SYMMETRIC INDEFINITE PRECONDITIONERS FOR SADDLE POINT PROBLEMS WITH APPLICATIONS TO PDE-CONSTRAINED OPTIMIZATION PROBLEMS

JOACHIM SCHÖBERL* AND WALTER ZULEHNER†

Abstract. We consider large scale sparse linear systems in saddle point form. A natural property of such indefinite 2-by-2 block systems is the positivity of the (1,1) block on the kernel of the (2,1) block. Many solution methods, however, require that the positivity of the (1,1) block is satisfied everywhere. To enforce the positivity everywhere, an augmented Lagrangian approach is usually chosen. However, the adjustment of the involved parameters is a critical issue. We will present a different approach that is not based on such an explicit augmentation technique. For the considered class of symmetric and indefinite preconditioners, assumptions are presented that lead to symmetric and positive definite problems with respect to a particular scalar product. Therefore, conjugate gradient acceleration can be used.

An important class of applications are optimal control problems. It is typical for such problems that the cost functional contains an extra regularization parameter. For control problems with elliptic state equations and distributed control a special indefinite preconditioner for the discretized problem is constructed which leads to convergence rates of the preconditioned conjugate gradient method that are not only independent of the mesh size but also independent of the regularization parameter. Numerical experiments are presented for illustrating the theoretical results.

Key words. saddle point problems, indefinite preconditioners, KKT systems, conjugate gradient methods, PDE-constrained optimization problems, optimal control problems

AMS subject classifications. 65F10, 15A12, 49M15

1. Introduction.

In this paper we consider large scale sparse linear systems of equations in saddle point form

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

where A is a real, symmetric and positive semi-definite n-by-n matrix, B is a real m-by-n matrix with full rank $m \le n$, and B^T denotes the transposed matrix of B. Such systems typically result from the discretization of mixed variational problems for systems of partial differential equations (in short: PDEs), see Brezzi and Fortin [6], in particular, from the discretization of optimization problems with PDE-constraints. A natural property of such a problem is that A is positive definite on the kernel of B, i.e.:

$$(Aw, w) > 0 \quad \text{for all } w \in \ker B \text{ with } w \neq 0, \tag{1.2}$$

where (x, w) denotes the Euclidean scalar product. This condition guarantees in combination with the full rank of B that the matrix

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

is non-singular.

Under the assumptions stated above, the system (1.1) can be interpreted as the Karush-Kuhn-Tucker (KKT) conditions which characterize the solution x of the following constrained optimization problem, see, e.g., Fletcher [9],

$$\text{Minimize} \quad J(x) \equiv \frac{1}{2}(Ax, x) - (f, x) \quad \text{subject to the constraints} \quad Bx = g$$

[*] Center for Computational Engineering Science, RWTH Aachen, D-52074 Aachen, Germany, (schoeberl@mathcces.rwth-aachen.de).

[†] Institute of Computational Mathematics, Johannes Kepler University, A-4040 Linz, Austria (zulehner@numa.uni-linz.ac.at). The work was supported in part by the Austrian Science Foundation (FWF) under the grant SFB F013/F1309.

with associated Lagrangian parameter p.

As long as the matrix A is positive definite not only on $\ker B$ but on the whole space $\mathbb{R}^n$, the (negative) Schur complement $S = BA^{-1}B^T$ is well-defined. Then several approaches for an efficient solution procedure have been proposed. Most of them can be viewed as preconditioned Richardson methods for (1.1) typically accelerated by a Krylov subspace method, see Saad, van der Vorst [15] for a review of iterative methods for linear systems. The discussed preconditioners for $\mathcal{K}$ are 2-by-2 block matrices $\hat{\mathcal{K}}$ depending on a preconditioner $\hat{A}$ for approximating A and a preconditioner $\hat{S}$ which is either interpreted as approximation of the Schur complement S or as approximation of the so-called inexact Schur complement $H = B\hat{A}^{-1}B^T$. Typical classes of such preconditioners are block diagonal preconditioners, see, e.g., Rusten and Winther [14], Silvester and Wathen [16], block triangular preconditioners (originating from the classical Uzawa method [1]), see, e.g., Elman, Golub [8], Bramble, Pasciak, Vassilev [5], symmetric indefinite preconditioners, see, e.g., Dyn, Ferguson [7], Bank, Welfert, Yserentant [2], and symmetric positive definite block (but not block-diagonal) preconditioners, see Vassilevski, Lazarov [18]. Depending on the properties of the preconditioned systems Krylov subspace methods either for symmetric indefinite or for non-symmetric systems like MINRES or GMRES were proposed. In Bramble and Pasciak [4] a block triangular preconditioner was used in order to obtain a preconditioned system which is symmetric and positive definite and, therefore, can be solved by the conjugate gradient method, which is usually considered as the best or at least the best-understood Krylov subspace method. The block triangular preconditioner in [4] requires a symmetric and positive definite approximation $\hat{A}$ with $\hat{A} < A$. For a much more detailed discussion of available methods for saddle point problems we refer to the review article by Benzi, Golub and Liesen [3] .

In this paper, however, we will focus on systems, where A is positive definite in a stable way (to be specified later) only on $\ker B$, a typical situation for certain classes of optimization problems with PDE-constraints. One strategy to enforce the definiteness on the whole space $\mathbb{R}^n$ is the augmented Lagrangian approach, where the matrix A and the vector f in (1.1) are replaced by a matrix of the form $A_W = A + B^TWB$ and a vector $f_W = f + B^TWg$, respectively, with an appropriate matrix W, see e.g. Fortin and Glowinski [10]. This does not change the solution of the problem, and the new (1,1) block A_W becomes positive definite if W is properly chosen, e.g, if it is positive definite, and all methods from above applied to the augmented system could be used, in principle. It is, however, a delicate issue to choose the matrix W in order to obtain good convergence properties, see the discussions in Golub and Greif [11], Golub, Greif and Varah [12].

Here we will take a different approach and discuss preconditioners $\hat{\mathcal{K}}$ for the original system matrix $\mathcal{K}$ (without augmentation), which, nevertheless, work also well, in the case that A is positive definite only on the kernel of B. Under appropriate assumptions it will be shown that the preconditioned matrix $\hat{\mathcal{K}}^{-1}\mathcal{K}$ is even symmetric and positive definite in some appropriate scalar product. Therefore, conjugate gradient acceleration can be applied. In contrast to Bramble and Pasciak [4] this new technique requires a symmetric and positive definite approximation $\hat{A}$ with $\hat{A} > A$, which is easier to achieve and can be applied also if A itself is only positive definite on the kernel of B.

An important field of applications are PDE-constrained optimization problems, in particular, optimal control problems, see, e.g, Tröltzsch [17]. It is typical for optimal control problems that the cost functional contains an extra regularization parameter. If discretized by an appropriate finite element method, the resulting KKT system is of the form (1.1), where the matrices A and B depend on the underlying subdivision, say with mesh size h, and on the regularization parameter, say ν. For optimal control problems with elliptic state equations and distributed control a special symmetric indefinite preconditioner will be constructed and convergence rate estimates are given which are robust in h as well as in ν.

The paper is organized as follows: In Section 2 the considered class of preconditioners is introduced and analyzed. Section 3 describes how the algebraic conditions for the preconditioners are linked to the conditions of the theorem of Brezzi for mixed variational problems, and a general framework for constructing the preconditioners is sketched. In Section 4 a problem from optimal control is discussed and preconditioners are constructed which are robust with respect to the mesh size as well as to the involved regularization parameter. Numerical experiments are presented in

Section 5, followed by some concluding remarks.

Throughout the paper the following notations are used: $M < N$ ($N > M$) iff $N - M$ is positive definite, and $M \leq N$ ($N \geq M$) iff $N - M$ is positive semi-definite, for symmetric matrices M and N. For a symmetric and positive definite matrix M the associated scalar product $(v, w)_M$ and norm $\|v\|_M$ are given by

$$(v, w)_M = (Mv, w) \quad \text{and} \quad \|v\|_M = (v, v)_M^{1/2},$$

where (v, w) (without index) denotes the Euclidean scalar product. The Euclidean norm of a vector v is denoted by $\|v\|$ (without index).

2. A class of symmetric and indefinite preconditioners. A well-known class of preconditioners is given by

$$\hat{\mathcal{K}} = \begin{pmatrix} \hat{A} & B^T \\ B & B\hat{A}^{-1}B^T - \hat{S} \end{pmatrix},$$

where $\hat{A}$ and $\hat{S}$ are symmetric and positive definite matrices, see Bank, Welfert and Yserentant [2]. More precisely, we will assume that $\hat{A}$ and $\hat{S}$ are preconditioners, i.e., efficient evaluations of $\hat{A}^{-1}s$ and $\hat{S}^{-1}t$ are available for given vectors s and t.

We have the following factorization

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

which implies that $\hat{\mathcal{K}}$ is non-singular and that the solution of a linear system

$$\hat{\mathcal{K}} \begin{pmatrix} w \\ q \end{pmatrix} = \begin{pmatrix} s \\ t \end{pmatrix}$$

reduces to the consecutive solution of the following three linear systems:

$$\begin{aligned} \hat{A}\hat{w} &= s, \\ \hat{S}\, q &= B\hat{w} - t, \\ \hat{A}w &= s - B^T q. \end{aligned}$$

So, one application of the preconditioner $\hat{\mathcal{K}}$ requires two applications of the preconditioner $\hat{A}$ and one application of the preconditioner $\hat{S}$.

In Bank, Welfert and Yserentant [2] and later in Zulehner [19], this preconditioner has been analyzed for the case that A is positive definite. One important part of the analysis easily carries over to the case, considered here:

Theorem 2.1. *Assume that $A \geq 0$, condition (1.2) is satisfied, and* $\operatorname{rank} B = m$. *Let $\hat{A} > 0$ and $\hat{S} > 0$.*

1. If

$$\hat{A} \geq A \quad \text{and} \quad \hat{S} \leq B\hat{A}^{-1}B^T, \tag{2.1}$$

then all eigenvalues of $\hat{\mathcal{K}}^{-1}\mathcal{K}$ are real and positive.

2. If

$$\hat{A} > A \quad \text{and} \quad \hat{S} < B\hat{A}^{-1}B^T, \tag{2.2}$$

then $\hat{\mathcal{K}}^{-1}\mathcal{K}$ is symmetric and positive definite with respect to the scalar product

$$\left(\begin{pmatrix} x \\ p \end{pmatrix}, \begin{pmatrix} w \\ q \end{pmatrix} \right)_{\mathcal{D}} = ((\hat{A} - A)x, w) + ((B\hat{A}^{-1}B^T - \hat{S})p, q). \tag{2.3}$$

Proof. Apply Theorem 5.2 from Zulehner [19] to the regularized matrices $A + \varepsilon I$ and $\hat{A} + \varepsilon I$ for $\varepsilon > 0$, take the limit $\varepsilon \to 0$ and observe that $\mathcal{K}$ is non-singular. □

Estimates for the extreme eigenvalues of $\hat{\mathcal{K}}^{-1}\mathcal{K}$ were derived in Zulehner [19] under the assumption that A is positive definite on the whole space. However, the estimate for the smallest eigenvalue degenerates, if directly applied to the case considered here. In this paper this gap will be closed.

First of all, we have to discuss reasonable assumptions on $\hat{A}$ and $\hat{S}$, which measure the quality of these preconditioners. Comparing the matrix $\mathcal{K}$ and the preconditioner $\hat{\mathcal{K}}$ it seems to be natural to consider $\hat{A}$ as an approximation to A at least on $\ker B$ and to consider $\hat{S}$ as an approximation to the so-called inexact Schur complement H, given by

$$H = B^T \hat{A}^{-1} B.$$

Therefore, we assume that constants $\alpha > 0$ and $\beta > 0$ exist such that

$$(Aw, w) \geq \alpha\, (\hat{A}w, w) \quad \text{for all } w \in \ker B$$

and

$$B\hat{A}^{-1}B^T \leq \beta\, \hat{S}.$$

Observe, that we will still require condition (2.1), therefore $\alpha \leq 1$ and $\beta \geq 1$. The closer α and β are to 1 the better, we expect, the preconditioner $\hat{\mathcal{K}}$ will be.

Now we have

THEOREM 2.2. *Assume that $A \geq 0$, condition (1.2) is satisfied, and $\operatorname{rank} B = m$. Let $\hat{A} > 0$ and $\hat{S} > 0$ with*

$$(Aw, w) \geq \alpha\, (\hat{A}w, w) \quad \textit{for all } w \in \ker B \quad \textit{and} \quad \hat{A} \geq A, \tag{2.4}$$

and

$$\hat{S} \leq B\hat{A}^{-1}B^T \leq \beta\, \hat{S}. \tag{2.5}$$

Then

$$\lambda_{max}(\hat{\mathcal{K}}^{-1}\mathcal{K}) \leq \beta + \sqrt{\beta^2 - \beta} = \beta\,(1 + \sqrt{1 - 1/\beta})$$

and

$$\lambda_{min}(\hat{\mathcal{K}}^{-1}\mathcal{K}) \geq \frac{1}{2}\left[2 + \alpha - 1/\beta - \sqrt{(2 + \alpha - 1/\beta)^2 - 4\alpha}\right] \geq \alpha\, \left[\frac{2}{\sqrt{1 - 1/\beta} + \sqrt{5 - 1/\beta}}\right]^2 > 0.$$

Proof. The upper bound directly follows from Theorem 5.2 in Zulehner [19], again by considering the regularized matrices $A + \varepsilon I$ and $\hat{A} + \varepsilon I$ for $\varepsilon > 0$ with $\varepsilon \to 0$.

For the lower bound we consider an eigenvalue λ of the matrix $\hat{\mathcal{K}}^{-1}\mathcal{K}$:

$$\mathcal{K}\begin{pmatrix} x \\ p \end{pmatrix} = \lambda\, \hat{\mathcal{K}}\begin{pmatrix} x \\ p \end{pmatrix},$$

which is equivalent to the eigenvalue problem

$$\mathcal{K}\begin{pmatrix} x \\ p \end{pmatrix} = \mu\, \mathcal{D}\begin{pmatrix} x \\ p \end{pmatrix}$$

with

$$\lambda = \frac{\mu}{1 + \mu} \quad \text{and} \quad \mathcal{D} = \hat{\mathcal{K}} - \mathcal{K} = \begin{pmatrix} \hat{A} - A & 0 \\ 0 & B\hat{A}^{-1}B^T - \hat{S} \end{pmatrix},$$

or, in an equivalent variational form:

$$\begin{aligned} (Ax, w) + (Bw, p) &= \mu\, ((\hat{A} - A)x, w) && \text{for all } w \in \mathbb{R}^n, \\ (Bx, q) &= \mu\, ((B\hat{A}^{-1}B^T - \hat{S})p, q) && \text{for all } q \in \mathbb{R}^m. \end{aligned}$$

Now, two cases are distinguished: Firstly, for the case $\mu \le 0$, it follows that $\lambda = \mu/(1+\mu) > 1$, since λ must be positive by Theorem 2.1. (The case $\mu = -1$ can be excluded, since $\hat{\mathcal{K}}$ is nonsingular.) So, in this case, the eigenvalues λ are bounded from below by 1.

Next, we consider the remaining case $\mu > 0$: Let

$$W = \ker B, \quad W^{\perp} = \{x \in \mathbb{R}^n : (\hat{A}x, w) = 0 \text{ for all } w \in W\}.$$

Then there is a unique representation of x of the following form:

$$x = x_1 + x_2 \quad \text{with } x_1 \in W \text{ and } x_2 \in W^{\perp}.$$

Now the variational form reads:

$$\begin{aligned} (Ax_1, w_1) + (Ax_2, w_1) &= \mu\, \Big[((\hat{A} - A)x_1, w_1) - (Ax_2, w_1)\Big] && \text{for all } w_1 \in W, \\ (Ax_1, w_2) + (Ax_2, w_2) + (Bw_2, p) &= \mu\, \Big[-(Ax_1, w_2) + ((\hat{A} - A)x_2, w_2)\Big] && \text{for all } w_2 \in W^{\perp}, \\ (Bx_2, q) &= \mu\, ((B\hat{A}^{-1}B^T - \hat{S})p, q) && \text{for all } q \in \mathbb{R}^m. \end{aligned}$$

From the first equation we obtain for $w_1 = x_1$:

$$\alpha\, (x_1, x_1)_{\hat{A}} \le (Ax_1, x_1) = \mu\, ((\hat{A} - A)x_1, x_1) - (\mu + 1)(Ax_2, x_1).$$

Using

$$\begin{aligned} |(Aw_2, w_1)| = |((\hat{A} - A)w_2, w_1)| &\le ((\hat{A} - A)w_1, w_1)^{1/2}((\hat{A} - A)w_2, w_2)^{1/2} \\ &\le \sqrt{1-\alpha}\, \|w_1\|_{\hat{A}}\, \|w_2\|_{\hat{A}} \quad \text{for all } w_1 \in W,\ w_2 \in W^{\perp}, \end{aligned}$$

it follows that

$$\alpha\, (x_1, x_1)_{\hat{A}} \le \mu\, (1 - \alpha)\, (x_1, x_1)_{\hat{A}} + (\mu + 1)\sqrt{1-\alpha}\, \|x_1\|_{\hat{A}} \|x_2\|_{\hat{A}},$$

which implies

$$\alpha\, \|x_1\|_{\hat{A}} \le \mu\, (1 - \alpha)\, \|x_1\|_{\hat{A}} + (\mu + 1)\sqrt{1-\alpha}\, \|x_2\|_{\hat{A}}.$$

From the second equation we obtain

$$\sup_{w_2 \in W^{\perp}} \frac{(Bw_2, p)}{\|w_2\|_{\hat{A}}} = \sup_{w_2 \in W^{\perp}} \frac{-(\mu+1)(Ax_1, w_2) + ((\mu(\hat{A} - A) - A)x_2, w_2)}{\|w_2\|_{\hat{A}}}.$$

Using

$$|(Ax_1, w_2)| = |(Aw_2, x_1)| \le \sqrt{1-\alpha}\, \|x_1\|_{\hat{A}}\, \|w_2\|_{\hat{A}}$$

and

$$\begin{aligned} |(\mu(\hat{A} - A) - A)x_2, w_2)| &= |(\hat{A}^{-1}[\mu(\hat{A} - A) - A])x_2, w_2)_{\hat{A}}| \\ &\le \|\hat{A}^{-1}[\mu(\hat{A} - A) - A]\|_{\hat{A}} \|x_2\|_{\hat{A}}\, \|w_2\|_{\hat{A}} \end{aligned}$$

with

$$\|\hat{A}^{-1}[\mu(\hat{A} - A) - A]\|_{\hat{A}} \le \mu\, \|\hat{A}^{-1}(\hat{A} - A)\|_{\hat{A}} + \|\hat{A}^{-1}A\|_{\hat{A}} \le \mu + 1,$$

it follows that

$$\sup_{w_2 \in W^\perp} \frac{(Bw_2, p)}{\|w_2\|_{\hat{A}}} \le (\mu + 1)\sqrt{1-\alpha}\,\|x_1\|_{\hat{A}} + (\mu+1)\,\|x_2\|_{\hat{A}}.$$

From the third equation we obtain

$$\sup_{0 \ne q} \frac{(Bx_2, q)}{\|q\|_H} = \sup_{0 \ne q} \frac{\mu\,((B\hat{A}^{-1}B^T - \hat{S})p, q)}{\|q\|_H} \le \mu\,(1 - 1/\beta)\,\|p\|_H.$$

Observe that, for the left-hand sides of the last two inequalities, we have the following well-known representation:

$$\sup_{w_2 \in W^\perp} \frac{(Bw_2, p)}{\|w_2\|_{\hat{A}}} = \sup_{w \in \mathbb{R}^n} \frac{(Bw, p)}{\|w\|_{\hat{A}}} = (B\hat{A}^{-1}B^T p, p)^{1/2} = \|p\|_H$$

and

$$\sup_{0 \ne q \in \mathbb{R}^m} \frac{(Bx_2, q)}{\|q\|_H} = (B^T H^{-1} B x_2, x_2)^{1/2} = (\hat{A}^{-1} B^T H^{-1} B x_2, x_2)_{\hat{A}}^{1/2} = (x_2, x_2)_{\hat{A}}^{1/2} = \|x_2\|_{\hat{A}},$$

since $P = \hat{A}^{-1}B^T H^{-1} B$ is a projection onto $W^\perp$, so $Px_2 = x_2$ for $x_2 \in W^\perp$.

Hence, in summary,

$$\begin{pmatrix} \alpha & -\sqrt{1-\alpha} & 0 \\ -\sqrt{1-\alpha} & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \|x_1\|_{\hat{A}} \\ \|x_2\|_{\hat{A}} \\ \|p\|_H \end{pmatrix} \le \mu \begin{pmatrix} 1-\alpha & \sqrt{1-\alpha} & 0 \\ \sqrt{1-\alpha} & 1 & 0 \\ 0 & 0 & 1 - 1/\beta \end{pmatrix} \begin{pmatrix} \|x_1\|_{\hat{A}} \\ \|x_2\|_{\hat{A}} \\ \|p\|_H \end{pmatrix},$$

or, in short:

$$Ke \le \mu\, De \tag{2.6}$$

with

$$K = \begin{pmatrix} \alpha & -\sqrt{1-\alpha} & 0 \\ -\sqrt{1-\alpha} & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \; D = \begin{pmatrix} 1-\alpha & \sqrt{1-\alpha} & 0 \\ \sqrt{1-\alpha} & 1 & 0 \\ 0 & 0 & 1 - 1/\beta \end{pmatrix}, \; e = \begin{pmatrix} \|x_1\|_{\hat{A}} \\ \|x_2\|_{\hat{A}} \\ \|p\|_H \end{pmatrix}.$$

Since K^{-1} is non-negative element-wise, it follows

$$e \le \mu\, K^{-1} De.$$

Elementary calculations show that

$$\nu_+ = \frac{1}{2\alpha}\left[2 - \alpha - 1/\beta + \sqrt{(2 - \alpha - 1/\beta)^2 + 4\alpha(1 - 1/\beta)}\right]$$

is a non-negative eigenvalue of $K^{-1}D$ with component-wise non-negative left eigenvector l_+^T, given by

$$l_+^T = \left(\sqrt{1-\alpha}, 1, \alpha\nu_+ - 1 + \alpha\right).$$

Then

$$l_+^T e \le \mu\nu_+ l_+^T e.$$

Obviously, $l_+^T e \ge 0$. One can easily show that $\nu_+ > 0$ and $l_+^T e > 0$: $\nu_+ = 0$ implies $\alpha = \beta = 1$, then (2.6) implies $e = 0$. In a similar way the case $l_+^T e = 0$ can be excluded.

Therefore, after dividing by $l_+^T e > 0$, we obtain

$$\mu \geq \frac{1}{\nu_+}.$$

Consequently,

$$\begin{aligned}
\lambda = \frac{\mu}{1+\mu} \geq \frac{1}{1+\nu_+} &= \frac{1}{2}\left[2 + \alpha - 1/\beta - \sqrt{(2+\alpha-1/\beta)^2 - 4\alpha}\right] \\
&= \frac{2\alpha}{2+\alpha-1/\beta+\sqrt{(2+\alpha-1/\beta)^2-4\alpha}} \\
&\geq \frac{2\alpha}{3-1/\beta+\sqrt{(3-1/\beta)^2-4}} = \alpha\left[\frac{2}{\sqrt{1-1/\beta}+\sqrt{5-1/\beta}}\right]^2 > 0.
\end{aligned}$$

This lower bound is obviously smaller than 1, which was the lower bound for the first case $\mu \leq 0$. This completes the proof. □

By slightly strengthening the conditions (2.4) and (2.5) to

$$(Aw, w) \geq \alpha\,(\hat{A}w, w) \quad \text{for all } w \in \ker B \quad \text{and} \quad \hat{A} > A \tag{2.7}$$

and

$$\hat{S} < B\hat{A}^{-1}B^T \leq \beta\,\hat{S}, \tag{2.8}$$

the scalar product (2.3) is well-defined, and, by Theorem 2.1, the standard conjugate gradient method can be applied to the preconditioned system

$$\hat{\mathcal{K}}^{-1}\mathcal{K}\begin{pmatrix} x \\ p \end{pmatrix} = \hat{\mathcal{K}}^{-1}\begin{pmatrix} f \\ g \end{pmatrix} \tag{2.9}$$

with respect to the scalar product (2.3).

The actual construction of the preconditioners $\hat{A}$ and $\hat{S}$ is usually done in two steps. First, some preliminary candidates $\hat{A}_0$ and $\hat{S}_0$ are chosen which approximate the matrices A and $B\hat{A}_0^{-1}B^T$. In the second step, these candidates are properly scaled: $\hat{A} = (1/\sigma)\,\hat{A}_0$ and $\hat{S} = (\sigma/\tau)\,\hat{S}_0$, where the positive parameters σ and τ must be chosen such that (2.2) are satisfied, i.e.:

$$\frac{1}{\sigma}\,\hat{A}_0 > A \quad \text{and} \quad \frac{1}{\tau}\,\hat{S}_0 < B\hat{A}_0^{-1}B^T.$$

So, the correct choice of the parameters σ and τ require some rough information of the size of the largest eigenvalue of A relative to $\hat{A}_0$, which is, in general, quite easy to obtain, and of the size of the smallest eigenvalue of $B\hat{A}_0^{-1}B^T$ relative to $\hat{S}_0$, which, in general, is more costly, but which is available here from the analysis for the problem discussed in Section 4. The values of α and β in (2.7) and (2.8) are not needed for the construction, but only for the analysis.

It is well-known, e.g. Hackbusch [13], that the error $e^{(k)}$ for the k-th iterate $(x^{(k)}, p^{(k)})^T$ measured in the corresponding energy norm can be estimated by

$$e^{(k)} \leq \frac{2q^k}{1+q^{2k}}\,e^{(0)} \quad \text{with} \quad q = \frac{\sqrt{\kappa(\hat{\mathcal{K}}^{-1}\mathcal{K})} - 1}{\sqrt{\kappa(\hat{\mathcal{K}}^{-1}\mathcal{K})} + 1},$$

where $\kappa(\hat{\mathcal{K}}^{-1}\mathcal{K})$ denotes the relative condition number:

$$\kappa(\hat{\mathcal{K}}^{-1}\mathcal{K}) = \frac{\lambda_{\max}(\hat{\mathcal{K}}^{-1}\mathcal{K})}{\lambda_{\min}(\hat{\mathcal{K}}^{-1}\mathcal{K})}.$$

From Theorem 2.2 the following upper bound for the relative condition number follows:

$$\begin{aligned}\kappa(\hat{\mathcal{K}}^{-1}\mathcal{K}) &\leq \frac{2(\beta+\sqrt{\beta^2-\beta})}{2+\alpha-1/\beta-\sqrt{(2+\alpha-1/\beta)^2-4\alpha}} \equiv \kappa(\alpha,\beta) \\ &\leq \frac{\beta}{\alpha}\,(1+\sqrt{1-1/\beta})\left[\frac{\sqrt{1-1/\beta}+\sqrt{5-1/\beta}}{2}\right]^2 .\end{aligned}$$

This shows that the convergence rate q can be bounded by α and β only. If the preconditioners are chosen such that α and β are independent of certain parameters like the mesh size h of some discretization or some involved regularization parameter ν, then the convergence rate is also robust with respect to such parameters.

Furthermore, for $\alpha \to 1$ and $\beta \to 1$, the lower and upper bounds for the eigenvalues in Theorem 2.2 both approach 1, (implying that all eigenvalues of the preconditioned matrix $\hat{\mathcal{K}}^{-1}\mathcal{K}$ approach 1,) leading to a relative condition number approaching 1 and a convergence factor q approaching 0.

In the limit case $\alpha = 1$ and $\beta = 1$ one can easily derive the following representations for the preconditioners from the conditions (2.4) and (2.5):

$$\hat{A} = A + B^T\, W\, B \quad \text{and} \quad \hat{S} = B\hat{A}^{-1}B$$

for some matrix $W \geq 0$. Then, we obtain for $\hat{\mathcal{K}}$:

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

From the considerations above it follows in this case that all eigenvalue of $\hat{\mathcal{K}}^{-1}\mathcal{K}$ must be equal to 1. Moreover, it can easily be shown that

$$\left[I - \hat{\mathcal{K}}^{-1}\mathcal{K}\right]^2 = 0.$$

So, the corresponding preconditioned Richardson method terminates at the solution after two steps.

In a simplified way one could describe the proposed strategy as follows: Good preconditioners $\hat{A}$ can be interpreted as good approximations to some augmented matrix $A + B^TWB$, but we do not change the matrix A itself in the system matrix $\mathcal{K}$. This seems to be only a slight variant to the augmented Lagrangian approach, where first A itself is replaced by $A+B^TWB$ in $\mathcal{K}$. However, the actual construction of the preconditioner is not based on selecting first some augmentation matrix W and then preconditioning the augmented matrix. Instead, as it will be detailed in the next section, the construction is guided by the analysis of an underlying (infinite-dimensional) variational problem, whose discretization leads to the discussed large scale linear systems of equations in saddle point form.

3. Application to mixed variational problems. Consider an (infinite-dimensional) mixed variational problem of the following form:
Find $x \in X$ and $p \in Q$ such that

$$\begin{aligned} a(x,w) \,+\, b(w,p) &= \langle F, w\rangle \quad \text{for all } w \in X,\\ b(x,q) \qquad\qquad &= \langle G, q\rangle \quad \text{for all } q \in Q. \end{aligned}$$

Here, X and Q are real Hilbert spaces, $a : X \times X \longrightarrow \mathbb{R}$ and $b : X \times Q \longrightarrow \mathbb{R}$ are bilinear forms, $F : X \longrightarrow \mathbb{R}$ and $G : Q \longrightarrow \mathbb{R}$ are continuous linear functionals, and $\langle F, w\rangle$ ($\langle G, q\rangle$) denotes the evaluation of F (G) at the element w (q).

The existence and uniqueness of a solution to this mixed variational problem is well-established (theorem of Brezzi, see Brezzi and Fortin [6]) under the following conditions:

1. The bilinear form a is bounded:

$$a(x,w) \le \|a\|\,\|x\|_X \|w\|_X \quad \text{for all } x, w \in X.$$

2. The bilinear form a is coercive on $\ker B = \{w \in X : b(w,q) = 0 \text{ for all } q \in Q\}$: There exists a constant $\alpha_0 > 0$ such that

$$a(w,w) \ge \alpha_0\,\|w\|_X^2 \quad \text{for all } w \in \ker B.$$

3. The bilinear form b is bounded:

$$\sup_{0 \ne w \in X} \frac{b(w,q)}{\|w\|_X} \le \|b\|\,\|q\|_Q \quad \text{for all } q \in Q.$$

4. The bilinear form b satisfies the inf-sup condition: There exists a constant $k_0 > 0$ such that

$$\sup_{0 \ne w \in X} \frac{b(w,q)}{\|w\|_X} \ge k_0\,\|q\|_Q \quad \text{for all } q \in Q.$$

Under the additional assumptions that

5. the bilinear form a is symmetric on X:

$$a(x,w) = a(w,x) \quad \text{for all } x, w \in X, \text{ and}$$

6. the bilinear form a is non-negative on X:

$$a(w,w) \ge 0 \quad \text{for all } w \in X,$$

the theorem of Brezzi implies the equivalence of the mixed variational problem to the following constrained optimization problem:
Find $x \in X$ such that

$$J(x) = \min_{w \in X_g} J(w) \tag{3.1}$$

with

$$J(w) = \frac{1}{2} a(w,w) - \langle F, w\rangle$$

and

$$X_g = \{w \in X : b(w,q) = \langle G, q\rangle \text{ for all } q \in Q\}.$$

For discretizing the infinite-dimensional problem the spaces X and Q are replaced by finite-dimensional subspaces $X_h \subset X$ and $Q_h \subset Q$, which results in the following finite-dimensional variational problem:
Find $x_h \in X_h$ and $p_h \in Q_h$ such that

$$\begin{aligned} a(x_h, w_h) + b(w_h, p_h) &= \langle F, w_h\rangle && \text{for all } w_h \in X_h,\\ b(x_h, q_h) &= 0 && \text{for all } q_h \in Q_h. \end{aligned}$$

By introducing suitable basis functions in X_h and Q_h, we finally obtain the following saddle point problem in matrix-vector notation:

$$\begin{aligned} A_h \underline{x}_h + B_h^T \underline{p}_h &= \underline{f}_h,\\ B_h \underline{x}_h &= \underline{g}_h, \end{aligned}$$

where $\underline{x}_h$ and $\underline{p}_h$ denote the corresponding vectors of coefficients with respect to these basis functions.

We assume that the conditions of the theorem of Brezzi are also satisfied in X_h and Q_h. This is trivial for the first and the third condition. The second and fourth condition must be proven for the particular equations and elements. To simplify the notation the same symbols are used to denote the constants.

These conditions read in matrix-vector notations:

$$A_h \le \|a\| \, \underline{X}_h, \tag{3.2}$$

$$(A_h, \underline{w}_h, \underline{w}_h) \ge \alpha_0 \, (\underline{X}_h \underline{w}_h, \underline{w}_h) \quad \text{for all } \underline{w}_h \in \ker B_h, \tag{3.3}$$

$$B_h \underline{X}_h^{-1} B_h^T \le \|b\|^2 \, \underline{Q}_h, \tag{3.4}$$

$$B_h \underline{X}_h^{-1} B_h^T \ge k_0^2 \, \underline{Q}_h. \tag{3.5}$$

Here, $\underline{X}_h$ and $\underline{Q}_h$ denote the matrices representing the scalar products $(x, q)_X$ and $(p, q)_Q$ as bilinear forms on X_h and Q_h, respectively:

$$(x_h, w_h)_X = (\underline{X}_h \underline{x}_h, \underline{w}_h), \quad (p_h, q_h)_Q = (\underline{Q}_h \underline{p}_h, \underline{q}_h).$$

For the third and fourth condition we used the well-known representation

$$\sup_{0 \ne w_h \in X_h} \frac{b(w_h, q_h)}{\|w_h\|_X} = (B_h \underline{X}_h^{-1} B_h^T \underline{q}_h, \underline{q}_h)^{1/2}.$$

Comparing with the conditions (2.7) and (2.8) it is reasonable to choose for $\hat{A}_h$ a properly scaled preconditioner of $\underline{X}_h$, and for $\hat{S}_h$ a properly scaled preconditioner of $\underline{Q}_h$, say

$$\hat{A}_h = \frac{1}{\sigma} \hat{X}_h \quad \text{and} \quad \hat{S}_h = \frac{\sigma}{\tau} \hat{Q}_h \tag{3.6}$$

for some real parameters $\sigma > 0$ and $\tau > 0$ and with preconditioners $\hat{X}_h$ and $\hat{Q}_h$ satisfying, e.g., the spectral estimates

$$(1 - q_X) \, \hat{X}_h \le \underline{X}_h \le \hat{X}_h \quad \text{and} \quad (1 - q_Q) \, \hat{Q}_h \le \underline{Q}_h \le \hat{Q}_h. \tag{3.7}$$

The constants $q_X, q_Q \in [0, 1)$ describe the quality of these preconditioners. The smaller these constants are the better the preconditioners are.

Combining all estimates we easily obtain the following result:

LEMMA 3.1. *Assume that (3.2) - (3.7) hold. Then the conditions (2.7) and (2.8) are satisfied with*

$$\alpha = \sigma \, (1 - q_X) \, \alpha_0 \quad \textit{and} \quad \beta = \tau \, \|b\|^2,$$

if the parameters σ and τ are chosen such that

$$\sigma < \frac{1}{\|a\|} \quad \textit{and} \quad \tau > \frac{1}{(1 - q_X)(1 - q_Q) k_0^2}.$$

Proof. We have

$$A_h \le \|a\| \, \underline{X}_h \le \|a\| \, \hat{X}_h = \sigma \, \|a\| \, \hat{A}_h < \hat{A}_h$$

if $\sigma < 1/\|a\|$. Next

$$(A_h \underline{w}_h, \underline{w}_h) \ge \alpha_0 \, (\underline{X}_h \underline{w}_h, \underline{w}_h) \ge (1 - q_X) \, \alpha_0 \, (\hat{X}_h \underline{w}_h, \underline{w}_h) = \alpha \, (\hat{A}_h \underline{w}_h, \underline{w}_h)$$

with $\alpha = \sigma\,(1 - q_X)\,\alpha_0$. Next

$$B_h \hat{A}_h^{-1} B_h^T = \sigma\, B_h \hat{X}_h^{-1} B_h^T \le \sigma\, B_h \underline{X}_h^{-1} B_h^T \le \sigma\, \|b\|^2\, \underline{Q}_h \le \sigma\, \|b\|^2\, \hat{Q}_h = \beta\, \hat{S}_h$$

with $\beta = \tau\, \|b\|^2$. Finally

$$\begin{aligned} B_h \hat{A}_h^{-1} B_h^T = \sigma\, B_h \hat{X}_h^{-1} B_h^T &\ge \sigma\,(1 - q_X)\, B_h \underline{X}_h^{-1} B_h^T \ge \sigma\,(1 - q_X)\, k_0^2\, \underline{Q}_h \\ &\ge \sigma\,(1 - q_X)\,(1 - q_Q)\, k_0^2\, \hat{Q}_h = \tau\,(1 - q_X)\,(1 - q_Q)\, k_0^2\, \hat{S}_h > \hat{S}_h \end{aligned}$$

if $\tau > 1/[(1 - q_X)(1 - q_Q)k_0^2]$. □

Good and efficient preconditioners $\hat{X}_h$ and $\hat{Q}_h$ are usually available, as it will be shown for a particular problem in the next section. Therefore, the quantities q_X and q_Q are typically small, say 0.1.

Roughly speaking, the parameter σ has to be sufficiently small, while the parameter τ has to be sufficiently large in order to guarantee the conditions (2.7) and (2.8). On the other hand, in order to obtain a small upper bound $\kappa(\alpha, \beta)$ for the condition number of the preconditioned matrix $\hat{\mathcal{K}}^{-1}\mathcal{K}$, α should be as large as possible and β should be as small as possible, i.e.: σ should be as large as possible and τ should be as small as possible. This, of course, requires at least a rough quantitative knowledge of the constants $\|a\|$ and k_0, which are involved in the choice of σ and τ.

Next, we will study a particular problem from optimal control, where the parameters $\|a\|$, α_0, $\|b\|$, and k_0 are known:

4. A problem from optimal control.

Let $\Omega \subset \mathbb{R}^d$ be an open and bounded set. We consider the following optimization problem with PDE-constraints:

Find the state $y \in H^1(\Omega)$ and the control $u \in L^2(\Omega)$ such that

$$J(y, u) = \min_{(z,v) \in H^1(\Omega) \times L^2(\Omega)} J(z, v),$$

subject to the state equation with distributed control u

$$\begin{aligned} -\Delta y + y &= u \quad \text{in } \Omega, \\ \frac{\partial y}{\partial n} &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where the cost functional is given by

$$J(y, u) = \frac{1}{2} \int_\Omega (y - y_d)^2 \, dx + \frac{\nu}{2} \int_\Omega u^2 \, dx.$$

More precisely, we prescribe the state equation in weak form:

$$\int_\Omega \nabla y \cdot \nabla q \, dx + \int_\Omega y\, q \, dx = \int_\Omega u\, q \, dx \quad \text{for all } q \in H^1(\Omega).$$

Let $X = Y \times U$ with $Y = H^1(\Omega)$, $U = L^2(\Omega)$ and $Q = H^1(\Omega)$. With $x = (y, u) \in X$, $w = (z, v) \in X$ and $q \in Q$ we introduce the following bilinear forms and linear functionals:

$$\begin{aligned} a(x, w) &= \int_\Omega y\, z \, dx + \nu \int_\Omega u\, v \, dx, \\ b(w, q) &= \int_\Omega \nabla z \cdot \nabla q \, dx + \int_\Omega z\, q \, dx - \int_\Omega v\, q \, dx, \\ \langle F, w \rangle &= \int_\Omega y_d\, z \, dx, \\ \langle G, q \rangle &= 0. \end{aligned}$$

With this setting the optimization problem is of the standard form (3.1).

The conditions of the theorem of Brezzi can easily be verified for the Hilbert spaces $X = Y \times U$ and Q introduced above and equipped with the standard scalar products $(y,z)_{H^1(\Omega)}$ in Y, $(u,v)_{L^2(\Omega)}$ in U and $(p,q)_{H^1(\Omega)}$ in Q. Then, however, the parameters $\|a\|$, α_0, $\|b\|$ and k_0 depend on the regularization parameter ν, eventually resulting in convergence rates also depending on ν.

With a different scaling of the scalar products in Y, U and Q we obtain parameters $\|a\|$, α_0, $\|b\|$ and k_0 independent of ν, eventually leading to preconditioners with convergence rates robust in ν: In particular, we consider the following new scalar products $(y,z)_Y$ in $Y = H^1(\Omega)$, $(u,v)_U$ in $U = L^2(\Omega)$ and $(p,q)_Q$ in $Q = H^1(\Omega)$:

$$(y,z)_Y = (y,z)_{L^2(\Omega)} + \sqrt{\nu}\,(y,z)_{H^1(\Omega)}, \quad (u,v)_U = \nu\,(u,v)_{L^2(\Omega)}$$

and

$$(p,q)_Q = \frac{1}{\nu}\,(p,q)_{L^2(\Omega)} + \frac{1}{\sqrt{\nu}}\,(p,q)_{H^1(\Omega)}$$

and we set $(x,w)_X = (y,z)_Y + (u,v)_U$ for $x = (y,u), w = (z,v) \in X = Y \times U$. Observe that the corresponding new norms are equivalent to the standard norms in these spaces for fixed $\nu > 0$.

With these definitions of the scalar products the following properties can be verified:

Lemma 4.1.

1. *The bilinear form a is bounded:*

$$a(x,w) \le \|x\|_X \|w\|_X \quad \textit{for all } x, w \in X.$$

2. *The bilinear form a is coercive on* $\ker B$*:*

$$a(w,w) \ge \alpha_0\,\|w\|_X^2 \quad \textit{for all } w \in \ker B \quad \textit{with } \alpha_0 = \frac{2}{3}.$$

3. *The bilinear form b is bounded:*

$$\sup_{0 \ne w \in X} \frac{b(w,q)}{\|w\|_X} \le \|q\|_Q \quad \textit{for all } q \in Q.$$

4. *The bilinear form b satisfies the inf-sup condition:*

$$\sup_{0 \ne w \in X} \frac{b(w,q)}{\|w\|_X} \ge k_0\,\|q\|_Q \quad \textit{with } k_0 = \sqrt{\frac{3}{4}}.$$

Proof. (1) is trivial since a is symmetric and $a(w,w) \le \|w\|_X^2$. For (2) take $w = (z,v) \in \ker B$. Then:

$$(z,q)_{H^1(\Omega)} = (v,q)_{L^2(\Omega)} \quad \text{for all } q \in H^1(\Omega).$$

In particular, it follows for $q = z$:

$$\|z\|_{H^1(\Omega)}^2 = (v,z)_{L^2(\Omega)} \le \|v\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)},$$

which implies

$$\|w\|_X^2 = \|z\|_Y^2 + \|v\|_U^2 \le \|z\|_{L^2(\Omega)}^2 + \sqrt{\nu}\,\|z\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \nu\,\|v\|_{L^2(\Omega)}^2.$$

Then

$$a(w,w) \ge \alpha_0\,\|w\|_X^2$$

is certainly satisfied if

$$a(w,w) \ge \alpha_0\,\left[\|z\|_{L^2(\Omega)}^2 + \sqrt{\nu}\,\|z\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \nu\,\|v\|_{L^2(\Omega)}^2\right],$$

which is equivalent to

$$(1-\alpha_0)\,\|z\|^2_{L^2(\Omega)} - \alpha_0\,\sqrt{\nu}\,\|z\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + (1-\alpha_0)\,\nu\,\|v\|^2_{L^2(\Omega)} \geq 0.$$

This is obviously the case for $\alpha_0 = 2/3$, since

$$\frac{1}{3}\,\|z\|^2_{L^2(\Omega)} - \frac{2}{3}\,\sqrt{\nu}\,\|z\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \frac{1}{3}\,\nu\,\|v\|^2_{L^2(\Omega)} = \frac{1}{3}\,\left[\|z\|_{L^2(\Omega)} - \sqrt{\nu}\|v\|_{L^2(\Omega)}\right]^2.$$

To show (3) and (4) we start with the following formula:

$$\begin{aligned}\sup_{0\neq w\in X}\frac{b(w,q)^2}{\|w\|_X^2} &= \sup_{0\neq(z,v)\in Y\times U}\frac{\left[(z,q)_{H^1(\Omega)} - (v,q)_{L^2(\Omega)}\right]^2}{\|z\|_Y^2 + \|v\|_U^2}\\ &= \sup_{0\neq z\in Y}\frac{(z,q)^2_{H^1(\Omega)}}{\|z\|_Y^2} + \sup_{0\neq v\in U}\frac{(v,q)^2_{L^2(\Omega)}}{\|v\|_U^2}\\ &= \sup_{0\neq z\in Y}\frac{(z,q)^2_{H^1(\Omega)}}{\|z\|_Y^2} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)}.\end{aligned}$$

Then (3) easily follows from the estimates

$$\begin{aligned}\sup_{0\neq z\in Y}\frac{(z,q)^2_{H^1(\Omega)}}{\|z\|_Y^2} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)} &\leq \sup_{0\neq z\in Y}\frac{\|z\|^2_{H^1(\Omega)}\|q\|^2_{H^1(\Omega)}}{\|z\|_Y^2} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)}\\ &= \sup_{0\neq z\in Y}\frac{\|z\|^2_{H^1(\Omega)}\|q\|^2_{H^1(\Omega)}}{\|z\|^2_{L^2(\Omega)} + \sqrt{\nu}\,\|z\|^2_{H^1(\Omega)}} + \frac{1}{\nu}\|q\|^2_{L^2(\Omega)}\\ &\leq \frac{1}{\sqrt{\nu}}\|q\|^2_{H^1(\Omega)} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)} = \|q\|_Q^2.\end{aligned}$$

For (4) observe that:

$$\begin{aligned}\sup_{0\neq z\in Y}\frac{(z,q)^2_{H^1(\Omega)}}{\|z\|_Y^2} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)} &\geq \frac{\|q\|^4_{H^1(\Omega)}}{\|q\|_Y^2} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)}\\ &= \frac{\|q\|^4_{H^1(\Omega)}}{\|q\|^2_{L^2(\Omega)} + \sqrt{\nu}\,\|q\|^2_{H^1(\Omega)}} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)}.\end{aligned}$$

Then the inf-sup condition

$$\sup_{0\neq w\in X}\frac{b(w,q)}{\|w\|_X} \geq k_0\,\|q\|_Q$$

is certainly satisfied if

$$\frac{\|q\|^4_{H^1(\Omega)}}{\|q\|^2_{L^2(\Omega)} + \sqrt{\nu}\,\|q\|^2_{H^1(\Omega)}} + \frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)} \geq k_0^2\,\|q\|_Q^2 = k_0^2\,\left[\frac{1}{\nu}\,\|q\|^2_{L^2(\Omega)} + \frac{1}{\sqrt{\nu}}\,\|q\|^2_{H^1(\Omega)}\right],$$

which is equivalent to

$$(1-k_0^2)\,\|q\|^4_{H^1(\Omega)} + (1-2k_0^2)\frac{1}{\sqrt{\nu}}\,\|q\|^2_{L^2(\Omega)}\|q\|^2_{H^1(\Omega)} + (1-k_0^2)\frac{1}{\nu}\,\|q\|^4_{L^2(\Omega)} \geq 0.$$

This is obviously the case for $k_0^2 = 3/4$ since

$$\frac{1}{4}\,\|q\|^4_{H^1(\Omega)} - \frac{1}{2}\frac{1}{\sqrt{\nu}}\,\|q\|^2_{L^2(\Omega)}\|q\|^2_{H^1(\Omega)} + \frac{1}{4}\frac{1}{\nu}\,\|q\|^4_{L^2(\Omega)} = \frac{1}{4}\left[\|q\|^2_{H^1(\Omega)} - \frac{1}{\sqrt{\nu}}\,\|q\|^2_{L^2(\Omega)}\right]^2.$$

□

By the theorem of Brezzi it now follows that the optimization problem is equivalent to the following mixed variational problem:
Find $x \in H^1(\Omega) \times L^2(\Omega)$ and $p \in H^1(\Omega)$ such that

$$\begin{aligned} a(x,w) + b(w,p) &= \langle F, x\rangle \quad \text{for all } w \in H^1(\Omega) \times L^2(\Omega), \\ b(x,q) \qquad\qquad &= 0 \qquad\quad\;\; \text{for all } q \in H^1(\Omega). \end{aligned}$$

For the spaces $Y_h = U_h = Q_h$ we choose, as an example, the space of piecewise linear and continuous functions on a simplicial subdivision of Ω. By introducing the standard nodal basis, we finally obtain the following saddle point problem in matrix-vector notation:

$$\begin{aligned} A_h \underline{x}_h + B_h^T \underline{p}_h &= \underline{f}_h, \\ B_h \underline{x}_h \qquad\quad &= 0, \end{aligned}$$

with

$$A_h = \begin{pmatrix} M_h & 0 \\ 0 & \nu\, M_h \end{pmatrix} \quad \text{and} \quad B_h = \begin{pmatrix} K_h & -M_h \end{pmatrix},$$

where M_h denotes the mass matrix representing the $L^2(\Omega)$ inner product on Y_h, and K_h denotes the stiffness matrix representing the bilinear form (on Y) of the state equation, here $(\nabla y, \nabla q)_{L^2(\Omega)} + (y,q)_{L^2(\Omega)}$, on Y_h.

For the matrices $\underline{X}_h$ and $\underline{Q}_h$ representing the scalar products $(x,w)_X = (y,z)_Y + (u,v)_U$ and $(p,q)_Q$ on X_h and Q_h we obtain

$$\underline{X}_h = \begin{pmatrix} \underline{Y}_h & 0 \\ 0 & \nu\, M_h \end{pmatrix} \quad \text{and} \quad \underline{Q}_h = \frac{1}{\nu} \underline{Y}_h$$

with

$$\underline{Y}_h = \sqrt{\nu}\, K_h + (\sqrt{\nu} + 1)\, M_h.$$

Observe that $\underline{Y}_h$ is the stiffness matrix representing the bilinear form $\sqrt{\nu}\,(\nabla y, \nabla q)_{L^2(\Omega)} + (\sqrt{\nu} + 1)\,(y,q)_{L^2(\Omega)}$ on Y_h, which is of the same type as the bilinear form (on Y) of the state equation, but with modified coefficients.

It is easy to see that Lemma 4.1 remains valid with the same constants if Y, U, Q are replaced by the finite-dimensional spaces Y_h, U_h, Q_h, as long as $Y_h = Q_h \subset U_h$.

As discussed before, it is reasonable to use a (properly scaled) preconditioner for $\underline{X}_h$ to approximate $\hat{A}_h$, and to use a (properly scaled) preconditioner for $\underline{Q}_h$ to approximate $\hat{S}_h$. For $\underline{Y}_h$, which appears in the first diagonal block of $\underline{X}_h$ and in $\underline{Q}_h$, we use, e.g., a standard multigrid preconditioner $\hat{Y}_h$ for the second-order elliptic differential operator represented by the bilinear form $\sqrt{\nu}\,(\nabla y, \nabla q)_{L^2(\Omega)} + (\sqrt{\nu}+1)\,(y,q)_{L^2(\Omega)}$. For the well-conditioned matrix M_h, which appears in the second diagonal block of $\underline{X}_h$ a simple preconditioner $\hat{M}_h$, e.g. a few steps of a symmetric Gauss-Seidel iteration, is used. So, eventually we set

$$\hat{A}_h = \frac{1}{\sigma}\, \hat{X}_h = \frac{1}{\sigma} \begin{pmatrix} \hat{Y}_h & 0 \\ 0 & \nu\, \hat{M}_h \end{pmatrix} \quad \text{and} \quad \hat{S}_h = \frac{\sigma}{\tau}\, \frac{1}{\nu}\, \hat{Y}_h \tag{4.1}$$

with real parameters $\sigma > 0$ and $\tau > 0$.

In summary, the preconditioner

$$\hat{\mathcal{K}}_h = \begin{pmatrix} \hat{A}_h & B_h^T \\ B_h & B_h \hat{A}_h^{-1} B_h^T - \hat{S}_h \end{pmatrix}$$

for the matrix

$$\mathcal{K}_h = \begin{pmatrix} A_h & B_h^T \\ B_h & 0 \end{pmatrix}$$

is given by (4.1), where $\hat{Y}_h$ is a preconditioner for the second-order elliptic differential operator represented by the bilinear form $\sqrt{\nu}\,(\nabla y, \nabla q)_{L^2(\Omega)}+(\sqrt{\nu}+1)\,(y,q)_{L^2(\Omega)}$ and a simple preconditioner $\hat{M}_h$ for the mass matrix.

It is reasonable to assume that

$$(1-q_X)\,\hat{Y}_h \le \underline{Y}_h \le \hat{Y}_h \quad \text{and} \quad (1-q_X)\,\hat{M}_h \le M_h \le \hat{M}_h, \tag{4.2}$$

for some small value $q_X \in [0,1)$. The factor q_X describes the quality of the preconditioners $\hat{Y}_h$ and $\hat{M}_h$.

Then the discussion of the last section shows that the conditions (2.7) and (2.8) are satisfied with

$$\alpha = \sigma\,(1-q_X)\,\frac{2}{3} \quad \text{and} \quad \beta = \tau$$

for parameters σ and τ satisfying

$$\sigma < 1 \quad \text{and} \quad \tau > \frac{4}{3(1-q_X)^2}.$$

In particular, assuming that $q_X \approx 0$, we can except $\alpha \approx 2/3$ and $\beta \approx 4/3$ for $\sigma \approx 1$ and $\tau \approx 4/3$, leading to a rough estimate of the condition number $\kappa \approx \kappa(2/3, 4/3) \approx 4$, which implies a convergence factor $q \approx 1/3$ for the conjugate gradient method.

5. Numerical Experiments.

We consider the optimal control problem from the previous section on the unit cube $\Omega = (0,1)^3$ and with homogeneous data $y_d \equiv 0$. Starting from an initial mesh of 24 tetrahedra (starting level $l = 1$) we obtain a hierarchy of nested meshes by uniform refinement up to some final level $l = L$. On each tetrahedral mesh piecewise linear and continuous finite elements are used for $Y_h = U_h = Q_h$.

The discretized mixed problem is solved on the finest mesh (level $l = L$) by using the conjugate gradient method for the preconditioned system (2.9) with the scalar product (2.3) as described before. For the preconditioner we used the proposed symmetric block preconditioner, where $\hat{Y}_h$ is one V-cycle of the multigrid method with m_1 forward Gauss-Seidel steps for pre-smoothing and m_1 backward Gauss-Seidel steps for post-smoothing (in short $V(m_1, m_1)$) for the second-order elliptic differential operator represented by the bilinear form $\sqrt{\nu}\,(\nabla y, \nabla q)_{L^2(\Omega)} + (\sqrt{\nu}+1)\,(y,q)_{L^2(\Omega)}$. For $\hat{M}_h$ we use m_2 steps of the symmetric Gauss-Seidel method (in short $SGS(m_2)$).

Starting values $\underline{x}_h^{(0)}$ and $\underline{p}_h^{(0)}$ are generated randomly. The exact solution of the problem is the trivial solution $\underline{x}_h = 0$ and $\underline{p}_h = 0$. The quality of an approximation $(\underline{x}_h^{(k)}, \underline{p}_h^{(k)})$ is measured either by the energy norm $e^{(k)}$ of the error, which here is given by

$$e^{(k)} = \left\| \begin{pmatrix} \underline{x}_h^{(k)} \\ \underline{p}_h^{(k)} \end{pmatrix} \right\|_{\mathcal{D}_h \hat{\mathcal{K}}_h^{-1} \mathcal{K}_h}$$

or the residual $r^{(k)}$:

$$r^{(k)} = \left\| \mathcal{K}_h \begin{pmatrix} \underline{x}_h^{(k)} \\ \underline{p}_h^{(k)} \end{pmatrix} \right\|.$$

Figure 5.1 shows a typical convergence history (number of iterations versus $e^{(k)}/e^{(0)}$ and $r^{(k)}/r^{(0)}$) for level $L = 5$ (number of unknowns 3 × 17.985) and regularization parameter $\nu = 1$

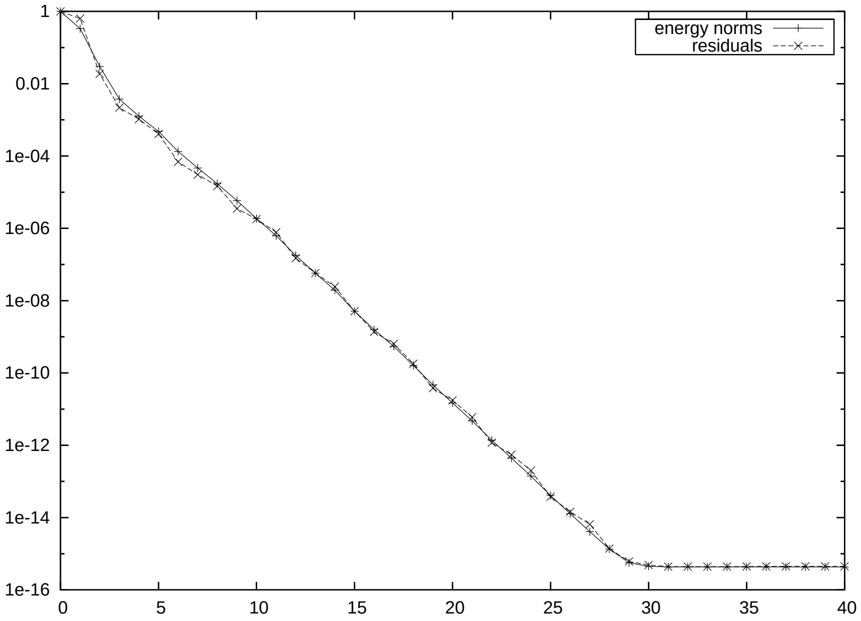

Fig. 5.1. *Convergence history: number of iterations versus relative accuracy.*

using a $V(3,3)$-cycle for $\hat{Y}_h$ and $SGS(3)$ for $\hat{M}_h$ and parameters $\sigma = 0.9$ and $\tau = 1.1/k_0^2$ with $k_0^2 = 3/4$.

Table 5.1 shows that the number of iterations does not depend on the level of refinement. L denotes the level of refinement, n the total number of all unknowns $\underline{y}_h$, $\underline{u}_h$ and $\underline{p}_h$, and k the number of iterations needed to satisfy the stopping rule

$$r^{(k)} \le \varepsilon\, r^{(0)} \quad \text{with } \varepsilon = 10^{-8}.$$

Table 5.1
Dependence of the number of iterations on the mesh size for fixed $\nu = 1$.

level L	number of unknown n	iterations k
3	1.107	14
4	7.395	15
5	53.955	15
6	412.035	16
7	3.200.227	15

Table 5.2 shows that the number of iterations does not depend on the regularization parameter ν either. The results are given for refinement level $L = 5$.

Table 5.2
Dependence of the number of iterations on ν for fixed refinement level $L = 5$.

ν	iterations k
10^{-4}	15
10^{-2}	14
1	15
10^{2}	14
10^{4}	15

6. Concluding remarks. Comparing the matrix $\mathcal{K}_h$ and the preconditioner $\hat{\mathcal{K}}_h$, a first remarkable observation is that the mass matrix M_h (representing the L^2 inner product on Y_h) in the first diagonal block of A_h is preconditioned by a preconditioner for a second-order elliptic differential operator. Of course, such a preconditioner cannot be a good preconditioner for M_h on the whole space Y_h, but it is a good preconditioner on the kernel of B_h, as it was shown. And this suffices for the convergence analysis.

A more straight forward alternative would be to use some lumped mass matrix for preconditioning M_h or even to use M_h itself, because it is well-conditioned, and, therefore, easy to invert. Then, however, the resulting inexact Schur can be interpreted as a discretized fourth-order elliptic differential operator, for which it is much harder to find an efficient preconditioner. With our choice of the preconditioner for the mass matrix, the inexact Schur complement remains a discretized second order differential operator of the same complexity as the discretized second-order differential operator of the state equation, for which an efficient preconditioner is usually available.

So in this context, it pays off to invest (a little) more on preconditioning the mass matrix by a (properly scaled) Laplace-type preconditioner instead of some simple preconditioner. This would normally be considered as a very obscure strategy. Here, however, it is a very natural thing to do. It just reflects the standard conditions of the theorem of Brezzi.

A second remarkable observation concerns the discussed problem from optimal control. For the considered case of distributed control, it was shown theoretically and confirmed experimentally that the proposed preconditioner leads to convergence rates not only robust with respect to the mesh size h but also robust with respect to the regularization parameter ν.

REFERENCES

[1] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.

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

[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical Solution of Saddle Point Problems.*, Acta Numerica, 14 (2005), pp. 1–137.

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

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

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, 1991.

[7] N. Dyn and W. E. Ferguson, *The numerical solution of equality-constrained quadratic programming problems.*, Math. Comput., 41 (1983), pp. 165–170.

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

[9] R. Fletcher, *Practical methods of optimization. Vol. 2: Constrained optimization.*, Chichester etc.: John Wiley & Sons, 1981.

[10] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Application to the Numerical Solution of Boundary–Value Problems*, North–Holland, Amsterdam, 1983.

[11] G. H. Golub and C. Greif, *On solving block-structured indefinite linear systems.*, SIAM J. Sci. Comput., 24 (2003), pp. 2076–2092.

[12] G. H. Golub, C. Greif, and J. M. Varah, *An algebraic analysis of a block diagonal preconditioner for saddle point problems.*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 779–792.

[13] W. Hackbusch, *Iterative Solutions of Large Sparse Systems of Equations*, Springer Verlag, New York, 1994.

[14] T. Rusten and R. Winther, *A preconditioned iterative method for saddle-point problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887 – 904.

[15] Y. Saad and H. A. van der Vorst, *Iterative solution of linear systems in the 20th century.*, J. Comput. Appl. Math., 123 (2000), pp. 1–33.

[16] D. Silvester and A. Wathen, *Fast iterative solutions of stabilized Stokes systems. Part II: Using block diagonal preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352 – 1367.

[17] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen*, Wiesbaden: Vieweg, 2005.

[18] P. Vassilevski and R. Lazarov, *Preconditioning mixed finite element saddle-point elliptic problems*, Numer. Linear Algebra Appl., 3 (1996), pp. 1 – 20.

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