/2}\, \overline{M}_A^{-1}\, A^{1/2})^{-1/2} A^{1/2}\, \overline{M}_A^{-1} & \\ & -M_S^{-1/2} \end{pmatrix} \begin{pmatrix} I & \\ -B\, M_A^{-1} & I \end{pmatrix} \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \\
&\quad \begin{pmatrix} I & -M_A^{-1} B^T \\ & I \end{pmatrix} \begin{pmatrix} A^{-1/2}(A^{1/2}\, \overline{M}_A^{-1}\, A^{1/2})^{1/2} & \\ & M_S^{-1/2} \end{pmatrix} \\
&= \begin{pmatrix} (A^{1/2}\, \overline{M}_A^{-1}\, A^{1/2})^{1/2} A^{-1/2} & \\ & M_S^{-1/2} \end{pmatrix} \\
&\quad \begin{pmatrix} A & (I - A\, M_A^{-1}) B^T \\ -B(I - M_A^{-1} A) & C + B(2 M_A^{-1} - M_A^{-1} A\, M_A^{-1}) B^T \end{pmatrix} \\
&\quad \begin{pmatrix} A^{-1/2}(A^{1/2}\, \overline{M}_A^{-1}\, A^{1/2})^{1/2} & \\ & M_S^{-1/2} \end{pmatrix}.
\end{aligned}
$$

Let X be an orthogonal matrix that diagonalizes $A^{1/2} M_A^{-1} A^{1/2}$, i.e., such that $X^T A^{1/2} M_A^{-1} A^{1/2} X = \Lambda$, where Λ 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 $G = M_S^{1/2} B A^{-1/2} X$ and $\widetilde{C} = M_S^{-1/2} C\, M_S^{-1/2}$. Because the numerical range is invariant under unitary similarity transformations, $W(Z) = W(I - \widehat{K})$ with

$$
\begin{aligned}
\widehat{K} &= \begin{pmatrix} X^T & \\ & I \end{pmatrix} (I - Z) \begin{pmatrix} X & \\ & I \end{pmatrix} \\
&= \begin{pmatrix} \overline{\Lambda} & \overline{\Lambda}^{1/2}(I - \Lambda) G^T \\ -G(I - \Lambda)\overline{\Lambda}^{1/2} & \widetilde{C} + G(2\Lambda - \Lambda^2) G^T \end{pmatrix} = \begin{pmatrix} \widehat{A} & \widehat{B}^T \\ -\widehat{B} & \widehat{C} \end{pmatrix}.
\end{aligned}
$$
(4.11)

We may apply Theorem 4.1 to this matrix. Moreover, the same matrix was already analyzed during the proof of Theorem 4.3 in [18], yielding bounds on most quantities involved:

$$
\begin{aligned}
\lambda_{\min}(\widehat{A}) &\geq 1 - \rho_A^2, & \lambda_{\min}(\widehat{C}) &\geq (1 - \rho_A^2)\,\underline{\nu}, \\
\lambda_{\max}(\widehat{A}) &\leq 1, & \lambda_{\max}(\widehat{C}) &\leq \overline{\nu}, \\
\lambda_{\max}(S_{\widehat{A}}) &\leq 1, & \lambda_{\max}(S_{\widehat{C}}) &\leq \overline{\nu},
\end{aligned}
$$

and

$$
\|\widehat{B}\|^2 \leq \begin{cases} \overline{\nu}\, \rho_A^2 (1 - \rho_A^2) & \text{if } \rho_A^2 < \tfrac{1}{2}, \\ \tfrac{1}{4}\, \overline{\nu} & \text{otherwise.} \end{cases}
$$

This straightforwardly provides the required result (4.5) if we can further show that $\left\| \widehat{C}^{-1/2} \widehat{B} \widehat{A}^{-1/2} \right\| = \left\| \widehat{C}^{-1/2} G(I - \Lambda) \right\| \leq \rho_A (1 - \rho_A^2)^{-1/2}$. In this view, observe that, for any $\mathbf{z}$,

$$\frac{\mathbf{z}^* G(I-\Lambda)^2 G^T \mathbf{z}}{\mathbf{z}^* \widehat{C}\, \mathbf{z}} \leq \frac{\mathbf{z}^* G(I-\Lambda)^2 G^T \mathbf{z}}{\mathbf{z}^* G\left(2\,\Lambda - \Lambda^2\right) G^T \mathbf{z}}$$

$$\leq \max_{\lambda \in [\underline{\mu},\,\overline{\mu}]} \frac{(1-\lambda)^2}{2\,\lambda - \lambda^2}$$

$$= \frac{\rho_A^2}{1 - \rho_A^2}.$$

It remains thus only to prove (4.7). Here we use (4.4) of Theorem 4.1, yielding

$$\left(\max_{z\, \in\, \mathcal{F}(\underline{\alpha},\overline{\alpha},\beta,\gamma,\delta)} |z|\right)^2 \;\leq\; \max\left(1 - \tfrac{\beta(2-\beta)}{1+\gamma^2},\; 1 - 2\,\underline{\alpha} + \underline{\alpha}^2(1+\gamma^2)\right)$$

$$\leq\; \max\Big(1 - (1+\rho_S)(1-\rho_S)(1-\rho_A^2),$$
$$1 - 2(1-\rho_A^2)(1-\rho_S) + \tfrac{(1-\rho_A^2)^2(1-\rho_S)^2}{1-\rho_A^2}\Big)$$

$$= \rho_A^2 + \rho_S^2 - \rho_A^2\rho_S^2. \qquad \square$$

Perhaps the best previous norm estimate for this preconditioner is the one that would be obtained by combining the results in [6, 26] for the inexact Uzawa algorithm with Theorem 2.2 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 [18, Corollary 4.5]: $\rho(T) \leq \max(\rho_A, \rho_S)$. On the other hand, Theorem A.1 in the 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 (4.7) 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 Uzawa preconditioner, but most of the previous proof can be reused.

THEOREM 4.3 (analysis of inexact block factorization preconditioner). *Let K be a matrix of the form* (1.2) *such that A is an $n \times n$ SPD matrix and C is an $m \times m$ 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* (1.5) *where M_A and M_S are SPD matrices. Let $\underline{\mu}, \overline{\mu}, \underline{\nu}, \overline{\nu}, \rho_A, \rho_S$ be defined by* (1.11), (1.12), (1.13), (1.14), *and assume $\rho_A, \rho_S < 1$.*

Let $\mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$ be the region (4.3) *of the complex plane described by the following parameters:*

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

There holds, for any positive integer m,

$$(4.12) \qquad \|(I - P^{-1}K)^m\|_E \leq \xi \left(\max_{z \in \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)} |z| \right)^m,$$

where

$$(4.13) \qquad \xi = \frac{\max\left(\sqrt{\overline{\mu}}, \sqrt{\overline{\nu}}\right) \left(\overline{\mu} \left(3 + \sqrt{5}\right) + 2\left(\overline{\mu} - 1\right)^2\right)}{\min\left(\sqrt{\underline{\mu}}, \sqrt{\underline{\nu}}\right)}.$$

Moreover, if $\overline{\mu} \leq 1 + \frac{\rho_S}{1 + \rho_S}$,

$$(4.14) \quad \hat{\rho} = \max_{z \in \mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)} |z| \leq \max\left(\sqrt{\rho_A^2 + \rho_S^2 - \rho_A^2 \rho_S^2}, \frac{\rho_A \sqrt{2}}{\sqrt{1 + \rho_A}}\right)$$

$$= \max\left(\sqrt{1 - (1 - \rho_A^2)(1 - \rho_S^2)}, \sqrt{1 - (1 - \rho_A)\frac{1 + 2\rho_A}{1 + \rho_A}}\right) < 1.$$

Proof. The proof follows that of Theorem 4.2, in which we have just to exchange everywhere $\overline{M}_A$ for M_A, including in the definition of F:

$$F = \begin{pmatrix} \left(A^{1/2} M_A^{-1} A^{1/2}\right)^{-1/2} & \\ & M_S^{1/2} S^{-1/2} \end{pmatrix}.$$

Avoiding repeating identical steps, we have then

$$\|F\|^2 = \left(\lambda_{\min}\left(F^{-T} F^{-1}\right)\right)^{-1} = \left(\min\left(\underline{\mu}, \underline{\nu}\right)\right)^{-1}$$

and

$$\|F^{-1}\|^2 = \lambda_{\max}\left(F^{-T} F^{-1}\right) = \max\left(\overline{\mu}, \overline{\nu}\right)$$

instead of (4.9) and (4.10), while the other norm estimates are unchanged. This provides the value of ξ in (4.13).

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

$$\widehat{K} = \begin{pmatrix} \Lambda & \Lambda^{1/2}\left(I - \Lambda\right)G^T \\ -G\left(I - \Lambda\right)\Lambda^{1/2} & \widetilde{C} + G\left(2\Lambda - \Lambda^2\right)G^T \end{pmatrix} = \begin{pmatrix} \widehat{A} & \widehat{B}^T \\ -\widehat{B} & \widehat{C} \end{pmatrix}.$$

Again, the proof of Theorem 4.3 in [18] provides for this matrix bounds on $\lambda_{\min}(\widehat{A})$, $\lambda_{\min}(\widehat{C})$, $\lambda_{\max}(\widehat{A})$, $\lambda_{\max}(\widehat{C})$, $\lambda_{\max}(S_{\widehat{A}})$, $\lambda_{\max}(S_{\widehat{C}})$, and $\|\widehat{B}\|$ which straightforwardly justify the given expressions for $\underline{\alpha}, \overline{\alpha}, \beta, \delta$. Finally, $\widehat{C}^{-1/2} \widehat{B} \widehat{A}^{-1/2}$ takes here exactly the same expression as in the proof of Theorem 4.2, justifying to set $\delta = \frac{\rho_A}{1 - \rho_A^2}$ here as well.

Regarding (4.14), first note that if $\overline{\mu} \leq 1 + \frac{\rho_S}{1 + \rho_S}$, then $\beta \leq 1 + \rho_S$ as in Theorem 4.2. If, in addition, $\min\left(1 - \rho_A, (1 - \rho_A^2)(1 - \rho_S)\right) = (1 - \rho_A^2)(1 - \rho_S)$, the parameters $\underline{\alpha}, \beta$ and δ have all identical bounds, yielding

$$\max\left(1 - \frac{\beta(2 - \beta)}{1 + \gamma^2}, 1 - 2\underline{\alpha} + \underline{\alpha}^2(1 + \gamma^2)\right) \leq \rho_A^2 + \rho_S^2 - \rho_A^2 \rho_S^2$$

with exactly the same arguments. On the other hand, if $1 - \rho_A < (1 - \rho_A^2)(1 - \rho_S)$, we find

$$\max\left(1 - \frac{\beta(2-\beta)}{1+\gamma^2},\ 1 - 2\underline{\alpha} + \underline{\alpha}^2(1+\gamma^2)\right) \leq \max\left(1 - (1+\rho_S)(1-\rho_S)(1-\rho_A^2),\right.$$
$$\left. 1 - 2(1-\rho_A) + \frac{(1-\rho_A)^2}{1-\rho_A^2}\right)$$
$$= \max\left(\rho_A^2 + \rho_S^2 - \rho_A^2\rho_S^2,\ \frac{2\rho_A^2}{1+\rho_A}\right).$$

The inequality (4.14) 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 (1.9) with different weights for both terms in the right-hand side, the analysis in [2] proves monotone convergence with factor $\rho = \max(\rho_A, 2\rho_S/(1-\rho_S))$. It is therefore better than ours if ρ_S is small enough but fails for larger ρ_S. This is improved in [24] and further in [26] 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 [26] fails to prove convergence if $\rho_A = \rho_S = \frac{1}{2}$ or if $\rho_S = \frac{\sqrt{2}}{2}$ and $\overline{\mu} = 1 + \frac{\rho_S}{1+\rho_S} = \sqrt{2}$. (Observe that in the latter case our condition to have (4.14) holds.)

Nevertheless, this comparison shows that our analysis is suboptimal for small ρ_S, at least when $C = 0$. This is further confirmed by Theorem A.1 in the 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 (4.14) 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 A.1).

Regarding the condition $\overline{\mu} \leq 1 + \frac{\rho_S}{1+\rho_S}$ to have (4.14), 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 $\hat{\rho}$ below 1. For instance, if $\rho_A = \rho_S = \frac{1}{2}$, using the inequalities associated with $\underline{\alpha}, \overline{\alpha}, \gamma$, and δ, we get $\hat{\rho} \leq \frac{\sqrt{13}}{4}$. 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 $\overline{\mu} \leq 1 + \frac{\rho_S}{1+\rho_S}$ does not hold; see the numerical results in the next section.

4.4. Inexact Uzawa and block triangular preconditioners. When $\overline{\mu} \leq 1$, it is straightforward to obtain a convergence proof for both preconditioners (1.6) and (1.7), by combining Theorem 4.2 with Theorem 2.2 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 4.2, in order to obtain a better estimate for the constant ξ in (4.15).

THEOREM 4.4 (analysis of inexact block triangular and inexact Uzawa preconditioners). *Let K be a matrix of the form* (1.2) *such that A is an $n \times n$ SPD matrix and C is an $m \times m$ 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 $\underline{\mu}, \overline{\mu}, \underline{\nu}, \overline{\nu}, \rho_A, \rho_S$ be defined by* (1.11), (1.12), (1.13), (1.14), *and assume $\rho_A, \rho_S < 1$, and $\overline{\mu} \leq 1$.*

Let $\mathcal{F}(\underline{\alpha}, \overline{\alpha}, \beta, \gamma, \delta)$ be the region (4.3) *of the complex plane described by the following parameters:*

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

There holds, for both preconditioners P defined by (1.6) *and* (1.7),

$$(4.15) \qquad \|(I - P^{-1}K)^m\|_E\ \leq\ \xi\left(\max_{z\,\in\,\mathcal{F}(\underline{\alpha},\overline{\alpha},\beta,\gamma,\delta)} |z|\right)^{m-1},$$

where m is any positive integer and

$$(4.16) \qquad \xi = \left(\frac{2\,\overline{\nu}\,(1 + \overline{\nu})\,(2 + \overline{\nu}(1 + \sqrt{5}))}{\underline{\nu}}\right)^{1/2}\ \max\left(1,\ \sqrt{\tfrac{\rho_A}{1 - \rho_A}}\right).$$

Moreover,

(4.17)

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

Proof. For the clarity of the proof, we adopt the notation of Theorem 2.2, that is, we exchange M_A for $\overline{M}_A$ in the definition of the preconditioner and let M_A be defined by (2.8). This entails that with $\lambda_{\max}(\overline{M}_A^{-1} A) = 1$ and $\lambda_{\min}(\overline{M}_A^{-1} A) = 1 - \rho_A$, 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 (1.11) and (1.13)).

We thus apply Theorem 2.2 and use (2.12) to denote the inexact Uzawa preconditioner while defining $\widehat{T}$ by (2.11). We obtain, with L, F, and Z as in the proof of Theorem 4.2

$$\begin{aligned}\left(I - P_{\text{uza}}^{-1}K\right)^m &= T_{(0,M_S^{-1})}\ T_{(M_A^{-1},0)}\ \widehat{T}^{\,m-1}\ T_{(M_A^{-1},0)} \\ &= T_{(0,M_S^{-1})}\ T_{(M_A^{-1},0)}\ L^{-T}E^{-1/2}F^{-1}\ Z^{m-1}\ F\,E^{1/2}L^T\ T_{(M_A^{-1},0)}.\end{aligned}$$

The norm of Z can be analyzed exactly as in the proof of Theorem 4.2, only taking care to exchange ρ_A for $\sqrt{\rho_A}$ in the obtained results. This yields (4.15) and (4.17) if we can prove

$$\left\|E^{1/2}T_{(0,M_S^{-1})}\ T_{(M_A^{-1},0)}\ L^{-T}E^{-1/2}F^{-1}\right\|\ \left\|F\,E^{1/2}L^T\ T_{(M_A^{-1},0)}E^{-1/2}\right\|\ \leq\ \tfrac{\xi}{2}.$$

Setting

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

this will be done by showing that

$$\begin{aligned}(4.18) \quad &\left\|E^{1/2}R^{-1}\right\|\ \left\|R\,T_{(0,M_S^{-1})}\ R^{-1}\right\|\ \left\|R\,T_{(M_A^{-1},0)}\ L^{-T}E^{-1/2}F^{-1}\right\| \\ &\qquad\qquad\qquad \left\|F\,E^{1/2}L^T\ T_{(M_A^{-1},0)}R^{-1}\right\|\ \left\|R\,E^{-1/2}\right\|\ \leq\ \tfrac{\xi}{2}.\end{aligned}$$

We proceed similarly for the inexact block triangular preconditioner: applying Theorem 2.2 yields

$$\left(I - P_{\mathrm{bt}}^{-1}K\right)^m = T_{(M_A^{-1},0)} \;\; \widehat{T}^{\,m-1} \;\; T_{(M_A^{-1},0)} \, T_{(0,M_S^{-1})}$$
$$= T_{(M_A^{-1},0)} \;\; L^{-T}E^{-1/2}F^{-1} \, Z^{m-1} \, F\,E^{1/2}L^T \;\; T_{(M_A^{-1},0)} \, T_{(0,M_S^{-1})},$$

which provides the required results if there holds

$$\left\|E^{1/2}T_{(M_A^{-1},0)} \, L^{-T}E^{-1/2}F^{-1}\right\| \;\; \left\|F\,E^{1/2}L^T \, T_{(M_A^{-1},0)} \, T_{(0,M_S^{-1})} \, E^{-1/2}\right\| \;\le\; \tfrac{\xi}{2}.$$

Using $\;E^{1/2}T_{(M_A^{-1},0)} \, L^{-T}E^{-1/2}F^{-1} = \left(E^{1/2}R^{-1}\right)\left(R\,T_{(M_A^{-1},0)} \, L^{-T}E^{-1/2}F^{-1}\right)\;$ and

$$F\,E^{1/2}L^T \, T_{(M_A^{-1},0)} \, T_{(0,M_S^{-1})} \, E^{-1/2}$$
$$= \left(F\,E^{1/2}L^T \, T_{(M_A^{-1},0)} \, R^{-1}\right) \left(R\,T_{(0,M_S^{-1})} \, R^{-1}\right) \left(R\,E^{-1/2}\right),$$

this is again proved by (4.18).

We are thus left with the analysis of all terms in the left-hand side of (4.18). We find

$$\left\|E^{1/2}R^{-1}\right\| \;\; \left\|R\,E^{-1/2}\right\| \;\le\; \sqrt{\tfrac{\overline{\nu}}{\underline{\nu}}}.$$

Further,

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

implying

$$R\,T_{(0,M_S^{-1})} \, R^{-1}\left(R\,T_{(0,M_S^{-1})} \, R^{-1}\right)^T = \begin{pmatrix} I & \\ & (I - M_S^{-1/2}C\,M_S^{-1/2})^2 \end{pmatrix}$$
$$+ \begin{pmatrix} I & \\ & M_S^{-1/2}B\,A^{-1/2} \end{pmatrix}\begin{pmatrix} 0 & I \\ I & I \end{pmatrix}\begin{pmatrix} I & \\ & A^{-1/2}B^T M_S^{-1/2} \end{pmatrix}.$$

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

$$\left\| \begin{pmatrix} 0 & I \\ I & I \end{pmatrix} \right\| = \tfrac{1+\sqrt{5}}{2}.$$

Thus we have proved

$$\left\|R\,T_{(0,M_S^{-1})} \, R^{-1}\right\| \;\le\; \left(1 + \overline{\nu}\,\tfrac{1+\sqrt{5}}{2}\right)^{1/2}.$$

Setting $\Delta = A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}$, we proceed with the analysis of

$$R\,T_{(M_A^{-1},0)} \, L^{-T}E^{-1/2}F^{-1}$$
$$= \begin{pmatrix} A^{1/2} & \\ & M_S^{1/2} \end{pmatrix} \begin{pmatrix} I - M_A^{-1}A & -M_A^{-1}B^T \\ & I \end{pmatrix} \begin{pmatrix} I & -M_A^{-1}B^T \\ & I \end{pmatrix}$$
$$\begin{pmatrix} A^{-1/2}\left(A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{1/2} & \\ & M_S^{-1/2} \end{pmatrix}$$
$$= \begin{pmatrix} (I - A^{1/2}M_A^{-1}A^{1/2})\Delta^{1/2} & A^{1/2}(M_A^{-1}A\,M_A^{-1} - 2\,M_A^{-1})B^T M_S^{-1/2} \\ & I \end{pmatrix}.$$

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

$$(\overline{\nu} + 1)\,I - R\,T_{(M_A^{-1},0)}\,L^{-T}E^{-1/2}F^{-1}\left(R\,T_{(M_A^{-1},0)}\,L^{-T}E^{-1/2}F^{-1}\right)^T$$
$$= \begin{pmatrix} (\overline{\nu}+1)\,I - (I-\Delta)\Delta - \Delta\,A^{-1/2}B^T M_S^{-1}B^T A^{-1/2}\Delta & \Delta\,A^{-1/2}B^T M_S^{-1/2} \\ M_S^{-1/2}B^T A^{-1/2}\Delta & \overline{\nu}\,I \end{pmatrix}$$

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

$$(\overline{\nu}+1)\,I - (I-\Delta)\Delta - \left(1 + \tfrac{1}{\overline{\nu}}\right)\Delta\,A^{-1/2}B^T M_S^{-1}B^T A^{-1/2}\Delta$$
$$= (\overline{\nu}+1)\,\Delta\left(I - \tfrac{1}{\overline{\nu}}A^{-1/2}B^T M_S^{-1}B^T A^{-1/2}\right)\Delta + (I-\Delta) + \overline{\nu}\left(I - \Delta^2\right)$$

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

$$\lambda_{\max}\left(A^{-1/2}B^T M_S^{-1}B^T A^{-1/2}\right) = \lambda_{\max}\left(M_S^{-1}B\,A^{-1}B^T\right) \;\leq\; \lambda_{\max}\left(M_S^{-1}S\right) \;\leq\; \overline{\nu}.$$

We have thus proved

$$\left\|R\,T_{(M_A^{-1},0)}\,L^{-T}E^{-1/2}F^{-1}\right\| \;\leq\; (1+\overline{\nu})^{1/2}$$

and it remains to analyze $F\,E^{1/2}L^T\,T_{(M_A^{-1},0)}R^{-1}$. This is more straightforward because

$$F\,E^{1/2}L^T\,T_{(M_A^{-1},0)}R^{-1} = \begin{pmatrix} \left(A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{-1/2}A^{1/2} & \\ & M_S^{1/2} \end{pmatrix}\begin{pmatrix} I & M_A^{-1}A \\ & I \end{pmatrix}$$
$$\begin{pmatrix} I - M_A^{-1}A & -M_A^{-1}B^T \\ & I \end{pmatrix}\begin{pmatrix} A^{-1/2} & \\ & M_S^{-1/2} \end{pmatrix}$$
$$= \begin{pmatrix} \left(A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{-1/2}\left(I - A^{1/2}M_A^{-1}A^{1/2}\right) & \\ & I \end{pmatrix}.$$

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

$$\left\|\left(A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)^{-1}\left(I - A^{1/2}M_A^{-1}A^{1/2}\right)^2\right\| = \left\|A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\left(I - A^{1/2}\,\overline{M}_A^{-1}\,A^{1/2}\right)\right\|$$
$$= \max_{\xi\in[1-\rho_A,1]}\left|\tfrac{1}{\xi} - 1\right|,$$

that is, $\|F\,E^{1/2}L^T\,T_{(M_A^{-1},0)}R^{-1}\| \;\leq\; \max(1, \sqrt{\tfrac{\rho_A}{1-\rho_A}})$. Combining this and the above inequalities yields (4.18), concluding the proof. □

In [6, 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 (4.17), using a rather exotic norm and assuming $C = 0$, $\overline{\nu} \leq 1$, and $\overline{\mu} \leq 1$. Using another preconditioner dependent norm, the same bound is recovered in [26] (see Remark 4.4 in this work), as a particular case of a more general result compatible with $\overline{\mu}, \overline{\nu} > 1$. However, it is clear from the discussion in [18] that the convergence estimate rapidly deteriorates when $\overline{\mu}$ increases above 1. This is confirmed with Theorem A.1 in the

appendix, which proves that $\rho(T) \geq \overline{\mu} - 1 + \sqrt{\mu(\overline{\mu} - 1)}$ in the limit case $m = n$, $C = 0$, and $M_S = S$. One may then see that convergence is lost as soon as $\overline{\mu} \geq 4/3$.

On the other hand, staying with the condition $\overline{\mu} \leq 1$, the best spectral radius estimate for the inexact Uzawa and block triangular preconditioners is given in [18, Corollary 4.5] (combining results from there and Theorem 2 in [23]): $\rho(T) \leq \max(\sqrt{\rho_A}, \rho_S)$. Here again (see Theorem A.1), 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 (4.17) is optimal in these cases 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 results with two examples.

In Example 1, we consider the linear systems 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 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 (1.1) has to be solved at each time step. Here we consider a sample obtained setting $\mathbf{b}_A$ to a random vector of unit norm, while $\mathbf{b}_C = \mathbf{0}$.

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 [16] (see [15, 17] 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.

In Example 2, we consider the linear system resulting from the finite element discretization of the two-dimensional stationary Stokes problem referred to as Example 5.1.2 in [11]. We use the IFISS software [22] to generate the matrix and the right-hand side corresponding to stabilized Q_1/P_0 finite elements on an underlying 16×48 grid. Since the problem is stationary, the two diagonal blocks forming A are negative discrete Laplacians (without additional term). Hence for M_A we consider only the AGMG preconditioner. In this example, the Schur complement is tougher to accurately approximate. It is known to be spectrally equivalent to the pressure mass matrix, which is not easy to invert but is well conditioned [11]. A standard approach then takes for M_S the diagonal (or lumped) mass matrix, i.e., the diagonal matrix with the same row-sum as the true mass matrix.

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 5.1, where ω_A and ω_S refer to the overrelaxation factor applied when using M_A and M_S (respectively), that is, ω_A and ω_S are the inverse of the scaling factor applied to M_A and M_S. Indeed, without rescaling or (over)relaxation (ω_A, $\omega_S = 1$), all largest eigenvalues are equal to 1, except $\overline{\nu}$ in Example 2. Hence there is room to decrease the corresponding spectral radius by rescaling the preconditioner, and the provided values of ω_A and ω_S correspond to a rough optimization of ρ_A and ρ_S.

TABLE 5.1

Parameters for the examples as a function of the applied overrelaxation factor ω_A or ω_S.

	ω_A	$\underline{\mu}$	$\overline{\mu}$	ρ_A	ρ_A^2	ω_S	$\underline{\nu}$	$\overline{\nu}$	ρ_S	ρ_S^2
Example 1										
SGS for A	1.0	0.07	1.0	0.93	0.86	1.0	0.28	1.0	0.72	0.51
	1.7	0.12	1.7	0.88	0.77	1.5	0.43	1.5	0.57	0.33
AGMG for A	1.0	0.54	1.0	0.46	0.21	1.0	0.28	1.0	0.72	0.51
	1.3	0.70	1.3	0.30	0.09	1.5	0.43	1.5	0.57	0.33
Example 2	1.0	0.46	1.0	0.54	0.29	1.0	0.025	1.93	0.975	0.95
	1.3	0.57	1.3	0.40	0.16					

The convergence curves for the three main methods investigated here are reported in Figures 5.1–5.3. One sees that the symmetrized inexact Uzawa preconditioner is the winner or among the winners in all cases. Moreover, as predicted by our bound (4.7), improving ρ_A with rescaling enhances the convergence when ρ_A dominates ρ_S (Figure 5.1) and has little effect otherwise (Figures 5.2 and 5.3). Conversely, improving ρ_S with rescaling enhances the convergence when ρ_S dominates ρ_A (Figure 5.2) and has little effect otherwise (Figure 5.1).

Regarding the inexact block factorization preconditioner, it is generally as fast as the symmetrized inexact Uzawa preconditioner when our condition $\overline{\mu} \leq 1 + \frac{\rho_S}{1+\rho_S}$ to have (4.14) holds (all pictures except the right pictures in Figure 5.1). Otherwise, it may be competitive (Figure 5.1, top right picture), but convergence may also be lost (Figure 5.1, bottom right picture).

As expected rescaling M_A spoils the convergence of the inexact Uzawa algorithm, which is severely deteriorated in one case (Figure 5.2, top right picture) and lost in all other cases (other right pictures). Hence the restriction $\overline{\mu} \leq 1$ 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 Uzawa preconditioner if and only if ρ_S^2 dominates ρ_A. This is what happens in Figure 5.3, left picture, and in Figure 5.2, top left picture. In this latter case, observe, however, that when rescaling M_S (bottom left picture), ρ_S^2 still dominates ρ_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 Uzawa 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 their respective potentialities.

The symmetrized inexact Uzawa preconditioner (2.5) (see section 2 for its definition as a segregated iteration) can be recommended unconditionally, as it converges with a near optimal convergence rate as soon as $\rho_A < 1$ and $\rho_S < 1$.

The inexact block factorization preconditioner (1.5) has about the same cost per iteration and is often as fast, but it needs to satisfy a scaling condition on $\lambda_{\max}(M_A^{-1}A)$ 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 ρ_A.

The inexact Uzawa algorithm and the closely related block triangular preconditioners (1.6), (1.7) need also to satisfy a scaling condition to define a convergent iteration. Although our assumption $\lambda_{\max}(M_A^{-1}A) \leq 1$ is likely somewhat stricter
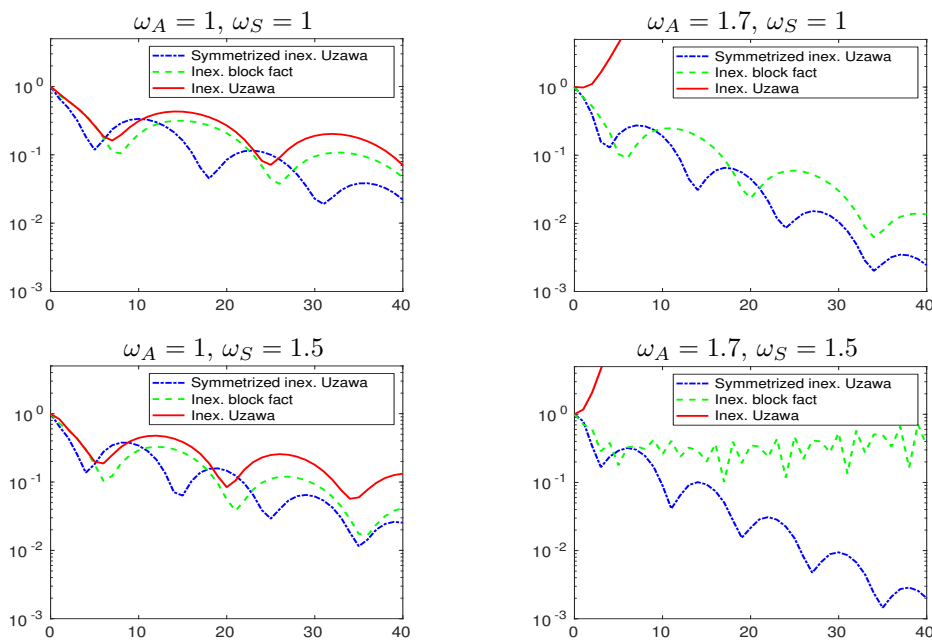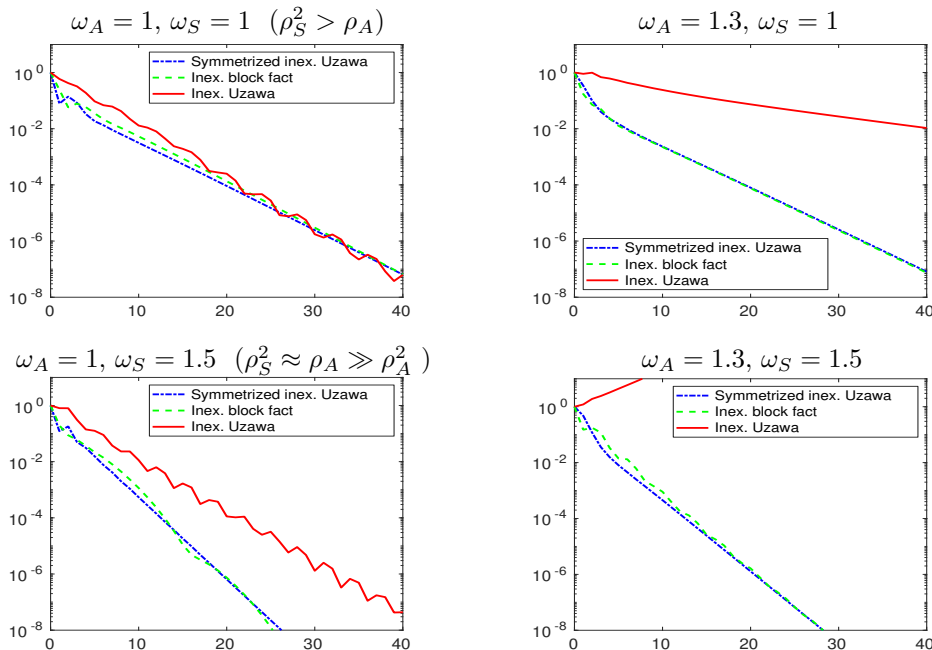
FIG. 5.1. *Residual norm as a function of the number of iterations for Example* 1 *with SGS preconditioner for A* ($\rho_A > \rho_S$ *in all cases, whereas* $\overline{\mu} = 1$ *for both left pictures and* $\overline{\mu} > 1 + \frac{\rho_S}{1+\rho_S} > 1$ *for both right pictures*).

FIG. 5.2. *Residual norm as a function of the number of iterations for Example* 1 *with AGMG preconditioner for A* ($\rho_S > \rho_A$ *in all cases, whereas* $\overline{\mu} = 1$ *for both left pictures and* $1 < \overline{\mu} < 1 + \frac{\rho_S}{1+\rho_S}$ *with* $\rho_S^2 \gg \rho_A^2$ *for both right pictures*).
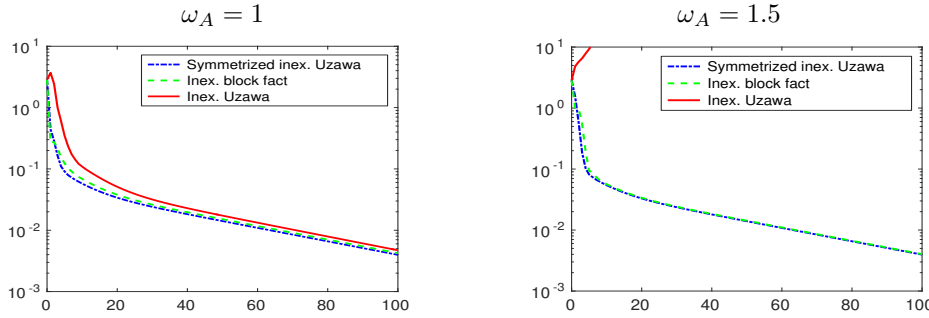
FIG. 5.3. *Residual norm as a function of the number of iterations for Example 2 ($\rho_S^2 \gg \rho_A$ in both cases, whereas $\overline{\mu} = 1$ for the left picture and $1 < \overline{\mu} < 1 + \frac{\rho_S}{1 + \rho_S}$ for the right picture).*

than necessary, it is clear that 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 Uzawa preconditioner requires two. They can thus be cost-effective and are in fact recommended if $\rho_S^2 \gg \rho_A$. This includes the cases where the Schur complement is tougher to approximate. Opposite to this, if $\rho_A \gg \rho_S$, the symmetrized inexact Uzawa preconditioner is likely better despite its higher cost, because its faster convergence can further be enhanced by optimizing ρ_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 A. Exact spectral radius in some limit cases.

THEOREM A.1. *Let K be a matrix of the form (1.2) such that A is an $n \times n$ SPD matrix and C is an $m \times m$ 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 $M_S = B A^{-1} B^T = S$, then, for the inexact block factorization preconditioner (1.5) and the symmetrized inexact Uzawa preconditioner (2.5), there holds*

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

whereas, for the inexact block triangular preconditioner (1.6) and the inexact Uzawa preconditioner (1.7) one has

$$\text{(A.2)} \qquad \rho\left(I - P^{-1}K\right) = \max\left(1 - \underline{\mu}, \ \overline{\mu} - 1 + \sqrt{\mu(\overline{\mu} - 1)}\right).$$

(2) *If $M_A = A$, for all four preconditioners defined by (1.5), (1.6), (1.7), (2.5), one has*

$$\text{(A.3)} \qquad \rho\left(I - P^{-1}K\right) = \rho_S.$$

Proof. To prove (A.1), we can consider that $I - T$ is similar to the matrix $\widehat{K}$ seen in the proof of Theorem 4.2 (equation (4.11)), where Λ is diagonal with diagonal entries equal to the eigenvalues of $M_A^{-1}A$, where $\overline{\Lambda} = \Lambda$ for the inexact block factorization preconditioner (1.5) and $\overline{\Lambda} = 2\Lambda - \Lambda^2$ for the symmetrized

inexact Uzawa preconditioner (2.5), where $\widetilde{C} = M_S^{-1/2} C M_S^{-1/2}$ is here the zero matrix and where $G = M_S^{1/2} B A^{-1/2} X$ with X orthogonal is here orthogonal because $M_S = S = B A^{-1} B^T$, that is, T is similar to

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

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. (1.5): $(1-\lambda)\begin{pmatrix} 1 & -\sqrt{\lambda} \\ \sqrt{\lambda} & 1 - \lambda \end{pmatrix};$ Prec. (2.5): $(1-\lambda)\begin{pmatrix} 1 - \lambda & -\sqrt{2\lambda - \lambda^2} \\ \sqrt{2\lambda - \lambda^2} & 1 - \lambda \end{pmatrix}.$

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 (A.1).

For (A.2), we first note that by Theorem 2.2, the spectral radius of the iteration matrix is necessarily the same for the inexact block triangular and the inexact Uzawa preconditioners (see also [18, Theorem 3.1]). Hence we may analyze the latter only, for which we find, using $I = S^{-1} B A^{-1} B$,

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

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

Regarding (A.3), we first observe that, when $M_A = A$, there is no more difference between the inexact block factorization preconditioner (1.5) and the symmetrized inexact Uzawa preconditioner (2.5). Moreover, $T_{(A^{-1},0)} = \left(T_{(A^{-1},0)}\right)^2$; hence one sees with Theorem 2.2 that the symmetrized inexact Uzawa 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

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

yielding straightforwardly the required result. □

REFERENCES

[1] S. Badia and R. Codina, *Algebraic pressure segregation methods for the incompressible Navier-Stokes equations*, Arch. Comput. Methods Eng., 15 (2007), pp. 1–52.

[2] R. E. Bank, B. D. Welfert, and H. Yserentant, *A class of iterative methods for solving saddle point problems*, Numer. Math., 56 (1990), pp. 645–666.

[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[4] M. Benzi and V. Simoncini, *On the eigenvalues of a class of saddle point matrices*, Numer. Math., 103 (2006), pp. 173–196.

[5] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.

[6] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.

[7] J. Cahouet and J.-P. Chabard, *Some fast 3D finite element solvers for the generalized Stokes problem*, Internat. J. Numer. Methods Fluids, 8 (1988), pp. 869–895.

[8] L. Chen and Y. Wu, *Convergence Analysis for A Class of Iterative Methods for Solving Saddle Point Systems*, arXiv:1710.03409v3[math.NA], 2018.

[9] D. Drzisga, L. John, U. Rde, B. Wohlmuth, and W. Zulehner, *On the analysis of block smoothers for saddle point problems*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 932–960.

[10] H. Elman, V. Howle, J. Shadid, R. Shuttleworth, and R. Tuminaro, *A taxonomy and comparison of parallel block multi-level preconditioners for the incompressible Navier-Stokes equations*, J. Comput. Phys., 227 (2008), pp. 1790–1808.

[11] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers*, Oxford University Press, Oxford, UK, 2005.

[12] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[13] A. Greenbaum, V. Pták, and Z. Strakoŝ, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.

[14] C. Hirsch, *Numerical Computation of Internal and External Flows, 2nd Ed.*, Elsevier, Amsterdam, 2003.

[15] A. Napov and Y. Notay, *An algebraic multigrid method with guaranteed convergence rate*, SIAM J. Sci. Comput., 34 (2012), pp. A1079–A1109.

[16] Y. Notay, *AGMG software and documentation.* http://agmg.eu.

[17] Y. Notay, *An aggregation-based algebraic multigrid method*, Electron. Trans. Numer. Anal., 37 (2010), pp. 123–146.

[18] Y. Notay, *A new analysis of block preconditioners for saddle point problems*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 143–173.

[19] Y. Notay, *Analysis of two-grid methods: The nonnormal case*, Tech. report GANMN 18–01, Université Libre de Bruxelles, Brussels, Belgium, 2018. http://homepages.ulb.ac.be/~ynotay/.

[20] S. V. Patankar and D. B. Spalding, *A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows*, Int. J. Heat Mass Transf., 15 (1972), pp. 1787–1806.

[21] J. Pestana and A. J. Wathen, *Natural preconditioning and iterative methods for saddle point systems*, SIAM Rev., 57 (2015), pp. 71–91.

[22] D. Silvester, H. Elman, and A. Ramage, *Incompressible Flow and Iterative Solver Software (IFISS) Version 3.2*, http://www.manchester.ac.uk/ifiss/ (2012).

[23] V. Simoncini, *Block triangular preconditioners for symmetric saddle-point problems*, Appl. Numer. Math., 49 (2004), pp. 63–80.

[24] Z. Tong and A. Sameh, *On an iterative method for saddle point problems*, Numer. Math., 79 (1998), pp. 643–646.

[25] H. Uzawa, *Iterative methods for concave programming*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stansford, CA, 1958, pp. 154–165.

[26] W. Zulehner, *Analysis of iterative methods for saddle point problems: A unified approach*, Math. Comp., 71 (2002), pp. 479–505.