Uzawa-type Methods for Block-Structured Indefinite Linear Systems

Walter Zulehner[*,†]
Institute of Computational Mathematics
Johannes Kepler University, A-4040 Linz, Austria

Abstract

In this paper solution techniques for a class of block-structured symmetric and indefinite systems of linear equations are considered. In particular, we discuss preconditioned conjugate gradient methods with inexact Uzawa preconditioners built from given preconditioners of a sequence of Schur complements. Depending on the quality of the Schur complement preconditioners sharp estimates are derived for the spectrum of the preconditioned system matrix, which determine the convergence properties of the conjugate gradient method.

AMS subject classification. 65N22, 65F10, 65N30

Key words. Indefinite systems of linear equations, saddle point problems, Uzawa-type methods, conjugate gradient method

1 Introduction

In this paper we will discuss iterative methods for solving block-tridiagonal symmetric and indefinite linear systems

$$\mathcal{K}u = f \tag{1}$$

with a block-structured vector of unknowns $u = (u_1, u_2, \ldots, u_n)^T \in \mathbb{R}^N$, $u_i \in \mathbb{R}^{N_i}$ for $i = 1, 2, \ldots, n$, $N = N_1 + N_2 \cdots + N_n$, a given block-structured right hand side $f = (f_1, f_2, \ldots, f_n)^T \in \mathbb{R}^N$, and a block matrix

$$\mathcal{K} = \begin{pmatrix} A_1 & B_1^T & & \\ B_1 & -A_2 & \ddots & \\ & \ddots & \ddots & B_{n-1}^T \\ & & B_{n-1} & (-1)^{n-1} A_n \end{pmatrix}, \tag{2}$$

[*]Partially supported by the Austrian Science Fund (FWF) under grant SFB F013/F1309.
[†]E-mail: zulehner@numa.uni-linz.ac.at

where the number of unknowns N is large, while the number of blocks $n > 1$ is typically a small positive integer like 2, 3 or 4. Throughout the paper we will assume that

$$A_1 = A_1^T > 0, \quad A_i = A_i^T \geq 0 \quad \text{for all } i = 2, \ldots, n$$

and that

$$S_i > 0 \qquad \text{for all } i = 1, 2, \ldots, n,$$

where $(S_i)_{i=1,2,\ldots,n}$ denotes the sequence of Schur complements, recursively given by

$$S_{i+1} = A_{i+1} + B_i S_i^{-1} B_i^T \quad \text{for } i = 1, \ldots, n-1$$

with initial setting $S_1 = A_1$. Strictly speaking, S_i is usually called the Schur complement only for odd indices i, otherwise it is the negative Schur complement. Here and in the sequel we use the notations $M > N$ and $M \geq N$, respectively, for symmetric matrices M, N iff $M - N$ is positive definite and $M - N$ is positive semi-definite, respectively, and B^T to denote the transpose of the matrix B.

The most frequently studied case of a block-structured indefinite system is the case $n = 2$: Finite element methods for mixed variational problems often lead to symmetric and indefinite linear systems of a natural 2-by-2 block-structure

$$\begin{pmatrix} A_1 & B_1^T \\ B_1 & -A_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{3}$$

for the vector of unknowns $u = (u_1, u_2)^T \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ with

$$A_1 > 0, \quad A_2 \geq 0, \quad \text{and} \quad S_2 = A_2 + B_1 A_1^{-1} B_1^T > 0,$$

see, for example, [2].

Symmetric and indefinite linear systems of a natural 3-by-3 block-structure

$$\begin{pmatrix} A_1 & B_1^T & 0 \\ B_1 & -A_2 & B_2^T \\ 0 & B_2 & A_3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \tag{4}$$

with

$$A_1 > 0, \quad A_2 \geq 0, \quad A_3 \geq 0$$

and

$$S_2 = A_2 + B_1 A_1^{-1} B_1^T > 0 \quad \text{and} \quad S_3 = A_3 + B_2 S_2^{-1} B_2^T > 0$$

result, for example, from the so-called dual-dual formulation for coupling boundary and finite element methods, see [3]. Another example of a variational problem leading to systems of this 3-by-3 block-structure is the Hu-Washizu principle in elasticity, see [8].

Recently, Langer and Steinbach [6] introduced the boundary element tearing and interconnecting (BETI) method for second-order elliptic equations, which quite naturally leads to a 4-by-4 block-structured system with analogous properties. For the BETI method also a natural 3-by-3 block-structured formulation in appropriate sub-spaces is available, see [5].

As a general framework for efficient solvers to (1) we will consider preconditioned Richardson methods accelerated by Krylov-space techniques. As it will be seen later, efficient preconditioners for the matrix (2) can be constructed using efficient preconditioners of the Schur complements.

There are simple techniques to reduce the number n of blocks, leading to systems which are better understood: For example, by eliminating u_1 from (3) we obtain a symmetric and positive definite system for u_2 only:

$$S_2 u_2 = B_1 A_1^{-1} f_1 - f_2.$$

By eliminating variables the system is, in general, less indefinite, so to speak. Elimination of u_1 for systems (3) results in a (positive) definite system, for systems (4) in a standard indefinite saddle point problem of the form (3), and so on. However, the application of Richardson-type methods for the reduced system requires that matrix-vector products for the corresponding Schur complements must be available, a prize, which might prevent us from using this approach.

Besides eliminating variables, another way of reducing the number of blocks is to reorder the block-columns and block-rows. For example, if we first collect the blocks with odd indices followed by the blocks with even indices, we obtain the block matrix

$$\begin{pmatrix} A_1 & & & B_1^T & & \\ & A_3 & & B_2 & B_3^T & \\ & & \ddots & & \ddots & \ddots \\ B_1 & B_2^T & & -A_2 & & \\ & B_3 & \ddots & & -A_4 & \\ & & \ddots & & & \ddots \end{pmatrix}$$

with a natural 2-by-2 block-structure, where the first large diagonal block is positive semi-definite and the second large one is negative semi-definite. However, if one of the matrices A_i with odd indices greater than one is singular (e.g., 0, as it is often the case in applications), the whole first block becomes singular, and, perhaps more significantly, even if this block remains definite we change the Schur complement by the reordering, for which it might be harder to find an appropriate preconditioner than before the reordering.

So, whether it is advantageous to reduce the system to a system of a smaller number of blocks depends on the availability of efficient evaluations of the matrix-vector products for the corresponding Schur complements and on the availability of efficient preconditioners for Schur complements. One should not exclude a priori some of the different possible formulations of a problem, and, therefore, we feel that there is a need for understanding the general case. With this paper we want to provide an analysis for the general case as a basis for comparing the efficiency of different approaches.

The focus will be on Uzawa-type methods for preconditioning and on the conjugate gradient method for accelerating the method. The main reference for this approach is

Bramble and Pasciak [1], where it was shown for (3) that, under appropriate assumptions, the inexact Uzawa method leads to a preconditioner such that the preconditioned system matrix becomes symmetric and positive definite with respect to a particular scalar product, and, therefore, can be accelerated by the conjugate gradient method. See also [9], where some of the estimates were improved. An extension of the work by Bramble and Pasciak to systems of the form (4) (two-fold saddle point problems) can be found in Gatica and Heuer [4].

The purpose of this paper is to extended the construction of inexact Uzawa preconditioners as well as their analysis to the general case $n \geq 2$ and provide estimates for the eigenvalues of the preconditioned system matrix, which are partly sharper than known estimates from literature. It will turn out that this will be beneficial even for the most interesting special cases of indefinite systems of the form (3) and (4). Compared to former results for these special cases the proofs are more transparent by stressing the inherently recursive nature of the arguments.

The paper is organized as follows: In Section 2 the inexact Uzawa method is discussed. Conditions are specified which ensure that the preconditioned system matrix becomes symmetric and positive definite. Section 3 deals with estimates for the eigenvalues of the preconditioned system matrix based on lower and upper spectral bounds for the Schur complement preconditioners. Finally, in Section 4 some remarks concerning the implementation of the conjugate gradient acceleration are given.

For numerical experiments we refer to the companion papers [6] and [5], were results are reported for different formulations of the boundary element tearing and interconnecting (BETI) method for second-order elliptic equations. Formulations leading either to positive definite problems, or standard saddle point problems, or 2-fold saddle point problems are compared. It is demonstrated that either formulation has its own advantage depending on the range of the number of unknowns.

2 Inexact Uzawa methods

With

$$\mathcal{L} = \begin{pmatrix} S_1 & & & \\ B_1 & -S_2 & & \\ & \ddots & \ddots & \\ & & B_{n-1} & (-1)^{n-1} S_n \end{pmatrix} \quad \text{and} \quad \mathcal{S} = \begin{pmatrix} S_1 & & & \\ & -S_2 & & \\ & & \ddots & \\ & & & (-1)^{n-1} S_n \end{pmatrix}$$

it is easy to see that the following block LU-decomposition holds:

$$\mathcal{K} = \mathcal{L}\, \mathcal{S}^{-1}\, \mathcal{L}^T = \mathcal{L}\, \mathcal{U}$$

with

$$\mathcal{U} = \mathcal{S}^{-1}\,\mathcal{L}^T = \begin{pmatrix} I & S_1^{-1}B_1^T & & \\ & I & \ddots & \\ & & \ddots & (-1)^{n-2}S_{n-1}^{-1}B_{n-1}^T \\ & & & I \end{pmatrix}.$$

If, in $\mathcal{U}$, the off-diagonal blocks are neglected, and if, in $\mathcal{L}$, the matrices $A_1 = S_1$, S_2, …, S_n in the diagonal blocks are replaced by approximations (preconditioners) $\hat{A}_1 = \hat{S}_1$, $\hat{S}_2$, …, $\hat{S}_n$ we obtain the approximation (preconditioner)

$$\hat{\mathcal{L}} = \begin{pmatrix} \hat{S}_1 & & & \\ B_1 & -\hat{S}_2 & & \\ & \ddots & \ddots & \\ & & B_{n-1} & (-1)^{n-1}\,\hat{S}_n \end{pmatrix}$$

for $\mathcal{K}$, which could also be recursively defined by $\hat{\mathcal{L}} = \hat{\mathcal{L}}_n$ with

$$\hat{\mathcal{L}}_{i+1} = \begin{pmatrix} \hat{\mathcal{L}}_i & 0 \\ \mathsf{B}_i & (-1)^i\,\hat{S}_{i+1} \end{pmatrix} \quad \text{for } i = 1, \ldots, n-1,$$

where

$$\mathsf{B}_i = \begin{pmatrix} 0 & \cdots & 0 & B_i \end{pmatrix} = B_i\,\mathsf{E}_i, \quad \mathsf{E}_i = \begin{pmatrix} 0 & \cdots & 0 & I \end{pmatrix}$$

and initial setting $\hat{\mathcal{L}}_1 = \hat{A}_1$. Here, and in the sequel, we use symbols like $\mathcal{K}$, $\mathcal{L}$, …, to denote block matrices, symbols like B, E, …, to denote block vectors and symbols like S, B, …, to denote individual blocks.

This is not yet the final definition of the preconditioner. It will turn out that additional relaxation parameters $\tau_i > 0$ for $i = 1, \ldots, n-1$ are needed leading to block triangular preconditioners $\hat{\mathcal{L}}$ recursively defined by $\hat{\mathcal{L}} = \hat{\mathcal{L}}_n$, where $\hat{\mathcal{L}}_1 = \hat{A}_1$ and

$$\hat{\mathcal{L}}_{i+1} = \begin{pmatrix} \tau_i\,\hat{\mathcal{L}}_i & 0 \\ \mathsf{B}_i & (-1)^i\,\hat{S}_{i+1} \end{pmatrix} \quad \text{for } i = 1, \ldots, n-1.$$

Throughout the paper it is assumed that

$$\hat{S}_i = \hat{S}_i^T > 0 \quad \text{for } i = 1, 2, \ldots, n.$$

In an analogous recursive way we can write $\mathcal{K} = \mathcal{K}_n$ with

$$\mathcal{K}_{i+1} = \begin{pmatrix} \mathcal{K}_i & \mathsf{B}_i^T \\ \mathsf{B}_i & (-1)^i\,A_{i+1} \end{pmatrix} \quad \text{for } i = 1, \ldots, n-1,$$

and the initial setting $\mathcal{K}_1 = A_1$.

The corresponding preconditioned Richardson method then reads:

$$u^{(k+1)} = u^{(k)} + \hat{\mathcal{L}}^{-1}(f - \mathcal{K}u^{(k)}) = \mathcal{M}u^{(k)} + \hat{\mathcal{L}}^{-1}f$$

with the iteration matrix

$$\mathcal{M} = I - \hat{\mathcal{L}}^{-1}\mathcal{K} = \hat{\mathcal{L}}^{-1}(\hat{\mathcal{L}} - \mathcal{K}),$$

where I denotes the identity matrix.

This simple factorization in two non-symmetric matrices is not very helpful for the analysis. The following lemma provides an alternative factorization of the iteration matrix, on which the analysis of the method will be based:

Lemma 1. *Let $\mathcal{M}_j = I - \hat{\mathcal{L}}_j^{-1}\mathcal{K}_j$ for $j = 1, 2, \ldots, n$. Then*

$$\tau_i\, \mathcal{M}_{i+1} = \mathcal{P}_{i+1}^{-1}\ \mathcal{N}_{i+1}\ \mathcal{Q}_{i+1}$$

with

$$\mathcal{P}_{i+1} = \begin{pmatrix} \hat{S}_1 & & & \\ & \hat{S}_2 & & \\ & & \ddots & \\ & & & \hat{S}_{i+1} \end{pmatrix} = \begin{pmatrix} \mathcal{P}_i & \\ & \hat{S}_{i+1} \end{pmatrix}, \quad \mathcal{Q}_{i+1} = \begin{pmatrix} [I - \mathcal{M}_i] - \tau_i\, I & 0 \\ 0 & I \end{pmatrix}$$

and

$$\mathcal{N}_{i+1} = \begin{pmatrix} -\mathcal{P}_i & (-1)^i\, \mathsf{E}_i^T B_i^T \\ (-1)^i\, B_i \mathsf{E}_i & \tau_i\,(\hat{S}_{i+1} - A_{i+1}) - B_i \hat{S}_i^{-1} B_i^T \end{pmatrix}$$

for $i = 1, \ldots, n-1$.

The proof is straight forward and, therefore, omitted.

Assume now that the iteration matrix $\mathcal{M}_i$ is symmetric with respect to some scalar product $(u, v)_i$:

$$(\mathcal{M}_i u, v)_i = (u, \mathcal{M}_i v)_i \quad \text{for all } u, v, \tag{5}$$

that the preconditioned system matrix $\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i = I - \mathcal{M}_i$ is positive definite with respect to the same scalar product $(u, v)_i$:

$$([I - \mathcal{M}_i]u, u)_i > 0 \quad \text{for all } u, \tag{6}$$

and that

$$(\mathsf{E}_i^T u_i, v)_i = (\hat{S}_i u_i, \mathsf{E}_i v)_{\ell_2} \quad \text{for all } u_i, v, \tag{7}$$

where $(u, v)_{\ell_2}$ denotes the Euclidean scalar product. The third condition says that the scalar product $(u, v)_i$ of two block vectors u and v reduces to the simple scalar product $(\hat{S}_i u_i, v_i)_{\ell_2}$ of their last blocks, if the first $i-1$ blocks of one of the factors, say u, vanish.

Observe that these conditions are easily verified for $i = 1$ and the scalar product

$$(u, v)_1 = (\hat{A}_1 u, v)_{\ell_2} = (\hat{S}_1 u, v)_{\ell_2}. \tag{8}$$

We will now construct a scalar product $(u, v)_{i+1}$ for block vectors with $i+1$ blocks.

In a first step we introduce the following auxiliary scalar product:

$$(u, v)_* = (u', v')_i + (\hat{S}_{i+1} u_{i+1}, v_{i+1})_{\ell_2}$$

with $u' = (u_1, \ldots, u_i)^T$, $v' = (v_1, \ldots, v_i)^T$, $u = (u', u_{i+1})^T$ and $v = (v', v_{i+1})^T$.

Then we have:

Lemma 2. *Assume that (5), (6) and (7) hold. Then $\mathcal{P}_{i+1}^{-1}\mathcal{N}_{i+1}$ and $\mathcal{Q}_{i+1}$ are symmetric with respect to the scalar product $(u, v)_*$. If, additionally, the parameter $\tau_i > 0$ is chosen such that*

$$([I - \mathcal{M}_i]u, u)_i > \tau_i\, (u, u)_i \quad \text{for all } u \neq 0, \tag{9}$$

and such parameters exist, then $\mathcal{Q}_{i+1}$ is even positive definite with respect to the scalar product $(u, v)_*$, for all $i = 1, \ldots, n-1$.

Proof. From the definition of $\mathcal{P}_{i+1}$ and $\mathcal{N}_{i+1}$ we obtain

$$\begin{aligned}
(\mathcal{P}_{i+1}^{-1}\mathcal{N}_{i+1} u, v)_* &= \Big(-u' + (-1)^i\, \mathsf{E}_i^T \hat{S}_i^{-1} B_i^T u_{i+1}, v'\Big)_i \\
&\quad + \Big((-1)^i\, B_i \mathsf{E}_i u' + \Big[\tau_i\, (\hat{S}_{i+1} - A_{i+1}) - B_i \hat{S}_i^{-1} B_i^T\Big]\, u_{i+1}, v_{i+1}\Big)_{\ell_2} \\
&= -(u', v')_i + \Big(\Big[\tau_i\, (\hat{S}_{i+1} - A_{i+1}) - B_i \hat{S}_i^{-1} B_i^T\Big]\, u_{i+1}, v_{i+1}\Big)_{\ell_2} \\
&\quad + (-1)^i\, \Big(\mathsf{E}_i^T \hat{S}_i^{-1} B_i^T u_{i+1}, v'\Big)_i + (-1)^i\, (B_i \mathsf{E}_i u', v_{i+1})_{\ell_2}\,.
\end{aligned}$$

By using (7) it follows that

$$\Big(\mathsf{E}_i^T \hat{S}_i^{-1} B_i^T u_{i+1}, v'\Big)_i = \big(B_i^T u_{i+1}, \mathsf{E}_i v'\big)_{\ell_2} = (u_{i+1}, B_i \mathsf{E}_i v')_{\ell_2}\,,$$

which immediately implies the symmetry of $\mathcal{P}_{i+1}^{-1}\mathcal{N}_{i+1}$.

The symmetry of $\mathcal{Q}_{i+1}$ under the assumption (5) is obvious. The existence of a positive parameter τ_i satisfying (9) directly follows from (6), the positive definiteness of $\mathcal{Q}_{i+1}$ is then trivial. □

Under the assumptions (5), (9), and (7) the new scalar product can be introduced for block vectors with $i+1$ blocks by

$$(u, v)_{i+1} = (\mathcal{Q}_{i+1} u, v)_* = (([I - \mathcal{M}_i] - \tau_i\, I)u', v')_i + (\hat{S}_{i+1} u_{i+1}, v_{i+1})_{\ell_2} \tag{10}$$

with $u' = (u_1, \ldots, u_i)^T$, $v' = (v_1, \ldots, v_i)^T$, $u = (u', u_{i+1})^T$ and $v = (v', v_{i+1})^T$.

We will show in the next lemma that analogous properties as (5), (6) and (7) hold for this new scalar product for block vectors with $i+1$ blocks.

In the proof of the lemma, the following notations will be used: For a 2-by-2 block matrix

$$\mathcal{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with a non-singular square matrix A_{11} its Schur complement $\mathrm{Schur}(\mathcal{A})$ is introduced by

$$\mathrm{Schur}(\mathcal{A}) = A_{22} - A_{21} A_{11}^{-1} A_{12}.$$

The following three simple properties (11), (12), and (13) will be helpful:

$$[\mathrm{Schur}(\mathcal{A})]^{-1} = \begin{pmatrix} 0 & I \end{pmatrix} \mathcal{A}^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix} \tag{11}$$

for non-singular matrices $\mathcal{A}$ with non-singular blocks A_{11}.

$$\mathrm{Schur}(\mathcal{L}\mathcal{A}) = L_{22}\ \mathrm{Schur}(\mathcal{A}) \tag{12}$$

for matrices

$$\mathcal{L} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \quad \mathcal{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with non-singular blocks L_{11} and A_{11}.

$$(\mathrm{Schur}(\mathcal{A})\, x_2, x_2)_{\ell_2} = \min_{x_1} \left(\mathcal{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right)_{\ell_2} \tag{13}$$

for symmetric matrices $\mathcal{A}$ with $A_{11} > 0$.

Lemma 3. *Assume that (5), (6), (7) hold and τ_i is chosen according to (9). Then $\mathcal{M}_{i+1}$ is symmetric and $\hat{\mathcal{L}}_{i+1}^{-1}\mathcal{K}_{i+1} = I - \mathcal{M}_{i+1}$ is symmetric and positive definite with respect to the scalar product $(u,v)_{i+1}$, and*

$$(\mathsf{E}_{i+1}^T u_{i+1}, v)_{i+1} = (\hat{S}_{i+1} u_{i+1}, \mathsf{E}_{i+1} v)_{\ell_2} \quad \textit{for all } u_{i+1}, v,$$

for all $i = 1, \ldots, n-1$.

Proof. We have

$$\tau_i\, (\mathcal{M}_{i+1} u, v)_{i+1} = (\mathcal{Q}_{i+1} \mathcal{P}_{i+1}^{-1} \mathcal{N}_{i+1} \mathcal{Q}_{i+1} u, v)_* = (\mathcal{P}_{i+1}^{-1} \mathcal{N}_{i+1} \mathcal{Q}_{i+1} u, \mathcal{Q}_{i+1} v)_*,$$

which shows the symmetry of $\mathcal{M}_{i+1}$, see Lemma 2.

Condition (9) guarantees that $\hat{\mathcal{L}}_{i+1}^{-1}\mathcal{K}_{i+1} = I - \mathcal{M}_{i+1}$ is symmetric with respect to the scalar product $(u,v)_{i+1}$ and, therefore, it has only real eigenvalues, see Lemma 2.

The scalar product $(u,v)_j$ for $j \le i+1$ can be represented by some symmetric and positive definite matrix, say $\mathcal{D}_j$:

$$(u,v)_j = (\mathcal{D}_j u, v)_{\ell_2}.$$

From (10) we obtain the recursion

$$\mathcal{D}_{i+1} = \begin{pmatrix} \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \tau_i\, \mathcal{D}_i & 0 \\ 0 & \hat{S}_{i+1} \end{pmatrix}. \tag{14}$$

Let $\Theta_j(\lambda)$ be the Schur complement of $\mathcal{D}_j\hat{\mathcal{L}}_j^{-1}\mathcal{K}_j - \lambda\,\mathcal{D}_j$, considered as 2-by-2 block matrix, where the first block includes the first $j-1$ blocks and the second block is the j-th block of the original decomposition. By using (12) it is easy to see that

$$\begin{aligned}\Theta_j(\lambda) &= \mathrm{Schur}(\mathcal{D}_j\hat{\mathcal{L}}_j^{-1}\mathcal{K}_j - \lambda\,\mathcal{D}_j) = \mathrm{Schur}(\mathcal{D}_j\hat{\mathcal{L}}_j^{-1}(\mathcal{K}_j - \lambda\,\hat{\mathcal{L}}_j)) \\ &= (-1)^{j-1}\,\mathrm{Schur}(\mathcal{K}_j - \lambda\,\hat{\mathcal{L}}_j).\end{aligned}$$

So

$$\mathrm{Schur}(\mathcal{K}_j - \lambda\,\hat{\mathcal{L}}_j) = (-1)^{j-1}\,\Theta_j(\lambda).$$

For $\lambda \le 0$, the matrix $\mathcal{D}_i\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i - \lambda\,\mathcal{D}_i = \mathcal{D}_i\hat{\mathcal{L}}_i^{-1}(\mathcal{K}_i - \lambda\,\hat{\mathcal{L}}_i)$ is positive definite, since the matrices $\mathcal{D}_i\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i$ and $\mathcal{D}_i$ are positive definite. Therefore, the matrix $\mathcal{K}_i - \lambda\,\hat{\mathcal{L}}_i$ must be non-singular and the Schur complement $\Theta_i(\lambda)$ is well-defined for $\lambda \le 0$ and positive definite:

$$\Theta_i(\lambda) > 0 \quad \text{for } \lambda \le 0. \tag{15}$$

Observe that

$$\mathcal{K}_{i+1} - \lambda\,\hat{\mathcal{L}}_{i+1} = \begin{pmatrix} \mathcal{K}_i - \tau_i\lambda\,\hat{\mathcal{L}}_i & \mathsf{E}_i^T B_i^T \\ (1-\lambda)B_i\mathsf{E}_i & (-1)^i\,A_{i+1} - \lambda\,(-1)^i\,\hat{S}_{i+1} \end{pmatrix}$$

with a non-singular sub-matrix $\mathcal{K}_i - \tau_i\lambda\,\hat{\mathcal{L}}_i$ for $\lambda \le 0$. From the identity (11) we then obtain

$$\begin{aligned}\Theta_{i+1}(\lambda) &= A_{i+1} - \lambda\,\hat{S}_{i+1} - (-1)^i(1-\lambda)\,B_i\mathsf{E}_i(\mathcal{K}_i - \tau_i\lambda\,\hat{\mathcal{L}}_i)^{-1}\mathsf{E}_i^T B_i^T \\ &= A_{i+1} - \lambda\,\hat{S}_{i+1} - (-1)^i(1-\lambda)\,B_i\left[\mathrm{Schur}(\mathcal{K}_i - \tau_i\lambda\,\hat{\mathcal{L}}_i)\right]^{-1} B_i^T \\ &= A_{i+1} - \lambda\,\hat{S}_{i+1} + (1-\lambda)\,B_i\Theta_i(\tau_i\lambda)^{-1}B_i^T.\end{aligned}$$

Using this formula and (15) it follows that $\Theta_{i+1}(\lambda) > 0$ for $\lambda \le 0$.

Since both the sub-matrix $\mathcal{K}_i - \tau_i\lambda\,\hat{\mathcal{L}}_i$ and the Schur complement $(-1)^i\,\Theta_{i+1}(\lambda)$ of the matrix $\mathcal{K}_{i+1} - \lambda\,\hat{\mathcal{L}}_{i+1}$ are non-singular, the matrix $\mathcal{K}_{i+1} - \lambda\,\hat{\mathcal{L}}_{i+1}$ itself is non-singular for $\lambda \le 0$, which shows that non-positive numbers λ cannot be eigenvalues.

The rest trivially follows from the construction of the scalar product. □

Condition (9) can also be written as $\tau_i < \lambda_{\min}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i)$. Therefore, we summarize the discussion and formulate the main result of this section in the following way:

Theorem 1. *If the parameters τ_i are chosen such that*

$$0 < \tau_i < \lambda_{min}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i) \quad \textit{for all } i = 1, \dots, n-1,$$

and this is possible, then $\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n$ is symmetric and positive definite with respect to the scalar product $(u,v)_n$, recursively given by (8) and (10).

3 Eigenvalue estimates

All statements in Section 2 remain valid even without the assumption

$$A_i \geq 0 \quad \text{for all } i = 2, \ldots, n.$$

However, for the forthcoming estimates, this assumption will be necessary.

Lower and upper bounds for the eigenvalues of $\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n$ are now derived under the assumptions

$$\underline{\sigma}_i \, \hat{S}_i \leq S_i \leq \overline{\sigma}_i \, \hat{S}_i, \tag{16}$$

for some lower and upper spectral bounds $\underline{\sigma}_i$, $\overline{\sigma}_i$ with $0 < \underline{\sigma}_i \leq \overline{\sigma}_i$ for $i = 1, 2, \ldots, n$.

We start the discussion with a special example of a block-tridiagonal indefinite matrix. This example serves two purposes: Firstly, it allows to introduce and discuss two families of rational functions $\underline{\theta}_n(\lambda)$ and $\overline{\theta}_n(\lambda)$ which are needed for the general case. And, secondly, this example will prove that the estimates for the general case are sharp.

3.1 An example

Consider the following block matrix:

$$\mathcal{K}_n = \begin{pmatrix} A_1 & B_1^T & & \\ B_1 & -A_2 & \ddots & \\ & \ddots & \ddots & B_{n-1}^T \\ & & B_{n-1} & (-1)^{n-1} A_n \end{pmatrix}$$

with

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_i = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for all } i = 2, \ldots, n$$

and

$$B_i = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{for all } i = 1, \ldots, n-1,$$

whose sequence of Schur complements is given by

$$S_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{for } i = 1, 2, \ldots, n.$$

The following Schur complement approximations are considered:

$$\hat{S}_i = \operatorname{diag}\left(\frac{1}{\underline{\sigma}_i}, \frac{1}{\overline{\sigma}_i}, \frac{1}{\overline{\sigma}_i} \right) \quad \text{for } i = 1, 2, \ldots, n.$$

Then Condition (16) is obviously satisfied.

Let $\hat{\mathcal{L}}_n$ be the corresponding block-triangular preconditioner, introduced in Section 2. From the analysis in Section 2 it immediately follows that a symmetric and positive definite matrix $\mathcal{D}_n$ can be constructed such that the matrix $\mathcal{D}_n \hat{\mathcal{L}}_n^{-1} \mathcal{K}_n$ is symmetric and positive definite, as long as the parameters τ_i are chosen such that

$$0 < \tau_i < \underline{\lambda}_i \quad \text{for all } i = 1, \ldots, n-1 \tag{17}$$

with

$$\underline{\lambda}_i = \lambda_{\min}(\hat{\mathcal{L}}_i^{-1} \mathcal{K}_i)$$

for all $i = 1, \ldots, n$. Additionally, we introduce

$$\overline{\lambda}_i = \lambda_{\max}(\hat{\mathcal{L}}_i^{-1} \mathcal{K}_i)$$

for all $i = 1, \ldots, n$.

We will next describe these extreme eigenvalues more explicitely as extreme zeroes of special rational functions: For this we first discuss the functions $\Theta_j(\lambda)$, see Section 2, given by

$$\Theta_j(\lambda) = \text{Schur}\left(\mathcal{D}_j \hat{\mathcal{L}}_j^{-1} \mathcal{K}_j - \lambda\, \mathcal{D}_j\right) = (-1)^{j-1}\ \text{Schur}\left(\mathcal{K}_j - \lambda\, \hat{\mathcal{L}}_j\right),$$

which satisfy the recursion

$$\Theta_{i+1}(\lambda) \quad = \quad A_{i+1} - \lambda\, \hat{S}_{i+1} + (1-\lambda)\, B_i \Theta_i(\tau_i \lambda)^{-1} B_i^T \quad \text{for } i = 1, \ldots, n-1$$

with the initial setting

$$\Theta_1(\lambda) = A_1 - \lambda\, \hat{A}_1.$$

Simple calculations show that, for our example, these functions $\Theta_j(\lambda)$ reduce to 3-by-3 diagonal matrices of the following form:

$$\Theta_j(\lambda) = \text{diag}\left(\underline{\theta}_j(\lambda), 1 - \frac{\lambda}{\overline{\sigma}_j}, \overline{\theta}_j(\lambda)\right) \tag{18}$$

with

$$\underline{\theta}_{i+1}(\lambda) = -\frac{\lambda}{\underline{\sigma}_{i+1}} + \frac{1-\lambda}{1 - \tau_i \lambda / \overline{\sigma}_i} \quad \text{for } i = 1, \ldots, n-1 \tag{19}$$

and initial setting $\underline{\theta}_1(\lambda) = 1 - \lambda/\underline{\sigma}_1$, and

$$\overline{\theta}_{i+1}(\lambda) = -\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{\overline{\theta}_i(\tau_i \lambda)} \quad \text{for } i = 1, \ldots, n-1 \tag{20}$$

with initial setting $\overline{\theta}_1(\lambda) = 1 - \lambda/\overline{\sigma}_1$.

Now we have

Lemma 4. *Assume that (17) is satisfied. Then:*

1. *The functions* $\underline{\theta}_i(\lambda)$ *and* $\overline{\theta}_i(\lambda)$ *are positive and monotonically decreasing on* $(-\infty, \underline{\lambda}_i)$ *and negative and monotonically decreasing on* $(\overline{\lambda}_i, \infty)$ *for all* $i = 1, 2, \ldots, n$.

2. *The following estimates hold for all* $i = 1, 2, \ldots, n$*:*

$$0 < \underline{\theta}_i(\lambda) \leq \overline{\theta}_i(\lambda) \leq 1 - \frac{\lambda}{\overline{\sigma}_i} \quad \text{for all } \lambda \in [0, \underline{\lambda}_i)$$

and

$$\max\left(1 - \frac{\lambda}{\overline{\sigma}_i}, \underline{\theta}_i(\lambda)\right) \leq \overline{\theta}_i(\lambda) < 0 \quad \text{for all } \lambda \in (\overline{\lambda}_i, \infty).$$

3. *The value* $\underline{\lambda}_i$ *is the smallest zero of* $\underline{\theta}_i(\lambda)$ *and* $\overline{\lambda}_i$ *is the largest zero of* $\overline{\theta}_i(\lambda)$ *for all* $i = 1, 2, \ldots, n$.

Proof. (a) For $\lambda_1 \leq \lambda_2 < \underline{\lambda}_i$ we have

$$\mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \lambda_1 \mathcal{D}_i \geq \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \lambda_2 \mathcal{D}_i > 0,$$

which implies the same ordering of the Schur complements, see (11):

$$\Theta_i(\lambda_1) \geq \Theta_i(\lambda_2) > 0.$$

Analogously, for $\overline{\lambda}_i < \lambda_1 \leq \lambda_2$ it follows that

$$0 > \Theta_i(\lambda_1) \geq \Theta_i(\lambda_2).$$

Because of (18) the functions $\underline{\theta}_i(\lambda)$ and $\overline{\theta}_i(\lambda)$ inherit these properties.

(b) For the rest of the proof a refined analysis of the monotonicity is needed: Observe that

$$\underline{\theta}_{i+1}(\lambda) = -\frac{\lambda}{\underline{\sigma}_{i+1}} + \underline{\psi}_{i+1}(\lambda) \quad \text{and} \quad \overline{\theta}_{i+1}(\lambda) = -\frac{\lambda}{\overline{\sigma}_{i+1}} + \overline{\psi}_{i+1}(\lambda)$$

with

$$\underline{\psi}_{i+1}(\lambda) = \frac{1 - \lambda}{1 - \tau_i \lambda / \overline{\sigma}_i} \quad \text{and} \quad \overline{\psi}_{i+1}(\lambda) = \frac{1 - \lambda}{\overline{\theta}_i(\tau_i \lambda)}.$$

Now

$$\mathcal{D}_{i+1} \hat{\mathcal{L}}_{i+1}^{-1} \mathcal{K}_{i+1} - \lambda \mathcal{D}_{i+1} = \mathcal{A}_{i+1}(\lambda) - \lambda \begin{pmatrix} 0 & 0 \\ 0 & \hat{S}_{i+1} \end{pmatrix} \tag{21}$$

with

$$\mathcal{A}_{i+1}(\lambda) = \mathcal{D}_{i+1} \hat{\mathcal{L}}_{i+1}^{-1} \mathcal{K}_{i+1} - \lambda \begin{pmatrix} \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \tau_i \mathcal{D}_i & 0 \\ 0 & 0 \end{pmatrix},$$

see (14). For the left upper block $A_{i+1}^{(1,1)}(\lambda)$ of $\mathcal{A}_{i+1}(\lambda)$ we obtain

$$A_{i+1}^{(1,1)}(\lambda) = \frac{1}{\tau_i} \left(\mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \tau_i \mathcal{D}_i\right) \mathcal{D}_i^{-1} \left(\mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i - \tau_i \lambda \mathcal{D}_i\right).$$

It follows that $A_{i+1}^{(1,1)}(\lambda) > 0$ for $\tau_i\lambda < \underline{\lambda}_i$. Obviously, $\mathcal{A}_{i+1}(\lambda)$ is monotonically decreasing in λ. Therefore, (13) ensures that $\Psi_{i+1}(\lambda) = \mathrm{Schur}(\mathcal{A}_{i+1}(\lambda))$ is also monotonically decreasing in λ on the intervals $(-\infty, \underline{\lambda}_i/\tau_i)$.

Analogously, it follows that $\Psi_{i+1}(\lambda)$ is monotonically decreasing in λ on the intervals $(\overline{\lambda}_i/\tau_i, \infty)$.

From (21) we obtain

$$\Theta_{i+1}(\lambda) = \Psi_{i+1}(\lambda) - \mathrm{diag}\left(\frac{\lambda}{\underline{\sigma}_{i+1}}, \frac{\lambda}{\overline{\sigma}_{i+1}}, \frac{\lambda}{\overline{\sigma}_{i+1}}\right)$$

which shows that

$$\Psi_{i+1}(\lambda) = \mathrm{diag}\left(\underline{\psi}_{i+1}(\lambda), 1, \overline{\psi}_{i+1}(\lambda)\right).$$

The functions $\underline{\psi}_{i+1}(\lambda)$ and $\overline{\psi}_{i+1}(\lambda)$ inherit the monotonicity of $\Psi_{i+1}(\lambda)$.

(c) The rest is shown by induction in i: The statements are trivially true for $i = 1$. Assume now that

$$0 < \underline{\theta}_i(\lambda) \leq \overline{\theta}_i(\lambda) \leq 1 - \frac{\lambda}{\overline{\sigma}_i} \quad \text{for all } \lambda \in [0, \underline{\lambda}_i), \tag{22}$$

and

$$\max\left(1 - \frac{\lambda}{\overline{\sigma}_i}, \underline{\theta}_i(\lambda)\right) \leq \overline{\theta}_i(\lambda) < 0 \quad \text{for all } \lambda \in (\overline{\lambda}_i, \infty), \tag{23}$$

and $\underline{\lambda}_i$ is the smallest zero of $\underline{\theta}_i(\lambda)$, and $\overline{\lambda}_i$ is the largest zero of $\overline{\theta}_i(\lambda)$.

An immediate consequence of (22) and (23) is:

$$0 < \tau_i < \underline{\lambda}_i \leq \overline{\sigma}_i \leq \overline{\lambda}_i.$$

Hence, $1 - \tau_i\lambda/\overline{\sigma}_i > 1 - \lambda$ for all $\lambda \in (0, 1]$. Therefore, the function $\underline{\theta}_{i+1}(\lambda)$ has no pole in $[0, 1]$. From $\underline{\theta}_{i+1}(1) = -1/\underline{\sigma}_{i+1} < 0$, it follows for the smallest eigenvalue: $\underline{\lambda}_{i+1} < 1 < \underline{\lambda}_i/\tau_i$.

Since

$$\overline{\psi}_{i+1}(\lambda) \leq \overline{\psi}_{i+1}(0) = 1 \quad \text{for all } \lambda \in [0, \underline{\lambda}_{i+1}),$$

it follows that

$$\overline{\theta}_{i+1}(\lambda) = -\frac{\lambda}{\overline{\sigma}_{i+1}} + \overline{\psi}_{i+1}(\lambda) \leq -\frac{\lambda}{\overline{\sigma}_{i+1}} + 1 \quad \text{for all } \lambda \in [0, \underline{\lambda}_{i+1}).$$

Since

$$\overline{\theta}_i(\tau_i\lambda) \leq 1 - \frac{\tau_i\lambda}{\overline{\sigma}_i} \quad \text{for all } \lambda \in (0, \underline{\lambda}_i/\tau_i),$$

see (22), it follows that

$$-\frac{\lambda}{\underline{\sigma}_{i+1}} + \frac{1 - \lambda}{1 - \tau_i\lambda/\overline{\sigma}_i} \leq -\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1 - \lambda}{\overline{\theta}_i(\tau_i\lambda)} \quad \text{for all } \lambda \in [0, \underline{\lambda}_{i+1}).$$

With (19) and (20) these inequalities can be written as

$$\underline{\theta}_{i+1}(\lambda) \leq \overline{\theta}_{i+1}(\lambda) \leq -\frac{\lambda}{\overline{\sigma}_{i+1}} + 1 \quad \text{for all } \lambda \in [0, \underline{\lambda}_{i+1}). \tag{24}$$

Since $1 - \lambda < 1 - \overline{\lambda}_i/\tau_i < 0$ and $\overline{\theta}_i(\tau_i\lambda) < 0$ for all $\lambda > \overline{\lambda}_i/\tau_i$, it follows from (20) that

$$\lim_{\lambda \to \overline{\lambda}_i/\tau_i +} \overline{\theta}_{i+1}(\lambda) = +\infty.$$

Then, we know for the largest eigenvalue: $\overline{\lambda}_{i+1} > \overline{\lambda}_i/\tau_i > 1$.

Since

$$\overline{\psi}_{i+1}(\lambda) \geq \overline{\psi}_{i+1}(+\infty) = \frac{\overline{\sigma}_i}{\tau_i} > 1 \quad \text{for all } \lambda \in (\overline{\lambda}_i/\tau_i, \infty),$$

it follows

$$-\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{\overline{\theta}_i(\tau_i\lambda)} \geq -\frac{\lambda}{\overline{\sigma}_{i+1}} + 1 \quad \text{for all } \lambda \in (\overline{\lambda}_{i+1}, \infty).$$

Since

$$\overline{\theta}_i(\tau_i\lambda) \geq 1 - \frac{\tau_i\lambda}{\overline{\sigma}_i} \quad \text{for all } \lambda \in (\overline{\lambda}_i/\tau_i, \infty),$$

it follows

$$-\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{\overline{\theta}_i(\tau_i\lambda)} \geq -\frac{\lambda}{\underline{\sigma}_{i+1}} + \frac{1-\lambda}{1 - \tau_i\lambda/\overline{\sigma}_i} \quad \text{for all } \lambda \in (\overline{\lambda}_{i+1}, \infty).$$

With (19) and (20) these inequalities can be written as

$$\overline{\theta}_{i+1}(\lambda) \geq \underline{\theta}_{i+1}(\lambda) \quad \text{and} \quad \overline{\theta}_{i+1}(\lambda) \geq -\frac{\lambda}{\overline{\sigma}_{i+1}} + 1.$$

Since $\underline{\lambda}_{i+1}$ is an eigenvalue of $\hat{\mathcal{L}}_{i+1}^{-1}\mathcal{K}_{i+1}$, the matrix $\mathcal{K}_{i+1} - \underline{\lambda}_{i+1}\,\hat{\mathcal{L}}_{i+1}$ is singular. However, its left upper sub-matrix $\mathcal{K}_i - \tau_i\underline{\lambda}_{i+1}\,\hat{\mathcal{L}}_i$ is non-singular since $\tau_i\underline{\lambda}_{i+1} < \underline{\lambda}_i$. This implies that its Schur complement must be singular, i.e.: $\Theta(\underline{\lambda}_{i+1})$ is singular.

But then, because of (18) and the ordering of the diagonal elements of $\Theta(\underline{\lambda}_{i+1})$, see (24), $\underline{\lambda}_{i+1}$ must be a zero of $\underline{\theta}_{i+1}(\lambda)$. Since $\underline{\theta}_{i+1}(\lambda) > 0$ for all $\lambda < \underline{\lambda}_{i+1}$, $\underline{\lambda}_{i+1}$ must be the smallest zero of $\underline{\theta}_{i+1}(\lambda)$.

An analogous argument applies to $\overline{\lambda}_{i+1}$. □

3.2 General block-tridiagonal symmetric and indefinite matrices

We now turn to the general case. With the help of the functions $\underline{\theta}_i(\lambda)$ and $\overline{\theta}_i(\lambda)$, just introduced, estimates for general indefinite block matrices can be derived:

Lemma 5. *Assume that $\mathcal{K}_n$ is a general block-tridiagonal symmetric and indefinite matrix and $\hat{\mathcal{L}}_n$ is the preconditioner of Section 2 for which the spectral estimates (16) are satisfied and the parameters τ_i are chosen such that (17) holds.*

Then, with the notations of Section 2, the matrix-valued functions $\Theta_i(\lambda)$ are well-defined and we have:

1. *The following estimates hold:*

$$0 < \underline{\theta}_i(\lambda)\, S_i \le \Theta_i(\lambda) \le \left[1 - \frac{\lambda}{\overline{\sigma}_i}\right] S_i \quad \textit{for all } \lambda \in [0, \underline{\lambda}_i)$$

and

$$\Theta_i(\lambda) \le \overline{\theta}_i(\lambda)\, S_i < 0 \quad \textit{for all } \lambda \in (\overline{\lambda}_i, \infty),$$

for all $i = 1, 2, \ldots, n$.

2. *The following eigenvalue bounds hold:*

$$\underline{\lambda}_i \le \lambda_{min}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i) \quad \textit{and} \quad \overline{\lambda}_i \ge \lambda_{max}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i)$$

for all $i = 1, 2, \ldots, n$.

Proof. For $i = 1$ the inequalities are a simple consequence of the spectral estimates for $\hat{A}_1 = \hat{S}_1$. For general $i \le n$ the inequalities are shown by induction:

Assume that

$$0 < \underline{\theta}_i(\lambda)\, S_i \le \Theta_i(\lambda) \le \left[1 - \frac{\lambda}{\overline{\sigma}_i}\right] S_i \quad \text{for all } \lambda \in [0, \underline{\lambda}_i),$$

and

$$\Theta_i(\lambda) \le \overline{\theta}_i(\lambda)\, S_i < 0 \quad \text{for all } \lambda \in (\overline{\lambda}_i, \infty),$$

and

$$\underline{\lambda}_i \le \lambda_{\min}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i) \quad \text{and} \quad \overline{\lambda}_i \ge \lambda_{\max}(\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i).$$

Then, for $\lambda \in [0, \underline{\lambda}_{i+1})$, we have $\lambda < \underline{\lambda}_{i+1} < 1 < \underline{\lambda}_i/\tau_i$, and $\underline{\psi}_{i+1}(\lambda) = (1-\lambda)/(1 - \tau_i\lambda/\overline{\sigma}_i) \le \underline{\psi}_{i+1}(0) = 1$, see the proof of Lemma 4. Therefore,

$$\begin{aligned}
\Theta_{i+1}(\lambda) &= A_{i+1} + (1-\lambda)\, B_i \Theta_i(\tau_i\lambda)^{-1} B_i^T - \lambda\, \hat{S}_{i+1} \\
&\ge A_{i+1} + \frac{1-\lambda}{1 - \tau_i\lambda/\overline{\sigma}_i}\, B_i S_i^{-1} B_i^T - \frac{\lambda}{\underline{\sigma}_{i+1}}\, S_{i+1} \\
&\ge \frac{1-\lambda}{1 - \tau_i\lambda/\overline{\sigma}_i}\, (A_{i+1} + B_i S_i^{-1} B_i^T) - \frac{\lambda}{\underline{\sigma}_{i+1}}\, S_{i+1} \\
&= \left[-\frac{\lambda}{\underline{\sigma}_{i+1}} + \frac{1-\lambda}{1 - \tau_i\lambda/\overline{\sigma}_i}\right] S_{i+1} = \underline{\theta}_{i+1}(\lambda)\, S_{i+1}.
\end{aligned}$$

In the same manner as in the proof of Lemma 4 one shows that $(1-\lambda)/\underline{\theta}_i(\tau_i\lambda)$ is monotonically decreasing on $(-\infty, \underline{\lambda}_i/\tau_i)$ and $(1-\lambda)/\underline{\theta}_i(\tau_i\lambda) \le 1$ on $[0, \underline{\lambda}_i/\tau_i)$. Then, for $\lambda \in [0, \underline{\lambda}_{i+1})$, we have

$$\begin{aligned}
\Theta_{i+1}(\lambda) &= A_{i+1} + (1-\lambda)\, B_i \Theta_i(\tau_i\lambda)^{-1} B_i^T - \lambda\, \hat{S}_{i+1} \\
&\le A_{i+1} + \frac{1-\lambda}{\underline{\theta}_i(\tau_i\lambda)}\, B_i S_i^{-1} B_i^T - \frac{\lambda}{\overline{\sigma}_{i+1}}\, S_{i+1} \\
&\le A_{i+1} + B_i S_i^{-1} B_i^T - \frac{\lambda}{\overline{\sigma}_{i+1}}\, S_{i+1} = \left[1 - \frac{\lambda}{\overline{\sigma}_{i+1}}\right] S_{i+1}.
\end{aligned}$$

For $\lambda \in (\overline{\lambda}_{i+1}, \infty)$, we have $\lambda > \overline{\lambda}_{i+1} > \overline{\lambda}_i/\tau_i > 1$, and $\overline{\psi}_{i+1}(\lambda) = (1-\lambda)/\overline{\theta}_i(\tau_i\lambda) \leq \overline{\psi}_{i+1}(0) = 1$, see the proof of Lemma 4, and, therefore,

$$\begin{aligned}
\Theta_{i+1}(\lambda) &= A_{i+1} + (1-\lambda)\, B_i \Theta_i(\tau_i\lambda)^{-1} B_i^T - \lambda\, \hat{S}_{i+1} \\
&\leq A_{i+1} + \frac{1-\lambda}{\overline{\theta}_i(\tau_i\lambda)}\, B_i S_i^{-1} B_i^T - \frac{\lambda}{\overline{\sigma}_{i+1}}\, S_{i+1} \\
&\leq \frac{1-\lambda}{\overline{\theta}_i(\tau_i\lambda)}\, (A_{i+1} + B_i S_i^{-1} B_i^T) - \frac{\lambda}{\overline{\sigma}_{i+1}}\, S_{i+1} \\
&= \left[-\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{\overline{\theta}_i(\tau_i\lambda)} \right] S_{i+1} = \overline{\theta}_{i+1}(\lambda)\, S_{i+1}.
\end{aligned}$$

Now, let $\lambda \in [0, \underline{\lambda}_{i+1})$. Then $\tau_i\lambda < \tau_i\underline{\lambda}_{i+1} < \underline{\lambda}_i$. Therefore, both the sub-matrix $\mathcal{K}_i - \tau_i\lambda\, \hat{\mathcal{L}}_i$ and the Schur complement $(-1)^i\, \Theta_{i+1}(\lambda)$ of the matrix $\mathcal{K}_{i+1} - \lambda\, \hat{\mathcal{L}}_{i+1}$ are non-singular, which implies that the matrix $\mathcal{K}_{i+1} - \lambda\, \hat{\mathcal{L}}_{i+1}$ itself is non-singular for $\lambda \in [0, \underline{\lambda}_{i+1})$. Analogously, one shows that $\mathcal{K}_{i+1} - \lambda\, \hat{\mathcal{L}}_{i+1}$ is non-singular for $\lambda \in (\overline{\lambda}_{i+1}, \infty)$. This implies statement (2) for the index $i+1$. $\square$

In particular we obtain for $i = n$:

Theorem 2. *Assume that the spectral estimates (16) are satisfied. If the parameters τ_i are chosen such that (17) is satisfied, then*

$$0 < \underline{\lambda}_n \leq \lambda_{min}(\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n) \leq \lambda_{max}(\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n) \leq \overline{\lambda}_n,$$

where $\underline{\lambda}_n$ denotes the smallest zero of $\underline{\theta}_n(\lambda)$, given by (19), and $\overline{\lambda}_n$ denotes the largest zero of $\overline{\theta}_n(\lambda)$, recursively given by (20).

Remark 1. *For the discussed special example of a block-tridiagonal symmetric and indefinite block matrix we have $\underline{\lambda}_n = \lambda_{min}(\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n)$ and $\overline{\lambda}_n = \lambda_{max}(\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n)$, see Lemma 4, which shows that the bounds are sharp.*

Remark 2. *A closed formula for $\overline{\lambda}_{i+1}$ is not available for larger i, while $\underline{\lambda}_{i+1}$ is explicitely given by*

$$\underline{\lambda}_{i+1} = \frac{1}{2\tau_i}\left(\overline{\sigma}_i(\underline{\sigma}_{i+1}+1) - \sqrt{\overline{\sigma}_i^2(\underline{\sigma}_{i+1}+1)^2 - 4\tau_i\overline{\sigma}_i\underline{\sigma}_{i+1}} \right)$$

for $i \geq 1$.

For the further analysis the following estimates of the eigenvalues are useful:

Corollary 1. *Assume that the spectral estimates (16) are satisfied. If the parameters τ_i are chosen such that (17) is satisfied, then*

$$\overline{\lambda}_{i+1} \leq \frac{\hat{\kappa}_i}{2}\left(1 + \overline{\sigma}_{i+1} + \sqrt{(1+\overline{\sigma}_{i+1})^2 - \frac{4\overline{\sigma}_{i+1}}{\hat{\kappa}_i}} \right) \leq \hat{\kappa}_i\, (1+\overline{\sigma}_{i+1})$$

and

$$\frac{1}{\underline{\lambda}_{i+1}} \;\leq\; \frac{1}{2}\left(1+\frac{1}{\underline{\sigma}_{i+1}}+\sqrt{\left(1+\frac{1}{\underline{\sigma}_{i+1}}\right)^2-\frac{4}{\hat{\kappa}_i\,\underline{\sigma}_{i+1}}}\right) \leq 1+\frac{1}{\underline{\sigma}_{i+1}}$$

with $\hat{\kappa}_i = \overline{\lambda}_i/\tau_i$.

Proof. For $\lambda \in (\overline{\lambda}_i, \infty)$ we have

$$\overline{\theta}_i'(\lambda) = -\frac{1}{\overline{\sigma}_i} + \overline{\psi}_i'(\lambda) \leq -\frac{1}{\overline{\sigma}_i} \leq -\frac{1}{\overline{\lambda}_i}$$

which implies

$$\overline{\theta}_i(\lambda) \leq -\frac{1}{\overline{\lambda}_i}(\lambda - \overline{\lambda}_i) = 1 - \frac{\lambda}{\overline{\lambda}_i}.$$

Hence

$$\overline{\theta}_{i+1}(\lambda) \leq -\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{1-\tau_i\lambda/\overline{\lambda}_i} = -\frac{\lambda}{\overline{\sigma}_{i+1}} + \frac{1-\lambda}{1-\lambda/\hat{\kappa}_i}.$$

Therefore, the largest zero $\overline{\lambda}_{i+1}$ of the function $\overline{\theta}_{i+1}(\lambda)$ on the left hand side is less or equal to the largest zero of the function on the right hand side, which gives the first inequality for $\overline{\lambda}_{i+1}$.

For $\underline{\lambda}_{i+1}$ we have

$$\frac{1}{\underline{\lambda}_{i+1}} = \frac{1}{2}\left(1+\frac{1}{\underline{\sigma}_{i+1}}+\sqrt{\left(1+\frac{1}{\underline{\sigma}_{i+1}}\right)^2-\frac{4\tau_i}{\overline{\sigma}_i\,\underline{\sigma}_{i+1}}}\right).$$

Then the first inequality for $\underline{\lambda}_{i+1}$ easily follows from $\overline{\sigma}_i \leq \overline{\lambda}_i$.

The remaining inequalities are trivial. ☐

Remark 3. *It is easy to see that, for the case $i = 1$, the first inequalities in Corollary 1 actually become identities.*

Remark 4. *A simple consequence of Corollary 1 is the estimate*

$$\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}} \leq \hat{\kappa}_i\,(1+\overline{\sigma}_{i+1})\left(1+\frac{1}{\underline{\sigma}_{i+1}}\right).$$

If

$$1 + \underline{\sigma}_{i+1}\overline{\sigma}_{i+1} \leq O(\overline{\sigma}_{i+1}),$$

then it immediately follows

$$\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}} \leq O\left(\hat{\kappa}_i\,\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}\right).$$

3.3 Optimal scaling

It is easy to see that $1/\underline{\lambda}_{i+1}$ as well as $\overline{\lambda}_{i+1}$ are monotonically decreasing in τ_i. So the parameter τ_i should be chosen close to its limiting value $\underline{\lambda}_i$ in order to reduce the relative condition number $\overline{\lambda}_{i+1}/\underline{\lambda}_{i+1}$. More precisely, we will assume that there is a constant $c > 1$ (and close to 1) such that

$$\tau_i < \underline{\lambda}_i \leq c\, \tau_i. \tag{25}$$

This implies

$$\hat{\kappa}_i \leq c\, \frac{\overline{\lambda}_i}{\underline{\lambda}_i}. \tag{26}$$

See [1] for a discussion concerning robustness if τ_i approaches $\underline{\lambda}_i$ in the case $i = 1$.

The other question is how to scale the preconditioners $\hat{S}_{i+1}$ in order to reduce the relative condition number $\overline{\lambda}_{i+1}/\underline{\lambda}_{i+1}$. Actually, in stead of discussing $\overline{\lambda}_{i+1}/\underline{\lambda}_{i+1}$ directly, we use the upper bounds from Corollary 1:

$$\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}} \leq f(1/\underline{\sigma}_{i+1}, \sqrt{\hat{\kappa}_i}) f(\overline{\sigma}_{i+1}, \sqrt{\hat{\kappa}_i})$$

with

$$f(x, a) = \frac{1}{2}\left(a(1 + x) + \sqrt{a^2(1 + x)^2 - 4x} \right).$$

Replacing the preconditioners $\hat{S}_{i+1}$ by $\rho\, \hat{S}_{i+1}$ changes the spectral bounds $\underline{\sigma}_{i+1}$ and $\overline{\sigma}_{i+1}$ to $\underline{\sigma}_{i+1}/\rho$ and $\overline{\sigma}_{i+1}/\rho$ and we obtain the upper bound

$$\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}} \leq f\left(\frac{\rho}{\underline{\sigma}_{i+1}}, \sqrt{\hat{\kappa}_i} \right) f\left(\frac{\overline{\sigma}_{i+1}}{\rho}, \sqrt{\hat{\kappa}_i} \right).$$

We will consider the scaling factor ρ as optimal if this upper bound is minimal.

Since the function $g(\xi, a) = \ln f(e^\xi, a)$ is convex with respect to $\xi \in \mathbb{R}$ for all $a \geq 1$, it follows that

$$f(x^2, a) f(y^2, a) \geq f(xy, a)^2 \quad \text{for all } x, y \geq 0.$$

In particular,

$$f\left(\frac{\rho}{\underline{\sigma}_{i+1}}, \sqrt{\hat{\kappa}_i} \right) f\left(\frac{\overline{\sigma}_{i+1}}{\rho}, \sqrt{\hat{\kappa}_i} \right) \geq f\left(\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}}, \sqrt{\hat{\kappa}_i} \right)^2,$$

where the equality sign holds if

$$\rho = \sqrt{\underline{\sigma}_{i+1}\, \overline{\sigma}_{i+1}}.$$

Therefore,

$$\min_{\rho > 0} f\left(\frac{\rho}{\underline{\sigma}_{i+1}}, \sqrt{\hat{\kappa}_i} \right) f\left(\frac{\overline{\sigma}_{i+1}}{\rho}, \sqrt{\hat{\kappa}_i} \right) = f\left(\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}}, \sqrt{\hat{\kappa}_i} \right)^2,$$

which characterizes the optimal scaling of the preconditioners $\hat{S}_{i+1}$. In other words, the preconditioners $\hat{S}_{i+1}$ are optimally scaled (i.e. $\rho = 1$) if

$$\underline{\sigma}_{i+1}\,\overline{\sigma}_{i+1} = 1. \tag{27}$$

In this case we obtain the following upper bound:

$$\sqrt{\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}}} \le f\left(\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}}, \sqrt{\hat{\kappa}_i}\right).$$

Using $f(x,a) \le ax + \sqrt{a^2-1}$ for $x \ge 1$ we can simplify this bound to

$$\sqrt{\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}}} \le \sqrt{\hat{\kappa}_i}\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}} + \sqrt{\hat{\kappa}_i - 1}. \tag{28}$$

Combining (26) and (28) we finally obtain:

Corollary 2. *Assume that the spectral estimates (16) with (27) are satisfied. If the parameters τ_i are chosen such that (25) is satisfied, then*

$$\sqrt{\frac{\overline{\lambda}_{i+1}}{\underline{\lambda}_{i+1}}} \le \sqrt{c}\sqrt{\frac{\overline{\lambda}_i}{\underline{\lambda}_i}}\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}} + \sqrt{c\,\frac{\overline{\lambda}_i}{\underline{\lambda}_i} - 1} \le \sqrt{c}\sqrt{\frac{\overline{\lambda}_i}{\underline{\lambda}_i}}\left(\sqrt{\frac{\overline{\sigma}_{i+1}}{\underline{\sigma}_{i+1}}} + 1\right).$$

Remark 5. *The second estimate sufficiently describes the correct asymptotic behavior for large condition numbers $\overline{\lambda}_i/\underline{\lambda}_i$, while the first estimate better reflects the correct asymptotic behavior for small condition numbers $\overline{\lambda}_i/\underline{\lambda}_i$.*

As a short summary of the analysis it can be claimed that the (upper bound of the) condition number $\overline{\lambda}_n/\underline{\lambda}_n$ of the preconditioned system matrix $\hat{\mathcal{L}}_n^{-1}\mathcal{K}_n$ is of the same order of magnitude as the product of the condition numbers $\overline{\sigma}_i/\underline{\sigma}_i$ of the preconditioned Schur complements $\hat{S}_i^{-1}S_i$:

$$\frac{\overline{\lambda}_n}{\underline{\lambda}_n} = O\left(\frac{\overline{\sigma}_1}{\underline{\sigma}_1}\frac{\overline{\sigma}_2}{\underline{\sigma}_2}\cdots\frac{\overline{\sigma}_n}{\underline{\sigma}_n}\right),$$

if the preconditioner is reasonably scaled, see Remark 4 and Corollary 2.

4 CG acceleration

For the inexact Uzawa method it was shown that the preconditioned matrix $\hat{\mathcal{L}}^{-1}\mathcal{K}$ is symmetric and positive definite with respect to a scalar product generated by some matrix $\mathcal{D}$. Therefore, the conjugate gradient method can be applied. The convergence rate of the preconditioned conjugate gradient method depends on the relative condition number $\lambda_{\max}(\hat{\mathcal{L}}^{-1}\mathcal{K})/\lambda_{\min}(\hat{\mathcal{L}}^{-1}\mathcal{K})$, for which estimates were derived in Section 3.

In its standard form the method would require the evaluation of the matrix-vector product $\mathcal{D}v$, which is not efficiently available in typical applications. Therefore, we consider the following variant, which is better suited:

Algorithm:

Compute $s^{(0)} = \hat{\mathcal{L}}^{-1} r^{(0)}$ and $t^{(0)} = \mathcal{D}\hat{\mathcal{L}}^{-1} r^{(0)}$ with $r^{(0)} = f - \mathcal{K}\, u^{(0)}$ for some initial guess $u^{(0)}$.

For $k = 0, 1, 2, \ldots$:

$$\begin{array}{rcl} \rho_k &=& (t^{(k)}, s^{(k)})_{\ell_2} \\ p^{(k)} &=& \begin{cases} s^{(0)} & \text{with} \quad k = 0, \\ s^{(k)} + \beta_k\, p^{(k-1)} \text{ with } \beta_k = \rho_k/\rho_{k-1} & \text{with} \quad k \geq 1, \end{cases} \\ v^{(k)} &=& \hat{\mathcal{L}}^{-1}\, q^{(k)} \text{ and } w^{(k)} \;=\; \mathcal{D}\hat{\mathcal{L}}^{-1}\, q^{(k)} \text{ with } q^{(k)} \;=\; \mathcal{K}\, p^{(k)}, \\ u^{(k+1)} &=& u^{(k)} + \alpha_k\, p^{(k)} \text{ with } \alpha_k = \rho_k/(w^{(k)}, p^{(k)})_{\ell_2}, \\ s^{(k+1)} &=& s^{(k)} - \alpha_k\, v^{(k)}, \\ t^{(k+1)} &=& t^{(k)} - \alpha_k\, w^{(k)}. \end{array}$$

Each step of this algorithm requires the simultaneous evaluation of

$$v = \hat{\mathcal{L}}^{-1} q \quad \text{and} \quad w = \mathcal{D}\hat{\mathcal{L}}^{-1} q,$$

where the evaluation of q itself requires a matrix-vector product $\mathcal{K}\, p$.

With

$$\mathcal{D}_{i+1} = \begin{pmatrix} \mathcal{D}_i[\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i - \tau_i\, I] & 0 \\ 0 & \hat{S}_{i+1} \end{pmatrix} \quad \text{and} \quad \tau_i\, \hat{\mathcal{L}}_{i+1}^{-1} = \begin{pmatrix} \hat{\mathcal{L}}_i^{-1} & 0 \\ (-1)^{i-1}\, \hat{S}_{i+1}^{-1} B_i \mathsf{E}_i \hat{\mathcal{L}}_i^{-1} & \tau_i\, (-1)^i\, \hat{S}_{i+1}^{-1} \end{pmatrix}$$

it follows that

$$\begin{array}{rcl} \tau_i\, \mathcal{D}_{i+1}\hat{\mathcal{L}}_{i+1}^{-1} &=& \begin{pmatrix} \mathcal{D}_i[\hat{\mathcal{L}}_i^{-1}\mathcal{K}_i - \tau_i\, I] & 0 \\ 0 & (-1)^i\, \hat{S}_{i+1} \end{pmatrix} \begin{pmatrix} \hat{\mathcal{L}}_i^{-1} & 0 \\ -\hat{S}_{i+1}^{-1} B_i \mathsf{E}_i \hat{\mathcal{L}}_i^{-1} & \tau_i\, \hat{S}_{i+1}^{-1} \end{pmatrix} \\ &=& \begin{pmatrix} \mathcal{D}_i \hat{\mathcal{L}}_i^{-1}[\mathcal{K}_i \hat{\mathcal{L}}_i^{-1} - \tau_i\, I] & 0 \\ -(-1)^i\, B_i \mathsf{E}_i \hat{\mathcal{L}}_i^{-1} & \tau_i\, (-1)^i\, I \end{pmatrix} \\ &=& \begin{pmatrix} [\tau_i^{-1}\, \mathcal{K}_i \hat{\mathcal{L}}_i^{-T} - I]\tau_i\, \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} & 0 \\ -(-1)^i\, B_i \mathsf{E}_i \hat{\mathcal{L}}_n^{-1} & \tau_i\, (-1)^i\, I \end{pmatrix} \end{array}$$

with

$$\mathcal{D}_1 = \hat{A}_1, \quad \hat{\mathcal{L}}_1 = \hat{A}_1, \quad \mathcal{K}_1 = A_1.$$

Therefore, the simultaneous evaluation of

$$v = \hat{\mathcal{L}}_{i+1}^{-1} q \quad \text{and} \quad w = \mathcal{D}_{i+1}\hat{\mathcal{L}}_{i+1}^{-1} q$$

leads to the following recursive procedure:

$$\begin{aligned}
s &= \hat{\mathcal{L}}_i^{-1} q' \quad \text{and} \quad t \;=\; \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} q', \\
r &= \tau_i^{-1}\, \hat{\mathcal{L}}_i^{-T} t, \\
v' &= \tau_i^{-1}\, s, \\
w' &= \mathcal{K}_i\, r - t, \\
w_{i+1} &= (-1)^i\, (q_{i+1} - B_i v_i), \\
v_{i+1} &= (-1)^i\, \hat{S}_{i+1}^{-1} w_{i+1}.
\end{aligned}$$

Finally, the relaxation parameters have to be chosen such that

$$\tau_i\, \mathcal{D}_i < \mathcal{D}_i \hat{\mathcal{L}}_i^{-1} \mathcal{K}_i \quad \text{for all } i = 1, \ldots, n.$$

This can be done by the gradient method for $\hat{\mathcal{L}}_i^{-1} \mathcal{K}_i$ preconditioned by $\mathcal{D}_i$, see. e.g., [7]. So the same operations as before are required.

References

[1] James H. Bramble and Joseph E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comput.*, 51(183):387–388, 1988.

[2] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods.* New York etc.: Springer-Verlag, 1991.

[3] Gabriel N. Gatica and Norbert Heuer. A dual-dual formulation for the coupling of mixed FEM and BEM in hyperelasticity. *SIAM J. Numer. Anal.*, 38(2):380–400, 2000.

[4] Gabriel N. Gatica and Norbert Heuer. Conjugate gradient method for dual-dual mixed formulations. *Math. Comput.*, 71(240):1455–1472, 2002.

[5] Ulrich Langer, Günther Of, Olaf Steinbach, and Walter Zulehner. Inexact Data–Sparse Boundary Element Tearing and Interconnecting Methods. SFB-Report, SFB F013, Johannes Kepler University Linz, Austria, 2005.

[6] Ulrich Langer and Olaf Steinbach. Boundary element tearing and interconnecting methods. *Computing*, 71(3):205–228, 2003.

[7] Aleksandr A. Samarskij and Evgenii S. Nikolaev. *Numerical Methods for Grid Equations. Volume II: Iterative Methods.* Basel etc.: Birkhäuser Verlag, 1989.

[8] Kyuichiro Washizu. *Variational Methods in Elasticity and Plasticity.* Oxford etc.: Pergamon Press, 1968.

[9] Walter Zulehner. Analysis of iterative methods for saddle point problems: A unified approach. *Math. Comput.*, 71(238):479–505, 2002.