MATHEMATICS OF COMPUTATION
Volume 71, Number 238, Pages 479–505
S 0025-5718(01)01324-2
Article electronically published on May 14, 2001

ANALYSIS OF ITERATIVE METHODS FOR SADDLE POINT PROBLEMS: A UNIFIED APPROACH

WALTER ZULEHNER

Abstract. In this paper two classes of iterative methods for saddle point problems are considered: inexact Uzawa algorithms and a class of methods with symmetric preconditioners. In both cases the iteration matrix can be transformed to a symmetric matrix by block diagonal matrices, a simple but essential observation which allows one to estimate the convergence rate of both classes by studying associated eigenvalue problems. The obtained estimates apply for a wider range of situations and are partially sharper than the known estimates in literature. A few numerical tests are given which confirm the sharpness of the estimates.

1. Introduction

In this paper we consider systems of linear equations of the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{1.1}$$

where A is a symmetric, positive definite $n \times n$-matrix, B is a $m \times n$-matrix with full rank $m \leq n$, and B^T denotes the transposed matrix of B.

Linear systems of the form (1.1) correspond to the Kuhn-Tucker conditions for linearly constrained quadratic programming problems or saddle point problems. Such systems typically result from mixed or hybrid finite element approximations of second-order elliptic problems, elasticity problems or the Stokes equations (see e.g. Brezzi, Fortin [5]) and from Lagrange multiplier methods (see e.g. Fortin, Glowinski [7]).

Under the assumptions mentioned above the coefficient matrix

$$\mathcal{K} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$$

is nonsingular and the (negative) Schur complement

$$C = BA^{-1}B^T$$

is symmetric and positive definite.

Received by the editor March 3, 1998 and, in revised form, February 11, 1999 and May 30, 2000.

2000 *Mathematics Subject Classification.* Primary 65N22, 65F10.

Key words and phrases. Indefinite systems, iterative methods, preconditioners, saddle point problems, Uzawa algorithm.

The system (1.1) can be reformulated in the following way:

$$\begin{aligned} Au &= f - B^T p, \\ Cp &= BA^{-1}f - g. \end{aligned}$$

This gives rise to the following theoretical solution approach: Starting with the second equation p can be computed. Then the first equation allows the computation of u. For large scale problems, however, the exact solution of systems of the form $Au = b$ and $Cp = c$ is not possible, in general. At best only approximate solvers for this type of equations are available. The question now is how to construct an efficient method out of the two building blocks of approximate solvers for the systems $Au = b$ and $Cp = c$.

A classical method of this type is the Uzawa algorithm [1], which relies on an exact solver for $Au = b$ and a Jacobi-like iteration for $Cp = c$. Several modifications have been suggested to avoid the exact solution of $Au = b$, reaching from Jacobi-like iterations (Arrow-Hurwics algorithm [1]) to multigrid methods (see Verfürth [12]). Methods of this type can be summarized as inexact Uzawa algorithms, whose convergence properties have been investigated, e.g., by Queck [10], Elman and Golub [6] and more recently by Bramble, Pasciak and Vassilev [4]. A second class of methods has been introduced and analyzed by Bank, Welfert and Yserentant [2]. There, an additional correction step for u is suggested leading to a symmetric preconditioner of $\mathcal{K}$.

The obvious advantage of methods of this type, sometimes called segregated methods, is that they rely only on efficient solvers for separated problems for u and p, which are often quite well understood. They can be implemented very easily as combination of available algorithms.

The aim of this paper is to present a convergence analysis for inexact Uzawa algorithms and for the class of methods with symmetric preconditioners on a common theoretical basis. The analysis has been strongly influenced by the theory in [4]. The main tool is summarized in Lemma 3.1 which gives lower and upper bounds for the eigenvalues of a generalized eigenvalue problem associated with the iteration matrices. The upper bounds yield bounds for the convergence rate of inexact Uzawa algorithms, which generalizes the results in [3], where only preconditioners of A were considered which underestimate A, and in [4], where only preconditioners of A were considered which overestimate A. Some of the estimates in these papers are improved. The lower bounds can be used to estimate the convergence of the class of methods with symmetric preconditioners. This approach differs completely from the technique in [2]. The consequences drawn from this new approach sharpens some of the consequences of the theory presented in [2], and special cases are identified which allow the application of an acceleration by a conjugate gradient technique for the total iteration, similar to the results in [3].

The paper is organized as follows. In Section 2 the two classes of iterative methods are shortly reviewed and, for either class, a factorization of the iteration matrix is derived. Section 3 contains the scaling technique and the necessary eigenvalue estimates, which are applied to the two classes of iterative methods in the subsequent Sections 4 and 5, respectively. Each of these two last sections is divided into three parts. The subsections on special cases present the results under assumptions which guarantee a spectrum of real numbers for the iteration matrices, the second part deals with the general case with full freedom with respect to the chosen norm, and, finally, the third part shows the implications if restrictions on the norms are to

be considered, which are typical in cases of variable preconditioners or acceleration techniques. Finally, Section 6 contains the results of a few numerical tests.

2. Basic iterative methods

Two classes of iterative methods are considered here. For each class a product representation of the iteration matrix is shown.

2.1. **Inexact Uzawa algorithms.** We start by considering the following class of methods:

$$\begin{aligned} \hat{A}(u_{k+1}-u_k) &= f - Au_k - B^T p_k, \\ \hat{C}(p_{k+1}-p_k) &= Bu_{k+1} - g, \end{aligned}$$

where $\hat{A}$ and $\hat{C}$ are symmetric positive definite matrices. This includes the classical Uzawa algorithm ($\hat{A} = A$, $\hat{C} = \gamma I$), and the classical Arrow-Hurwicz algorithm ($\hat{A} = \alpha I$, $\hat{C} = \gamma I$).

The method can be seen as a preconditioned Richardson method

$$\hat{\mathcal{K}}_1 \begin{pmatrix} u_{k+1}-u_k \\ p_{k+1}-p_k \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} - \mathcal{K} \begin{pmatrix} u_k \\ p_k \end{pmatrix}$$

with preconditioner

$$\hat{\mathcal{K}}_1 = \begin{pmatrix} \hat{A} & 0 \\ B & -\hat{C} \end{pmatrix}.$$

Let (u_*, p_*) be the exact solution of (1.1). For the error

$$\Delta u_k = u_k - u_*, \qquad \Delta p_k = p_k - p_*$$

we have

$$\begin{pmatrix} \Delta u_{k+1} \\ \Delta p_{k+1} \end{pmatrix} = \mathcal{M}_1 \begin{pmatrix} \Delta u_k \\ \Delta p_k \end{pmatrix},$$

where $\mathcal{M}_1$ denotes the iteration matrix of the inexact Uzawa algorithm, given by

$$\mathcal{M}_1 = I - \hat{\mathcal{K}}_1^{-1}\mathcal{K} = \begin{pmatrix} \hat{A}^{-1}(\hat{A}-A) & -\hat{A}^{-1}B^T \\ \hat{C}^{-1}B\hat{A}^{-1}(\hat{A}-A) & I - \hat{C}^{-1}B\hat{A}^{-1}B^T \end{pmatrix}.$$

It is easy to see that

$$\mathcal{M}_1 = -\begin{pmatrix} \hat{A}^{-1} & \hat{A}^{-1}B^T\hat{C}^{-1} \\ \hat{C}^{-1}B\hat{A}^{-1} & \hat{C}^{-1}B\hat{A}^{-1}B^T\hat{C}^{-1} - \hat{C}^{-1} \end{pmatrix} \begin{pmatrix} A-\hat{A} & 0 \\ 0 & \hat{C} \end{pmatrix} \tag{2.1}$$

$$= -\begin{pmatrix} \hat{A}^{-1} & 0 \\ 0 & \hat{C}^{-1} \end{pmatrix} \begin{pmatrix} \hat{A} & B^T \\ B & B\hat{A}^{-1}B^T - \hat{C} \end{pmatrix} \begin{pmatrix} \hat{A}^{-1}(A-\hat{A}) & 0 \\ 0 & I \end{pmatrix} \tag{2.2}$$

Remark 2.1. If we reverse the order in dealing with the equations, the following iterative process is obtained:

$$\begin{aligned} \hat{C}(p_{k+1}-p_k) &= Bu_k - g, \\ \hat{A}(u_{k+1}-u_k) &= f - Au_k - B^T p_{k+1}. \end{aligned}$$

The method can be seen as a preconditioned Richardson method with preconditioner

$$\hat{\mathcal{K}}_1^T = \begin{pmatrix} \hat{A} & B^T \\ 0 & -\hat{C} \end{pmatrix}.$$

It is easy to see that this iteration reduces to the previously discussed inexact Uzawa algorithm starting with (u_0, p_1), where p_1 is given by

$$\hat{C}(p_1 - p_0) = Bu_0 - g.$$

So, iteration methods of this type are included in the discussion of inexact Uzawa methods.

2.2. **A class of symmetric preconditioners.** The factorization

$$\mathcal{K} = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & -C \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & I \end{pmatrix}$$

motivates the use of another preconditioner

$$\hat{\mathcal{K}}_2 = \begin{pmatrix} \hat{A} & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \hat{A}^{-1} & 0 \\ 0 & -\hat{C} \end{pmatrix} \begin{pmatrix} \hat{A} & B^T \\ 0 & I \end{pmatrix}$$

(see [2]). This is equivalent to the iterative procedure

$$\begin{aligned} \hat{A}(\hat{u}_{k+1} - u_k) &= f - Au_k - B^T p_k, \\ \hat{C}(p_{k+1} - p_k) &= B\hat{u}_{k+1} - g, \\ \hat{A}(u_{k+1} - \hat{u}_{k+1}) &= -B^T(p_{k+1} - p_k). \end{aligned}$$

This method can be viewed as an inexact Uzawa algorithm with an additional correction step for u. It can also be interpreted as a correction method in the following sense: In a first step a preliminary approximation $\hat{u}_{k+1}$ is determined from

$$\hat{A}(\hat{u}_{k+1} - u_k) = f - Au_k - B^T p_k$$

in the same way as in the inexact Uzawa algorithm. In order to satisfy the equation $Bu = g$ one uses the ansatz

$$\hat{A}(u_{k+1} - \hat{u}_{k+1}) = -B^T(p_{k+1} - p_k). \tag{2.3}$$

Ideally speaking, this would lead to the correction equation

$$B\hat{A}^{-1}B^T \delta p_{k+1} = B\hat{u}_{k+1} - g.$$

The second step of the iterative method

$$\hat{C}(p_{k+1} - p_k) = B\hat{u}_{k+1} - g$$

can now be interpreted as approximating the solution of the correction equation starting from the initial approximation 0 for the correction δp_{k+1}. In this sense $\hat{C}$ approximates the inexact Schur complement $B\hat{A}^{-1}B^T$. Finally, the correction of u is done according to (2.3).

Simple calculation shows that the iteration matrix is now given by

$$\begin{aligned} \mathcal{M}_2 &= I - \hat{\mathcal{K}}_2^{-1}\mathcal{K} = \hat{\mathcal{K}}_2^{-1}(\hat{\mathcal{K}}_2 - \mathcal{K}) \\ &= -\begin{pmatrix} \hat{A} & B^T \\ B & B\hat{A}^{-1}B^T - \hat{C} \end{pmatrix}^{-1} \begin{pmatrix} A - \hat{A} & 0 \\ 0 & \hat{C} - B\hat{A}^{-1}B^T \end{pmatrix}. \end{aligned} \tag{2.4}$$

Summarizing both iteration methods, we have seen that

$$\mathcal{M}_i = \mathcal{P}_i \mathcal{N}_i \mathcal{Q}_i$$

for $i = 1, 2$, where $\mathcal{P}_i$ and $\mathcal{Q}_i$ are block diagonal matrices, and $\mathcal{N}_i$ is a symmetric matrix. The matrices are given for $i = 1$ either by (2.1) or (2.2), for $i = 2$ by (2.4).

In the next section the spectrum of such a matrix is investigated.

3. Scaling and spectral estimates

The results presented here apply to both classes of iteration methods. So, for a moment, we drop the subscripts and write

$$\mathcal{M} = \mathcal{P}\mathcal{N}\mathcal{Q}$$

with block diagonal matrices $\mathcal{P}$, $\mathcal{Q}$ and a symmetric matrix $\mathcal{N}$.

Let $\mathcal{D}$ and $\mathcal{E}$ be two further block matrices, which are also symmetric and positive definite. Then

$$\bar{\mathcal{M}} = \mathcal{D}^{1/2}\mathcal{M}\mathcal{D}^{-1/2} = \left[\mathcal{D}^{1/2}\mathcal{P}\mathcal{E}^{1/2}\right]\left[\mathcal{E}^{-1/2}\mathcal{N}\mathcal{E}^{-1/2}\right]\left[\mathcal{E}^{1/2}\mathcal{Q}\mathcal{D}^{-1/2}\right] = \bar{\mathcal{P}}\bar{\mathcal{N}}\bar{\mathcal{Q}}.$$

Here and in the sequel we use the following notation: For a real function $f\colon \mathbb{R} \to \mathbb{R}$ and a symmetric matrix M the matrix $f(M)$ is defined in the usual way via the spectral decomposition of M.

We discuss three different situations:

Special cases: If $\mathcal{D}$ and $\mathcal{E}$ can be chosen such that

$$\bar{\mathcal{P}} = \bar{\mathcal{Q}} = I, \tag{3.1}$$

then it follows that $\bar{\mathcal{M}} = \bar{\mathcal{N}}$, which implies that

1. $\mathcal{M}$ is symmetric with respect to the scalar product $\langle ., . \rangle_{\mathcal{D}}$, given by
$$\langle x, y \rangle_{\mathcal{D}} = \langle \mathcal{D}x, y \rangle,$$
where $\langle ., . \rangle$ denotes the ordinary Euclidean scalar product;
2. the spectrum of $\mathcal{M}$ and spectrum of $\bar{\mathcal{N}}$ coincide:
$$\sigma(\mathcal{M}) = \sigma(\bar{\mathcal{N}})$$
(here $\sigma(.)$ denotes the spectrum); and
3. the convergence rate of the iteration method is given by
$$\rho(\mathcal{M}) = \|\mathcal{M}\|_{\mathcal{D}} = \rho(\bar{\mathcal{N}}).$$
Here, $\rho(.)$ denotes the spectral radius and $\|.\|_{\mathcal{D}}$ the norm associated to the scalar product $\langle ., . \rangle_{\mathcal{D}}$.

It is easy to see that condition (3.1) can be satisfied if $\mathcal{P} = I$ and $\mathcal{Q}$ is symmetric and positive definite. Then the transformation matrices are given by

$$\mathcal{D} = \mathcal{Q}, \qquad \mathcal{E} = \mathcal{Q}^{-1}.$$

The above-mentioned conditions on $\mathcal{P}$ and $\mathcal{Q}$ correspond to the following special cases for the considered iteration methods:

Special case for the inexact Uzawa algorithm, see (2.1):

$$\hat{A} < A.$$

Here and in the sequel, we write $M < N$, respectively $M \leq N$, for symmetric matrices M and N if and only if $N - M$ is positive definite, respectively positive semidefinite.

Special case for the iteration method with symmetric preconditioners:

$$\hat{A} < A \quad \text{and} \quad \hat{C} > B\hat{A}^{-1}B^T.$$

There is a second special case for this method: If $\mathcal{D}$ and $\mathcal{E}$ are chosen such that

$$\bar{\mathcal{P}} = I \quad \text{and} \quad \bar{\mathcal{Q}} = -I, \tag{3.2}$$

then $\bar{\mathcal{M}} = -\bar{\mathcal{N}}$ with similar implications as before. Condition (3.2) can be satisfied if $\mathcal{P} = I$ and $\mathcal{Q}$ is symmetric and negative definite. Then the transformation matrices are given by

$$\mathcal{D} = -\mathcal{Q}, \qquad \mathcal{E} = -\mathcal{Q}^{-1}.$$

This corresponds to the special case

$$\hat{A} > A \quad \text{and} \quad \hat{C} < B\hat{A}^{-1}B^T$$

for the iteration method with symmetric preconditioners.

In each of these special cases the convergence analysis reduces to the eigenvalue problem

$$\mathcal{N}x = \nu\,\mathcal{E}x, \tag{3.3}$$

respectively to the eigenvalue problem $\mathcal{N}x = -\nu\,\mathcal{E}x$.

The general case: If one can find transformation matrices $\mathcal{D}$ and $\mathcal{E}$ with

$$\|\bar{\mathcal{P}}\| \le 1, \quad \|\bar{\mathcal{Q}}\| \le 1 \tag{3.4}$$

($\|.\|$ denotes the spectral norm), then we can still conclude that

$$\|\mathcal{M}\|_{\mathcal{D}} \le \rho(\bar{\mathcal{N}}).$$

Condition (3.4) leads to

$$\mathcal{P}\mathcal{E}\mathcal{P}^T \le \mathcal{D}^{-1} \quad \text{and} \quad \mathcal{Q}\mathcal{D}^{-1}\mathcal{Q}^T \le \mathcal{E}^{-1}.$$

So, a necessary condition for $\mathcal{E}$ is

$$\mathcal{Q}\mathcal{P}\mathcal{E}\mathcal{P}^T\mathcal{Q}^T \le \mathcal{E}^{-1}.$$

This is equivalent to

$$|\mathcal{E}^{1/2}\mathcal{Q}\mathcal{P}\mathcal{E}^{1/2}| \le I \tag{3.5}$$

if $\mathcal{Q}\mathcal{P}$ is symmetric, which is the case for all applications in this paper.

Now, if $\mathcal{E}$ is chosen such that condition (3.5) is satisfied and if we set

$$\mathcal{D} = \left(\mathcal{P}\mathcal{E}\mathcal{P}^T\right)^{-1},$$

one easily sees that condition (3.4) is satisfied, too.

Restrictions on the norm: For some situations, like the analysis of iteration methods with variable preconditioners or nonlinear iterations, it is necessary to (partially) prescribe the transformation matrix $\mathcal{D}$. Then it is still possible to satisfy (3.4): One easily sees that (3.4) follows if

$$\mathcal{E}^{-1} \ge \mathcal{P}^T\mathcal{D}\mathcal{P} \quad \text{and} \quad \mathcal{E}^{-1} \ge \mathcal{Q}\mathcal{D}^{-1}\mathcal{Q}^T.$$

This discussion shows the importance of the generalized eigenvalue problem (3.3) for the analysis of the iteration methods. The next lemma is fundamental for estimating the eigenvalues of (3.3).

Lemma 3.1. *Let $\hat{A}$ and Q be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ and S symmetric, positive definite $m \times m$-matrices.*

Assume that there are real numbers $\rho_1 \le \rho_2 \le 0 < \rho_3 \le \rho_4 < \rho_5 \le \rho_6$ with

$$\varphi(\rho_1) \ge 0, \quad \varphi(\rho_2) \le 0, \quad \varphi(\rho_5) \ge 0, \quad \varphi(\rho_6) \le 0, \tag{3.6}$$

where

$$\varphi(\mu) = \mu\, B(\mu\, \hat{A} - \hat{A}Q^{-1}\hat{A})^{-1}B^T - \mu\, S - \hat{C},$$

and

$$\rho_3\, Q \le \hat{A} \le \rho_4\, Q.$$

Then we have: If λ is an eigenvalue of the generalized eigenvalue problem

$$\begin{pmatrix} \hat{A} & B^T \\ B & B\hat{A}^{-1}B^T - \hat{C} \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} Qu \\ Sp \end{pmatrix}, \tag{3.7}$$

then

$$\lambda \in [\rho_1, \rho_2] \cup [\rho_3, \rho_4] \cup [\rho_5, \rho_6].$$

Proof. Let $(u, p) \in \mathbb{R}^n \times \mathbb{R}^m$ with $(u, p) \ne (0, 0)$ and $\lambda \in \mathbb{R}$ satisfy

$$\begin{aligned} \hat{A}u + B^T p &= \lambda\, Qu, \\ Bu + (B\hat{A}^{-1}B^T - \hat{C})p &= \lambda\, Sp. \end{aligned}$$

Assume that $\lambda \notin [\rho_3, \rho_4]$. Then $\lambda\, Q - \hat{A}$ is either positive or negative definite and $p \ne 0$. In this case we obtain from the first equation

$$u = (\lambda\, Q - \hat{A})^{-1}B^T p.$$

Then the second equation yields

$$B(\lambda\, Q - \hat{A})^{-1}B^T p + (B\hat{A}^{-1}B^T - \hat{C})p = \lambda\, Sp,$$

or, equivalently,

$$\lambda\, B(\lambda\, \hat{A} - \hat{A}Q^{-1}\hat{A})^{-1}B^T p = (\lambda\, S + \hat{C})p.$$

Multiplying this equation by p^T from the left yields

$$\langle \varphi(\lambda)p, p \rangle = 0$$

with

$$\varphi(\mu) = \mu\, B(\mu\, \hat{A} - \hat{A}Q^{-1}\hat{A})^{-1}B^T - \mu\, S - \hat{C}.$$

It is easy to see that $\langle \varphi(\mu)p, p \rangle$ is strictly decreasing in μ on each interval outside the set $[\rho_3, \rho_4]$.

From the assumptions (3.6) it follows that

$$\langle \varphi(\rho_1)p, p \rangle \ge 0, \quad \langle \varphi(\rho_2)p, p \rangle \le 0, \quad \langle \varphi(\rho_5)p, p \rangle \ge 0, \quad \langle \varphi(\rho_6)p, p \rangle \le 0.$$

Then the estimates are direct consequences of the monotony of $\langle \varphi(\mu)p, p \rangle$. □

Remark 3.2. Additionally, one easily verifies that

$$\langle \varphi(0)p, p \rangle < 0 \quad \text{and} \quad \langle \varphi(-\rho_0)p, p \rangle \ge 0,$$

where ρ_0 is a positive real number satisfying the condition

$$\hat{C} \le \rho_0\, S.$$

This shows that it is always possible to choose ρ_2, such that $\rho_2 < 0$, and ρ_1, such that $-\rho_0 \le \rho_1$.

If, additionally, B has full rank $m \le n$, then

$$\langle \varphi(-\rho_0)p, p\rangle > 0,$$

i.e., ρ_1 can even be chosen, such that $-\rho_0 < \rho_1$.

Remark 3.3. For the special case $\hat{A} = A$, $\hat{C} = BA^{-1}B^T$ this lemma immediately yields the spectral estimates by Iliash, Rossi and Toivanen [9], provided spectral inequalities of the form

$$\eta_1 Q \le A \le \eta_2 Q$$

and

$$\theta_1 S \le BA^{-1}B^T \le \theta_2 S$$

are assumed. Especially for $\hat{A} = Q = A$ and $\hat{C} = S = BA^{-1}B^T$ we obtain the exact bounds $\rho_1 = \rho_2 = (1-\sqrt{5})/2$, $\rho_3 = \rho_4 = 1$ and $\rho_5 = \rho_6 = (1+\sqrt{5})/2$, a simple but interesting observation due to Yu. Kuznetsov, 1990. Spectral estimates of this kind are used to analyze block diagonal preconditioners (see e.g. Silvester and Wathen [11]). So, as a by-product of the convergence theory presented here, the known results on iteration methods with block diagonal preconditioners are reproduced. However, compared to literature, no new information is gained.

4. Convergence results for inexact Uzawa algorithms

The analysis of inexact Uzawa algorithms is based on the setting

$$\mathcal{P} = I, \quad \mathcal{Q} = \begin{pmatrix} A - \hat{A} & 0 \\ 0 & \hat{C} \end{pmatrix}$$

for the special case (see (2.1)), and on the setting

$$\mathcal{P} = \begin{pmatrix} \hat{A}^{-1} & 0 \\ 0 & \hat{C}^{-1} \end{pmatrix}, \quad \mathcal{Q} = \begin{pmatrix} \hat{A}^{-1}(A - \hat{A}) & 0 \\ 0 & I \end{pmatrix}$$

for the general case (see (2.2)).

First, we discuss the

4.1. **Special case: $\hat{A} < A$.** From the discussion in Section 3 we obtain the following scaling matrices in this case:

$$\mathcal{D} = \begin{pmatrix} A - \hat{A} & 0 \\ 0 & \hat{C} \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} (A - \hat{A})^{-1} & 0 \\ 0 & \hat{C}^{-1} \end{pmatrix},$$

leading to the norm $\|.\|_{\mathcal{D}}$, given by the scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{D}} = \langle (A - \hat{A})u, v\rangle + \langle \hat{C}p, q\rangle. \tag{4.1}$$

The generalized eigenvalue problem (3.3) leads to a generalized eigenvalue problem of the form (3.7) with

$$\lambda = -\nu, \quad Q = \hat{A}(A - \hat{A})^{-1}\hat{A}, \quad S = \hat{C}. \tag{4.2}$$

For the preconditioners $\hat{A}$ and $\hat{C}$ the following spectral inequalities are assumed: There is a constant $\alpha_2 \in \mathbb{R}$ with $1 < \alpha_2$, such that

$$\hat{A} < A \leq \alpha_2 \hat{A}, \tag{4.3}$$

and there are constants $\gamma_1, \gamma_2 \in \mathbb{R}$ with $0 < \gamma_1 \leq \gamma_2$, such that

$$\gamma_1 \hat{C} \leq BA^{-1}B^T \leq \gamma_2 \hat{C}. \tag{4.4}$$

A simple consequence of Lemma 3.1 is

Theorem 4.1. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (4.3) *and* (4.4). *Then we have:*

1. *The iteration matrix $\mathcal{M}_1$ of the inexact Uzawa algorithm is symmetric with respect to the scalar product* (4.1).
2. $\sigma(\mathcal{M}_1) \subset [\rho_1, \rho_2] \subset (-\infty, 1)$ *with*

$$\begin{aligned} \rho_1 &= \frac{2-(1+\gamma_2)\alpha_2}{2} - \sqrt{\frac{[2-(1+\gamma_2)\alpha_2]^2}{4} + \alpha_2 - 1}, \\ \rho_2 &= \frac{2-(1+\gamma_1)\alpha_2}{2} + \sqrt{\frac{[2-(1+\gamma_1)\alpha_2]^2}{4} + \alpha_2 - 1}. \end{aligned}$$

3. *If $\alpha_2(2+\gamma_2) < 4$, then $\rho(\mathcal{M}_1) = \|\mathcal{M}_1\|_{\mathcal{D}} < 1$.*
4. *$\hat{\mathcal{K}}_1^{-1}\mathcal{K}$ is symmetric and positive definite with respect to the scalar product* (4.1) *and $\sigma(\hat{\mathcal{K}}_1^{-1}\mathcal{K}) \subset [1-\rho_2, 1-\rho_1] \subset (0, \infty)$.*

Proof. With the notation of Section 3 we have: $\bar{\mathcal{M}}_1 = \bar{\mathcal{N}}_1$, where $\bar{\mathcal{N}}_1$ is symmetric. This shows the symmetry of $\mathcal{M}_1$ in the corresponding scalar product (4.1).

A number ν is an eigenvalue of $\bar{\mathcal{N}}_1$ if and only if $\lambda = -\nu$ is an eigenvalue of the generalized eigenvalue problem (3.7) with the setting (4.2).

The matrix function $\varphi(\mu)$ of Lemma 3.1 is given by

$$\varphi(\mu) = \mu B[(\mu+1)\hat{A} - A]^{-1}B^T - (\mu+1)\hat{C}.$$

Assume that $-1 \leq \mu \leq 0$. Then it follows from (4.4) that $\varphi(\mu) \geq 0$ if

$$\mu\,[(\mu+1)\hat{A} - A]^{-1} \geq \frac{\mu+1}{\gamma_1} A^{-1},$$

or, equivalently,

$$\mu\,\gamma_1 A \leq (\mu+1)\,[(\mu+1)\hat{A} - A].$$

Set $\bar{A} = \hat{A}^{-1/2} A \hat{A}^{-1/2}$. Then this inequality becomes

$$\mu\,\gamma_1 \bar{A} \leq (\mu+1)\,[(\mu+1)I - \bar{A}],$$

which is satisfied if and only if

$$\mu\,\gamma_1\,\bar{a} \leq (\mu+1)\,(\mu+1-\bar{a}) \tag{4.5}$$

for all eigenvalues $\bar{a}$ of $\bar{A}$. From (4.3) we know that $\sigma(\bar{A}) \subset (1, \alpha_2]$. It is easy to see that (4.5) is satisfied for all $\bar{a} \in (1, \alpha_2]$ if and only if it is satisfied for the extreme value $\bar{a} = \alpha_2$:

$$\mu\,\gamma_1\,\alpha_2 \leq (\mu+1)\,(\mu+1-\alpha_2),$$

which is equivalent to $\mu \leq \mu_{21}^-$, where μ_{21}^- is the negative root of the quadratic equation

$$\mu\,\gamma_1\,\alpha_2 = (\mu+1)\,(\mu+1-\alpha_2).$$

One easily verifies that $-1 < \mu_{21}^- \leq 0$. Therefore, we have $\varphi(\mu_{21}^-) \geq 0$, which, by Lemma 3.1, implies the bound $\lambda \geq \mu_{21}^-$ for the eigenvalues of (3.7) and the bound $\nu = -\lambda \leq -\mu_{21}^- = \rho_2$ for the eigenvalues of $\mathcal{M}_1$.

Observe that $\hat{A} \leq (\alpha_2 - 1)\,Q$. Then, for $\mu > \alpha_2 - 1$, it follows from (4.4) that $\varphi(\mu) \leq 0$ if

$$\mu\,[(\mu+1)\,\hat{A} - A]^{-1} \leq \frac{\mu+1}{\gamma_2}\,A^{-1},$$

or, equivalently,

$$\mu\,\gamma_2\,A \leq (\mu+1)\,[(\mu+1)\,\hat{A} - A].$$

In the same way as before, it can be shown that this inequality is satisfied if

$$\mu\,\gamma_2\,\alpha_2 \leq (\mu+1)\,(\mu+1-\alpha_2),$$

which is equivalent to $\mu \geq \mu_{22}^+$, where μ_{22}^+ is the positive root of the quadratic equation

$$\mu\,\gamma_2\,\alpha_2 = (\mu+1)\,(\mu+1-\alpha_2).$$

One easily verifies that $\mu_{22}^+ > \alpha_2 - 1$. Therefore, we have $\varphi(\mu_{22}^+) \leq 0$, which, by Lemma 3.1, implies the bound $\lambda \leq \mu_{22}^+$ for the eigenvalues of (3.7) and the bound $\nu = -\lambda \geq -\mu_{22}^+ = \rho_1$ for the eigenvalues of $\mathcal{M}_1$.

Statement 3 easily follows from 2, and 4 is an immediate consequence of 1 and 2. □

Remark 4.2. Statement 4 of Theorem 4.1 justifies the use of a conjugate gradient acceleration for inexact Uzawa algorithms, which was the key observation in [3]. The last part of Theorem 4.1 provides a conditioning estimate:

$$(1-\rho_2)\,\langle x,x\rangle_{\mathcal{D}} \leq \langle \hat{\mathcal{K}}_1^{-1}\mathcal{K}x,x\rangle_{\mathcal{D}} \leq (1-\rho_1)\,\langle x,x\rangle_{\mathcal{D}}.$$

Compared to the basic estimate in Theorem 1 in [3], which deals only with the special case $\hat{C} = BA^{-1}B^T$, the upper bounds agree, while the lower bound given here is sharper. Translating our results into the terminology of [3], Theorem 1 in [3] deals with the special case $\gamma_1 = \gamma_2 = 1$ and uses the notations $\hat{A} = A_0$, $\alpha_2 = 1/(1-\alpha)$ with $\alpha < 1$. Then, the lower bound in our theory becomes:

$$1 - \rho_2 = \alpha_2 - \sqrt{(\alpha_2-1)^2 + \alpha_2 - 1} = \frac{1-\sqrt{\alpha}}{1-\alpha}.$$

The coinciding upper bound in the notation of [3] reads:

$$1 - \rho_1 = \alpha_2 + \sqrt{(\alpha_2-1)^2 + \alpha_2 - 1} = \frac{1+\sqrt{\alpha}}{1-\alpha}.$$

An example in [3] shows the sharpness of these bounds, in particular, the sharpness of the lower bound given here.

As far as quantitative conditioning estimates are concerned, Theorem 1 in [3] covers only the case of using the exact Schur complement. In the general case, where the exact Schur complement is replaced by a preconditioner $\hat{C}$ one could, of course, derive conditioning estimates by comparing the overall preconditioning

matrix with the theoretic preconditioner of Theorem 1 in [3] (see Remark 2 in [3]). Generally speaking, this approach would lead to less sharp estimates than our approach, which directly provides conditioning estimates without the use of an intermediate theoretic preconditioner.

4.2. **The general case.** Now we assume the following spectral inequalities for the preconditioner $\hat{A}$: There are constants $\alpha_1, \alpha_2 \in \mathbb{R}$ with $0 < \alpha_1 \leq 1 \leq \alpha_2$, such that

$$\alpha_1 \hat{A} \leq A \leq \alpha_2 \hat{A}. \tag{4.6}$$

Additionally, we assume that at least one of the inequalities $\alpha_1 \leq 1$ or $\alpha_1 \hat{A} \leq A$ and at least one of the inequalities $1 \leq \alpha_2$ or $A \leq \alpha_2 \hat{A}$ hold strictly. Furthermore, we introduce

$$q_\alpha = \max(1 - \alpha_1, \alpha_2 - 1),$$

which is an upper bound of the convergence rate $\|I - \hat{A}^{-1}A\|_{\hat{A}}$ of the Richardson method for an equation of the form $Au = b$, preconditioned with $\hat{A}$.

As before, for the preconditioner $\hat{C}$ the following spectral inequalities are assumed: There are constants $\gamma_1, \gamma_2 \in \mathbb{R}$ with $0 < \gamma_1 \leq \gamma_2$, such that

$$\gamma_1 \hat{C} \leq BA^{-1}B^T \leq \gamma_2 \hat{C}. \tag{4.7}$$

According to the discussion in Section 3 we look for a scaling matrix

$$\mathcal{E} = \begin{pmatrix} E_u & 0 \\ 0 & E_p \end{pmatrix},$$

such that

$$|\mathcal{E}^{1/2} \mathcal{Q} \mathcal{P} \mathcal{E}^{1/2}| \leq I,$$

or, in details,

$$|E_u^{1/2} \hat{A}^{-1}(A - \hat{A})\hat{A}^{-1} E_u^{1/2}| \leq I \quad \text{and} \quad |E_p^{1/2} \hat{C}^{-1} E_p^{1/2}| \leq I.$$

Roughly speaking, the larger $\mathcal{E}$ is with respect to the ordering $\leq$ of matrices, introduced in Section 3, the smaller are the eigenvalues in (3.3). In this sense E_u and E_p should be as large as possible.

The optimal choice for E_p is obviously given by

$$E_p = \hat{C}.$$

For E_u we set

$$E_u = \hat{A}^{1/2} f(\bar{A}) \hat{A}^{1/2}, \tag{4.8}$$

where $\bar{A} = \hat{A}^{-1/2} A \hat{A}^{-1/2}$ and $f\colon (0, \infty) \to (0, \infty]$ is an arbitrary function.

Then the condition on E_u reads: All eigenvalues λ of the generalized eigenvalue problem

$$\hat{A}^{-1}(A - \hat{A})\hat{A}^{-1} u = \lambda E_u^{-1} u$$

or, equivalently,

$$(\bar{A} - I)u = \lambda f(\hat{A})^{-1} u$$

have to satisfy the condition: $|\lambda| \leq 1$.

From (4.6) we know: $\sigma(\bar{A}) \subset [\alpha_1, \alpha_2]$. The eigenvalues λ result from the eigenvalues $\bar{a}$ of $\bar{A}$ by the formula

$$\lambda = (\bar{a} - 1) f(\bar{a}).$$

The condition $|\lambda| \leq 1$ is certainly satisfied if

$$|x-1|f(x) \leq 1 \quad \text{for all } x \in [\alpha_1, \alpha_2].$$

This condition holds for the function f, given by

$$f(x)^{-1} = (1-\alpha_1)\frac{\alpha_2 - x}{\alpha_2 - \alpha_1} + (\alpha_2 - 1)\frac{x - \alpha_1}{\alpha_2 - \alpha_1} = \alpha_0\, x + \hat{\alpha} \tag{4.9}$$

with

$$\alpha_0 = \frac{\alpha_1 + \alpha_2 - 2}{\alpha_2 - \alpha_1}, \quad \hat{\alpha} = \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{\alpha_2 - \alpha_1}.$$

Observe that

$$|x-1|f(x) = 1 \quad \text{for } x = \alpha_i,$$

except for the case $\alpha_i = 1$, $i = 1, 2$. So, f is as large as possible at least at the boundary points of the interval which contains the spectrum of $\bar{A}$.

The additional conditions on α_1 and α_2 guarantee that E_u is well defined and positive definite.

The resulting matrix $\mathcal{D}$ is given by

$$\mathcal{D} = \begin{pmatrix} \hat{\alpha}\,\hat{A} + \alpha_0\, A & 0 \\ 0 & \hat{C} \end{pmatrix}$$

with associated scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{D}} = \left\langle \left(\hat{\alpha}\,\hat{A} + \alpha_0\, A\right) u, v \right\rangle + \langle \hat{C}p, q \rangle.$$

Then a simple consequence of Lemma 3.1 is

Theorem 4.3. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (4.6) *and* (4.7).

Then we have with the notations of Section 2 and 3:

1. $\sigma(\bar{\mathcal{N}}_1) \subset [\min(-q_\alpha, \rho_1, \rho_3), \max(\rho_2, \rho_4)]$ *with*

$$\begin{aligned}
q_\alpha &= \max(1-\alpha_1, \alpha_2 - 1), \\
\rho_1 &= \frac{2-(1+\gamma_2)\alpha_2}{2} - \sqrt{\frac{[2-(1+\gamma_2)\alpha_2]^2}{4} + \alpha_2 - 1}, \\
\rho_2 &= \frac{2-(1+\gamma_1)\alpha_2}{2} + \sqrt{\frac{[2-(1+\gamma_1)\alpha_2]^2}{4} + \alpha_2 - 1}, \\
\rho_3 &= \frac{(1-\gamma_2)\alpha_1}{2} - \sqrt{\frac{(1-\gamma_2)^2\alpha_1^2}{4} + 1 - \alpha_1}, \\
\rho_4 &= \frac{(1-\gamma_1)\alpha_1}{2} + \sqrt{\frac{(1-\gamma_1)^2\alpha_1^2}{4} + 1 - \alpha_1}.
\end{aligned}$$

2. $\|\mathcal{M}_1\|_{\mathcal{D}} \leq \max(q_\alpha, -\rho_1, -\rho_3, \rho_2, \rho_4)$.
3. *If* $\alpha_2(2+\gamma_2) < 4$, *then* $\|\mathcal{M}_1\|_{\mathcal{D}} < 1$.

Proof. With the notation of Section 3 we have

$$\|\mathcal{M}_1\|_{\mathcal{D}} \leq \rho(\bar{\mathcal{N}}_1)$$

and ν is an eigenvalue of $\bar{\mathcal{N}}_1$ if and only if $\lambda = -\nu$ is an eigenvalue of the generalized eigenvalue problem (3.7) with the setting $S = \hat{C}$, $Q = E_u$ (see (4.8), (4.9)).

The matrix function $\varphi(\mu)$ of Lemma 3.1 is given by

$$\varphi(\mu) = \mu\, B\hat{A}^{-1/2}\left(\mu\, I - f(\bar{A})^{-1}\right)^{-1}\hat{A}^{-1/2}B^T - (\mu+1)\,\hat{C}.$$

Assume that $-1 \le \mu \le 0$. Then it follows from (4.7) that $\varphi(\mu) \ge 0$ if

$$\mu\,\hat{A}^{-1/2}\left(\mu\, I - f(\bar{A})^{-1}\right)^{-1}\hat{A}^{-1/2} \ge \frac{\mu+1}{\gamma_1}\,A^{-1},$$

or, equivalently, if

$$\mu\,\gamma_1\,\bar{A} \le (\mu+1)\left(\mu\, I - f(\bar{A})^{-1}\right).$$

This means in terms of the eigenvalues $\bar{a}$ of $\bar{A}$:

$$\mu\,\gamma_1\,\bar{a} \le (\mu+1)\left(\mu - f(\bar{a})^{-1}\right). \tag{4.10}$$

Since $\sigma(\bar{A}) \subset [\alpha_1, \alpha_2]$, it suffices to fulfill (4.10) for all $\bar{a} \in [\alpha_1, \alpha_2]$. It is easy to see that (4.10) is satisfied for all $\bar{a} \in [\alpha_1, \alpha_2]$, if and only if it is satisfied for the extreme value α_1 and α_2:

$$\mu\,\gamma_1\,\alpha_1 \le (\mu+1)\,(\mu - 1 + \alpha_1),$$

and

$$\mu\,\gamma_1\,\alpha_2 \le (\mu+1)\,(\mu - \alpha_2 + 1),$$

which is equivalent to $\mu \le \mu_{11}^-$ and $\mu \le \mu_{21}^-$, where μ_{11}^- is the negative root of the quadratic equation

$$\mu\,\gamma_1\,\alpha_1 = (\mu+1)\,(\mu - 1 + \alpha_1)$$

and μ_{21}^- is the negative root of the quadratic equation

$$\mu\,\gamma_1\,\alpha_2 = (\mu+1)\,(\mu - \alpha_2 + 1).$$

One easily verifies that $-1 < \mu_{11}^- \le 0$ and $-1 < \mu_{21}^- \le 0$.

Therefore, we have $\varphi(\min(\mu_{11}^-, \mu_{21}^-)) \ge 0$, which, by Lemma 3.1, implies the bound $\lambda \ge \min(\mu_{11}^-, \mu_{21}^-)$ for the eigenvalues of (3.7) and the bound $\nu = -\lambda \le \max(\rho_2, \rho_4)$ for the eigenvalues of $\bar{\mathcal{N}}_1$ with $\rho_2 = -\mu_{21}^-$ and $\rho_4 = -\mu_{11}^-$.

Observe that $q_\alpha\, Q \ge \hat{A}$. Then, for $\mu > q_\alpha$, one obtains completely analogously to above that $\varphi(\mu) \le 0$ if $\mu \ge \max(\mu_{22}^+, \mu_{12}^+)$, where μ_{12}^+ is the positive root of the quadratic equation

$$\mu\,\gamma_2\,\alpha_1 = (\mu+1)\,(\mu - 1 + \alpha_1)$$

and μ_{22}^+ is the positive root of the quadratic equation

$$\mu\,\gamma_2\,\alpha_2 = (\mu+1)\,(\mu - \alpha_2 + 1).$$

Therefore, we have $\lambda \le \max(q_\alpha, \mu_{22}^+, \mu_{12}^+)$ for the eigenvalues of (3.7) and $\nu = -\lambda \ge \min(-q_\alpha, \rho_1, \rho_3)$ for the eigenvalues of $\bar{\mathcal{N}}_1$ with $\rho_1 = -\mu_{22}^+$ and $\rho_3 = -\mu_{12}^+$.

The statements 2 and 3 easily follow from 1. □

Remark 4.4. This theorem contains the special case $\alpha_1 < 1$, $\alpha_2 = 1$, $\gamma_1 < 1$ and $\gamma_2 = 1$, discussed in [4]. For this case we have

$$\rho_1 = 0, \quad \rho_2 = 1 - \gamma_1, \quad \rho_3 = -\sqrt{1 - \alpha_1}$$

and

$$\max(q_\alpha, -\rho_1, -\rho_3, \rho_2, \rho_4) = \rho_4,$$

which gives the estimate

$$\|\mathcal{M}_1\|_{\mathcal{D}} \leq \frac{(1-\gamma_1)\alpha_1}{2} + \sqrt{\frac{(1-\gamma_1)^2\alpha_1^2}{4} + 1 - \alpha_1},$$

or in terms of the notations $\alpha_1 = 1-\delta$, $\gamma_1 = 1-\gamma$ used in [4]:

$$\|\mathcal{M}_1\|_{\mathcal{D}} \leq \frac{\gamma(1-\delta)}{2} + \sqrt{\frac{\gamma^2(1-\delta)^2}{4} + \delta}.$$

This agrees with the results in [4].

4.3. **The general case with $\hat{A}$-independent norm.** The norms introduced so far for analyzing the convergence depend on the preconditioner $\hat{A}$. If the preconditioner is allowed to change during the iteration, these norms change, too. A convergence analysis based on these changing norms is very complicated if not impossible. In this situation it is more advisable to use a norm which is independent of $\hat{A}$. (The case of variable preconditioners for A also includes the case of using an inner iteration for solving an equation of the form $Au = b$, such as a conjugate gradient method. In Lemma 9 in [2] it is shown how to represent such an inner iteration by a variable preconditioner.)

Therefore we are now looking for estimates in a norm $\|.\|_{\mathcal{D}}$, given by the scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{D}} = \alpha_0 \langle Au, v\rangle + \langle \hat{C}p, q\rangle,$$

with some scalar factor $\alpha_0 > 0$. So, in the terminology of Section 3, we prescribe the scaling matrix $\mathcal{D}$ by

$$\mathcal{D} = \begin{pmatrix} \alpha_0 A & 0 \\ 0 & \hat{C} \end{pmatrix}$$

up to the factor $\alpha_0 > 0$.

Then, according to the discussion in Section 3, $\mathcal{E}$ must satisfy the inequalities

$$\mathcal{E}^{-1} \geq \begin{pmatrix} \alpha_0 \hat{A}^{-1} A \hat{A}^{-1} & 0 \\ 0 & \hat{C}^{-1} \end{pmatrix}$$

and

$$\mathcal{E}^{-1} \geq \begin{pmatrix} \frac{1}{\alpha_0} \hat{A}^{-1}(A-\hat{A})A^{-1}(A-\hat{A})\hat{A}^{-1} & 0 \\ 0 & \hat{C}^{-1} \end{pmatrix}.$$

If we set, as before,

$$\mathcal{E} = \begin{pmatrix} E_u & 0 \\ 0 & E_p \end{pmatrix}$$

with $E_p = \hat{C}$ and $E_u = \hat{A}^{1/2} f(\bar{A}) \hat{A}^{1/2}$, these inequalities reduce to

$$\alpha_0 \, f(\bar{A}) \, \bar{A} \leq I \quad \text{and} \quad \frac{1}{\alpha_0} \, f(\bar{A}) \, (\bar{A} + \bar{A}^{-1} - 2\, I) \leq I.$$

For this it suffices to choose f such that

$$f(x) \, \max\left(\alpha_0 \, x, \frac{1}{\alpha_0} \, \left(x + \frac{1}{x} - 2\right)\right) \leq 1 \quad \text{for all } x \in [\alpha_1, \alpha_2].$$

This condition is satisfied for the function f, given by

$$f(x)^{-1} = q_1 \frac{\alpha_2 - x}{\alpha_2 - \alpha_1} + q_2 \frac{x - \alpha_1}{\alpha_2 - \alpha_1}, \tag{4.11}$$

with

$$\begin{aligned} q_1 &= \max\left(\alpha_0\,\alpha_1, \frac{1}{\alpha_0}\,\left(\alpha_1 + \frac{1}{\alpha_1} - 2\right)\right), \\ q_2 &= \max\left(\alpha_0\,\alpha_2, \frac{1}{\alpha_0}\,\left(\alpha_2 + \frac{1}{\alpha_2} - 2\right)\right). \end{aligned}$$

Observe that

$$q_1 \geq 1 - \alpha_1 \quad \text{and} \quad q_2 \geq \alpha_2 - 1.$$

Then a simple consequence of Lemma 3.1 is

Theorem 4.5. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (4.6) *and* (4.7). *Then we have with the notations of Section 2 and 3:*

1. $\sigma(\bar{\mathcal{N}}_1) \subset [\min(-\widetilde{q}_A, \widetilde{\rho}_1, \widetilde{\rho}_3), \max(\widetilde{\rho}_2, \widetilde{\rho}_4)]$ *with*

$$\begin{aligned} \widetilde{q}_A &= \max(q_1, q_2), \\ \widetilde{\rho}_1 &= \frac{1 - q_2 - \gamma_2\alpha_2}{2} - \sqrt{\frac{[1 - q_2 - \gamma_2\alpha_2]^2}{4} + q_2}, \\ \widetilde{\rho}_2 &= \frac{1 - q_2 - \gamma_1\alpha_2}{2} + \sqrt{\frac{[1 - q_2 - \gamma_1\alpha_2]^2}{4} + q_2}, \\ \widetilde{\rho}_3 &= \frac{1 - q_1 - \gamma_2\alpha_1}{2} - \sqrt{\frac{[1 - q_1 - \gamma_2\alpha_1]^2}{4} + q_1}, \\ \widetilde{\rho}_4 &= \frac{1 - q_1 - \gamma_1\alpha_1}{2} + \sqrt{\frac{[1 - q_1 - \gamma_1\alpha_1]^2}{4} + q_1}. \end{aligned}$$

2. $\|\mathcal{M}_1\|_{\mathcal{D}} \leq \max(\widetilde{q}_A, -\widetilde{\rho}_1, -\widetilde{\rho}_3, \widetilde{\rho}_2, \widetilde{\rho}_4)$.
3. *If* $\alpha_2(2+\gamma_2) < 4$ *and* $4(1-\alpha_1)^2\alpha_2 < \alpha_1(2-\alpha_1\gamma_2)(2-\alpha_2\gamma_2)$, *then* $\|\mathcal{M}_1\|_{\mathcal{D}} < 1$ *for values of* α_0 *with*

$$\max\left(\frac{2(1-\alpha_1)^2}{\alpha_1(2-\alpha_1\gamma_2)}, \frac{2(\alpha_2-1)^2}{\alpha_2(2-\alpha_2\gamma_2)}\right) < \alpha_0 < \frac{2-\alpha_2\gamma_2}{2\alpha_2}.$$

Proof. The proof follows along the lines of the proof of the last theorem. The only difference is the definition of the function $f(x)$ and the quadratic equations.

The upper bound $\nu \leq \max(\widetilde{\rho}_2, \widetilde{\rho}_4)$ for the eigenvalues of $\bar{\mathcal{N}}_1$ with $\widetilde{\rho}_2 = -\widetilde{\mu}_{21}^-$ and $\widetilde{\rho}_4 = -\widetilde{\mu}_{11}^-$ is given by the negative root $\widetilde{\mu}_{11}^-$ of the quadratic equation

$$\mu\,\gamma_1\,\alpha_1 = (\mu + 1)\,(\mu - q_1)$$

and the negative root $\widetilde{\mu}_{21}^-$ of the quadratic equation

$$\mu\,\gamma_1\,\alpha_2 = (\mu + 1)\,(\mu - q_2).$$

The lower bound $\nu \geq \min(-\widetilde{q}_A, \widetilde{\rho}_1, \widetilde{\rho}_3)$ for the eigenvalues of $\bar{\mathcal{N}}_1$ with $\widetilde{\rho}_1 = -\widetilde{\mu}_{21}^+$ and $\widetilde{\rho}_3 = -\widetilde{\mu}_{12}^+$ is given by the positive root $\widetilde{\mu}_{12}^+$ of the quadratic equation

$$\mu\,\gamma_2\,\alpha_1 = (\mu + 1)\,(\mu - q_1),$$

the positive root $\widetilde{\mu}_{22}^+$ of the quadratic equation

$$\mu\,\gamma_2\,\alpha_2 = (\mu + 1)\,(\mu - q_2),$$

and the constant $\widetilde{q}_A$, which ensures $\widetilde{q}_A\, Q \geq \hat{A}$. □

Remark 4.6. In [4] the special case

$$\alpha_1 = 1-\delta, \quad \alpha_2 = 1+\delta, \quad \gamma_1 = 1-\gamma, \quad \gamma_2 = 1 \tag{4.12}$$

was studied. It could be shown in a very short and elegant way that $\|\mathcal{M}_1\|_{\mathcal{D}} < 1$ if

$$\delta < \frac{1-\gamma}{3-\gamma} \tag{4.13}$$

for the scaling factor

$$\alpha_0 = \frac{\delta}{1+\delta}.$$

From the theorem given here a convergence rate estimate of the form $\|\mathcal{M}_1\|_{\mathcal{D}} < 1$ already follows if

$$\delta < \frac{1}{3} \tag{4.14}$$

for scaling factors α_0 with

$$\frac{2\delta^2}{1-\delta^2} < \alpha_0 < \frac{1-\delta}{2(1+\delta)}.$$

Observe that Condition (4.14) is weaker than Condition (4.13).

So, in this paper, convergence could be shown in an $\hat{A}$-independent norm under Condition (4.14), which is weaker than Condition (4.13) used in [4]. Furthermore, observe that Condition (4.14) coincides with the convergence condition 3 in Theorem 4.1 under the assumptions (4.12). This is remarkable, because the norm was allowed to depend on $\hat{A}$ in Theorem 4.1, while now an $\hat{A}$-independent norm was required, which is more restrictive.

5. Convergence results for symmetric preconditioners

The analysis for iteration methods with symmetric preconditioners corresponds to the setting

$$\mathcal{P} = I, \quad \mathcal{Q} = \begin{pmatrix} A-\hat{A} & 0 \\ 0 & \hat{C} - B\hat{A}^{-1}B^T \end{pmatrix}.$$

Again we start by discussing the

5.1. **Special case: $\hat{A} < A$ and $\hat{C} > B\hat{A}^{-1}B^T$.** In this case we use the scaling

$$\mathcal{D} = \begin{pmatrix} A-\hat{A} & 0 \\ 0 & \hat{C} - B\hat{A}^{-1}B^T \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} (A-\hat{A})^{-1} & 0 \\ 0 & \left(\hat{C} - B\hat{A}^{-1}B^T\right)^{-1} \end{pmatrix},$$

leading to the norm $\|.\|_{\mathcal{D}}$, given by the scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle = \langle (A-\hat{A})u, v\rangle + \langle (\hat{C} - B\hat{A}^{-1}B^T)p, q\rangle. \tag{5.1}$$

The generalized eigenvalue problem (3.3) leads to a generalized eigenvalue problem of the form (3.7) with

$$\lambda = -\frac{1}{\nu}, \quad Q = A-\hat{A}, \quad S = \hat{C} - B\hat{A}^{-1}B^T. \tag{5.2}$$

For the preconditioners the following spectral inequalities are assumed. There is a constant $\alpha_2 \in \mathbb{R}$ with $\alpha_2 > 1$, such that

$$\hat{A} < A \leq \alpha_2 \hat{A}, \tag{5.3}$$

and there is a constant $\beta_1 \in \mathbb{R}$ with $\beta_1 < 1$, such that

$$\beta_1 \hat{C} \leq B\hat{A}^{-1}B^T < \hat{C}. \tag{5.4}$$

Then a simple consequence of Lemma 3.1 is

Theorem 5.1. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (5.3) *and* (5.4). *Then we have:*

1. *The iteration matrix $\mathcal{M}_2$ of the iteration method with symmetric preconditioner is symmetric with respect to the scalar product* (5.1).
2. $\sigma(\mathcal{M}_2) \subset [1-\alpha_2, \rho_2] \subset (-\infty, 1)$ *with*
$$\rho_2 = \frac{(2-\alpha_2)(1-\beta_1)}{2} + \sqrt{\frac{(2-\alpha_2)^2(1-\beta_1)^2}{4} + (\alpha_2 - 1)(1-\beta_1)}.$$
3. *If $\alpha_2 < 2$, then $\rho(\mathcal{M}_2) = \|\mathcal{M}_1\|_{\mathcal{D}} < 1$.*
4. *$\hat{\mathcal{K}}_2^{-1}\mathcal{K}$ is symmetric and positive definite with respect to the scalar product* (5.1) *and $\sigma(\hat{\mathcal{K}}_2^{-1}\mathcal{K}) \subset [1-\rho_2, \alpha_2] \subset (0, \infty)$.*

Proof. With the notation of Section 3 we have: $\bar{\mathcal{M}}_2 = \bar{\mathcal{N}}_2$, where $\bar{\mathcal{N}}_2$ is symmetric. This shows the symmetry of $\mathcal{M}_1$ in the corresponding scalar product (4.1).

We consider the eigenvalue problem (3.7) with the settings (5.2).

The matrix function $\varphi(\mu)$ of Lemma 3.1 has the form

$$\varphi(\mu) = \mu\, B \left[\mu\, \hat{A} - \hat{A}(A-\hat{A})^{-1}\hat{A}\right]^{-1} B^T + \mu\, B\hat{A}^{-1}B^T - (\mu+1)\,\hat{C}.$$

Assume that $\mu \leq -1$. Then it follows from (5.4) that $\varphi(\mu) \leq 0$ if

$$\mu\, \left[\mu\, \hat{A} - \hat{A}(A-\hat{A})^{-1}\hat{A}\right]^{-1} + \mu\, \hat{A}^{-1} \leq \frac{\mu+1}{\beta_1}\, \hat{A}^{-1},$$

or, equivalently,

$$\mu\, \beta_1\, \hat{A} \geq (\mu + 1 - \mu\, \beta_1)(\mu \hat{A} - \hat{A}(A-\hat{A})^{-1}\hat{A}).$$

This means for the transformed matrix $\bar{A} = \hat{A}^{-1/2} A \hat{A}^{-1/2}$:

$$\mu\, \beta_1\, I \geq (\mu + 1 - \mu\, \beta_1)(\mu\, I - (\bar{A} - I)^{-1}).$$

For $\mu \geq -1/(1-\beta_1)$, this inequality is satisfied if and only if

$$\mu\, \beta_1 \geq (\mu + 1 - \mu\, \beta_1) \left(\mu - \frac{1}{\alpha_2 - 1}\right).$$

The negative μ_{21}^- root of the quadratic equation

$$\mu\, \beta_1 = [1 + \mu\, (1-\beta_1)] \left(\mu - \frac{1}{\alpha_2 - 1}\right)$$

is given by

$$\frac{1}{\mu_{21}^-} = -\frac{(2-\alpha_2)(1-\beta_1)}{2} - \sqrt{\frac{(2-\alpha_2)^2(1-\beta_1)^2}{4} + (\alpha_2 - 1)(1-\beta_1)}$$

and lies in the interval $(-1/(1-\beta_1), -1)$. Therefore, $\varphi(\mu_{21}^-) \leq 0$. Then, by Lemma 3.1, it follows that $\lambda \leq \mu_{21}^- < -1$ for negative eigenvalues λ and $\nu = -1/\lambda \leq \rho_2$ with $\rho_2 = -1/\mu_{21}^-$ for the eigenvalues ν of $\bar{\mathcal{N}}_2$.

For estimating positive eigenvalues λ from below by some value ρ we need the condition

$$\rho(A - \hat{A}) \leq \hat{A}$$

or, equivalently,

$$\rho\,(\bar{A} - I) \leq I,$$

which is satisfied if and only if

$$\rho \leq \frac{1}{\alpha_2 - 1}.$$

So, $\lambda \geq 1/(\alpha_2 - 1)$ for positive eigenvalues, and, consequently, $\nu = -1/\lambda \geq 1 - \alpha_2$ for the eigenvalues ν of $\bar{\mathcal{N}}_2$.

Statement 3 easily follows from 2, and 4 is an immediate consequence of 1 and 2. □

The second special case can be analyzed completely analogously:

5.2. **Special case: $\hat{A} > A$ and $\hat{C} < B\hat{A}^{-1}B^T$.** In this case we use the norm $\|.\|_{\mathcal{D}}$, given by the scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle = \langle (\hat{A} - A)u, v\rangle + \langle (B\hat{A}^{-1}B^T - \hat{C})p, q\rangle. \tag{5.5}$$

For the preconditioners the following spectral inequalities are assumed. There is a constant $\alpha_1 \in \mathbb{R}$ with $\alpha_1 < 1$, such that

$$\alpha_1\,\hat{A} \leq A < \hat{A}, \tag{5.6}$$

and there is a constant $\beta_2 \in \mathbb{R}$ with $\beta_2 > 1$, such that

$$\hat{C} < B\hat{A}^{-1}B^T \leq \beta_2\,\hat{C}. \tag{5.7}$$

Then a simple consequence of Lemma 3.1 is

Theorem 5.2. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (5.6) *and* (5.7). *Then we have:*

1. *The iteration matrix $\mathcal{M}_2$ of the iteration method with symmetric preconditioner is symmetric with respect to the scalar product* (5.5).
2. $\sigma(\mathcal{M}_2) \subset [-\rho_1, 1 - \alpha_1] \subset (-\infty, 1)$ *with*
$$\rho_1 = \frac{(2-\alpha_1)(\beta_2 - 1)}{2} + \sqrt{\frac{(2-\alpha_1)^2(\beta_2-1)^2}{4} + (1-\alpha_1)(\beta_2 - 1)}.$$
3. *If $\beta_2 < 1 + 1/(3 - 2\alpha_1)$, then $\rho(\mathcal{M}_2) = \|\mathcal{M}_2\|_{\mathcal{D}} < 1$.*
4. *$\hat{\mathcal{K}}_2^{-1}\mathcal{K}$ is symmetric and positive definite with respect to the scalar product* (5.5) *and $\sigma(\hat{\mathcal{K}}_2^{-1}\mathcal{K}) \subset [\alpha_1, 1 + \rho_1] \subset (0, \infty)$.*

Proof. The proof, which is completely analogous to the proof of the previous theorem, is omitted. □

5.3. **The general case.** For the preconditioner $\hat{A}$ the following spectral inequalities are assumed: There are constants $\alpha_1, \alpha_2 \in \mathbb{R}$ with $0 < \alpha_1 \leq 1 \leq \alpha_2$, such that

$$\alpha_1 \hat{A} \leq A \leq \alpha_2 \hat{A}. \tag{5.8}$$

Additionally, it is assumed that $\alpha_1 < \alpha_2$. As before, we introduce

$$q_\alpha = \max(1 - \alpha_1, \alpha_2 - 1).$$

For the preconditioner $\hat{C}$ we assume that there are constants $\beta_1, \beta_2 \in \mathbb{R}$ with $0 < \beta_1 \leq 1 \leq \beta_2$, such that

$$\beta_1 \hat{C} \leq B\hat{A}^{-1}B^T \leq \beta_2 \hat{C}. \tag{5.9}$$

Additionally, we assume that at least one of the inequalities $\beta_1 \leq 1$ or $\beta_1 \hat{C} \leq B\hat{A}^{-1}B^T$ and at least one of the inequalities $1 \leq \beta_2$ or $B\hat{A}^{-1}B^T \leq \beta_2 \hat{C}$ hold strictly.

According to the discussion in Section 3 we are looking for a scaling matrix

$$\mathcal{E} = \begin{pmatrix} E_u & 0 \\ 0 & E_p \end{pmatrix},$$

such that

$$|\mathcal{E}^{1/2} \mathcal{Q}\mathcal{P}\mathcal{E}^{1/2}| \leq I,$$

or, in details

$$|E_u^{1/2}(A - \hat{A})E_u^{1/2}| \leq I \quad \text{and} \quad |E_p^{1/2}(\hat{C} - B\hat{A}^{-1}B^T)E_p^{1/2}| \leq I.$$

Roughly speaking, the larger $\mathcal{E}$ is with respect to the ordering $\leq$ of matrices, introduced in Section 3, the smaller are the eigenvalues in (3.3). In this sense E_u and E_p should be as large as possible.

For E_u we set

$$E_u = \frac{1}{q_\alpha} \hat{A}^{-1}. \tag{5.10}$$

It is easy to see that E_u satisfies the required estimate.

For E_p we set

$$E_p = \hat{C}^{-1/2} g(\bar{C}) \hat{C}^{-1/2}, \tag{5.11}$$

where $\bar{C} = \hat{C}^{-1/2} B\hat{A}^{-1}\hat{C}^{-1/2}$ and $g\colon (0, \infty) \to (0, \infty]$ is an arbitrary function.

Then the condition on E_p reads: All eigenvalues λ of the generalized eigenvalue problem

$$(\hat{C} - B\hat{A}^{-1}B^T)p = \lambda E_p^{-1} p$$

or, equivalently,

$$(I - \bar{C})p = \lambda\, g(\bar{C})^{-1}$$

have to satisfy the condition: $|\lambda| \leq 1$.

From (5.9) we know $\sigma(\bar{C}) \subset [\beta_1, \beta_2]$. The eigenvalues λ result from the eigenvalues $\bar{c}$ of $\bar{C}$ by the formula

$$\lambda = (1 - \bar{c})g(\bar{c}).$$

The condition $|\lambda| \leq 1$ is certainly satisfied if

$$|x - 1|g(x) \leq 1 \quad \text{for all } x \in [\beta_1, \beta_2].$$

This condition holds for the affine function

$$g(x)^{-1} = (1-\beta_1)\frac{\beta_2 - x}{\beta_2 - \beta_1} + (\beta_2 - 1)\frac{x - \beta_1}{\beta_2 - \beta_1} = \beta_0\, x + \hat{\beta} \tag{5.12}$$

with

$$\beta_0 = \frac{\beta_1 + \beta_2 - 2}{\beta_2 - \beta_1}, \quad \hat{\beta} = \frac{\beta_1 + \beta_2 - 2\beta_1\beta_2}{\beta_2 - \beta_1}.$$

Observe that

$$|x-1|g(x) = 1 \quad \text{for } x = \beta_i,$$

for $i = 1, 2$, except for the case $\beta_i = 1$. So, g is as large as possible at least at the boundary points of the interval which contains the spectrum of $\bar{C}$.

The resulting matrix $\mathcal{D}$ is given by

$$\mathcal{D} = \begin{pmatrix} q_\alpha\, \hat{A} & 0 \\ 0 & \hat{\beta}\, \hat{C} + \beta_0\, B\hat{A}^{-1}B^T \end{pmatrix}$$

with associated scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{D}} = q_\alpha \left\langle \hat{A}u, v \right\rangle + \left\langle \left(\hat{\beta}\, \hat{C} + \beta_0\, B\hat{A}^{-1}B^T \right) p, q \right\rangle.$$

With these settings the generalized eigenvalue problem (3.3) leads to a generalized eigenvalue problem of the form (3.7) with

$$\lambda = -\frac{1}{\nu}, \quad Q = q_\alpha\, \hat{A}, \quad S = \hat{\beta}\, \hat{C} + \beta_0\, B\hat{A}^{-1}B^T. \tag{5.13}$$

Then a simple consequence of Lemma 3.1 is

Theorem 5.3. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying (5.8) and (5.9).*

Then we have with the notations of Sections 2 and 3:

1. $\sigma(\bar{\mathcal{N}}_2) \subset [-q_\alpha, \max(\rho_1, \rho_2)]$ *with*

$$\begin{aligned} q_\alpha &= \max(1-\alpha_1, \alpha_2 - 1), \\ \rho_1 &= \frac{(1-q_\alpha)(1-\beta_1)}{2} + \sqrt{\frac{(1-q_\alpha)^2(1-\beta_1)^2}{4} + q_\alpha(1-\beta_1)}, \\ \rho_2 &= \frac{(1+q_\alpha)(\beta_2-1)}{2} + \sqrt{\frac{(1+q_\alpha)^2(\beta_2-1)^2}{4} + q_\alpha(\beta_2-1)}. \end{aligned}$$

2. $\|\mathcal{M}_2\|_{\mathcal{D}} \le \max(q_\alpha, \rho_1, \rho_2)$.
3. *If $\alpha_2 < 2$ and $\beta_2 < 1 + 1/(1+2q_\alpha)$, then $\|\mathcal{M}_2\|_{\mathcal{D}} < 1$.*

Proof. Consider the eigenvalue problem (3.7) with the settings (5.13).

The matrix function $\varphi(\mu)$ of Lemma 3.1 has the form

$$\varphi(\mu) = \left(\frac{\mu}{\mu - 1/q_\alpha} - \mu\, \beta_0 \right) B\hat{A}^{-1}B^T - (\mu\, \hat{\beta} + 1)\, \hat{C}.$$

It is clear that $\varphi(\mu) \le 0$ if and only if

$$\left(\frac{\mu}{\mu - 1/q_\alpha} - \mu\, \beta_0 \right) \bar{C} - (\mu\, \hat{\beta} + 1)\, I \le 0.$$

From (5.9) we know $\sigma(\bar{C}) \subset [\beta_1, \beta_2]$. So it suffices to have

$$\left(\frac{\mu}{\mu - 1/q_\alpha} - \mu\,\beta_0\right)\,\beta_1 - (\mu\,\hat{\beta} + 1) \le 0$$

and

$$\left(\frac{\mu}{\mu - 1/q_\alpha} - \mu\,\beta_0\right)\,\beta_2 - (\mu\,\hat{\beta} + 1) \le 0.$$

For $\mu \le 0$ this leads to

$$\mu\,\beta_1 \ge [1 + \mu\,(1 - \beta_1)]\left(\mu - \frac{1}{q_\alpha}\right) \quad \text{and} \quad \mu\,\beta_2 \ge [1 + \mu\,(\beta_2 - 1)]\left(\mu - \frac{1}{q_\alpha}\right).$$

This means $\mu \ge \max(\mu_1^-, \mu_2^-)$, where μ_1^- is the negative root of the quadratic equation

$$\mu\,\beta_1 = [1 + \mu\,(1 - \beta_1)]\left(\mu - \frac{1}{q_\alpha}\right),$$

given by

$$\frac{1}{\mu_1^-} = -\frac{(1 - q_\alpha)(1 - \beta_1)}{2} - \sqrt{\frac{(1 - q_\alpha)^2(1 - \beta_1)^2}{4} + q_\alpha(1 - \beta_1)},$$

and μ_2^- is the negative root of the quadratic equation

$$\mu\,\beta_2 = [1 + \mu\,(\beta_2 - 1)]\left(\mu - \frac{1}{q_\alpha}\right),$$

given by

$$\frac{1}{\mu_1^-} = -\frac{(1 + q_\alpha)(\beta_2 - 1)}{2} - \sqrt{\frac{(1 + q_\alpha)^2(\beta_2 - 1)^2}{4} + q_\alpha(\beta_2 - 1)}.$$

So, from Lemma 3.1 we obtain the upper bound $\lambda \le \max(\mu_1^-, \mu_2^-)$ for negative eigenvalues λ.

Then, for eigenvalues ν of $\bar{\mathcal{N}}_2$, the upper bound $\nu \le \max(\rho_1, \rho_2)$ with $\rho_1 = -1/\mu_1^-$ and $\rho_2 = -1/\mu_2^-$ follows.

From $Q = q_\alpha\,\hat{A}$ we obtain the lower bound $\lambda \ge 1/q_\alpha$ for positive eigenvalues λ, which gives the lower bound $\nu \ge -q_\alpha$ for the eigenvalues of $\bar{\mathcal{N}}_2$.

The statements 2 and 3 follow directly from 1. □

Remark 5.4. In [2] the special case

$$\alpha_1 = 1 - \alpha, \quad \alpha_2 = 1 + \alpha, \quad \beta_1 = 1 - \beta, \quad \beta_2 = 1 + \beta$$

was studied. In this case Theorem 5.3 gives the bound

$$\|\mathcal{M}_2\|_{\mathcal{D}} \le \max\left(\alpha, \frac{(1 + \alpha)\beta}{2} + \sqrt{\frac{(1 + \alpha)^2\beta^2}{4} + \alpha\beta}\right),$$

from which convergence follows for $\alpha < 1$ and $\beta < 1/(1 + 2\alpha)$.

Compare this to the estimate (in a different norm)

$$\|\mathcal{M}_2\| \le \max\left(\alpha, \frac{2\beta}{1 - \beta}\right),$$

given in [2], from which convergence only follows under the more restrictive conditions $\alpha < 1$ and $\beta < 1/3$.

Our estimate can also be written in the form

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \begin{cases} \alpha & \text{if } \alpha \geq \dfrac{2\beta}{1-\beta}, \\ \dfrac{(1+\alpha)\beta}{2} + \sqrt{\dfrac{(1+\alpha)^2\beta^2}{4} + \alpha\beta} & \text{if } \alpha < \dfrac{2\beta}{1-\beta}. \end{cases}$$

It is easy to see that

$$\frac{(1+\alpha)\beta}{2} + \sqrt{\frac{(1+\alpha)^2\beta^2}{4} + \alpha\beta} < \frac{2\beta}{1-\beta} \quad \text{if} \quad \alpha < \frac{2\beta}{1-\beta},$$

from which we obtain the weaker estimate

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max\left(\alpha, \frac{2\beta}{1-\beta}\right),$$

comparable with the estimate in [2].

A second special case, namely

$$\alpha_1 = 1-\alpha, \quad \alpha_2 = 1+\alpha, \quad \beta_1 = 1-\beta, \quad \beta_2 = 1,$$

was discussed in [2]. Here Theorem 5.3 gives the bound

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max\left(\alpha, \frac{(1-\alpha)\beta}{2} + \sqrt{\frac{(1-\alpha)^2\beta^2}{4} + \alpha\beta}\right).$$

This estimate can also be written in the form

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \begin{cases} \alpha & \text{if } \alpha \geq \dfrac{2\beta}{1+\beta}, \\ \dfrac{(1-\alpha)\beta}{2} + \sqrt{\dfrac{(1-\alpha)^2\beta^2}{4} + \alpha\beta} & \text{if } \alpha < \dfrac{2\beta}{1+\beta}. \end{cases}$$

Since

$$\frac{(1-\alpha)\beta}{2} + \sqrt{\frac{(1-\alpha)^2\beta^2}{4} + \alpha\beta} < \frac{2\beta}{1+\beta} \quad \text{if} \quad \alpha < \frac{2\beta}{1+\beta},$$

we obtain the weaker estimate

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max\left(\alpha, \frac{2\beta}{1+\beta}\right),$$

comparable with the estimate in [2].

An example in [2] shows the sharpness of the bounds in Theorem 5.3.

5.4. **The general case with $\hat{C}$-independent norm.** The norms introduced so far for analyzing the convergence depend on the preconditioner $\hat{C}$. If the preconditioner for $B\hat{A}^{-1}B^T$ is allowed to change during the iteration, it is advisable to use a norm which is independent of $\hat{C}$. (This includes the case of using an inner iteration for solving an equation of the form $B\hat{A}^{-1}B^Tp = c$, such as a conjugate gradient method. We again refer to Lemma 9 in [2].)

Therefore we are now looking for estimates in a norm $\|.\|_{\mathcal{D}}$, given by the scalar product

$$\left\langle \begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{D}} = q_\alpha \langle \hat{A}u, v\rangle + \beta_0 \langle B\hat{A}^1B^Tp, q\rangle$$

with some scalar factor $\beta_0 > 0$. So, in the terminology of Section 3, we prescribe the scaling matrix $\mathcal{D}$ by

$$\mathcal{D} = \begin{pmatrix} q_\alpha \hat{A} & 0 \\ 0 & \beta_0 B\hat{A}^{-1}B^T \end{pmatrix}$$

up to the scalar factor β_0.

Here we choose

$$\mathcal{E} = \begin{pmatrix} q_\alpha \hat{A} & 0 \\ 0 & \hat{C}^{-1/2} g(\bar{C}) \hat{C}^{-1/2} \end{pmatrix}$$

with

$$g(x)^{-1} = s_1 \frac{\beta_2 - x}{\beta_2 - \beta_1} + s_2 \frac{x - \beta_1}{\beta_2 - \beta_1} \tag{5.14}$$

and

$$\begin{aligned} s_1 &= \max\left(\beta_0\,\beta_1, \frac{1}{\beta_0}\left(\beta_1 + \frac{1}{\beta_1} - 2\right)\right), \\ s_2 &= \max\left(\beta_0\,\beta_2, \frac{1}{\beta_0}\left(\beta_2 + \frac{1}{\beta_2} - 2\right)\right), \end{aligned}$$

and obtain

Theorem 5.5. *Let A, $\hat{A}$ be symmetric, positive definite $n \times n$-matrices, B a $m \times n$ matrix, $\hat{C}$ a symmetric, positive definite $m \times m$ matrix, satisfying* (5.8) *and* (5.9).

Then we have with the notations of Sections 2 and 3:

1. $\sigma(\bar{\mathcal{N}}_2) \subset [-q_\alpha, \max(\widetilde{\rho}_1, \widetilde{\rho}_2)]$ *with*

$$\begin{aligned} q_\alpha &= \max(1 - \alpha_1, \alpha_2 - 1), \\ \widetilde{\rho}_1 &= \frac{s_1 - q_\alpha(1 - \beta_1)}{2} + \sqrt{\frac{[s_1 - q_\alpha(1 - \beta_1)]^2}{4} + q_\alpha s_1}, \\ \widetilde{\rho}_2 &= \frac{s_2 + q_\alpha(\beta_2 - 1)}{2} + \sqrt{\frac{[s_2 + q_\alpha(\beta_2 - 1)]^2}{4} + q_\alpha s_2}. \end{aligned}$$

2. $\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max(q_\alpha, \widetilde{\rho}_1, \widetilde{\rho}_2)$.
3. *If $\alpha_2 < 2$ and $\beta_2 < 1 + 1/(1 + 2q_\alpha)$ and*

$$(1 - \beta_1)^2 \beta_2 (1 + q_\alpha)^2 < \beta_1 [1 + q_\alpha(1 - \beta_1)][1 - q_\alpha(\beta_2 - 1)],$$

then $\|\mathcal{M}_2\|_{\mathcal{D}} < 1$ for values of β_0 with

$$\max\left(\frac{(1 - \beta_1)^2(1 + q_\alpha)}{\beta_1[1 + q_\alpha(1 - \beta_1)]}, \frac{(\beta_2 - 1)^2(1 + q_\alpha)}{\beta_2[1 - q_\alpha(\beta_2 - 1)]}\right) < \beta_0 < \frac{1 - q_\alpha(\beta_2 - 1)}{\beta_2(1 + q_\alpha)}.$$

Proof. The proof follows along the lines of the proof of Theorem 4.5 and, therefore, is omitted. □

Remark 5.6. In [2] the special case

$$\alpha_1 = 1 - \alpha, \quad \alpha_2 = 1 + \alpha, \quad \beta_1 = 1 - \beta, \quad \beta_2 = 1 + \beta$$

was studied. In this case Theorem 5.5 gives the bound $\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max(q_\alpha, \widetilde{\rho}_1, \widetilde{\rho}_2)$ with

$$\begin{aligned} q_\alpha &= \alpha, \\ \widetilde{\rho}_1 &= \frac{s_1 - \alpha\beta}{2} + \sqrt{\frac{(s_1 - \alpha\beta)^2}{4} + \alpha s_1}, \\ \widetilde{\rho}_2 &= \frac{s_2 + \alpha\beta}{2} + \sqrt{\frac{(s_2 + \alpha\beta)^2}{4} + \alpha s_2}. \end{aligned}$$

It turns out that there is a parameter β_0 with $\max(q_\alpha, \widetilde{\rho}_1, \widetilde{\rho}_2) = \alpha$ if and only if $\alpha \geq 2\beta/(1-\beta)$.

Consider the more interesting case $\alpha < 2\beta/(1-\beta)$. For parameters $\beta_0 \in [\beta/(1+\beta), \beta/(1-\beta)]$ we have:

$$s_1 = \frac{1}{\beta_0}\frac{\beta^2}{1-\beta} \quad \text{and} \quad s_2 = \beta_0(1+\beta).$$

It is easy to see that

$$\widetilde{\rho}_1 < \widetilde{\rho}_2 \quad \text{for} \quad \beta_0 = \frac{2\beta}{1+\beta}$$

and

$$\widetilde{\rho}_2 < \widetilde{\rho}_1 < \frac{2\beta}{1-\beta} \quad \text{for} \quad \beta_0 = \frac{2\beta}{1-\beta}.$$

From this discussion we can conclude that

$$\|\mathcal{M}_2\|_{\mathcal{D}} \leq \max\left(\alpha, \frac{2\beta}{1-\beta}\right),$$

if we choose the parameter $\beta_0 = 2\beta/(1-\beta)$. Then convergence follows if $\alpha < 1$ and $\beta < 1/3$. This corresponds (up to the norm) to the estimates in [2].

A sharper estimate is obtained if $\beta_0 \in [\beta/(1+\beta), \beta/(1-\beta)]$ is chosen such that $\widetilde{\rho}_1 = \widetilde{\rho}_2$. The existence of such a parameter follows from the discussion above (the expression $\widetilde{\rho}_2 - \widetilde{\rho}_1$ changes the sign from $\beta_0 = 2\beta/(1+\beta)$ to $\beta_0 = 2\beta/(1-\beta)$. Then, according to Theorem 5.5, the convergence conditions read

$$\alpha < 1 \quad \text{and} \quad \beta^2(1+\beta)(1+\alpha)^2 < (1-\beta)(1-\alpha^2\beta^2).$$

Sufficient for these conditions are $\alpha < 1$ and $\beta < 1/3$. For small α the condition on β becomes less restrictive. So, e.g., for the limiting case $\alpha = 0$, β must be smaller than the positive root of the cubic equation $x^3 + x^2 + x = 1$, or, approximately, $\beta < 0.54$.

6. Numerical experiments

We consider the P_2-P_0 finite element approximation of the Stokes problem

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= f \quad \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in} \quad \Omega, \\ \mathbf{u} &= \mathbf{g} \quad \text{on} \quad \partial\Omega, \\ \int_\Omega p \, dx &= 0 \end{aligned}$$

TABLE 1. Spectral constants

level	α_1^0	α_2^0	β_1^0	β_2^0	γ_1^0	γ_2^0
4	0.578	1.000	0.218	0.975	0.258	0.999
5	0.568	1.000	0.199	0.978	0.238	1.001
6	0.563	1.000	0.187	0.979	0.225	1.002
7	0.561	1.000	0.179	0.979	0.216	1.002
8	0.561	1.000	0.175	0.979	0.210	1.002

in a bounded domain $\Omega \subset \mathbb{R}^2$. Here, $\mathbf{u}$ denotes the velocity and p the pressure of a fluid; $\mathbf{g}$ has to satisfy the compatibility condition

$$\int_{\partial\Omega} \mathbf{g}\, ds = 0.$$

In particular, we consider the classical driven cavity problem on the unit square, where a unit tangential velocity is prescribed at the top of the square (and 0 elsewhere) and $f = 0$.

The square is triangulated using a uniform mesh. The level 1 mesh consists of two triangles by connecting the lower left with the upper right corner of the square. The finest mesh considered is the level 8 mesh, obtained by successive uniform refinement, which consists of 32768 triangles. This corresponds to the following numbers of unknowns: $n = 130050$ and $m = 32767$ in the notation of Section 1.

As an a priori preconditioner A_0 for A we chose the classical V-cycle multigrid method with one forward Gauss-Seidel pre-smoothing step and one backward Gauss-Seidel post-smoothing step. As an a priori preconditioner C_0 for the Schur complement $BA^{-1}B^T$ we chose the identity premultiplied by the element area.

By the Lanczos method the following spectral inequalities were determined numerically:

$$\alpha_1^0\, A_0 \le A \le \alpha_2^0\, A_0$$

and

$$\beta_1^0\, C_0 \le BA_0^{-1}B^T \le \beta_2^0\, C_0 \quad \text{and} \quad \gamma_1^0\, C_0 \le BA^{-1}B^T \le \gamma_2^0\, C_0$$

with α_1^0, α_2^0, β_1^0, β_2^0, γ_1^0 and γ_2^0 given in Table 1.

From these estimates the preconditioner $\hat{A}$ is defined by

$$\hat{A} = 0.9\, \alpha_1^0\, A_0.$$

This guarantees that

$$\hat{A} < A \le \alpha_2\, \hat{A}$$

with $\alpha_2 = \alpha_2^0/(0.9\, \alpha_1^0)$.

The cg-accelerated inexact Uzawa algorithm was performed with the preconditioners $\hat{A}$ and C_0.

For the cg-accelerated iteration method with symmetric preconditioner an additional scaling is required for the Schur-complement preconditioner: $\hat{C}$ is defined by

$$\hat{C} = 1.1\, \frac{\beta_2^0}{0.9\, \alpha_1^0}\, C_0.$$

TABLE 2. cg-accelerated iteration methods

	Unknowns		Inexact Uzawa			Symm. precond.		
level	n	m	N	ρ	ρ_{th}	N	ρ	ρ_{th}
4	450	127	32	0.56	0.59	33	0.56	0.63
5	1922	511	32	0.56	0.61	34	0.57	0.65
6	7938	2047	33	0.57	0.62	34	0.57	0.66
7	32258	8191	34	0.58	0.63	35	0.59	0.67
8	130050	32767	34	0.58	0.63	35	0.59	0.67

This guarantees that

$$\beta_1 \hat{C} \leq B\hat{A}^{-1}B^T < \hat{C}$$

with $\beta_1 = \beta_1^0/(1.1\,\beta_2^0)$.

The cg-method applied to the matrix $\hat{\mathcal{K}}^{-1}\mathcal{K}$ with the scalar product $\langle .,. \rangle_{\mathcal{D}}$ was used, where $\mathcal{K}$, $\hat{\mathcal{K}}$ and $\mathcal{D}$ were chosen according to Subsections 4.1 and 5.1, with initial guesses $\mathbf{u} = 0$ and $p = 0$ and stopping rule

$$e_k = \|\hat{\mathcal{K}}^{-1}d_k\|_{\mathcal{D}} \leq \varepsilon\,\|\hat{\mathcal{K}}^{-1}d_0\|_{\mathcal{D}} = \varepsilon\,e_0$$

with $\varepsilon = 10^{-8}$. Here d_k denotes the defect of the k-th iterate with respect to original equation (1.1).

This corresponds to the classical cg-method applied to the matrix $\mathcal{A} = \mathcal{D}\hat{\mathcal{K}}^{-1}\mathcal{K}$ with the Euclidean scalar product and the preconditioner $\hat{\mathcal{A}} = \mathcal{D}$. The norm in the stopping rule is the $\mathcal{A}\hat{\mathcal{A}}^{-1}\mathcal{A}$-norm.

It is easy to see that the evaluation of $\hat{A}$ and $\hat{C}$ is not necessary for performing the cg-method. Only the action of $\hat{A}^{-1}$ and $\hat{C}^{-1}$ to a known vector is required.

Table 2 shows the results of the numerical experiments. For each mesh from level 4 to level 8 it contains the numbers n and m of unknowns, the iteration numbers N and the averaged convergence factor $\rho = (e_k/e_0)^{1/k}$ in comparison with (an upper bound of) the theoretical convergence factor $\rho_{th} = [2c^k/(1+c^{2k})]^{1/k}$ in the $\mathcal{A}$-norm, where $c = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ (see, e.g., Hackbusch [8]). Here κ is the upper bound of the condition number of $\hat{\mathcal{K}}^{-1}\mathcal{K}$ in the $\mathcal{D}$-norm, given in Theorems 4.1 and 5.1, respectively.

The numerical tests show a good agreement between the actual and the theoretical convergence rates, a consequence of the sharpness of the estimates. The convergence rates for the iteration method with symmetric preconditioner seem to indicate, as expected, the superiority of the cg-accelerated version given here to the tested method in [2], where a cg-acceleration was used only for the inner iteration of the p-equation. The iteration numbers for the inexact Uzawa algorithm and the iteration method with symmetric preconditioner are approximately the same. However, the computational costs for the iteration method with symmetric preconditioner is about twice as high as for the inexact Uzawa method.

Of course, all these comparisons must be considered with some caution. We are comparing with respect to different norms. Nevertheless, they seem to confirm the theoretical results.

Concluding remarks. It has been shown in this paper that several interesting classes of iteration methods for saddle point problems can be analyzed by the following general strategy: Transform the iteration matrix to a symmetric matrix,

perform an appropriate scaling and estimate the eigenvalues by Lemma 3.1. There is no problem in using this technique under assumptions different from the assumptions considered here. For example, one could analyze inexact Uzawa algorithms on the basis of spectral inequalities of the form (5.4) involving the inexact Schur complement or iteration methods with symmetric preconditioners on the basis of spectral inequalities of the form (4.4) involving the exact Schur complement. It is just a matter of patience and time and it requires only the solution of quadratic equations.

It might be also interesting that new cases have been identified, for which it can be shown that the spectrum of the iteration matrix is real. This widens the possibilities of accelerated methods.

Acknowledgments. The author is very grateful to Joachim Schöberl, Linz, for providing him with the numerical results. The numerical tests were performed with FEPP, a C++-coded finite element package by Joachim Schöberl.

References

1. K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in nonlinear programming*, Stanford University Press, Stanford, CA, 1958. MR **21:**7115
2. R. E. Bank, B. D. Welfert, and H. Yserentant, *A class of iterative methods for solving saddle point problems*, Numer. Math. **56** (1990), 645 – 666. MR **91b:**65035
3. J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp. **50** (1988), 1 – 17. MR **89m:**65097a
4. 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), 1072 – 1092. MR **98c:**65182
5. F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer-Verlag, 1991. MR **92d:**65187
6. H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal. **31** (1994), 1645 – 1661. MR **95f:**65065
7. M. Fortin and R. Glowinski, *Augmented Lagrangian methods: Applications to the numerical solution of boundary value problems*, North–Holland, Amsterdam, 1983. MR **85a:**49004
8. W. Hackbusch, *Iterative solutions of large sparse systems of equations*, Springer Verlag, New York, 1994. MR **94k:**65002
9. J. Iliash, T. Rossi, and J. Toivanen, *Two iterative methods for solving the Stokes problem*, Tech. Report 2, University of Jyväskylä, Department of Mathematics, Laboratory of Scientific Computing, 1993.
10. W. Queck, *The convergence factor of preconditioned algorithms of the Arrow-Hurwicz type*, SIAM J. Numer. Anal. **26** (1989), 1016 – 1030. MR **90m:**65071
11. D. Silvester and A. Wathen, *Fast iterative solutions of stabilized Stokes systems. Part II: Using general block preconditioners*, SIAM J. Numer. Anal. **31** (1994), 1352 – 1367. MR **95g:**65132
12. R. Verfürth, *A combined conjugate gradient–multigrid algorithm for the numerical solution fo the Stokes problem*, IMA J. Numer. Anal. **4** (1984), 441 – 455. MR **86f:**65200

Institute of Analysis and Computational Mathematics, Johannes Kepler University, A-4040 Linz, Austria

E-mail address: zulehner@numa.uni-linz.ac.at