Preconditioners for Some Matrices of Two-by-Two Block Form, with Applications, I

Owe Axelsson

Abstract Matrices of two-by-two block form with matrix blocks of equal order arise in various important applications, such as when solving complex-valued systems in real arithmetics, in linearized forms of the Cahn–Hilliard diffusive phase-field differential equation model and in constrained partial differential equations with distributed control. It is shown how an efficient preconditioner can be constructed which, under certain conditions, has a resulting spectral condition number of about 2. The preconditioner avoids the use of Schur complement matrices and needs only solutions with matrices that are linear combinations of the matrices appearing in each block row of the given matrix and for which often efficient preconditioners are already available.

Keywords Two-by-two block-structured matrices • Preconditioning • Complex-valued system • Cahn–Hilliard phase-field model • Optimal control • Distributed control

Mathematics Subject Classification (2010): 65F10, 65F35, 76T10, 49J20

1 Introduction

To motivate the study, we give first some examples of two-by-two block matrices where blocks of equal order, i.e. square blocks, appear. Although the matrices are of special type, as we shall see there are several important applications where

O. Axelsson (✉)
IT4 Innovations Department, Institute of Geonics AS CR, Ostrava, Czech Republic

King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: owe.axelsson@it.uu.se

O.P. Iliev et al. (eds.), *Numerical Solution of Partial Differential Equations: Theory, Algorithms, and Their Applications*, Springer Proceedings in Mathematics & Statistics 45, DOI 10.1007/978-1-4614-7172-1_3, © Springer Science+Business Media New York 2013

they arise. One such example is related to the solution of systems with complex-valued matrices. Complex-valued systems arise, for instance, when solving certain partial differential equations (PDE) appearing in electromagnetics and wave propagation; see [1]. Complex arithmetics requires more memory storage and may require more involved implementation. Therefore it is desirable to rewrite a complex-valued matrix system in a form that can be handled using real arithmetics.

Using straightforward derivations, for a complex-valued matrix $A+iB$, where A and B are real and A is nonsingular, it holds

$$(A+iB)(I-iA^{-1}B)=A+BA^{-1}B$$

so

$$(A+iB)^{-1}=(I-iA^{-1}B)(A+BA^{-1}B)^{-1}.$$

It follows that a complex-valued system

$$(A+iB)(\mathbf{x}+i\mathbf{y})=\mathbf{f}+i\mathbf{g},$$

where $\mathbf{x},\mathbf{y},\mathbf{f},\mathbf{g}$ are real vectors, can be solved by solving two real-valued systems with matrix $A+BA^{-1}B$ with right-hand sides $\mathbf{f}$ and $\mathbf{g}$ respectively, in addition to a matrix vector multiplication with B and two solutions of systems with the matrix A.

In many applications, $A+BA^{-1}B$ can be ill conditioned and costly to construct and solve systems with, in particular as it involves solutions with inner systems with the matrix A. Therefore, this approach is normally less efficient.

As has been shown in [2] (see also [1,3]), it may be better to rewrite the equation in real-valued form

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}=\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{1}$$

A matrix factorization shows that

$$\begin{bmatrix} A & 0 \\ B & A+BA^{-1}B \end{bmatrix}\begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}=\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix},$$

where I is the identity matrix. It is seen that here it suffices with one solution with matrix $A+BA^{-1}B$, in addition to two solves with A. However, we will show that the form (1) allows for an alternative solution method based on iteration and the construction of an efficient preconditioner that involves only two systems with matrices that are linear combinations of matrices A and B and that a corresponding iterative solution of (1) can substantially lower the computational expense. We shall show that such a preconditioner can be constructed for a matrix in the more general form

$$\mathcal{A}=\begin{bmatrix} A & -B^T \\ \beta^2 B & \alpha^2 A \end{bmatrix}, \tag{2}$$

where α,β are positive numbers. By the introduction of a new, scaled second variable vector $\mathbf{y} := \frac{1}{\alpha^2}\mathbf{y}$, the systems transform into the alternative form

$$\mathcal{A} = \begin{bmatrix} A & -aB^T \\ bB & A \end{bmatrix}, \tag{3}$$

where $a = \frac{1}{\alpha^2}, b = \beta^2$. This form arises in the two-phase version of the Cahn–Hilliard equation used to track interfaces between two fluids with different densities using a stationary grid; see [4, 5].

As we shall see in the sequel, a matrix in the form (2), with $\beta = 1$ arises also in optimization problems for PDE, with a distributed control function, that is, a control function defined in the whole domain of definition of the PDE. For an introduction to such problems, see [6, 7].

Problems of this kind appear in various applications in engineering and geosciences but also in medicine [8] and finance [9]. As a preamble to this topic, we recall that the standard form of a constrained optimization problem with a quadratic function takes the form

$$\min_{\mathbf{u}} \left\{ \frac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{f} \right\}$$

subject to the constraint $B\mathbf{u} = \mathbf{g}$. Here, $\mathbf{u}, \mathbf{f} \in \Re^n, \mathbf{g} \in \Re^m$, and A is a symmetric and positive definite (spd) matrix of order $n \times n$ and B has order $m \times n$, $m \leq n$. For the existence of a solution, if $m = n$ we must assume that $\dim \Re(B) < m$, where $\Re(B)$ denotes the range of B. The corresponding Lagrangian function with multiplier $\mathbf{p}$ and regularization term $-\alpha\mathbf{p}^T C\mathbf{p}$, where α is a small positive number and C is spd, takes the form

$$\mathfrak{L}(\mathbf{u}, \mathbf{p}) = \frac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{f} + \mathbf{p}^T(B\mathbf{u} - \mathbf{g}) - \frac{1}{2}\alpha\mathbf{p}^T C\mathbf{p}.$$

By the addition of the regularization term, the Lagrange multiplier vector $\mathbf{p}$ becomes unique.

The necessary first-order conditions for an optimal, saddle point solution lead to

$$\begin{bmatrix} A & B^T \\ B & -\alpha C \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{4}$$

Here, we can extend the matrix B with $n - m$ zero rows and the vector $\mathbf{g}$ with $n - m$ zero components, to make B of the same order as A. Similarly, C is extended. It is possible to let $C = A$. (Then the $n - m$ correspondingly added components of $\mathbf{p}$ become zero.) As we shall see, in optimal control problems with a distributed control, we get such a form with no need to add zero rows to B.

If we change the sign of $\mathbf{p}$, the corresponding matrix takes the form $\begin{bmatrix} A & -B^T \\ B & A \end{bmatrix}$, i.e. the same form as in (1). The matrix in (4) is indefinite. It can be preconditioned with a block-diagonal matrix, but it leads to eigenvalues on both sides of the origin, which slows down the convergence of the corresponding iterative acceleration method, typically of a conjugate gradient type, such as MINRES in [10]. In this paper we show that much faster convergence can be achieved if instead we precondition $\mathcal{A}$ with a matrix that is a particular perturbation of it, since this leads to positive eigenvalues and no Schur complements need to be handled. We consider then preconditioning of matrices of the form (2) or (3). Thereby we assume that A is symmetric and positive definite, or at least positive semidefinite and $ker(A) \cap ker(B) = \{\emptyset\}$, which will be shown to guarantee that $\mathcal{A}$ is nonsingular.

In Sect. 2 we present a preconditioner to this matrix, but given in the still more general form

$$\mathcal{A} = \begin{bmatrix} A & -aB_2 \\ bB_1 & A \end{bmatrix}, \tag{5}$$

where it is assumed that $H_i = A + \sqrt{ab}B_i$, $i = 1,2$ are regular. It involves only solutions with the matrices H_1 and H_2. Hence, no Schur complements needed to be handled arise here.

In Sect. 3 we perform an eigenvalue analysis of the preconditioning method. This result extends the applicability of the previous results, e.g. in [2] and [4]. Furthermore, the present proofs are sharper and more condensed.

In Sect. 4 we show that certain constrained optimal control problems for PDE with a distributed control can be written in the above two-by-two block form. The results in that section extend related presentations in [7].

Further development of the methods and numerical tests will be devoted to part II of this paper.

The notation $A \leq B$ for symmetric matrices A, B means that $A - B$ is positive semidefinite.

2 The Preconditioner and Its Implementation

Given a matrix in the form (2), we consider first a preconditioner to $\mathcal{A}$ in the form

$$\mathcal{B} = \begin{bmatrix} A & 0 \\ \beta^2 B & \tilde{\alpha}A + \beta B \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & A^{-1} \end{bmatrix} \begin{bmatrix} A & -B^T \\ 0 & \tilde{\alpha}A + \beta B^T \end{bmatrix} \tag{6}$$

where $\tilde{\alpha}$ is a positive preconditioning method parameter to be chosen. A computation shows that

$$\mathcal{B} = \mathcal{A} + \begin{bmatrix} 0 & 0 \\ 0 & (\tilde{\alpha}^2 - \alpha^2)A + \tilde{\alpha}\beta(B + B^T) \end{bmatrix}$$

We show now that an action of its inverse requires little computational work.

Proposition 1. *An action of the inverse of the form of the matrix $\mathcal{B}$ in (6) requires one solution of each of the matrices $A, \tilde{\alpha}A + \beta B$ and $A, \tilde{\alpha}A + \beta B^T$, in this order.*

Proof. To solve a system

$$\mathcal{B}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix},$$

solve first

$$\begin{bmatrix}A & 0\\ \beta^2 B & \tilde{\alpha}A+\beta B\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{x}}\\ \tilde{\mathbf{y}}\end{bmatrix}=\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix},$$

which requires a solution with A and $\tilde{\alpha}A+\beta B$. Solve then

$$\begin{bmatrix}A & -B^T\\ 0 & \tilde{\alpha}A+\beta B^T\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}A\tilde{\mathbf{x}}\\ A\tilde{\mathbf{y}}\end{bmatrix}=\begin{bmatrix}\mathbf{f}\\ A\tilde{\mathbf{y}}\end{bmatrix}$$

by solving

$$(\tilde{\alpha}A+\beta B)\mathbf{y}=A\tilde{\mathbf{y}},$$

$$\mathbf{z}:=A^{-1}B^T\mathbf{y}\quad \text{a}s$$

$$\mathbf{z}=\frac{1}{\beta}(\tilde{\mathbf{y}}-\tilde{\alpha}\mathbf{y})$$

to finally obtain

$$\mathbf{x}=\tilde{\mathbf{x}}+\mathbf{z}.$$ ■

In applications, often A is a mass matrix and B is a stiffness matrix. When A depends on heterogeneous material coefficients, the matrices $\tilde{\alpha}A+\beta B$ and $\tilde{\alpha}A+\beta B^T$ can be better conditioned than A. We show now that by applying the explicit expression for $\mathcal{B}^{-1}$, the separate solution with A in (6) can be avoided.

We find it convenient to show this first for preconditioners $\mathcal{B}$ applied to the matrix $\mathcal{A}$ in the form (3). Here,

$$\mathcal{B}=\begin{bmatrix}A & -aB^T\\ bB & A+\sqrt{ab}(B+B^T)\end{bmatrix}. \tag{7}$$

For its inverse the following proposition holds. For its proof, we assume first that A is spd.

Proposition 2. *Let A be spd. Then*

$$\begin{aligned}\mathcal{B}^{-1}&=\begin{bmatrix}A & -aB^T\\ bB & A+\sqrt{ab}(B+B^T)\end{bmatrix}^{-1}\\ &=\begin{bmatrix}H^{-1}+H^{-T}-H^{-T}AH^{-1} & \sqrt{\frac{a}{b}}(I-H^{-T}A)H^{-1}\\ -\sqrt{\frac{b}{a}}H^{-T}(I-AH^{-1}) & H^{-T}AH^{-1}\end{bmatrix},\end{aligned}$$

where $H=A+\sqrt{ab}B$, which is assumed to be nonsingular.

Proof. For the derivation of the expression for the inverse we use the form of the inverse of a general matrix in two-by-two block form. (However, clearly we can verify the correctness of the expression directly by computation of the matrix times its inverse. An alternative derivation can be based on the Schur–Banachiewicz form of the inverse.) Assume that A_{ii}, $i = 1,2$ are nonsingular. Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} S_1^{-1} & -A_{11}^{-1}A_{12}S_2^{-1} \\ -S_2^{-1}A_{21}A_{11}^{-1} & S_2^{-1} \end{bmatrix}.$$

Here, the Schur complements $S_i, i = 1,2$ equal

$$S_i = A_{ii} - A_{ij}A_{jj}^{-1}A_{ji},\ i,j = 1,2,\ i \neq j.$$

Further, $S_2^{-1}A_{21}A_{11}^{-1} = A_{22}^{-1}A_{21}S_1^{-1}$.

For the given matrix it holds

$$\begin{aligned} S_2 &= A + \sqrt{ab}(B + B^T) + abBA^{-1}B^T \\ &= (A + \sqrt{ab}B)A^{-1}(A + \sqrt{ab}B^T). \end{aligned}$$

Further,

$$\begin{aligned} &- A_{11}^{-1}A_{12}S_2^{-1} = aA^{-1}B^T(A + \sqrt{ab}B^T)^{-1}A(A + \sqrt{ab}B)^{-1} \\ &= \sqrt{\frac{a}{b}}A^{-1}((\sqrt{ab}B^T + A) - A)(A + \sqrt{ab}B^T)^{-1}A(A + \sqrt{ab}B)^{-1} \\ &= \sqrt{\frac{a}{b}}(H^{-1} - H^TAH^{-1}) = \sqrt{\frac{a}{b}}(I - H^{-T}A)H^{-1}. \end{aligned}$$

Similarly,

$$-A_{22}^{-1}A_{21}S_1^{-1} = -\sqrt{\frac{b}{a}}H^{-T}(I - AH^{-1}).$$

Finally, since the pivot block in the inverse matrix equals the inverse of the Schur complement, the corresponding equality holds for the pivot block in the matrix itself, that is,

$$A_{11} = (S_1^{-1} - A_{11}^{-1}A_{12}S_2^{-1}A_{21}A_{11}^{-1})^{-1}. \tag{8}$$

Therefore,

$$\begin{aligned} S_1^{-1} &= A_{11}^{-1} + A_{11}^{-1}A_{12}S_2^{-1}A_{21}A_{11}^{-1} \\ &= A^{-1}[A - (I - AH^{-T})A(I - H^{-1}A)]A^{-1} \\ &= H^{-1} + H^{-T} - H^{-T}AH^{-1} \end{aligned}$$

■

Remark 1. Incidently, relation (8) can be seen as a proof of the familiar Sherman–Morrison–Woodbury formula.

We show now that Proposition 2 implies that an action of the matrix $\mathcal{B}^{-1}$ needs only a solution with each of the matrices H and H^T. This result has appeared previously in [4], but the present proof is more condensed and more generally applicable. We will then show it for a matrix in the general form (5).

Guided by the result in Proposition 2, we give now the expression for the inverse of the preconditioner to a matrix in the form (5).

Proposition 3. *Let*

$$\mathcal{B} = \begin{bmatrix} A & -aB_2 \\ bB_1 & A+\sqrt{ab}(B_1+B_2) \end{bmatrix}$$

then

$$\mathcal{B}^{-1} = \begin{bmatrix} H_1^{-1}+H_2^{-1}-H_2^{-1}AH_1^{-1} & \sqrt{\frac{a}{b}}(I-H_2^{-1}A)H_1^{-1} \\ -\sqrt{\frac{b}{a}}H_2^{-1}(I-AH_1^{-1}) & H_2^{-1}AH_1^{-1} \end{bmatrix}$$

where $H_i = A+\sqrt{ab}B_i, i=1,2,$ which are assumed to be nonsingular.

Proof. We show first that $\mathcal{B}$ is nonsingular. If

$$\mathcal{B}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{9}$$

then $A\mathbf{x} = aB_2\mathbf{y}$ and

$$A\mathbf{y}+bB_1\mathbf{x}+\sqrt{ab}(B_1+B_2)\mathbf{y} = 0.$$

Then

$$(A+\sqrt{ab}B_1)\mathbf{y}+\sqrt{\frac{b}{a}}(\sqrt{ab}B_1\mathbf{x}+aB_2\mathbf{y}) = 0$$

or

$$(A+\sqrt{ab}B_1)(\sqrt{\frac{b}{a}}\mathbf{x}+\mathbf{y}) = 0.$$

Hence, $\mathbf{x} = -\sqrt{\frac{a}{b}}\mathbf{y}$, so $\sqrt{\frac{a}{b}}(A+\sqrt{ab}B_2)\mathbf{y} = 0$ or $\mathbf{y} = 0$, so (9) has only the trivial solution. The expression for $\mathcal{B}^{-1}$ follows by direct inspection. ■

Proposition 4. *Assume that $A+\sqrt{ab}B_i, i=1,2$ are nonsingular. Then $\mathcal{B}$ is nonsingular and a linear system with the preconditioner $\mathcal{B}$,*

$$\begin{bmatrix} A & -aB_2 \\ bB_1 & A+\sqrt{ab}(B_1+B_2) \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix}$$

can be solved with only one solution with $A+\sqrt{ab}B_1$ and one with $A+\sqrt{ab}B_2$.

Proof. It follows form Proposition 3 that an action of the inverse of $\mathcal{B}$ can be written in the form

$$\begin{bmatrix} A & -aB_2 \\ bB_1 & A+\sqrt{ab}(B_1+B_2) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} =$$

$$= \begin{bmatrix} H_1^{-1}\mathbf{f}_1 + H_2^{-1}\mathbf{f}_1 - H_2^{-1}AH_1^{-1}\mathbf{f}_1 + \sqrt{\tfrac{a}{b}}(I-H_2^{-1}A)H_1^{-1}\mathbf{f}_2 \\ -\sqrt{\tfrac{b}{a}}H_2^{-1}(I-AH_1^{-1})\mathbf{f}_1 + H_2^{-1}AH_1^{-1}\mathbf{f}_2 \end{bmatrix}$$

$$= \begin{bmatrix} H_2^{-1}\mathbf{f}_1 + \mathbf{g} - H_2^{-1}A\mathbf{g} \\ -\sqrt{\tfrac{b}{a}}H_2^{-1}\mathbf{f}_1 + \sqrt{\tfrac{b}{a}}H_2^{-1}A\mathbf{g} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{g} + H_2^{-1}(\mathbf{f}_1 - A\mathbf{g}) \\ -\sqrt{\tfrac{b}{a}}H_2^{-1}(\mathbf{f}_1 - A\mathbf{g}) \end{bmatrix} = \begin{bmatrix} \mathbf{g}+\mathbf{h} \\ -\sqrt{\tfrac{b}{a}}\mathbf{h} \end{bmatrix}$$

where

$$\mathbf{g} = H_1^{-1}(\mathbf{f}_1 + \sqrt{\frac{a}{b}}\mathbf{f}_2),\, \mathbf{h} = H_2^{-1}(\mathbf{f}_1 - A\mathbf{g}).$$

The computation can take place in the following order:

(i) Solve $H_1\mathbf{g} = \mathbf{f}_1 + \sqrt{\frac{a}{b}}\mathbf{f}_2$.
(ii) Compute $A\mathbf{g}$ and $\mathbf{f}_1 - A\mathbf{g}$.
(iii) Solve $H_2\mathbf{h} = \mathbf{f}_1 - A\mathbf{g}$.
(iv) Compute $\mathbf{x} = \mathbf{g}+\mathbf{h}$ and $\mathbf{y} = -\sqrt{\frac{b}{a}}\mathbf{h}$. ■

Remark 2. In some applications $H_1 = A + \sqrt{ab}B_1$, and $H_2 = A + \sqrt{ab}B_2$ may be better conditioned than A itself. Even if it is not, often software for these combinations exists.

3 Condition Number Bounds

To derive condition number bounds for the preconditioned matrix $\mathcal{B}^{-1}\mathcal{A}$, we consider two cases:

(i) $B_1 = B, B_2 = B^T, A$ is symmetric, A and $B+B^T$ are positive semidefinite, and

$$ker(A) \cap ker(B_i) = \{\emptyset\}, i = 1,2$$

(ii) A is symmetric and positive definite and certain conditions, to be specified later, hold for B_1 and B_2.

3.1 A Is Symmetric and Positive Semidefinite

Assume that conditions (i) hold. Then it follows that $A+\sqrt{ab}B$ and $A+\sqrt{ab}B^T$, and hence also $\mathcal{B}$, are nonsingular. We show first that then $\mathcal{A}$ is also nonsingular.

Proposition 5. *Let condition (i) hold. Then $\mathcal{A}$ is nonsingular.*

Proof. If

$$\begin{bmatrix} A & -aB^T \\ bB & A \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

then

$$\begin{aligned} &\mathbf{x}^*A\mathbf{x} - a\mathbf{x}^*B^T\mathbf{y} = \mathbf{0}, \\ &b\mathbf{y}^*B\mathbf{x} + \mathbf{y}^*A\mathbf{y} = \mathbf{0} \end{aligned}$$

so $\frac{1}{a}\mathbf{x}^*A\mathbf{x} + \frac{1}{b}\mathbf{y}^*A\mathbf{y} = 0$, where $\mathbf{x}^*$, $\mathbf{y}^*$ denote the complex conjugate vector.

Since A is positive semidefinite, it follows that $\mathbf{x},\mathbf{y} \in kerA$. But then $B^T\mathbf{y} = \mathbf{0}$ and $B\mathbf{x} = \mathbf{0}$, implying that $\mathbf{x},\mathbf{y} \in kerB$, so $\mathcal{A}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ has only the trivial solution. ■

Proposition 6. *Let $\mathcal{A} = \begin{bmatrix} A & aB^T \\ -bB & A \end{bmatrix}$, where a,b are nonzero and have the same sign and let $\mathcal{B} = \begin{bmatrix} A & aB^T \\ -bB & A+\sqrt{ab}(B+B^T) \end{bmatrix}$. If conditions (i) hold, then the eigenvalues of $\mathcal{B}^{-1}\mathcal{A}$, are contained in the interval $[\frac{1}{2},1]$.*

Proof. For the generalized eigenvalue problem

$$\lambda\mathcal{B}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathcal{A}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

it follows from Proposition 5 that $\lambda \neq 0$. It holds

$$\left(\frac{1}{\lambda} - 1\right)\mathcal{A}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{ab}\,(B+B^T)\mathbf{y} \end{bmatrix}$$

Here, $\lambda = 1$ if $\mathbf{y} \in ker(B+B^T)$. If $\lambda \neq 1$, then

$$A\mathbf{x} = -aB^T\mathbf{y}$$

and

$$\left(\frac{1}{\lambda} - 1\right)(\mathbf{y}^*A\mathbf{y} - b\mathbf{y}^*B\mathbf{x}) = \sqrt{ab}\mathbf{y}^*(B+B^T)\mathbf{y}$$

or

$$\left(\frac{1}{\lambda} - 1\right)(\mathbf{y}^*A\mathbf{y} + \frac{b}{a}\mathbf{x}^*A\mathbf{x}) = \sqrt{ab}\mathbf{y}^*(B+B^T)\mathbf{y}.$$

Since both A and $B+B^T$ are positive semidefinite, it follows that $\lambda \leq 1$.

Further it holds,

$$-\mathbf{y}^* A\mathbf{x} = a\mathbf{y}^* B^T\mathbf{y}$$

so

$$\left(\frac{1}{\lambda}-1\right)(a\mathbf{y}^* B^T\mathbf{y} + b\mathbf{x}^* B\mathbf{x}) = -\sqrt{ab}\,\mathbf{x}^*(B+B^T)\mathbf{y}$$

or

$$\left(\frac{1}{\lambda}-1\right)(a\mathbf{y}^*(B+B^T)\mathbf{y} + b\mathbf{x}^*(B+B^T)\mathbf{x}) = -2\sqrt{ab}\,\mathbf{x}^*(B+B^T)\mathbf{y}.$$

Since $B+B^T$ is positive semidefinite, $\mid \mathbf{x} \mid + \mid \mathbf{y} \mid \neq 0$, and a and b have the same sign, it follows that

$$\frac{1}{\lambda}-1 \leq \frac{2\sqrt{ab} \mid \mathbf{x}^*(B+B^T)\mathbf{y} \mid}{\mid a \mid \mathbf{y}^*(B+B^T)\mathbf{y}+ \mid \mathbf{b} \mid \mathbf{x}^*(B+B^T)\mathbf{x}} \leq 1,$$

that is, $\lambda \geq \frac{1}{2}$. ■

3.2 A Is Symmetric and Positive Definite

Assume now that A is symmetric and positive definite. Let $\mathcal{A}$ be defined in (5) and let $\tilde{B}_i = \sqrt{ab}A^{-1/2}B_iA^{-1/2}$, $i=1,2$. Assume that the eigenvalues of the generalized eigenvalue problem,

$$\mu(I+\tilde{B}_1\tilde{B}_2)\mathbf{z} = (\tilde{B}_1+\tilde{B}_2)\mathbf{z},\ \mathbf{z} \neq 0 \tag{10}$$

are real and $\mu_{\max} \geq \mu \geq \mu_{\min} > -1$.

Proposition 7. *Let $\mathcal{A}$ be defined in (5), let $\tilde{B}_i = \sqrt{ab}A^{-1/2}B_iA^{-1/2}$, $i=1,2$, and assume that $\tilde{B}_1+\tilde{B}_2$ is spd and (10) holds. Then the eigenvalues of $\mathcal{B}^{-1}\mathcal{A}$ are contained in the interval $\left[\frac{1}{1+\mu_{\max}}, \frac{1}{1+\mu_{\min}}\right]$.*

Proof. $\lambda\mathcal{B}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \mathcal{A}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ implies

$$(\lambda-1)\begin{bmatrix}A\mathbf{x} & -aB_2\mathbf{y}\\ A\mathbf{y}+bB_1\mathbf{x}+\sqrt{ab}(B_1+B_2)\mathbf{y}\end{bmatrix} = \begin{bmatrix}0\\ \sqrt{ab}(B_1+B_2)\mathbf{y}\end{bmatrix}.$$

Hence, a block-diagonal transformation with $\begin{bmatrix}A^{-1/2} & 0\\ 0 & A^{-1/2}\end{bmatrix}$ shows that

$$(\lambda-1)\begin{bmatrix}\tilde{\mathbf{x}}-\sqrt{\frac{a}{b}}\tilde{B}_2\tilde{\mathbf{y}}\\ \tilde{\mathbf{y}}+\sqrt{\frac{b}{a}}\tilde{B}_1\tilde{\mathbf{x}}+(\tilde{B}_1+\tilde{B}_2)\tilde{\mathbf{y}}\end{bmatrix}=\begin{bmatrix}0\\ -(\tilde{B}_1+\tilde{B}_2)\tilde{\mathbf{y}}\end{bmatrix},$$

where $\tilde{\mathbf{x}}=A^{1/2}\mathbf{x}, \tilde{\mathbf{y}}=A^{1/2}\mathbf{y}$.

If $\lambda\neq 1$, then

$$(1-\lambda)\left[I+\tilde{B}_1\tilde{B}_2\right]\tilde{\mathbf{y}}=\lambda(\tilde{B}_1+\tilde{B}_2)\tilde{\mathbf{y}},$$

Hence, by (10),

$$\frac{1}{\lambda}-1=\mu \quad \text{or}\lambda=\frac{1}{1+\mu},$$

which implies the stated eigenvalue bounds. ■

Corollary 1. *If $B_1=B, B_2=B^T$, and $I+\tilde{B}$ is nonsingular, then*

$$\frac{1}{2}\leq\lambda\leq\frac{1}{1+\mu_{\min}},$$

where $\mu_{\min}>-1$. If the symmetric part of B is positive semidefinite, then

$$\frac{1}{2}\leq\lambda\leq 1.$$

Proof. Since

$$(I-\tilde{B})(I-\tilde{B}^T)\geq 0$$

it follows that

$$I+\tilde{B}\tilde{B}^T\geq\tilde{B}+\tilde{B}^T$$

which implies $\mu\leq 1$ in (10). Similarly,

$$(I+\tilde{B})(I+\tilde{B}^T)\geq 0,$$

that is,

$$I+\tilde{B}\tilde{B}^T\geq-(\tilde{B}+\tilde{B}^T)$$

implies $\mu_{\min}\geq-1$. But $\mu_{\min}>-1$ since $I+\tilde{B}$, and hence $I+\tilde{B}^T$, are nonsingular. If $B+B^T\geq 0$, then $\mu_{\min}=0$. ■

Corollary 2. *If $B_1=B, B_2=B-\delta/\sqrt{ab}A$ for some real number δ, where B is spd and $2B>\delta/\sqrt{ab}A$, that is, $B_1+B_2=2B-\delta/\sqrt{ab}A$ is spd, then*

$$\frac{\sqrt{4-\delta^2}}{2+\sqrt{4-\delta^2}}\leq\lambda\leq 1.$$

Proof. Here, (10) takes the form

$$\mu(I+\tilde{B}^2-\delta\tilde{B})\tilde{\mathbf{z}}=(2\tilde{B}-\delta I)\tilde{\mathbf{z}},$$

where $\tilde{B}=\sqrt{ab}A^{-1/2}BA^{-1/2}$. Let β be an eigenvalue of $\tilde{B}$.

Then

$$\mu=\frac{2\beta-\delta}{1+\beta^2-\beta\delta}.$$

Since $\delta<2\beta$, it follows that $\mu>0$, that is, $\lambda\leq 1$. Further, a computation shows that μ takes its largest value when

$$(2\beta-\delta)^2=2(1+\beta^2-\beta\delta)$$

or

$$(2\beta-\delta)^2=2+\frac{1}{2}(2\beta-\delta)^2-\frac{\delta^2}{2},$$

that is when

$$2\beta-\delta=\sqrt{4-\delta^2}.$$

Then $\mu=2/\sqrt{4-\delta^2}$ and the statement follows from $\lambda=1/(1+\mu)$. ■

Remark 3. Matrices in the form as given in Corollary 2 appear in phase-field models; see, e.g. [4, 5]. For complex-valued systems, normally the coefficients are $a=b=1$. In other applications, such as those in Sects. 4.1 and 4.2, a form such as in Proposition 1 arises. One can readily transform from one form into the other.

Propositions 6 and 7 show that if A is spd and $B+B^T$ is positive semidefinite, then the condition number of the preconditioned matrix satisfies

$$\mathcal{K}(\mathcal{B}^{-1}\mathcal{A})\leq 1+\mu_{\max}\leq 2.$$

Using a preconditioning parameter, as in (6), we derive now a further-improved condition number bound under the assumption that matrix B is symmetric. We consider then the form (2) of matrix $\mathcal{A}$.

Proposition 8. *Let* $\mathcal{A}=\begin{bmatrix}A & -B^T\\ \beta^2B & \alpha^2A\end{bmatrix}$, *where* $\alpha>0,\beta>0$, *and let* $\mathcal{B}$ *be defined in (6). Assume that* A *and* B *are symmetric and that* A *is positive definite. Let* $\tilde{B}=\beta A^{-1/2}BA^{-1/2}$ *and assume that* $\tilde{B}$ *has eigenvalues* μ *in the interval* $[\mu_{\min},\mu_{\max}]$, *where* $0\leq|\mu_{\min}|<\mu_{\max}$, *and that* $\frac{\tilde{\alpha}}{\alpha}=|\tilde{\mu}_{\min}|+\sqrt{1+\tilde{\mu}_{\min}^2}$ *where* $\tilde{\mu}_{\min}=\mu_{\min}/\alpha$, $\tilde{\mu}_{\max}=\mu_{\max}/\alpha$. *Then the eigenvalues of* $\mathcal{B}^{-1}\mathcal{A}$ *satisfy*

$$\lambda(\mathcal{B}^{-1}\mathcal{A})=\frac{\alpha^2+\mu^2}{(\tilde{\alpha}+\mu)^2}.$$

For its condition number it holds

$$\min_{\tilde{\alpha}} \kappa(\mathcal{B}^{-1}\mathcal{A}) = \left(\frac{1-\delta}{1+\gamma}\right)^2 + (1+\tilde{\mu}_{\max}^2)\left(\frac{\gamma+\delta}{1+\delta}\right)^2,$$

where $\delta = \mid \mu_{\min} \mid / \mu_{\max}$ *and* $\gamma = \sqrt{(1+\tilde{\mu}_{\min}^2)/(1+\tilde{\mu}_{\max}^2)}$. *Here it holds*

$$\frac{\tilde{\alpha}}{\alpha} = \frac{\tilde{\alpha}_{opt}}{\alpha} = \frac{\mid \tilde{\mu}_{\min} \mid + \gamma \tilde{\mu}_{\max}^2}{1-\gamma}.$$

If B is positive semidefinite, then

$$\kappa(\mathcal{B}^{-1}A) \leq 1 + 1/\left(1 + \frac{1}{\sqrt{1+\tilde{\mu}_{\max}^2}}\right)^2,$$

where the upper bound is taken for

$$\frac{\tilde{\alpha}}{\alpha} = \frac{1}{\tilde{\mu}_{\max}} + \sqrt{1 + \frac{1}{\tilde{\mu}_{\max}^2}}.$$

Proof. Since both $\mathcal{A}$ and $\mathcal{B}$ are nonsingular, the eigenvalues λ of the generalized eigenvalue problem,

$$\lambda \mathcal{B} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathcal{A} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

are nonzero. Using (2) and (6), we find

$$\left(\frac{1}{\lambda} - 1\right) \mathcal{A} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \left[(\tilde{\alpha}^2 - \alpha^2)A + \tilde{\alpha}\beta(B+B^T)\right] \mathbf{y} \end{bmatrix}.$$

If $\mathbf{y} = \mathbf{0}$, then for all $\mathbf{x} \neq \mathbf{0}$ it follows that $\lambda = 1$. For $\lambda \neq 1$, it follows that $A\mathbf{x} = B^T\mathbf{y}$ and, since A is spd,

$$\left(\frac{1}{\lambda} - 1\right)\left(\beta^2 BA^{-1}B^T + \alpha^2 A\right)\mathbf{y} = \left[(\tilde{\alpha}^2 - \alpha^2)A + \tilde{\alpha}\beta(B+B^T)\right]\mathbf{y},$$

or

$$\frac{1}{\lambda}\left(\tilde{B}\tilde{B}^T + \alpha^2 I\right)\tilde{\mathbf{y}} = \left(\tilde{\alpha}^2 I + \tilde{B}\tilde{B}^T + \tilde{\alpha}(\tilde{B}+\tilde{B}^T)\right)\tilde{\mathbf{y}},$$

where $\tilde{B} = \beta A^{-1/2}BA^{-1/2}$ and $\tilde{\mathbf{y}} = A^{1/2}\mathbf{y}$. Since $\tilde{B}$ is symmetric, if $\tilde{B}\tilde{\mathbf{y}} = \mu\tilde{\mathbf{y}}$, $\tilde{\mathbf{y}} \neq 0$, i.e. μ is an eigenvalue of $\tilde{B}$, it follows that μ is real and

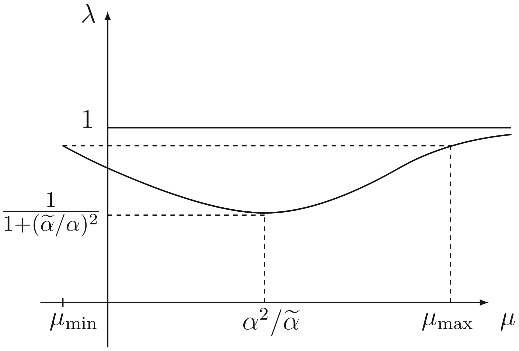

Fig. 1 $\lambda(\mu) = (\alpha^2+\mu^2)/(\tilde{\alpha}+\mu)^2$

$$\lambda = \lambda(\mu) = \frac{\alpha^2+\mu^2}{\tilde{\alpha}^2+\mu^2+2\tilde{\alpha}\mu} = \frac{\alpha^2+\mu^2}{(\tilde{\alpha}+\mu)^2}.$$

The eigenvalues vary as indicated in Fig. 1.

Consider first the case where there exists negative eigenvalues. To get $\lambda < 1$ for negative values of μ, we must choose $(\tilde{\alpha}+\mu)^2 > \alpha^2+\mu^2$, i.e. $\tilde{\alpha}^2+2\tilde{\alpha}\mu-\alpha^2 > 0$, that is,

$$\tilde{\alpha} > |\,\mu\,| + \sqrt{\mu^2+\alpha^2} \qquad \text{or}$$
$$\frac{\tilde{\alpha}}{\alpha} > |\,\tilde{\mu}_{\min}\,| + \sqrt{1+\tilde{\mu}_{\min}^2}.$$

The minimum value of $\lambda(\mu)$ can be found from

$$\lambda'(\mu) = \frac{2}{(\tilde{\alpha}+\mu)^3}\left(\tilde{\alpha}\mu-\alpha^2\right) = 0,$$

that is,

$$\min\lambda(\mu) = \lambda_{\min} = \lambda(\alpha^2/\tilde{\alpha}) = \frac{\alpha^2+\alpha^4/\tilde{\alpha}^2}{(\tilde{\alpha}+\alpha^2/\tilde{\alpha})^2} = \frac{1}{1+(\tilde{\alpha}/\alpha)^2}$$

To minimize the condition number, it can be seen (cf. Fig. 1) that we must choose $\tilde{\alpha}$ such that

$$\lambda(\mu_{\min}) = \lambda(\mu_{\max}),$$

that is,

$$\lambda_{\max} = \frac{\alpha^2+\mu_{\min}^2}{(\tilde{\alpha}-|\mu_{\min}|)^2} = \frac{\alpha^2+\mu_{\max}^2}{(\tilde{\alpha}+\mu_{\max})^2},$$

or

$$\frac{\tilde{\alpha}/\alpha - \tilde{\mu}_{\min}}{\tilde{\alpha}/\alpha + \tilde{\mu}_{\max}} = \gamma := \left(\frac{1+\tilde{\mu}_{\min}^2}{1+\tilde{\mu}_{\max}^2}\right)^{1/2}.$$

Here $\gamma < 1$, since by assumption $\mu_{\max} >\mid \mu_{\min} \mid$. Hence,

$$\frac{\tilde{\alpha}}{\alpha} = \frac{\tilde{\alpha}_{opt}}{\alpha} = \frac{\mid \tilde{\mu}_{\min} \mid + \gamma\tilde{\mu}_{\max}}{1-\gamma}$$

Then

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{1+\tilde{\mu}_{\max}^2}{\left(\frac{|\tilde{\mu}_{\min}|+\gamma\tilde{\mu}_{max}}{1-\gamma} + \tilde{\mu}_{\max}\right)^2}[1+\left(\frac{\mid \tilde{\mu}_{\min} \mid + \gamma\tilde{\mu}_{\max}}{1-\gamma}\right)^2]$$

$$= \frac{1+\tilde{\mu}_{\max}^2}{(\mid \tilde{\mu}_{\min} \mid + \tilde{\mu}_{\max})^2}\left[(1-\gamma)^2 + (\mid (\tilde{\mu}_{\min} \mid + \gamma\tilde{\mu}_{\max})^2\right].$$

It holds

$$(1-\gamma)^2 = \left(\frac{1-\gamma^2}{1+\gamma}\right)^2 = \frac{(\tilde{\mu}_{\max}^2 - \tilde{\mu}_{\min}^2)^2}{(1+\tilde{\mu}_{\max}^2)(1+\gamma)^2} = \frac{(\tilde{\mu}_{\max} + \mid \tilde{\mu}_{\min} \mid)^2(\tilde{\mu}_{\max} + \tilde{\mu}_{\min} \mid)^2}{(1+\tilde{\mu}_{\max}^2)(1+\gamma)^2}.$$

Hence,

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) = \left(\frac{1-\delta}{1+\gamma}\right)^2 + (1+\tilde{\mu}_{\max}^2)\left(\frac{\gamma+\delta}{1+\delta}\right)^2.$$

If B is positive semidefinite, then we let $\mu_{\min} = 0$ so $\delta = 0, \gamma = 1/\sqrt{1+\tilde{\mu}_{\max}^2}$ and

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le 1 + \frac{1}{(1+\frac{1}{\sqrt{1+\tilde{\mu}_{\max}^2}})^2}$$

which is taken for

$$\frac{\tilde{\alpha}}{\alpha} = \frac{1}{\tilde{\mu}_{\max}} + \sqrt{1+\frac{1}{\tilde{\mu}_{\max}^2}} \qquad \blacksquare$$

Remark 4. If $\mu_{\min} = 0$ then $\kappa(\mathcal{B}^{-1}\mathcal{A}) < 2$ and if $\mu_{\max} \to \infty$ then $\tilde{\alpha} \to \alpha$ and $\kappa(\mathcal{B}^{-1}\mathcal{A}) \to 2$. If $\tilde{\mu}_{\max} = 1$ then $\tilde{\alpha}/\alpha = 1+\sqrt{2}$ and

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le 1 + \frac{1}{(1+\frac{1}{\sqrt{2}})^2} \approx 1.34.$$

Remark 5. As is well known, when eigenvalue bounds of a preconditioned matrix, as in the case with $\mathcal{B}^{-1}\mathcal{A}$, are known, then one can replace the conjugate gradient (CG) with a Chebyshev acceleration method. This can be important, for instance, if one uses some domain decomposition method for massively parallel computations, as it avoids the global communication of inner products used in CG methods.

4 Distributed Optimal Control of Elliptic and Oseen Equations

Let Ω be a bounded domain in $\Re^d$, $d = 1, 2$ or 3, and let $\partial\Omega$ be its boundary which is assumed to be sufficiently smooth. Let $L^2(\Omega), H^1(\Omega)$ and $H_0^1(\Omega)$ denote the standard Lebesgue and Sobolev spaces of functions in Ω, where $H_0^1(\Omega)$ denotes functions with homogeneous Dirichlet boundary values at $\Gamma_0 \subset \partial\Omega$ where Γ_0 has a nonzero measure. Further, let $(\cdot,\cdot)$ and $\| \cdot \|$ denote the inner product and norm, respectively, in $L^2(\Omega)$, both for scalar and vector functions. Extending, but following [7], and based on [6], we consider now two optimal control problems. In [7] a block-diagonal preconditioner is used. Here we apply instead the preconditioner presented in Sect. 2.

4.1 An Elliptic State Equation

The problem is to find the state $u \in H_0^1(\Omega)$ and the control function $y \in L^2(\Omega)$ that minimizes the *cost function*

$$J(u,y) = \frac{1}{2} \| u - u_d \|^2 + \frac{\alpha}{2} \| y \|^2$$

subject to the *state equation*

$$\begin{cases} -\Delta u + (\mathbf{b} \cdot \nabla) u = y \text{ in } \Omega \\ \text{with boundary conditions} \\ u = 0 \text{ on } \Gamma_0 ; \nabla u \cdot \mathbf{n} = 0 \text{ on } \Gamma_1 = \partial\Omega \setminus \Gamma_0 . \end{cases} \tag{11}$$

Here $\mathbf{b}$ is a given, smooth vector. For simplicity, assume that $\mathbf{b} \cdot \mathbf{n} \mid_{\Gamma_1} = 0$. Further, u_d denotes a given, desired state (possibly obtained by measurements at some discrete points and then interpolated to the whole of Ω). The forcing term y acts as a control of the solution to the state equation. By including the control in the cost functional, the problem becomes well posed. The regularization parameter α, chosen a priori, is a positive parameter chosen sufficiently small to obtain a solution close to the desired state, but not too small and also not too large as this leads to ill conditioning. This is similar to the familiar Tikhonov regularization. The variational (weak) formulation of (11) reads

$$(\nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) = (y, v) \quad \forall v \in H_0^1(\Omega). \tag{12}$$

The Lagrangian formulation associated with the optimization problem takes the form

$$\mathcal{L}(u,y,p) = J(u,y) + (\nabla u, \nabla p) + (\mathbf{b} \cdot \nabla u, p) - (y, p),$$

where $p \in H_0^1(\Omega)$ is the Lagrange multiplier corresponding to the constraint (12). The weak formulation of the corresponding first-order necessary conditions,

$$\left(\frac{\partial \mathcal{L}}{\partial u}, v\right) = 0 \quad \forall v \in H_0^1(\Omega)$$

$$\left(\frac{\partial \mathcal{L}}{\partial y}, z\right) = 0 \quad \forall z \in L^2(\Omega)$$

$$\left(\frac{\partial \mathcal{L}}{\partial p}, q\right) = 0 \quad \forall q \in H_0^1(\Omega)$$

gives now the system of optimality equations:

$$\begin{cases} (u,v) + (\nabla v, \nabla p) + (\mathbf{b}\cdot\nabla v, p) & = (u_d, v) \ \forall v \in H_0^1(\Omega) \\ \alpha(y,z) - (z,p) & = 0 \qquad \forall z \in L^2(\Omega) \\ (\nabla u, \nabla q) + (\mathbf{b}\cdot\nabla u, q) - (y,q) & = 0 \qquad \forall q \in H_0^1(\Omega) \end{cases},$$

which defines the solution $(u,y) \in H_0^1(\Omega) \times L^2(\Omega)$ of the optimal control problem with Lagrange multiplier $p \in H_0^1(\Omega)$. From the second equation, it follows that the control function y is related to the Lagrange multiplier as $y = \frac{1}{\alpha}p$. Eliminating y and applying the divergence theorem, this leads to the reduced system

$$\begin{aligned} (u,v) + (\nabla v, \nabla p) - (\mathbf{b}\cdot\nabla p, v) \quad & = (u_d, v) \ \forall v \in H_0^1(\Omega) \\ (\nabla u, \nabla q) + (\mathbf{b}\cdot\nabla u, q) - \tfrac{1}{\alpha}(p,q) & = 0 \qquad \forall q \in H_0^1(\Omega). \end{aligned}$$

Since the problem is regularized, we may here use equal-order finite element approximations, for instance, piecewise linear basis functions on a triangular mesh (in 2D), for both the state variable u and the co-state variable p. This leads to a system of the form

$$\begin{bmatrix} M & K^T \\ K & -\alpha^{-1}M \end{bmatrix} \begin{bmatrix} u_h \\ p_h \end{bmatrix} = \begin{bmatrix} f_h \\ 0 \end{bmatrix},$$

where index h denotes the corresponding mesh parameter. Here M corresponds to a mass matrix and K, which has the same order as M, to the second-order elliptic operator with a first-order advection term.

By a change of sign of p_h, it can be put in the form

$$\begin{bmatrix} M & -K^T \\ K & \alpha^{-1}M \end{bmatrix} \begin{bmatrix} u_h \\ -p_h \end{bmatrix} = \begin{bmatrix} f_h \\ 0 \end{bmatrix}$$

and we can directly apply the preconditioner from Sects. 2 and 3, and the derived spectral condition number bounds. If

$$\int_{\Omega} \left(|\nabla u|^2 - \frac{1}{2}(\nabla \cdot \mathbf{b})u^2 \right) \geq 0,$$

i.e. if the operator is semi-coercive, then $K+K^T$ is positive semidefinite and it follows from Proposition 6 that the corresponding spectral condition number is bounded by 2, with eigenvalues in the interval $1/2 \leq \lambda \leq 1$.

Remark 6. In [7], a block-diagonal preconditioner,

$$\mathcal{D} = \begin{bmatrix} A+\alpha^{1/2}B & 0 \\ 0 & \alpha^{-1}A+\alpha^{-1/2}B \end{bmatrix},$$

is used for the saddle point matrix

$$\mathcal{A} = \begin{bmatrix} A & B \\ B & -\alpha^{-1}A \end{bmatrix},$$

where $B=B^T$ and A is symmetric and positive semidefinite, and $\ker(A)\cap\ker(B)=\{0\}$, so $A+\alpha^{1/2}B$ is symmetric and positive definite.

By assumptions made, from the generalized eigenvalue problem

$$A\mathbf{z} = \mu(A+\alpha^{1/2}B)\mathbf{z},$$

it follows that here $\mu \in [0,1]$ and it follows further readily that the preconditioned matrix $\mathcal{D}^{-1}\mathcal{A}$ has eigenvalues that satisfy

$$|\,\lambda\,| = \sqrt{\mu_i^2+(1-\mu_i)^2} \quad \text{for some } \mu_i \in [0,1],$$

that is, $1/\sqrt{2} \leq |\,\lambda\,| \leq 1$. Hence, the eigenvalues are located in the double interval:

$$I = [-1,-1/\sqrt{2}] \cup [1/\sqrt{2},1].$$

For such eigenvalues in intervals on both sides of the origin, an iterative method of conjugate gradient type, such as MINRES, needs typically the double number of iterations, as for eigenvalues in a single interval on one (positive) side of the origin, to reach convergence; see e.g. [11]. This can be seen from the polynomial approximation problem

$$\min_{x\in I,\ P_k\in\pi_k^0} |\,P_k(x)\,| \leq \varepsilon$$

where π_k^0 denotes the set of polynomials of degree k, normalized at the origin, i.e. $P_k(0)=1$.

Since the number of iterations increases as $O(\sqrt{\kappa})$, where $\kappa = |\,\lambda_{\max}\,|\,/\,|\,\lambda_{\min}\,|$ is the condition number, it follows that an indefinite interval condition number $\kappa=\sqrt{2}$ typically corresponds to a one-sided condition number of $4\sqrt{2}$.

The method proposed in the present paper has a condition number bounded by 2 and needs therefore a number of iterations about $\simeq \dfrac{\sqrt{2}}{2^{5/4}} = \dfrac{2^{1/4}}{2} \simeq 0.6$ times those for a corresponding block diagonal preconditioner. However, even if the block-diagonal preconditioning method requires more iterations, each iteration may be cheaper than in the method proposed in this paper. An actual comparison of the methods will appear.

4.2 Distributed Optimal Control of the Oseen Problem

In [7], Stokes equation is considered. Here, we extend the method to the Oseen equation and consider the velocity tracking problem for the stationary case, which reads as follows:

Find the velocity $\mathbf{u} \in H_0^1(\Omega)^d$; the pressure $p \in L_0^2(\Omega)$, where $L_0^2(\Omega) = \{q \in L^2(\Omega), \int_\Omega q\,dx = 1\}$; and the control function $\mathbf{f}$, which minimize the cost function

$$\mathcal{J}(\mathbf{u},\mathbf{f}) = \frac{1}{2}\|\mathbf{u} - \mathbf{u}_d\|^2 + \frac{1}{2}\alpha\|\mathbf{f}\|^2,$$

subject to state equation for an incompressible fluid velocity $\mathbf{u}$, such that

$$\begin{cases} -\Delta\mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \text{ in } \Omega \\ \qquad\qquad\qquad \nabla\cdot\mathbf{u} = 0 \text{ in } \Omega \end{cases}$$

and boundary conditions $\mathbf{u} = \mathbf{0}$ on $\partial\Omega_1$, $\mathbf{u}\cdot\mathbf{n} = 0$ on $\partial\Omega_2 = \partial\Omega \backslash \partial\Omega_1$, where $\mathbf{n}$ denotes the outward normal vector to the boundary $\partial\Omega$.

Here $\mathbf{u}_d$ is the desired solution and $\alpha > 0$ is a regularization parameter, used to penalize too large values of the control function. Further, $\mathbf{b}$ is a given, smooth vector. For simplicity we assume that $\mathbf{b} = \mathbf{0}$ on $\partial\Omega_1$ and $\mathbf{b}\cdot\mathbf{n} = 0$ on $\partial\Omega_2$.

In a Navier–Stokes problem, solved by a Picard iteration using the frozen coefficient framework, $\mathbf{b}$ equals the previous iterative approximation of $\mathbf{u}$, in which case normally $\nabla\cdot\mathbf{u} = 0$ in Ω. For simplicity, we assume that this holds here also, that is, $\nabla\cdot\mathbf{b} = 0$.

The variational form of the state equation reads as follows:

$$\begin{cases} (\nabla\mathbf{u},\nabla\tilde{\mathbf{u}}) + (\mathbf{b}\cdot\nabla\mathbf{u},\tilde{\mathbf{u}}) - (\nabla\tilde{\mathbf{u}},p) = (\mathbf{f},\tilde{\mathbf{u}}) \ \forall\tilde{\mathbf{u}} \in H_0^1(\Omega) \\ \qquad\qquad\qquad\qquad\qquad (\nabla\cdot\mathbf{u},\tilde{p}) = 0 \qquad \forall\tilde{p} \in L_0^2(\Omega) \end{cases}$$

The Lagrangian functional, corresponding to the optimization problem, is given by

$$\mathcal{L}(\mathbf{u},p,\mathbf{v},q,\mathbf{f}) = \mathcal{J}(\mathbf{u},\mathbf{f}) + (\nabla\mathbf{u},\nabla\mathbf{v}) + (\mathbf{b}\cdot\nabla\mathbf{u},\mathbf{v}) - (\nabla\cdot\mathbf{v},p) - (\nabla\cdot\mathbf{u},q) - (\mathbf{f},\mathbf{v})$$

where $\mathbf{v}$ is the Lagrange multiplier function for the state equation and q for its divergence constraint. Applying the divergence theorem, the divergence condition $\nabla \cdot \mathbf{b} = 0$ and the boundary conditions, we can write

$$\int_{\Omega} \mathbf{b} \cdot \nabla \tilde{\mathbf{u}} \cdot \mathbf{v} d\Omega = - \int_{\Omega} (\mathbf{b} \cdot \underline{\nabla} \mathbf{v}) \cdot \tilde{\mathbf{u}} d\Omega.$$

The five first-order necessary conditions for an optimal solution take then the form

$$\begin{array}{rll} (\mathbf{u},\tilde{\mathbf{u}}) + (\nabla \mathbf{v}, \nabla \tilde{\mathbf{u}}) - (\mathbf{b} \cdot \nabla \mathbf{v}, \tilde{\mathbf{u}}) - (\nabla \cdot \tilde{\mathbf{u}}, q) &= (\mathbf{u}_d, \tilde{\mathbf{u}}) & \forall \tilde{\mathbf{u}} \in H_0^1(\Omega)^d \\ (\nabla \cdot \mathbf{v}, \tilde{p}) &= 0 & \forall \tilde{p} \in L_0^2(\Omega) \\ (\nabla \mathbf{u}, \nabla \tilde{\mathbf{v}}) + (\mathbf{b} \cdot \nabla \mathbf{u}, \tilde{\mathbf{v}}) - (\nabla \cdot \tilde{\mathbf{v}}, p) - (\mathbf{f}, \tilde{\mathbf{v}}) &= 0 & \forall \tilde{\mathbf{v}} \in H_0^1(\Omega)^d \\ (\nabla \cdot \mathbf{u}, \tilde{q}) &= 0 & \forall \tilde{q} \in L_0^2(\Omega) \\ \alpha(\mathbf{f}, \tilde{\mathbf{f}}) - (\tilde{\mathbf{f}}, \mathbf{v}) &= 0 & \forall \tilde{\mathbf{f}} \in L^2(\Omega) \end{array} \tag{13}$$

Here $\mathbf{u}, p, \mathbf{f}$ are the solutions of the optimal control problem with $\mathbf{v}, q$ as Lagrange multipliers for the state equation, and $\tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \tilde{p}, \tilde{q}, \tilde{\mathbf{f}}$ denote corresponding test functions.

As in the elliptic control problem, the control function $\mathbf{f}$ can be eliminated, $\mathbf{f} = \alpha^{-1}\mathbf{v}$, resulting in the reduced system,

$$\begin{array}{rll} (\mathbf{u},\tilde{\mathbf{u}}) + (\nabla \mathbf{v}, \nabla \tilde{\mathbf{u}}) - (\mathbf{b} \cdot \nabla \mathbf{v}, \tilde{\mathbf{u}}) - (\nabla \cdot \tilde{\mathbf{u}}, q) &= (\mathbf{u}_d, \tilde{\mathbf{u}}) & \forall \tilde{\mathbf{u}} \in H_0^1(\Omega)^d \\ (\nabla \mathbf{u}, \nabla \tilde{\mathbf{v}}) + (\mathbf{b} \cdot \nabla \mathbf{u}, \tilde{\mathbf{v}}) - (\nabla \cdot \tilde{\mathbf{v}}, p) - \alpha^{-1}(\mathbf{v}, \tilde{\mathbf{v}}) &= 0 & \forall \tilde{\mathbf{v}} \in H_0^1(\Omega)^d \\ (\nabla \cdot \mathbf{v}, \tilde{p}) &= 0 & \forall \tilde{p} \in L_0^2(\Omega) \\ (\nabla \cdot \mathbf{u}, \tilde{q}) &= 0 & \forall \tilde{q} \in L_0^2(\Omega) \end{array}, \tag{14}$$

To discretize (14) we use an LBB-stable pair of finite element spaces for the pair $(\mathbf{u}, \mathbf{v})$ and (p, q). In [7] the Taylor–Hood pair with $\{Q2, Q2, Q1, Q1\}$ is used, namely, piecewise quadratic basis functions for $\mathbf{u}, \mathbf{v}$ and piecewise bilinear basis functions for p, q for a triangular mesh. The corresponding discrete system takes the form

$$\begin{bmatrix} M & -L+C & 0 & D^T \\ L+C & \alpha^{-1}M & D^T & 0 \\ 0 & D & 0 & 0 \\ D & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ -\mathbf{v} \\ p \\ q \end{bmatrix} = \begin{bmatrix} M\mathbf{u}_d \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{15}$$

where we have changed the sign of $\mathbf{v}$. Here D comes from the divergence terms. Further, M is the mass matrix and $L+C$ is the discrete operator, corresponding to the convection–diffusion term $-\Delta \mathbf{u} + \mathbf{b} \cdot \nabla \mathbf{u}$ and $-L+C$ to $\Delta \mathbf{v} + \mathbf{b} \cdot \nabla \mathbf{v}$, respectively. Due to the use of an inf–sup (LBB)-stable pairs of finite element spaces, the divergence matrix D has full rank.

As for saddle point problems of similar type, one can use either a grad–div stabilization or a div–grad stabilization. In the first case we add the matrix

$D^TW^{-1}D$ to M and $\alpha^{-1}D^TW^{-1}D$ to $\alpha^{-1}M$, respectively, possibly multiplied with some constant factor, where W is a weight matrix. If W is taken as the discrete Laplacian matrix, then $D^TW^{-1}D$ becomes a projection operator onto the orthogonal complement of the solenoidal vectors.

The other type of stabilization consists of perturbing the zero block matrix in (15) by $\varepsilon \begin{bmatrix} \Delta & 0 \\ 0 & \Delta \end{bmatrix}$, where ε is a small parameter, typically $\varepsilon = O(h^2)$ with h being the space discretization parameter. In that case there is no need to use LBB-stable elements; see, e.g. [12] for more details. In the present paper, however, we use LBB-stable elements and there is no need to use any additional regularization at all but consider instead the solution of the system with the Schur complement matrix system:

$$\begin{bmatrix} 0 & D \\ D & 0 \end{bmatrix} \begin{bmatrix} M & -L+C \\ L+C & \alpha^{-1}M \end{bmatrix}^{-1} \left(\begin{bmatrix} 0 & D^T \\ D^T & 0 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} M\mathbf{u}_d \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{16}$$

This system can be solved by inner–outer iterations. To compute the residuals, we must then solve inner systems with the matrix $\begin{bmatrix} M & -L+C \\ L+C & \alpha^{-1}M \end{bmatrix}$, which takes place in the way discussed earlier in Sect. 2. To recall, only systems with $M+\sqrt{\alpha}(L+C)$ and $M+\sqrt{\alpha}(L-C)$ have to be solved. Further, as is seen from (16), the corresponding systems which actually arise have the form $D[M+\sqrt{\alpha}(L+C)]^{-1}D^T$ and $D[M+\sqrt{\alpha}(L-C)]^{-1}D^T$. At least for not too large convection terms, related to the diffusion term, these systems are well conditioned and can be preconditioned with a mass matrix or a mass matrix minus a small multiple times the Laplacian.

To avoid the need to solve inner systems and for stronger convections, it may be better to use a block-triangular factorization of the matrix in (15). For the arising inner systems with $M+\sqrt{\alpha}(L+C)$ and $M+\sqrt{\alpha}(L-C)$, it can be efficient to use some off-the-shelf software, such as some algebraic multigrid (AMG) method; see [13, 14]. In [15] and [13] numerical tests are reported, showing that AGMG [13], as one choice of an AMG method, performs much better than some other possible methods.

The perturbations due to the use of inner iterations with stopping criteria lead in general to complex eigenvalues. A generalized conjugate gradient method of GMRES [16] type can be used. Such methods go under different names and have been referred to as nonlinear conjugate gradient, variable preconditioned conjugate gradient [17] and flexible GMRES [18]. Since, due to the accurate preconditioning, there are few iterations, the additional cost for having a full length Krylov subspace, involving all previous search directions, is not much heavier than if a conjugate gradient method with vectors, orthogonal with respect to a proper inner product and, hence, short recursions, is used.

We remark, however, that such a method has been constructed for indefinite matrices in [19], based on inner products, defined by the matrix

$$\mathcal{D} = \begin{bmatrix} \hat{M} - \hat{M}_0 & 0 \\ 0 & S_0 \end{bmatrix},$$

where $\hat{M}_0$ is an approximation of $\hat{M}$, such that $\hat{M}_0 < \hat{M}$ and $S_0 < \hat{B}\hat{M}^{-1}\hat{B}^T$ is an spd approximation of the Schur complement matrix for the two-by-two block system $\begin{bmatrix} \hat{M} & \hat{B}^T \\ \hat{B} & 0 \end{bmatrix}$. This makes the matrix $\begin{bmatrix} \hat{M}_0 & 0 \\ \hat{B} & -S_0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{M} & \hat{B}^T \\ \hat{B} & 0 \end{bmatrix}$ self-adjoint with respect to that inner product. The drawback of the method is the need to properly scale the approximation $\hat{M}_0$ to satisfy $\hat{M}_0 < \hat{M}$, and furthermore, $\hat{M}_0$ must be fixed, i.e. cannot be implicitly defined via variable inner iterations.

In our case, the corresponding preconditioning matrix defined in Sect. 2 satisfies $\hat{M}_0 > \hat{M}$, but there is no need to scale it. Furthermore, we may apply inner iterations for this preconditioner and also for the Schur complement matrix, hence the corresponding matrix $\hat{M}_0$ is in general not fixed so the above inner product method is not applicable.

The presentation of block-triangular factorization preconditioner and approximations of the arising Schur complement preconditioners with numerical tests will be devoted.

Acknowledgements This work was supported by the European Regional Development Fund in the IT4 Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

Discussions with Maya Neytcheva on implementation aspects of the method are gratefully acknowledged.

References

1. van Rienen, U.: Numerical Methods in Computational Electrodynamics. Linear Systems in Practical Applications. Springer, Berlin (1999)
2. Axelsson, O., Kucherov, A.: Real valued iterative methods for solving complex symmetric linear systems. Numer. Lin. Algebra Appl. **7**, 197–218 (2000)
3. Benzi, M., Bertaccini, D.: Block preconditioning of real-valued iterative algorithms for complex linear systems. IMA J. Numer. Anal. **28**, 598–618 (2008)
4. Axelsson, O., Boyanova, P., Kronbichler, M., Neytcheva, M., Wu, X.: Numerical and computational efficiency of solvers for two-phase problems. Comput. Math. Appl. **65**, 301–314 (2012). http://dx.doi.org./10.1016/j.camva.2012.05.020
5. Boyanova, P.: On numerical solution methods for block-structured discrete systems. Doctoral thesis, Department of Information Technology, Uppsala University, Sweden (2012). http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-173530
6. Lions, J.-L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Berlin (1971)
7. Zulehner, W.: Nonstandard norms and robust estimates for saddle-point problems. SIAM J. Matrix Anal. Appl. **32**, 536–560 (2011)
8. Arridge, S.R.: Optical tomography in medical imaging. Inverse Probl. **15**, 41–93 (1999)
9. Egger, H., Engl, H.W.: Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates. Inverse Probl. **21**, 1027–1045 (2005)

10. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12**, 617–629 (1975)
11. Axelsson, O., Barker, V.A.: Finite Element Solution of Boundary Value Problems. Theory and Computation. Academic, Orlando, FL (1984)
12. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Elements Methods. Springer, Berlin (1991)
13. Notay, Y.: The software package AGMG. http://homepages.ulb.ac.be/~ynotay/
14. Vassilevski, P.: Multilevel Block Factorization Preconditioners. Springer, New York (2008)
15. Notay, Y.: Aggregation-based algebraic multigrid for convection-diffusion equations. SIAM J. Sci. Comput. **34**, A2288–A2316 (2012)
16. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
17. Axelsson, O., Vassilevski, P.S.: A black box generalized conjugate gradient solver with inner iterations and variable-step preconditioning. SIAM J. Matrix Anal. Appl. **12**(4), 625–644 (1991)
18. Saad, Y.: A flexible inner-outer preconditioned GMRES algorithm. SIAM J. Sci. Comput. **14**, 461–469 (1993)
19. Bramble, J.H., Pasciak, J.E.: A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. Math. Comput. **50**, 1–17 (1988)