PRECONDITIONING ITERATIVE METHODS FOR THE OPTIMAL CONTROL OF THE STOKES EQUATIONS

TYRONE REES*, ANDREW J. WATHEN†

Abstract. Solving problems regarding the optimal control of partial differential equations (PDEs) – also known as PDE-constrained optimization – is a frontier area of numerical analysis. Of particular interest is the problem of flow control, where one would like to effect some desired flow by exerting, for example, an external force. The bottleneck in many current algorithms is the solution of the optimality system – a system of equations in saddle point form that is usually very large and ill-conditioned. In this paper we describe two preconditioners – a block-diagonal preconditioner for the minimal residual method and a block-lower triangular preconditioner for a non-standard conjugate gradient method – which can be effective when applied to such problems where the PDEs are the Stokes equations. We consider only distributed control here, although other problems – for example boundary control – could be treated in the same way. We give numerical results, and compare these with those obtained by solving the equivalent forward problem using similar techniques.

1. Introduction. Suppose that we have a flow that satisfies the Stokes equations in some domain Ω with some given boundary condition, and that we have some mechanism – for example, the application of a magnetic field – to change the forcing term on the right hand side of the PDE. Let $\widehat{\vec{v}}$ and $\hat{p}$ be given functions which are called the 'desired states'. Then the question is how do we choose the forcing term such that the velocity $\vec{v}$ and pressure p are as close as possible to $\widehat{\vec{v}}$ and $\hat{p}$, in some sense, while still satisfying the Stokes equations.

One way of formulating this problem is by minimizing a cost functional of tracking-type with the Stokes equations as a constraint, as follows:

$$\min_{v,p,u} \frac{1}{2}\|\vec{v} - \widehat{\vec{v}}\|^2_{L^2(\Omega)} + \frac{\delta}{2}\|p - \hat{p}\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|\vec{u}\|^2_{L^2(\Omega)} \tag{1.1}$$

$$\begin{aligned} \text{s.t.} -\nabla^2\vec{v} + \nabla p &= \vec{u} \quad \text{in } \Omega \\ \nabla\cdot\vec{v} &= 0 \quad \text{in } \Omega, \\ \vec{v} &= \vec{w} \quad \text{on } \partial\Omega. \end{aligned}$$

Here $\vec{u}$ denotes the forcing term on the right hand side, which is known as the control. In order for the problem to be well-posed we also include the control in the cost functional, together with a Tikhonov regularization parameter β, which is usually chosen a priori. A constant δ is added in front of the desired pressure to enable us to penalize the pressure. We would normally take $\hat{p} = 0$. We specify a Dirichlet boundary condition with $\vec{v}$ taking some value $\widehat{\vec{w}}$ – which may or may not be taken from the desired state – on the boundary.

There are two methods with which one can discretize this problem – we can either discretize the equations first and then optimize that system, or alternatively carry out the optimization first and then discretize the resulting optimality system. Since the Stokes equations are self-adjoint we will get the same discrete optimality system either way, provided the discretization methods are consistent between equations in

*Department of Computer Science, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada (tyronere@cs.ubc.ca)

†Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, United Kingdom (wathen@maths.ox.ac.uk)

the optimize-then-discretize technique. We will therefore only consider the discretize-then-optimize approach here.

Let $\{\vec{\phi}_j\},\ j = 1, \ldots, n_v + n_\partial$ and $\{\psi_k\},\ k = 1, \ldots, n_p$ be sets of finite element basis functions that form a stable mixed finite element discretization for the Stokes equations – see, for example, [11, Chapter 5] for further details – and let $\vec{v}_h = \sum_{i=1}^{n_v+n_\partial} V_i\vec{\phi}_i$ and $p_h = \sum_{i=1}^{n_p} P_i\psi_i$ be finite-dimensional approximations to $\vec{v}$ and p. Furthermore, let us also approximate the control from the velocity space, so $\vec{u}_h = \sum_{i=1}^{n_v} U_i\vec{\phi}_i$. The discrete Stokes equation is of the form

$$\begin{bmatrix} \underline{K} & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} Q_{\vec{v}} \\ 0 \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix},$$

where $\mathbf{v}$, $\mathbf{p}$ and $\mathbf{u}$ are the coefficient vectors in the expansions of $\vec{v}_h$, p_h and $\vec{u}_h$ respectively, $\underline{K} = [\int_\Omega \nabla\vec{\phi}_i : \nabla\vec{\phi}_j]$, $B = [-\int_\Omega \psi_k \nabla\cdot\vec{\phi}_j]$, $Q_{\vec{v}} = [\int_\Omega \vec{\phi}_i\cdot\vec{\phi}_j]$, $\mathbf{f} = [-\sum_{j=n_u+1}^{n_u+n_\partial} V_j \int_\Omega \nabla\vec{\phi}_i : \nabla\vec{\phi}_j]$ and $\mathbf{g} = [\sum_{j=n_u+1}^{n_u+n_\partial} V_j \int_\Omega \psi_i\nabla\cdot\vec{\phi}_j]$. Note that the coefficents V_j, $j = n_{u+1}, \ldots, n_u + n_\partial$ are fixed so that $\vec{v}_h$ interpolates the boundary data $\vec{w}$.

On discretizing, the cost functional becomes

$$\min \frac{1}{2}\mathbf{v}^T Q_{\vec{v}}\mathbf{v} - \mathbf{v}^T\mathbf{b} + \frac{\delta}{2}\mathbf{p}^T Q_p\mathbf{p} - \delta\mathbf{p}^T\mathbf{d} + \frac{\beta}{2}\mathbf{u}^T Q_{\vec{v}}\mathbf{u}$$

where $Q_p = [\int_\Omega \psi_i\psi_j]$, $\mathbf{b} = [\int_\Omega \widehat{\vec{v}}\vec{\phi}_i]$ and $\mathbf{d} = [\int_\Omega \widehat{p}\psi_i]$.

Let us introduce two vectors of Lagrange multipliers, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Then finding a critical point of the Lagrangian function gives the discrete optimality system of the form

$$\begin{bmatrix} Q_{\vec{v}} & 0 & 0 & \underline{K} & B^T \\ 0 & \delta Q_p & 0 & B & 0 \\ 0 & 0 & \beta Q_{\vec{v}} & -Q_{\vec{v}}^T & 0 \\ \underline{K} & B^T & -Q_{\vec{v}} & 0 & 0 \\ B & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{p} \\ \mathbf{u} \\ \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \delta\mathbf{d} \\ \mathbf{0} \\ \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{1.2}$$

It will be useful to relabel this system so that it becomes

$$\begin{bmatrix} \mathcal{Q} & 0 & \mathcal{K} \\ 0 & \beta Q_{\vec{v}} & -\widehat{Q}^T \\ \mathcal{K} & -\widehat{Q} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\xi} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{0} \\ \mathbf{h} \end{bmatrix}, \tag{1.3}$$

where $\mathcal{Q} = \texttt{blkdiag}(Q_{\vec{v}}, \delta Q_p)$, $\mathcal{K} = \begin{bmatrix} \underline{K} & B^T \\ B & 0 \end{bmatrix}$, $\widehat{Q} = [Q_{\vec{v}}\ 0]^T$ and the vectors $\mathbf{y}$, $\boldsymbol{\xi}$, $\mathbf{c}$ and $\mathbf{h}$ take their obvious definitions. For more detail on the practicalities of discretizing control problems of this type see, for example, Rees, Stoll and Wathen [23]. Finding an efficient method to solve this system will be the topic of the remainder of the paper.

In Section 2 we introduce two preconditioners that can be applied to this problem; one block diagonal, which we apply using the minimal residual method (MINRES) of Paige and Saunders [20], and one block lower triangular, which we use with the Conjugate Gradient method (CG) of Hestenes and Steifel [15] applied with a non-standard inner product. Both of these methods rely on good approximations to the $(1,1)$-block and the Schur complement, and we discuss suitable choices in Sections 2.3 and 2.4 respectively. Finally, in Section 3 we give numerical results.

2. Solution methods. The matrix in (1.3) is of saddle point form, that is

$$\mathcal{A} = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}, \tag{2.1}$$

where $A = \texttt{blkdiag}(\mathcal{Q}, \beta Q_{\vec{v}})$ and $C = [\mathcal{K} \; -\widehat{Q}]$. The matrix $\mathcal{A}$ is, in general, very large – the discrete Stokes equations are just one of it's components – yet is sparse. A good choice for solving such systems are iterative methods – in particular Krylov subspace methods. We will consider two such methods here: MINRES and Conjugate Gradients in a non-standard inner product, and extend the work of Rees, Dollar and Wathen [21] and Rees and Stoll [22] respectively to the case where the PDE is Stokes equations; significant complications arise here which are not present for simpler problems.

We comment that there are a large number of papers in the literature which deal with solving problems for the optimal control of PDEs. Below we comment on a few of these which share the philosophy of this paper. Most of these consider the model problem of the optimal control of Poisson's equation; it is not clear how easily they would be applied to the control of the Stokes equations and the additional difficulty this poses.

Schöberl and Zulehner [24] developed a preconditioner which is both optimal with respect to the problem size *and* with respect to the choice of regularization parameter, β. This method was recently generalized slightly by Herzog and Sachs [14]. A multigrid-based preconditioner has also been developed by Biros and Dogan [3] which has both h and β independent convergence properties, but it is not clear how their method would generalize to Stokes control. We note that the approximate reduced Hessian approximation used by Haber and Asher [13] and Biros and Ghattas [4] also leads to a preconditioner with $h-$independence. Other solution methods employing multigrid for this and similar classes of problems were described by Borzi [5], Asher and Haber [1] and Engel and Griebel [12].

2.1. Block diagonal preconditioners. It is well known that matrices of the form $\mathcal{A}$ are indefinite, and one choice of solution method for such systems is MINRES. For MINRES to be efficient for such a matrix we need to combine the method with a good preconditioner – i.e. a matrix $\mathcal{P}$ which is cheap to invert and which clusters the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$. One method that is often used – see [2, Section 10.1.1] and the references therein – is to look for a block diagonal preconditioner of the form

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix}.$$

Preconditioners of this form for the optimal control of Poisson's equation were discussed by Rees, Dollar and Wathen [21].

It is well known (see, for example, [11, Theorem 6.6]) that if A, A_0, $CA^{-1}C^T$ and S_0 are positive definite matrices such that there exist constants δ, Δ, ϕ and Φ such that the generalized Rayleigh quotients satisfy

$$\delta \leq \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T A_0 \mathbf{x}} \leq \Delta, \quad \phi \leq \frac{\mathbf{y}^T C A^{-1} C^T \mathbf{y}}{\mathbf{y}^T S_0 \mathbf{y}} \leq \Phi$$

for all vectors $\mathbf{x} \in \mathbb{R}^{2n_v+n_p}$ and $\mathbf{y} \in \mathbb{R}^{n_v+n_p}$, $\mathbf{x},\ \mathbf{y} \neq \mathbf{0}$, then the eigenvalues λ of

$\mathcal{P}^{-1}\mathcal{A}$ are real, and satisfy

$$\frac{\delta - \sqrt{\delta^2 + 4\Delta\Phi}}{2} \leq \lambda \leq \frac{\Delta - \sqrt{\Delta^2 + 4\phi\delta}}{2},$$
$$\delta \leq \lambda \leq \Delta,$$
or $$\frac{\delta + \sqrt{\delta^2 + 4\delta\phi}}{2} \leq \lambda \leq \frac{\Delta + \sqrt{\Delta^2 + 4\Phi\Delta}}{2}.$$

Therefore, if we can find matrices A_0 and S_0 that are cheap to invert and are good approximations to A and the Schur complement $CA^{-1}C^T$ in the sense defined above, then we will have a good preconditioner, since the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ will be in three distinct clusters bounded away from 0. In the ideal case where $A_0 = A$ and $S_0 = CA^{-1}C^T$ we have $\delta = \Delta = \phi = \Phi = 1$. Then the preconditioned system will have precisely three eigenvalues, $1, \frac{1+\sqrt{5}}{2}$ and $\frac{1-\sqrt{5}}{2}$, so MINRES would converge in three iterations [19].

2.2. Block lower-triangular preconditioners. Instead of MINRES we may want to use a conjugate gradient method to solve a saddle point problem of the form (2.1). Since (2.1) is not positive definite, the standard conjugate gradient algorithm cannot be used. However, the matrix

$$\begin{bmatrix} A_0 & 0 \\ C & -S_0 \end{bmatrix}^{-1} \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}$$

is self-adjoint with respect to the inner product defined by $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{H}} := \mathbf{u}^T \mathcal{H} \mathbf{v}$, where

$$\mathcal{H} = \begin{bmatrix} A - A_0 & 0 \\ 0 & S_0 \end{bmatrix},$$

provided that this defines an inner product – i.e. when $A - A_0$ and S_0 are positive definite. Therefore can we apply the conjugate gradient algorithm with this inner product, along with preconditioner

$$\mathcal{P} = \begin{bmatrix} A_0 & 0 \\ C & -S_0 \end{bmatrix}.$$

This method was first described by Bramble and Pasciak in [8], and has since generated a lot of interest – see, for example, [10, 16, 18, 24, 17, 27, 9]. This method was used in a control context by Rees and Stoll [22].

Convergence of this method again depends on the eigenvalue distribution of the preconditioned system – the clustering of the eigenvalues is given by, e.g., Rees and Stoll [22, Theorem 3.1], and the relevant result is stated below in Section 2.4. Note that in order to apply this preconditioner only solves with A_0 and S_0 are needed, hence an implicit approximation, for example multigrid, can be used; for more detail see e.g. Stoll [26].

One drawback of this method is that you need $A - A_0$ to be positive definite; this means that not just any approximation to A will do. This requirement usually results in having to find the eigenvalues of $A_0^{-1}A$ for a candidate A_0, and then adding an appropriate scaling γ so that $A > \gamma A_0$ – we will discuss this point further once we've described possible approximations A_0 in the following section.

2.3. Approximation of the (1,1) block. Suppose, for simplicity, that our domain $\Omega \subset \mathbb{R}^2$. If, as is usual, we use the same element space for all components in the velocity vector, and this has basis $\{\phi_i\}$. Then $Q_{\vec{v}} = \texttt{blkdiag}(Q_v, Q_v)$, where $Q_v = [\int_\Omega \phi_i \phi_j]$. Then the matrix A is just a block diagonal matrix composed of the mass matrices in the bases $\{\phi_i\}$ or $\{\psi_i\}$. Wathen [30] showed that for a general mass matrix, Q, if $D := \texttt{diag}(Q)$, then it is possible to calculate constants ξ and Ξ such that

$$\xi \leq \lambda(D^{-1}Q) \leq \Xi.$$

The constants depend on the elements used – for example, for $\mathbf{Q}_1$ elements $\xi = 1/4,\ \Xi = 9/4$ and for $\mathbf{Q}_2$ elements $\xi = 1/4,\ \Xi = 25/16$. The diagonal itself would therefore be a reasonable approximation to A.

However, as A is in a sense 'easy' to invert, it would help to have as good an approximation here as we can. Using the bounds described above we have all the information we need to use the relaxed Jacobi method accelerated by the Chebyshev-semi iteration. This is a method that is very cheap to use and, as demonstrated by Wathen and Rees in [31], is particularly effective in this case. In particular, since the eigenvalues of $D^{-1}Q$ are evenly distributed, there is very little difference between the convergence of this method and the (non-linear) conjugate gradient method preconditioned with D. Note that since the conjugate gradient algorithm is non-linear, we cannot use it as a preconditioner for a stationary Krylov subspace method such as MINRES, unless run to convergence. The Chebyshev semi-iteration, on the other hand, is a linear method. Suppose we use it to solve $Q\mathbf{x} = \mathbf{b}$ for some right hand side $\mathbf{b}$. Then we can write every iteration as $\mathbf{x}^{(m)} = T_m^{-1}\mathbf{b}$, for some matrix T_m implicitly defined by the method which is independent of $\mathbf{b}$.

Increasing m makes T_m a better approximation to Q in the sense defined above. The upper and lower eigenvalue bounds can be obtained analytically – for example, Table I in Rees and Stoll [22] gives the upper and lower bounds for each m from 1 to 20 for a $\mathbf{Q}_1$ discretization. Therefore, for the problem (1.2), if $\delta_m^v \leq \lambda\left(\left(T_m^v\right)^{-1} Q_v\right) \leq \Delta_m^v$ and $\delta_m^p \leq \lambda\left(\left(T_m^p\right)^{-1} Q_p\right) \leq \Delta_m^p$, then

$$\delta_m \leq \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T A_0 \mathbf{x}} \leq \Delta_m, \tag{2.2}$$

where $A_0 = \texttt{blkdiag}(T_m^v, T_m^v, T_m^p, \beta T_m^v, \beta T_m^v)$ and $\delta_m = \min(\delta_m^v, \delta_m^p)$ and $\Delta_m = \max(\Delta_m^v, \Delta_m^p)$, both independent of the mesh size, h. We therefore have an inexpensive way to make the bounds on $\lambda(A_0^{-1}A)$ as close to unity as required.

Note that, since we can work out these bounds accurately and inexpensively, the scaling parameter which needs to be calculated to ensure that $A - A_0$ is positive definite – which is a requirement for CG in a non-standard inner product 2.2 – can be easily chosen; see Rees and Stoll [22] for more details.

2.4. Approximation of the Schur complement. Now consider the Schur complement, $\frac{1}{\beta}\widehat{Q}Q_{\vec{v}}^{-1}\widehat{Q}^T + \mathcal{K}\mathcal{Q}^{-1}\mathcal{K} =: S$. The dominant term in this sum, for all but the smallest values of β, is $\mathcal{K}\mathcal{Q}^{-1}\mathcal{K}$ – the term that contains the PDE. Figure 2.1 shows the eigenvalue distribution for this approximation of S for a relatively coarse $\mathbf{Q}_2 - \mathbf{Q}_1$ discretization with $\beta = 0.01$. As we can see from the figure, the eigenvalues of $(\mathcal{K}\mathcal{Q}^{-1}\mathcal{K})^{-1}S$ are nicely clustered, and so we could expect good convergence of MINRES if we took S_0 as $\mathcal{K}\mathcal{Q}^{-1}\mathcal{K}$. The effect of varying β is described in, e.g., [29].

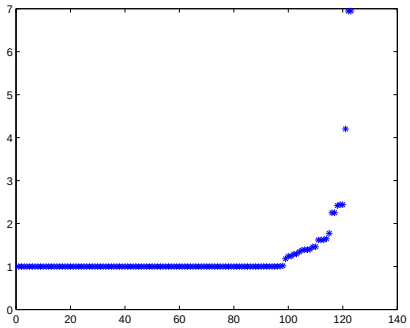


FIG. 2.1. *Eigenvalues of* $(\mathcal{K}\mathcal{Q}^{-1}\mathcal{K})^{-1}S$

However, a preconditioner must be easy to invert, and solving a system with $\mathcal{K}\mathcal{Q}^{-1}\mathcal{K}$ requires two solves with the discrete Stokes matrix, which is not cheap. We therefore would like some matrix, $\widetilde{\mathcal{K}}$, such that $\widetilde{\mathcal{K}}\mathcal{Q}^{-1}\widetilde{\mathcal{K}}$ approximates $\mathcal{K}\mathcal{Q}^{-1}\mathcal{K}$. Note that the mass matrices are not really significant in this context, and it is sufficient that $\widetilde{\mathcal{K}}\widetilde{\mathcal{K}}^T$ approximates $\mathcal{K}^2$. In order to achieve such an approximation, Braess and Peisker [7] show that it is *not* sufficient that $\widetilde{\mathcal{K}}$ approximates $\mathcal{K}$. Indeed, for the Stokes equations, Silvester and Wathen [25] showed that an ideal preconditioner is $\widehat{\mathcal{K}} = \texttt{blkdiag}(\underline{K}, M_p)$, where $\underline{K}$ is a multigrid cycle, but the eigenvalues of $(\widehat{\mathcal{K}}\widehat{\mathcal{K}}^T)^{-1}\mathcal{K}^2$ are not at all clustered, and the approximation of $\mathcal{K}^2$ is a poor one in this case. Suppose we wish to solve the equation $\mathcal{K}\mathbf{x} = \mathbf{b}$, for some right hand side vector $\mathbf{b}$. Braess and Peisker however go on to show that if we take an approximation $\mathcal{K}_m$ which is implicitly defined by an iteration $\mathbf{x}^{(m)} = \mathcal{K}_m^{-1}\mathbf{b}$, say, which converges to the solution $\mathbf{x}$ in the sense that

$$\|\mathbf{x}^{(m)} - \mathbf{x}\| \leq \eta_m \|\mathbf{x}\|,$$

then $\eta_m = \|\mathcal{K}_m^{-1}\mathcal{K} - I\|$, and one can show [7, Section 4]

$$(1-\eta)^2 \leq \frac{\mathbf{x}^T\mathcal{K}^2\mathbf{x}}{\mathbf{x}^T\mathcal{K}_m^T\mathcal{K}_m\mathbf{x}} \leq (1+\eta)^2. \tag{2.3}$$

Hence, approximation of $\mathcal{K}^2$ by $\mathcal{K}_m^T\mathcal{K}_m$ would be suitable in this case.

Note that MINRES cannot be used to approximate $\mathcal{K}$, unless run until convergence, since – like CG – MINRES is a Krylov subspace method, and hence nonlinear. We would therefore have to use a flexible outer method if we were to make use of an inner Krylov process as an approximation for the Stokes operator.

As before, consider a simple iteration of the form

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathcal{M}^{-1}\mathcal{K}\mathbf{r}^{(m)}, \tag{2.4}$$

where $\mathbf{r}^{(m)}$ is the residual at the m^{th} step, and with a block lower-triangular splitting matrix

$$\mathcal{M} := \begin{bmatrix} \underline{K}_0 & 0 \\ B & -Q_0 \end{bmatrix}, \tag{2.5}$$

where $\underline{K}_0$ approximates $\underline{K}$ and Q_0 approximates Q_p, which is itself spectrally equivalent to the Schur complement for the Stokes problem [11, Section 6.2]. By the result

of Braess and Peisker, we just need to show that this iteration converges – i.e. that $\rho(I - \mathcal{M}^{-1}\mathcal{K}) < 1$, where ρ denotes the spectral radius – in order that this defines a good approximation to the square. We ignore the one zero eigenvalue of $\underline{K}$ which is due to the hydrostatic pressure here, and in what follows, since if we start an iteration orthogonal to this kernel, we will remain orthogonal to the kernel [11, Section 2.3].

Consider two cases – $\underline{K} - \underline{K}_0$ positive definite, and $\underline{K} - \underline{K}_0$ indefinite. In the first case, it can be shown [22, Theorem 3.1], [32, Theorem 4.1] that if $\underline{K}_0$ and Q_0 are positive definite matrices such that

$$\upsilon \leq \frac{\mathbf{x}^T \underline{K} \mathbf{x}}{\mathbf{x}^T \underline{K}_0 \mathbf{x}} \leq \Upsilon, \quad \psi \leq \frac{\mathbf{y}^T B \underline{K}^{-1} B^T \mathbf{y}}{\mathbf{y}^T Q_0 \mathbf{y}} \leq \Psi, \tag{2.6}$$

then λ is real and positive, and moreover satisfies

$$\begin{aligned} \frac{(1+\psi)\Upsilon - \sqrt{(1+\psi)^2\Upsilon^2 - 4\psi\Upsilon}}{2} \leq & \lambda \leq \frac{(1+\Psi)\upsilon - \sqrt{(1+\Psi)^2\upsilon^2 - 4\Psi\upsilon}}{2} \\ \upsilon \leq & \lambda \leq \Upsilon \qquad \text{or} \\ \frac{(1+\psi)\upsilon + \sqrt{(1+\psi)^2\upsilon^2 - 4\psi\upsilon}}{2} \leq & \lambda \leq \frac{(1+\Psi)\Upsilon + \sqrt{(1+\Psi)^2\Upsilon^2 - 4\Psi\Upsilon}}{2}. \end{aligned}$$

We would like to put some numbers to these bounds in order to see what this means for a simple iteration based on a splitting with the block lower triangular matrix (2.5). It is well known that a multigrid iteration is a good approximation to the action of the inverse of $\underline{K}$, and we can scale such an iteration so that

$$1 \leq \frac{\mathbf{x}^T \underline{K} \mathbf{x}}{\mathbf{x} \underline{K}_0 \mathbf{x}} \leq \frac{1+\rho^m}{1-\rho^m},$$

where m is the number of V-cycles. A realistic value for ρ is 0.15 (see [11, pp. 294-295], for example), and experimentation shows $m = 2$ gives reasonable performance. Using Q_p for the Schur complement approximation we have

$$\gamma^2 \leq \frac{\mathbf{x}^T B \underline{K}^{-1} B^T \mathbf{x}}{\mathbf{x}^T Q_p \mathbf{x}} \leq \Gamma^2,$$

$\mathbf{x} \neq \mathbf{1}$, where for 2D Q1 elements, $\gamma^2 = 0.2$, $\Gamma^2 = 1$. Approximating this by 10 steps of the Chebyshev semi-iteration will weaken these bounds by a factor of 0.96 in the lower bound and $\Theta = 1.04$ in the upper. With these numbers, we have that $\lambda(\mathcal{M}^{-1}\mathcal{A}) \in [0.19, 1.29]$, and hence $\rho(I - \mathcal{M}^{-1}\mathcal{A}) = 0.81 < 1$. Therefore the simple iteration (2.4) with the splitting (2.5) will converge.

Although we've assumed $\underline{K} - \underline{K}_0 \geq 0$ in the analysis above, experiments show we still have good convergence properties even if this isn't true. In the case where $\underline{K} - \underline{K}_0$ is indefinite the situation is more complicated, as now the eigenvalues will in general be complex. Consider the generalized eigenvalue problem:

$$\begin{bmatrix} \underline{K} & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \lambda \begin{bmatrix} \underline{K}_0 & 0 \\ B & -Q_0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}. \tag{2.7}$$

We still assume that $\underline{K}$, $\underline{K}_0$ and Q_0 are positive definite. If $B\mathbf{x} = 0$, it is clear that $\lambda = \frac{\mathbf{x}^T \underline{K} \mathbf{x}}{\mathbf{x}^T \underline{K}_0 \mathbf{x}}$, and hence these eigenvalues must be real with

$$\upsilon \leq \lambda \leq \Upsilon.$$

Suppose now that $B\mathbf{x} \neq 0$. We can rearrange the second equation to give

$$\mathbf{y} = \frac{\lambda - 1}{\lambda} Q_0^{-1} B\mathbf{x},$$

and substituting this into the first equation and rearranging gives

$$\lambda = \lambda^2 \frac{\mathbf{x}^T \underline{K}_0 \mathbf{x}}{\mathbf{x}^T \underline{K} \mathbf{x}} + (1-\lambda) \frac{\mathbf{x}^T B^T Q_0^{-1} B\mathbf{x}}{\mathbf{x}^T \underline{K} \mathbf{x}}.$$

If we define

$$\kappa := \kappa(\mathbf{x}) = \frac{\mathbf{x}^T \underline{K} \mathbf{x}}{\mathbf{x}^T \underline{K}_0 \mathbf{x}}, \qquad \sigma := \sigma(\mathbf{x}) = \frac{\mathbf{x}^T B^T Q_0^{-1} B\mathbf{x}}{\mathbf{x}^T \underline{K} \mathbf{x}},$$

then we can write this as

$$\lambda^2/\kappa + (1-\lambda)\sigma - \lambda = 0,$$

or, alternatively,

$$\lambda^2 - (\sigma+1)\kappa\lambda + \sigma\kappa = 0.$$

Therefore the eigenvalues satisfy

$$\lambda = \frac{(\sigma+1)\kappa \pm \sqrt{(\sigma+1)^2\kappa^2 - 4\sigma\kappa}}{2}.$$

We know from above that if $\kappa = \frac{\mathbf{x}^T \underline{K} \mathbf{x}}{\mathbf{x}^T \underline{K}_0 \mathbf{x}} \geq 1$, then all the eigenvalues are real. Note that

$$(\sigma+1)^2\kappa^2 - 4\sigma\kappa = 0 \Rightarrow \kappa = 0 \text{ or } \kappa = \frac{4\sigma}{(1+\sigma)^2} \leq 1,$$

the last inequality being since $\frac{4\sigma}{(1+\sigma)^2}$ has a maximum value of 1 which occurs when $\sigma = 1$. This tells us that for $\kappa \in [0, \frac{4\sigma}{(1+\sigma)^2}]$, $\lambda \in \mathbb{C}$.

In this case,

$$\begin{aligned} \lambda &= \frac{(\sigma+1)\kappa \pm i\sqrt{4\sigma\kappa - (\sigma+1)^2\kappa^2}}{2} \\ \Rightarrow |\lambda|^2 &= \frac{(\sigma+1)^2\kappa^2 + 4\sigma\kappa - (\sigma+1)^2\kappa^2}{4} \\ &= \sigma\kappa. \end{aligned}$$

Therefore the complex eigenvalues satisfy

$$\sqrt{\upsilon\psi} \leq |\lambda| \leq \sqrt{\Psi}. \tag{2.8}$$

Moreover, $\mathrm{Re}(\lambda) = \frac{(\sigma+1)\kappa}{2} > 0$, so all the complex eigenvalues live in the right-hand plane. Also,

$$\begin{aligned} \frac{|\mathrm{Im}(\lambda)|}{\mathrm{Re}(\lambda)} &= \frac{\sqrt{4\sigma\kappa - (\sigma+1)^2\kappa^2}}{2} \cdot \frac{2}{(\sigma+1)\kappa} \\ &= \frac{\sqrt{4\sigma\kappa - (\sigma+1)^2\kappa^2}}{(\sigma+1)\kappa}. \end{aligned}$$

If we define

$$F(\sigma,\kappa) := \frac{\sqrt{4\sigma\kappa-(\sigma+1)^2\kappa^2}}{(\sigma+1)\kappa},$$

then

$$\frac{\partial F}{\partial \sigma} = \frac{2(\sigma-1)}{(\sigma+1)^2\sqrt{(4-2\kappa)\kappa\sigma-\kappa^2(\sigma^2+1)}},$$

so

$$\frac{\partial F}{\partial \sigma} = 0 \Rightarrow \sigma = 1.$$

This critical point is clearly a maximum. This means that, for any fixed κ, $F(\sigma,\kappa)$ has it's maximum at $\sigma = 1$. Therefore

$$\frac{|\mathrm{Im}(\lambda)|}{\mathrm{Re}(\lambda)} = F(\sigma,\kappa) \le \frac{\sqrt{\kappa-\kappa^2}}{\kappa} = \sqrt{\frac{1}{\kappa}-1} \le \sqrt{\frac{1}{\upsilon}-1}.$$

Therefore, putting this together with (2.8) above, the complex eigenvalues satisfy

$$\lambda \in \left\{ z = re^{i\theta} \in \mathbb{C} \ : \ \sqrt{\upsilon\psi} \le r \le \sqrt{\Psi},\ -\tan^{-1}(\sqrt{\upsilon^{-1}-1}) \le \theta \le \tan^{-1}(\sqrt{\upsilon^{-1}-1}) \right\}. \tag{2.9}$$

For $\kappa > 1$, the result given above for $\underline{K} - \underline{K}_0$ positive definite still hold, and we have $\lambda \in \mathbb{R}$ which satisfy

$$\frac{(\psi+1)\Upsilon - \sqrt{(\psi+1)^2\Upsilon^2-4\psi\Upsilon}}{2} \le \lambda \le \frac{(\Psi+1)\Upsilon + \sqrt{(\Psi+1)^2\Upsilon^2-4\Psi\Upsilon}}{2}.$$

What about $\kappa \in \left[\frac{4\sigma}{(1+\sigma)^2}, 1\right]$? In this case, too, the bounds above hold, since in the derivation of these bounds we required no information about δ, all that is assumed is that $\lambda \in \mathbb{R}$ – see [22, Theorem 3.1]. Verifying the inner bounds required that $\delta > 1$, so these do not carry over, but there is no such problem with the outer bounds. We have proved the following theorem:

THEOREM 2.1. *Let λ be an eigenvalue associated with the generalized eigenvalue problem*

$$\begin{bmatrix} \underline{K} & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \lambda \begin{bmatrix} \underline{K}_0 & 0 \\ B & -Q_0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

where $\underline{K}$, $\underline{K}_0$ and Q_0 are positive definite and satisfy (2.6). If $\lambda \in \mathbb{R}$, then it satisfies

$$\begin{aligned} \frac{(1+\psi)\Upsilon - \sqrt{(1+\psi)^2\Upsilon^2-4\psi\Upsilon}}{2} \le &\lambda \le \frac{(1+\Psi)\Upsilon + \sqrt{(1+\Psi)^2\Upsilon^2-4\Psi\Upsilon}}{2} \\ \text{or} \qquad \upsilon \le &\lambda \le \Upsilon, \end{aligned}$$

and if $\lambda \in \mathbb{C}$, then $\lambda = re^{i\theta}$, where r and θ satisfy

$$\sqrt{\upsilon\psi} \le r \le \sqrt{\Psi},\ -\tan^{-1}(\sqrt{\upsilon^{-1}-1}) \le \theta \le \tan^{-1}(\sqrt{\upsilon^{-1}-1}).$$

To get bounds for $\rho(I-\mathcal{M}^{-1}\mathcal{A})$ we have to be more careful because of the presence of the complex eigenvalues. Figure 2.2 is a relevant diagram for the situation here. All the complex eigenvalues will be contained in the unit circle if the line d labelled on the diagram is less than unity. By the cosine rule:

$$d^2 = 1 + \Psi - 2\sqrt{\Psi}\cos\theta,$$

where $\tan\theta = \sqrt{\upsilon^{-1} - 1}$. Therefore all the complex eigenvalues are in the unit circle if

$$\frac{\sqrt{\Psi}}{2} < \cos\theta.$$

Note that, using the same argument, the distance from the origin to the point where the circle of radius $\sqrt{\psi\upsilon}$ and centre -1 touches the ray that makes an angle θ with the x-axis is

$$\sqrt{1 + \psi\upsilon - 2\sqrt{\psi\upsilon}\cos\theta}.$$

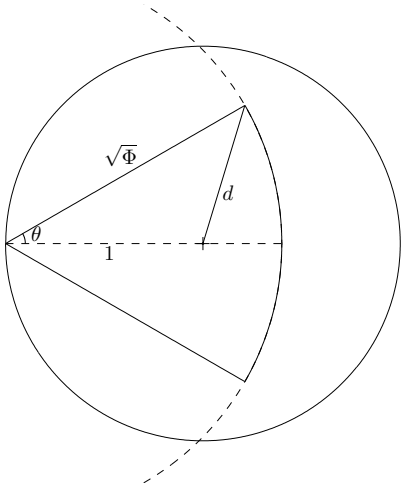


FIG. 2.2. *Diagram of the geometry containing the complex eigenvalues.* $\theta = \sqrt{\upsilon^{-1} - 1}$ *and* d *is the unknown length.*

There follows:

COROLLARY 2.2. *Suppose that the eigenvalues of the generalized eigenvalue problem (2.7) are as described in Theorem 2.1. Define*

$$\xi := \max\left\{1-\upsilon, \Upsilon - 1, 1 - \frac{(1+\psi)\Upsilon - \sqrt{(1+\psi)^2\Upsilon^2 - 4\psi\Upsilon}}{2}, \right.$$
$$\frac{(1+\Psi)\Upsilon + \sqrt{(1+\Psi)^2\Upsilon^2 - 4\Psi\Upsilon}}{2} - 1, \sqrt{1+\Psi-2\sqrt{\Psi}\cos\theta},$$
$$\left.\sqrt{1+\psi\upsilon - 2\sqrt{\psi\upsilon}\cos\theta}\right\}.$$

Then a simple iteration with splitting matrix

$$\mathcal{M} = \begin{bmatrix} \underline{K}_0 & 0 \\ B & -Q_0 \end{bmatrix}$$

will converge if $\xi < 1$, with the asymptotic convergence rate being ξ.

Zulehner also derived an approximation to the convergence factor [32, Theorem 4.3]. Note that Corollary 2.2 differs slightly from the result in Zulehner – this is because neither the result given here nor in [32] are sharp with regards to the complex eigenvalues. The two results are obtained in very different ways, and neither can be said to be a better approximation than the other one.

Figure 2.3 shows the bounds predicted above and the actual eigenvalues for a number of approximations to the matrix $\underline{K}$. This shows that we will get asymptotic convergence, but in practice we see good results from the first iteration. Also, the theory above is equally valid for the block *upper*-triangular approximation to the discrete Stokes matrix, whereas in practice we observe that it takes far more iterations with this upper-triangular splitting to converge.

Let us again return to the case where $\underline{K} - \underline{K}_0$ is positive definite. Then we know from Section 2.2 that

$$\mathcal{M}^{-1}\mathcal{K} = \begin{bmatrix} \underline{K}_0 & 0 \\ B & -Q_0 \end{bmatrix}^{-1} \begin{bmatrix} \underline{K} & B^T \\ B & 0 \end{bmatrix}$$

is self adjoint in the inner product defined by

$$\mathcal{H} = \begin{bmatrix} \underline{K} - \underline{K}_0 & 0 \\ 0 & Q_0 \end{bmatrix}.$$

If we define $\widehat{\mathcal{K}} := \mathcal{M}^{-1}\mathcal{K}$, then we have that $\widehat{\mathcal{K}}$ is $\mathcal{H}-$normal, i.e.

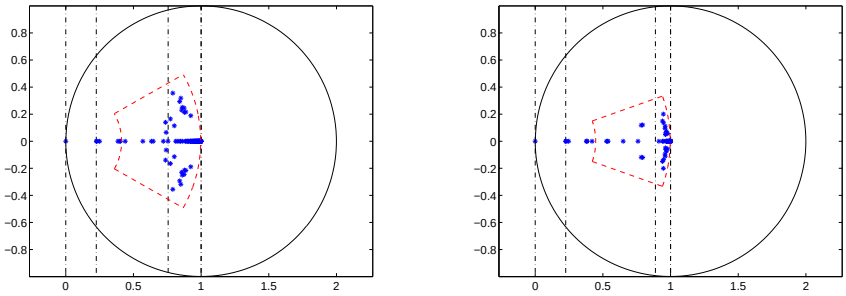
$$\widehat{\mathcal{K}}^\dagger\widehat{\mathcal{K}} = \widehat{\mathcal{K}}\widehat{\mathcal{K}}^\dagger,$$

where $\widehat{\mathcal{K}}^\dagger = \mathcal{H}^{-1}\widehat{\mathcal{K}}^T\mathcal{H}$. The iteration matrix $I - \mathcal{M}^{-1}\mathcal{K}$ is therefore $\mathcal{H}-$normal, and so

$$\|I - \mathcal{M}^{-1}\mathcal{K}\|_{\mathcal{H}} = \rho(I - \mathcal{M}^{-1}\mathcal{K}),$$

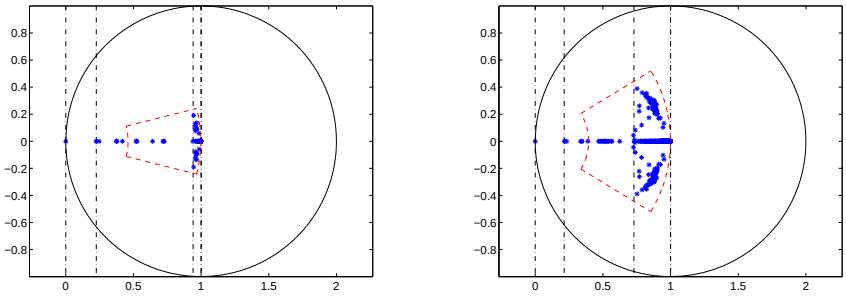
which tells us that

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathcal{H}} \leq \rho^k \|\mathbf{x}\|_{\mathcal{H}},$$



(a) h = 0.25, K_0 given by 1 AMG V-cycle with 1 pre- and 1 post-smoothing step

(b) h = 0.25, K_0 given by 1 AMG V-cycle with 2 pre- and 2 post-smoothing steps



(c) h = 0.25, K_0 given by 2 AMG V-cycles with 2 pre- and 2 post-smoothing steps

(d) h = 0.125, K_0 given by 1 AMG V-cycle with 1 pre- and 1 post-smoothing step

FIG. 2.3. **'s denote computed eigenvalues. Lines, from left to right, are at 0, $\frac{(\psi+1)\Upsilon-\sqrt{(\psi+1)^2\Upsilon^2-4\psi\Upsilon}}{2}$, υ, Υ and $\frac{(\Psi+1)\Upsilon+\sqrt{(\Psi+1)^2\Upsilon^2-4\Psi\Upsilon}}{2}$, (the last two virtually coincide here). Dashed region is the bounds of Theorem 2.1 for the complex eigenvalues. Also shown is the unit circle centred at $z = 1$.*

.

where $\rho = \rho(I - \mathcal{M}^{-1}\mathcal{K})$, the spectral radius of the iteration matrix. To apply the result of Braess and Peisker (2.3) we need a constant η_k such that the error converges the 2−norm, i.e.

$$\|\mathbf{x}_k - \mathbf{x}\|_2 \leq \eta_k \|\mathbf{x}\|_2.$$

We know that over a finite dimensional vector space all norms are equivalent, though the equivalence constants may be h−dependent for a discretized PDE problem. Thus there exist positive constants γ and Γ such that

$$\sqrt{\gamma}\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\mathcal{H}} \leq \sqrt{\Gamma}\|\mathbf{x}\|_2,$$

and hence

$$\begin{aligned}\|\mathbf{x}_k-\mathbf{x}\|_2 &\le \|\mathbf{x}_k-\mathbf{x}\|_{\mathcal{H}}/\sqrt{\gamma}\\ &\le \frac{\rho^m}{\sqrt{\gamma}}\|\mathbf{x}\|_{\mathcal{H}}\\ &\le \frac{\sqrt{\Gamma}\rho^m}{\sqrt{\gamma}}\|\mathbf{x}\|_2. \end{aligned} \tag{2.10}$$

We now need to know the values of the constants γ and Γ.

Recalling standard bounds for two dimensional finite element matrices – see e.g. Theorems 1.32 and 1.29 in [11] – we have that, under mild assumptions, there exist positive constants $d,\ D,\ c$ and C such that:

$$\begin{aligned} dh^2\mathbf{x}^T\mathbf{x} \le \mathbf{x}^T\underline{K}\mathbf{x} \le D\mathbf{x}^T\mathbf{x}\\ ch^2\mathbf{x}^T\mathbf{x} \le \mathbf{x}^TQ_p\mathbf{x} \le Ch^2\mathbf{x}^T\mathbf{x}. \end{aligned}$$

Then $\mathbf{x}^T\mathcal{H}\mathbf{x} \le \Gamma\mathbf{x}^T\mathbf{x}$ would mean that

$$\mathbf{y}^T(\underline{K}-\underline{K}_0)\mathbf{y}+\mathbf{z}^TQ_0\mathbf{z} \le \Gamma(\mathbf{y}^T\mathbf{y}+\mathbf{z}^T\mathbf{z}).$$

Therefore if we have constants Γ_1 and Γ_2 such that

$$\mathbf{y}^T(\underline{K}-\underline{K}_0)\mathbf{y} \le \Gamma_1\mathbf{y}^T\mathbf{y} \quad \text{and} \quad \mathbf{z}^TQ_0\mathbf{z} \le \Gamma_2\mathbf{z}^T\mathbf{z}$$

then we could take $\Gamma = \max(\Gamma_1,\Gamma_2)$.

First, note that from (2.6)

$$\begin{aligned} \mathbf{x}^T\underline{K}\mathbf{x} &\le \Upsilon\mathbf{x}^T\underline{K}_0\mathbf{x}\\ \Upsilon\mathbf{x}^T\underline{K}\mathbf{x}-(\Upsilon-1)\mathbf{x}^T\underline{K}\mathbf{x} &\le \Upsilon\mathbf{x}^T\underline{K}_0\mathbf{x}\\ \Upsilon(\mathbf{x}^T\underline{K}\mathbf{x}-\mathbf{x}^T\underline{K}_0\mathbf{x}) &\le (\Upsilon-1)\mathbf{x}^T\underline{K}\mathbf{x}\\ \mathbf{x}^T(\underline{K}-\underline{K}_0)\mathbf{x} &\le \frac{D(\Upsilon-1)}{\Upsilon}\mathbf{x}^T\mathbf{x}\\ \therefore\ \frac{\mathbf{x}^T(\underline{K}-\underline{K}_0)\mathbf{x}}{\mathbf{x}^T\mathbf{x}} &\le \frac{D(\Upsilon-1)}{\Upsilon}. \end{aligned}$$

Therefore

$$\Gamma_1 = \frac{(\Upsilon-1)D}{\Upsilon}.$$

Let $Q_0 = T_m^p$ represent m steps of the Chebyshev semi-iteration, as defined in Section 2.3, where

$$\delta_m^p \le \frac{\mathbf{x}^TQ_p\mathbf{x}}{\mathbf{x}^TT_m^p\mathbf{x}} \le \Delta_m^p.$$

Then

$$\begin{aligned}\frac{\mathbf{z}^TQ_0\mathbf{z}}{\mathbf{z}^T\mathbf{z}} &= \frac{\mathbf{z}^TT_m^p\mathbf{z}}{\mathbf{z}^T\mathbf{z}}\\ &= \frac{\mathbf{z}^TT_m^p\mathbf{z}}{\mathbf{z}^TQ_p\mathbf{z}}\cdot\frac{\mathbf{z}^TQ_p\mathbf{z}}{\mathbf{z}^T\mathbf{z}}\\ &\le \frac{C_ph^2}{\delta_m^p}. \end{aligned}$$

Therefore we can take $\Gamma_2 = C_p h^2$, and hence

$$\Gamma = \max\left(\frac{(\Upsilon - 1)D}{\Upsilon}, \frac{C_p h^2}{\delta_m^p}\right)$$

satisfies $\mathbf{x}^T \mathcal{H} \mathbf{x} \leq \Gamma \mathbf{x}^T \mathbf{x}$.

Now we turn our attention to a lower bound. Similarly to above, we take $\gamma = \min(\gamma_1, \gamma_2)$, where

$$\gamma_1 \mathbf{y}^T \mathbf{y} \leq \mathbf{y}^T (\underline{K} - \underline{K}_0)\mathbf{y} \quad \text{and} \quad \gamma_2 \mathbf{z}^T \mathbf{z} \leq \mathbf{z}^T Q_0 \mathbf{z}.$$

Again, we have from (2.6):

$$\begin{aligned} \upsilon \mathbf{y}^T \underline{K}_0 \mathbf{y} &\leq \mathbf{y}^T \underline{K} \mathbf{y} \\ &= \upsilon \mathbf{y}^T \underline{K} \mathbf{y} + (1 - \upsilon)\mathbf{y}^T \underline{K} \mathbf{y} \\ (\upsilon - 1)\mathbf{y}^T \underline{K} \mathbf{y} &\leq \upsilon \mathbf{y}^T (\underline{K} - \underline{K}_0)\mathbf{y} \\ \frac{(\upsilon - 1)dh^2}{\upsilon} &\leq \frac{\mathbf{y}^T (\underline{K} - \underline{K}_0)\mathbf{y}}{\mathbf{y}^T \mathbf{y}}. \end{aligned}$$

Again arguing as above,

$$\begin{aligned} \frac{\mathbf{z}^T Q_0 \mathbf{z}}{\mathbf{z}^T \mathbf{z}} &= \frac{\mathbf{z}^T T_m^p \mathbf{z}}{\mathbf{z}^T Q_p \mathbf{z}} \cdot \frac{\mathbf{x}^T Q_p \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \\ &\geq \frac{c_p h^2}{\Delta_m^p}. \end{aligned}$$

Therefore we can take

$$\gamma = \min\left(\frac{(\upsilon - 1)c_{\vec{u}} dh^2}{\upsilon}, \frac{c_p h^2}{\Delta_m^p}\right),$$

which satisfies $\gamma \mathbf{x}^T \mathbf{x} \leq \mathbf{x}^T \mathcal{H} \mathbf{x}$.

By equation (2.10) the contraction constant for convergence in the 2-norm is given by $\rho^m \sqrt{\Gamma}/\sqrt{\gamma}$. It is clear that $\sqrt{\gamma} = \nu h$, where ν is a constant. For the numerator, in general, we will have $\Gamma = \frac{(\Upsilon - 1)D}{\Upsilon}$, as h^2 is small. This would mean that

$$\frac{\sqrt{\Gamma}}{\sqrt{\gamma}} = \mathcal{O}\left(h^{-1}\right),$$

i.e. the contraction constant would be dependent upon h.

However, we have control over the value of Υ, as this measures the accuracy of the approximation to $\underline{K}$. Recall that $\underline{K}_0$ is a good approximation to $\underline{K}$ if Υ is close to unity. As $\underline{K}_0$ is a multigrid process we can make this parameter as close to 1 as required by simply taking more V-cycles, better smoothing, etc. If this approximation is good enough, and $\frac{(\Upsilon - 1)D}{\Upsilon}$ is smaller than $\frac{C_p h^2}{\delta_m^p}$, we will get a constant number of iterations, at least up to some value of h. Note that we have knowledge of all the parameters involved, so given a smallest required value of h – which one will know a priori – one can pick an approximation $\underline{K}_0$ which gives a reasonable method. The

quantity ρ^m also appears in the numerator, so convergence can be improved by taking more inexact Uzawa iterations.

Even though the above argument only holds when $\underline{K} - \underline{K}_0$ is positive definite, we see the same behaviour in practice for the general case. Since solving the approximation to $\mathcal{K}$ is particularly expensive here it is worth getting the approximation to the mass matrix, Q_0, as close to Q as possible. Therefore, in the results that follow we take Q_0 to be defined implicity by 20 steps of the Chebyshev semi-iteration applied to the appropriate mass matrix. The inexact Uzawa method can be improved with the introduction of a parameter τ in front of the approximation to the Schur complement [10]. In the inexact case the optimal parameter is hard to obtain, but a good approximation is $(\phi + \Phi)/2$, where $\lambda(S_0^{-1}S) \in [\phi, \Phi]$. For $\mathbf{Q}_1$ elements and a Dirichlet problem, $\lambda(Q_p^{-1}S) \in [0.2, 1]$ [11, p. 271], so we take our scaling parameter as $\tau = 3/5$. We therefore advocate a practical splitting matrix for inexact Uzawa of

$$\mathcal{M} = \begin{bmatrix} \underline{K}_0 & 0 \\ B & -\tau Q_0 \end{bmatrix}.$$

A matrix of the form

$$\mathcal{P} := \begin{bmatrix} A_0 & 0 \\ 0 & \mathcal{K}_m \mathcal{Q}^{-1} \mathcal{K}_m^T \end{bmatrix},$$

where A_0 is composed of Chebyshev approximations and $\mathcal{K}_m$ is a simple iteration based on the splitting matrix $\mathcal{M}$, should therefore be an effective preconditioner for the matrix $\mathcal{A}$.

3. Numerical Results. First, consider the following forward problem, which sets the boundary conditions that we will use for the control problem. This is a classic test problem in fluid dynamics called leaky cavity flow, and a discussion is given by Elman, Silvester and Wathen [11, Example 5.1.3].

EXAMPLE 3.1. *Let $\Omega = [0,1]^2$, and let $\vec{\mathbf{i}}$ and $\vec{\mathbf{j}}$ denote unit vectors in the direction of the x and y axis respectively. Let $\vec{v}$ and p satisfy the Stokes equations*

$$\begin{aligned} -\nabla^2 \vec{v} + \nabla p &= \vec{0} \quad \text{in } \Omega \\ \nabla \cdot \vec{v} &= 0 \quad \text{in } \Omega, \end{aligned}$$

and let $\vec{v} = \vec{0}$ on the boundary except for on $x = 1$, $0 \le y \le 1$, where $\vec{v} = -\vec{\mathbf{j}}$.

We discretize the Stokes problem using $\mathbf{Q}_2 - \mathbf{Q}_1$ elements and solve the resulting linear system using MINRES [20]. As a preconditioner we use the block diagonal matrix $\texttt{blkdiag}(\widehat{K}, T_{20})$, following Silvester and Wathen [25], where $\widehat{K}$ denotes one AMG V-cycle (using the HSL MI20 AMG routine [6] applied via a MATLAB interface) and T_{20}^{-1} is twenty steps of the Chebyshev semi-iteration applied with the pressure mass matrix. The problem was solved using MATLAB R2009b, and the number of iterations and the time taken for different mesh sizes is given in Table 3.1. The constant number of iterations independent of h and linear growth in CPU time (i.e. linear complexity of the solver) are well-understood for this problem – see [11, Chapter 6].

Figure 3.1 shows the streamlines and the pressure of the solution obtained. Note the small recirculations present in the lower corners – these are Moffatt eddies. Adding a forcing term that reduces these eddies will be the object of our control problem, Example 3.2.

h	size	CPU time (s)	Iterations
2^{-2}	187	0.015	25
2^{-3}	659	0.029	27
2^{-4}	2,467	0.076	28
2^{-5}	9,539	0.349	30
2^{-6}	37,507	1.504	30
2^{-7}	148,739	6.616	30

TABLE 3.1
Number of MINRES iterations and time taken to solve the forward problem in Example 3.1

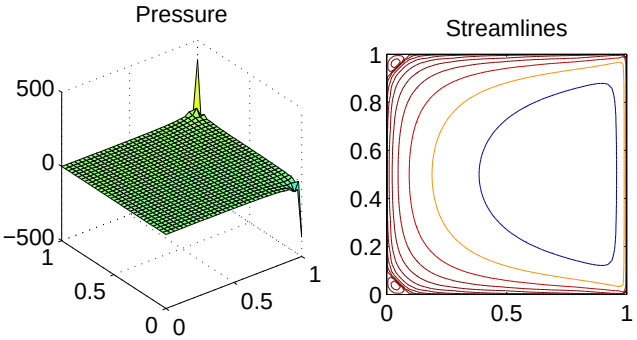


FIG. 3.1. *Solution of Example 3.1*

EXAMPLE 3.2. *Let $\Omega = [0,1]^2$, and consider an optimal control problem of the form (1.1), with Dirichlet boundary conditions as given in Example 3.1 (leaky cavity flow). Take the desired pressure as $\widehat{p} = 0$ and let $\widehat{v} = y\vec{\mathbf{i}} - x\vec{\mathbf{j}}$. The exponentially distributed streamlines of the desired velocity are shown in Figure 3.2.*

We discretize (1.1) using $\mathbf{Q}_2-\mathbf{Q}_1$ elements, also using $\mathbf{Q}_2$ elements for the control. Table 3.2 shows the results for solving the problem using MINRES, with right hand side as in Example 3.2 and with $\beta = 10^{-2}$ and $\delta = 1$. As a preconditioner we use the block diagonal preconditioner, with $\mathcal{K}$ approximated by m steps of the simple iteration with splitting matrix

$$\mathcal{M} = \begin{bmatrix} \underline{K}_0 & 0 \\ B & -S \end{bmatrix},$$

where $S = B\underline{K}^{-1}B^T$ is the exact Schur complement of the Stokes equation. $\underline{K}_0^{-1}$ is given by k HSL MI20 AMG V-cycles. This is not a practical preconditioner, since it includes the exact Schur complement of the Stokes matrix, but we can see clearly that if the approximation $\underline{K}_0$ is not good enough we do not – even in this idealized case – get an optimal preconditioner. This phenomenon is explained by the theory in Section 2.4. It is therefore vital that the approximation $\underline{K}_0$ is close enough to

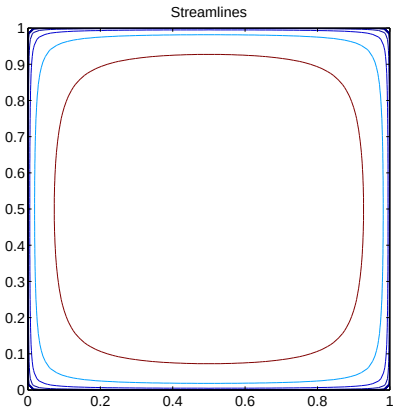


FIG. 3.2.

$\underline{K}$, in the sense defined in the previous section, in order to get an effective practical preconditioner.

TABLE 3.2
Comparison of solution methods for solving Example 3.2 using MINRES preconditioned with the block diagonal preconditioner with m steps of inexact Uzawa approximating $\mathcal{K}$ and k AMG V-cycles approximating $\underline{K}$.

h	size	Exact, m=1		m=1, k=1		m=1, k=2		m=1, k=3		m=1, k=4	
		time	its	time	its	time	its	time	its	time	its
2^{-2}	344	0.089	25	0.092	29	0.079	27	0.082	27	0.085	27
2^{-3}	1512	0.382	27	0.432	35	0.352	27	0.365	27	0.380	27
2^{-4}	6344	3.192	25	7.359	65	3.179	27	3.235	27	3.296	27
2^{-5}	25992	60.063	25	403.933	179	72.858	31	64.028	27	64.055	27
h	size	Exact, m=2		m=2, k=1		m=2, k=2		m=2, k=3		m=2, k=4	
		time	its	time	its	time	its	time	its	time	its
2^{-2}	344	0.073	21	0.100	27	0.099	25	0.096	23	0.101	23
2^{-3}	1512	0.408	23	0.429	29	0.400	25	0.423	25	0.450	25
2^{-4}	6344	3.466	23	3.954	31	3.347	25	3.193	23	3.319	23
2^{-5}	25992	57.284	21	98.885	39	65.489	25	60.051	23	61.398	23

As we saw in Section 2.4, a practical preconditioner can be obtained by replacing the exact Stokes Schur complement by the pressure mass matrix – or more generally, by something that approximates the pressure mass matrix. We take this to be 20 steps of the Chebyshev semi-iteration applied to the relevant matrix, as described in Section 2.3. Experimentation suggests that taking two steps of the inexact Uzawa method, in which $\underline{K}_0^{-1}$ is given by three HSL MI20 AMG V-cycles, will give a good preconditioner. In the results that follow we take $\beta = 10^{-2}$, $\delta = 1$ and solve to a tolerance of 10^{-6} in the appropriate norm.

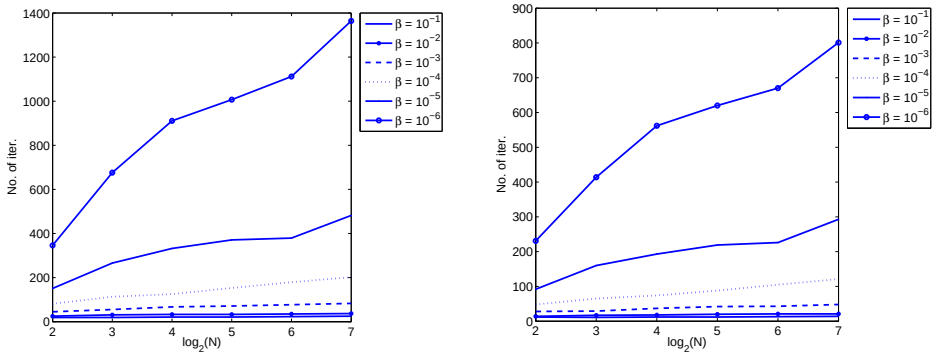
As we see from Table 3.3, the overall technique which we have described seems to be a good method for solving the Stokes control problem. Comparing the results here with those to solve the forward problem in Table 3.1 the iteration numbers aren't that much more, and they do not increase significantly with the mesh size; the solution times also scale roughly linearly. Solving the control problem using the block-triangular preconditioner is just over a factor of ten more expensive that solving a

TABLE 3.3
Comparison of solution methods for solving Example 3.2 using MINRES and BPCG preconditioned with the block diagonal and block lower triangular preconditioners respectively with 2 steps of inexact Uzawa approximating $\mathcal{K}$ and 3 AMG V-cycles approximating $\underline{K}$.

h	size	MINRES		BPCG		`backslash`
		time	its	time	its	time
2^{-2}	344	0.189	25	0.083	14	0.016
2^{-3}	1512	0.358	31	0.194	17	0.059
2^{-4}	6344	1.176	33	0.679	18	0.601
2^{-5}	25992	4.965	33	3.133	20	7.300
2^{-6}	105224	22.704	35	14.584	21	—

single forward problem for every grid size – an overhead that seems reasonable, given the increased complexity of the control problem in comparison to the forward problem.

Figures 3.3 and 3.4 show the number of iterations taken to solve this problem for different values of β and δ in (1.1) respectively. These show that – as we might expect from the theory – decreasing β and increasing δ increases the number of iterations required to solve the system using our methods. From the plots in Figures 3.5 and 3.6 it seems that the value $\delta = 1$ gives a pressure of the same order as the uncontrolled problem, the solution of which is shown in Figure 3.1. However, one can conceive of situations where we require a tighter bound on the pressure, and hence a higher value of δ.

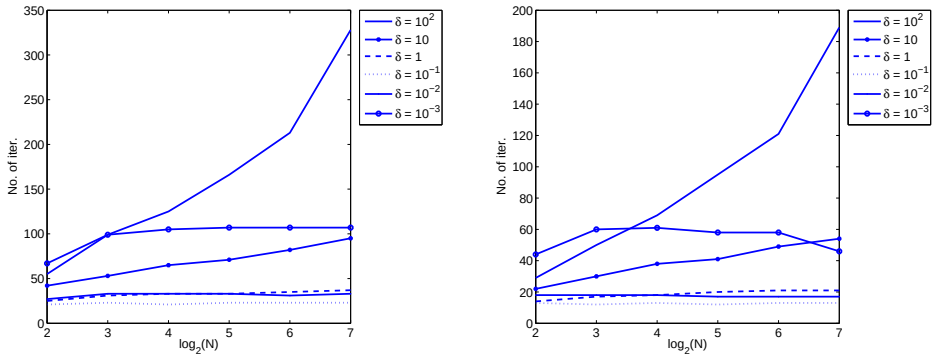


(a) MINRES (b) BPCG

FIG. 3.3. *Plot of problem size vs iterations needed for different β, where $\delta = 1$.*

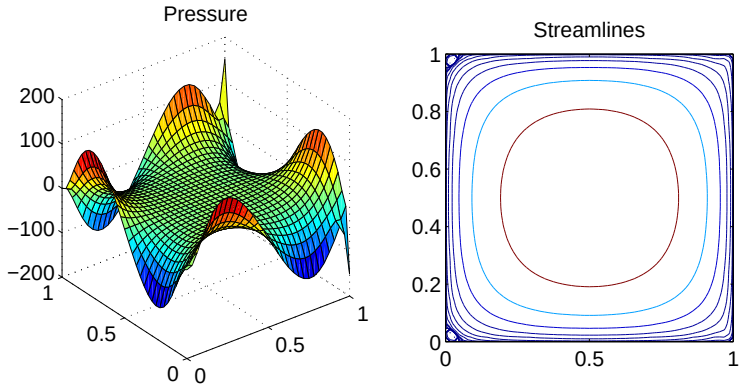
We have only presented a simple distributed control problem here. It is possible to solve other types of control problem using the same method – see [21] for a discussion in the simpler case of Poisson control. It is also possible to use this method together with bound constraints on the control – Stoll and Wathen [28] discuss this approach in consideration of the Poisson control problem.

4. Conclusions. In this paper we have presented two preconditioners – one for MINRES, and one for CG in a non-standard inner product – that can be used to solve problems in Stokes control. These both rely on effective approximations to the (1,1) block, which is composed of mass matrices, and to the Schur complement. We advocate using the Chebyshev semi-iteration used to accelerate a relaxed Jacobi iteration as an approximation to the (1,1) block, and an inexact Uzawa based approx-

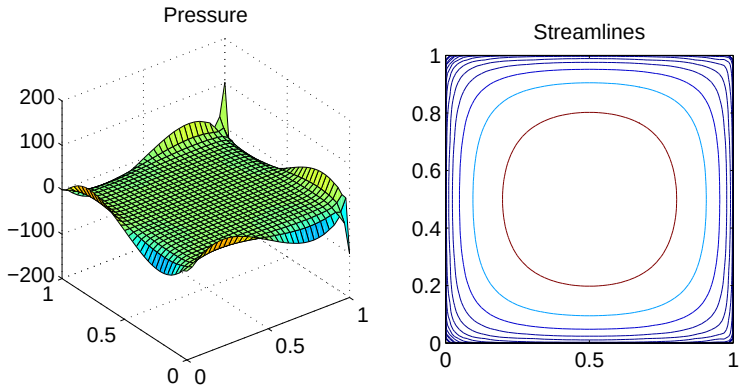


(a) MINRES (b) BPCG

FIG. 3.4. *Plot of problem size vs iterations needed for different δ, where $\beta = 10^{-2}$.*



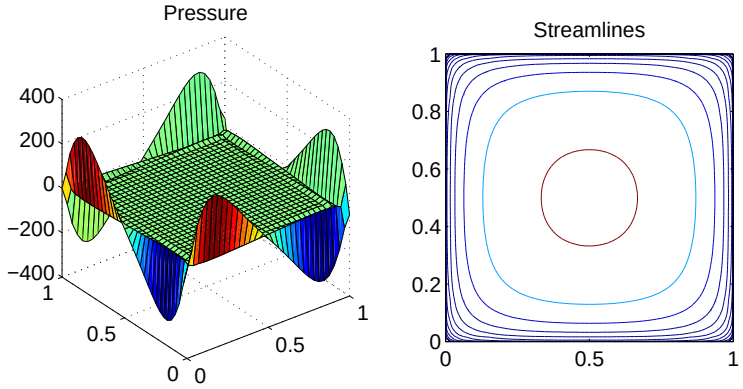
(a) Computed state for $\beta = 10^{-2}$, $\delta = 1$



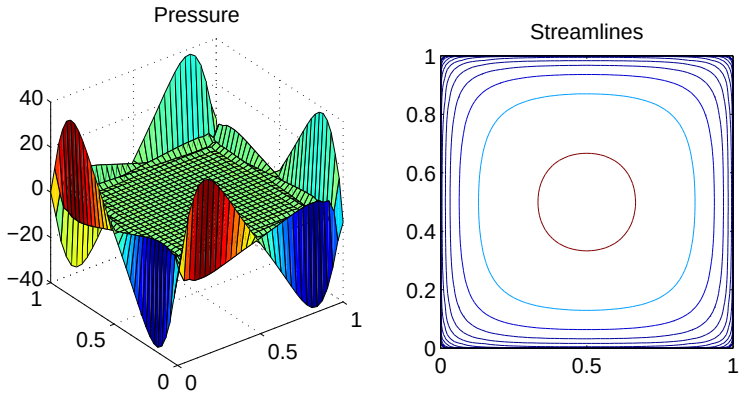
(b) Computed state for $\beta = 10^{-2}$, $\delta = 10$

FIG. 3.5. *Computed states for Example 3.2 in two dimensions, $\beta = 10^{-2}$.*

imation for the Schur complement. We have given some theoretical justification for the effectiveness of such preconditioners and have given some numerical results.



(a) Computed state for $\delta = 1$



(b) Computed state for $\delta = 10$

FIG. 3.6. *Computed states for Example 3.2 in two dimensions,* $\beta = 10^{-5}$.

We compared these results with those for solving the equivalent forward problem, and the iteration count is only marginally higher in the control case, and behaves in broadly the same way as the iterations taken the solve the forward problem as the mesh size decreases. These approximations therefore seem reasonable for problems of this type. Furthermore, the ideas presented here have the potential to be extended to develop preconditioners for a variety of problems, with the additional constraints and features that real-world applications require.

REFERENCES

[1] URI M. ASHER AND ELDAD HABER, *A multigrid method for distributed parameter estimation problems*, Electron. Trans. Numer. Anal., 15 (2003), pp. 1–17.

[2] MICHELE BENZI, GENE H. GOLUB, AND JÖRG LIESEN, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[3] GEORGE BIROS AND GÜNAY DOGAN, *A multilevel algorithm for inverse problems with elliptic PDE constraints*, Inverse Problems, 24 (2008), p. 034010 (18pp).

[4] G. BIROS AND O. GHATTAS, *Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. Part I: The Krylov-Schur solver*, SIAM J. Sci. Comput., 27 (2000).

[5] A. Borzì and V. Schulz, *Multigrid methods for PDE optimization*, SIAM Rev., 51 (2009), pp. 361–395.
[6] J. Boyle, M. D. Mihajlovic, and J. A. Scott, *HSL_MI20: an efficient amg preconditioner*, Tech. Report RAL-TR-2007-021, Department of Computational and Applied Mathematics, Rutherford Appleton Laboratory, 2007.
[7] D. Braess and D. Peisker, *On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning*, IMA J Numer. Anal., 6 (1986), pp. 393–404.
[8] 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.
[9] H. S. Dollar, Nicholas I. M. Gould, Martin Stoll, and Andrew J. Wathen, *Preconditioning saddle-point systems with applications in optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 249–270.
[10] H.C. Elman, *Multigrid and Krylov subspace methods for the discrete Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1995), pp. 755–770.
[11] Howard Elman, David Silvester, and Andy Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2005.
[12] M. Engel and M. Griebel, *A multigrid method for constrained optimal control problems*, tech. report, SFB-Preprint 406, Sonderforschungsbereich 611, Rheinische Friedrich-Wilhelms-Universität Bonn, 2008.
[13] E. Haber and U. Ascher, *Preconditioned all–at–once methods for large sparse parameter estimation problems*, Inverse Problems, 17 (2000), pp. 1847–1864.
[14] R. Herzog and E. Sachs, *Preconditioned conjugate gradient method for optimal control problems with control and state constraints*, Tech. Report Preprint-Number SPP1253-088, Deutsche Forschungsgemeinschaft, Priority Program 1253, 2009.
[15] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.
[16] Axel Klawonn, *Block-triangular preconditioners for saddle point problems with a penalty term*, SIAM J. Sci. Comput., 19 (1998), pp. 172–184.
[17] J. Liesen and B. N. Parlett, *On nonsymmetric saddle point matrices that allow conjugate gradient iterations*, Numer. Math., 108 (2008), pp. 605–624.
[18] A. Meyer and T. Steidten, *Improvement and experiments on the Bramble-Pasciak type CG for mixed problems in elasticity*, tech. report, TU Chemnitz, Germany, 2001.
[19] Malcolm F. Murphy, Gene H. Golub, and Andrew J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput., 21 (2000), pp. 1969–1972.
[20] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[21] Tyrone Rees, H. Sue Dollar, and Andrew J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 271–298.
[22] Tyrone Rees and Martin Stoll, *Block triangular preconditioners for PDE-constrained optimization*, Tech. Report 15/09, OCCAM, Oxford University Mathematical Institute, March 2009. (to appear in NLAA).
[23] Tyrone Rees, Martin Stoll, and Andrew J. Wathen, *All-at-once preconditioning in PDE-constrained optimization*, August 2009. (to appear in Kybernetika for special issue on Algorithmy meeting, Podbansk, Slovakia).
[24] Joachim Schöberl and Walter Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 752–773.
[25] D.J. Silvester and A.J. Wathen, *Fast iterative solution of stabilised Stokes systems Part II: Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.
[26] Martin Stoll, *Solving Linear Systems using the Adjoint*, PhD thesis, University of Oxford, 2009.
[27] Martin Stoll and Andy Wathen, *Combination preconditioning and the Bramble–Pasciak$^+$ preconditioner*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 582–608.
[28] ———, *Preconditioning for active set and projected gradient methods as semi-smooth Newton methods for PDE-constrained optimization with control constraints*, Tech. Report 09/25, Oxford Centre for Collaborative Applied Mathematics, 2009.
[29] H.S. Thorne, *Properties of linear systems in PDE-constrained optimization. Part i: Distributed control*, Technical Report RAL-TR-2009-017, Rutherford Appleton Laboratory, 2009.
[30] A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J Numer Anal, 7 (1987), pp. 449–457.

[31] A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electronic Transactions on Numerical Analysis, 34 (2009), pp. 125–135.

[32] Walter Zulehner, *Analysis of iterative methods for saddle point problems: a unified approach*, Math. Comput., 71 (2001), pp. 479–505.