SIAM J. MATRIX ANAL. APPL.
Vol. 33, No. 4, pp. 1126–1152

REGULARIZATION-ROBUST PRECONDITIONERS FOR TIME-DEPENDENT PDE-CONSTRAINED OPTIMIZATION PROBLEMS*

JOHN W. PEARSON†, MARTIN STOLL‡, AND ANDREW J. WATHEN†

Abstract. In this article, we motivate, derive, and test effective preconditioners to be used with the MINRES algorithm for solving a number of saddle point systems which arise in PDE-constrained optimization problems. We consider the distributed control problem involving the heat equation and the Neumann boundary control problem involving Poisson's equation and the heat equation. Crucial to the effectiveness of our preconditioners in each case is an effective approximation of the Schur complement of the matrix system. In each case, we state the problem being solved, propose the preconditioning approach, prove relevant eigenvalue bounds, and provide numerical results which demonstrate that our solvers are effective for a wide range of regularization parameter values, as well as mesh sizes and time-steps.

Key words. PDE-constrained optimization, saddle point systems, time-dependent PDE-constrained optimization, preconditioning, Krylov subspace solver

AMS subject classifications. Primary, 65F10, 65N22, 65F50; Secondary, 76D07

DOI. 10.1137/110847949

1. Introduction.

The development of fast iterative solvers for saddle point problems from a variety of applications is a subject attracting considerable attention in numerical analysis [12, 46, 57, 14]. As such problems become more complex, a natural objective in creating efficient solvers is to ensure that the computation time taken by the solver grows as close to linearly as possible with the mesh parameter of the discretized problem. In more detail, it is desirable that if the problem size doubles due to refinement of the mesh, then the computation time roughly doubles as well.

Recently, due to the development of efficient algorithms and increased computing power, the solution of optimal control problems with PDE constraints has become an increasingly active field [55, 27, 29]. The goal is to find efficient methods that solve the discretized problem with the objective in mind of creating preconditioners that again scale linearly with decreasing mesh size. The interested reader is referred to [48, 22, 40, 44, 50] and the references therein for steady (time-independent) problems and to [52, 53, 39, 51, 4] for unsteady (time-dependent) problems. There are also multigrid [20] approaches to both time-dependent and time-independent optimal control problems [25, 26, 54, 6, 7, 1, 19, 18].

Often, designing solvers that are insensitive to the mesh size is found to compromise the performance of the solver for small values of the regularization parameter inherent in PDE-constrained optimization problems, unless the approximation of the Schur complement of the matrix system is chosen carefully. Therefore, recently

*Received by the editors September 14, 2011; accepted for publication (in revised form) by M. Benzi June 28, 2012; published electronically October 11, 2012.
http://www.siam.org/journals/simax/33-4/84794.html

†Numerical Analysis Group, Mathematical Institute, 24–29 St Giles', Oxford, OX1 3LB, United Kingdom (john.pearson@worc.ox.ac.uk, wathen@maths.ox.ac.uk). The first author's work was supported by the Engineering and Physical Sciences Research Council (UK), grant EP/P505216/1.

‡Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany (stollm@mpi-magdeburg.mpg.de).

research has gone into developing preconditioners which are insensitive to the regularization as well as the mesh size; see [37, 48] for instance for such solvers for the Poisson control problem.

Here, we consider whether it is possible to build solvers for the time-dependent analogue of this problem, that is, the optimal control of the heat equation. We consider the distributed control problem and attempt to minimize a functional that is commonly used in the literature [55]. We also investigate solvers for the boundary control problem, first in the time-independent Poisson control case and then in the time-dependent heat equation control case. Further, we develop a solver for a distributed subdomain problem of this type.

This paper is structured as follows. In section 2, we outline some prerequisite saddle point theory, state the problems that we consider the iterative solution of, and describe a solver for the distributed Poisson control problem (originally detailed in [37]) that we base our methods on. In section 3, we motivate and derive the preconditioners that we apply for the problems stated, proving relevant eigenvalue bounds of the preconditioned Schur complements of the matrix systems when our recommended approximations are used. In section 4, we provide numerical results for a variety of test problems to demonstrate the effectiveness of our approaches, and in section 5 we make some concluding remarks.

2. Problems and discretization. This section is structured as follows. In section 2.1, we briefly detail elements of saddle point theory that we utilize throughout the remainder of this paper. In section 2.2, we describe work that has been undertaken on the (time-independent) distributed Poisson control problem and state the formulations of the time-dependent problem that we consider. In section 2.3, we describe the time-independent and time-dependent Neumann boundary control problems we consider in this paper.

2.1. Saddle point theory. The problems we discuss in this paper are all of *saddle point structure*, i.e., of the form

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \tag{2.1}$$

where $A \in \mathbb{R}^{m\times m}$ is symmetric and positive definite or semidefinite, $B \in \mathbb{R}^{p\times m}$ with $m \geq p$ and the matrix $\mathcal{A}$ is nonsingular. The properties and solution methods for such systems have been an active field of research for two decades. State-of-the-art numerical methods for solving saddle point problems can be found in [3, 12] and the references therein.

Throughout this paper, we consider block diagonal preconditioners for such saddle point systems of the form

$$\mathcal{P} = \begin{bmatrix} \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix},$$

which is symmetric and positive definite. To apply this preconditioner, we therefore require a good approximation $\widehat{A}$ to the $(1,1)$-block of the matrix system, A, and $\widehat{S}$ as an approximation to the (negative) *Schur complement*, $S := BA^{-1}B^T$. Note that in general we are only interested in the application of $\widehat{A}^{-1}$ and $\widehat{S}^{-1}$, which allows the use of multigrid [20] or algebraic multigrid (AMG) [45, 13] methods, for example.

Such a preconditioner is known to be effective because the spectrum of the matrix $\mathcal{P}^{-1}\mathcal{A}$ is given by

$$\lambda(\mathcal{P}^{-1}\mathcal{A}) = \left\{1, \frac{1}{2}(1 \pm \sqrt{5})\right\},$$

provided $\mathcal{P}^{-1}\mathcal{A}$ is nonsingular, when $\widehat{A} = A$, and $\widehat{S} = S$ (see [34] for details). In this case, an appropriate Krylov subspace method applied to the system (2.1) will converge in three iterations with this preconditioner. Throughout the remainder of this paper, we apply the MINRES algorithm of Paige and Saunders [35] to saddle point systems of the form $\mathcal{A}$, with preconditioner $\mathcal{P}$ as in (2.2).

Note that many other preconditioners are possible such as block triangular preconditioners [34, 8, 42, 49] or constraint preconditioners [11, 30, 59]. These usually have to be combined with different iterative solvers, either symmetric ones [8, 16] or nonsymmetric ones such as GMRES [47].

2.2. Distributed control problems. One of the most common problems employed in PDE-constrained optimization for the study of numerical techniques is the *distributed Poisson control problem* with Dirichlet boundary conditions [55]. This is written as

$$\begin{aligned} \min_{y,u} \ & \frac{1}{2}\|y - \bar{y}\|^2_{L_2(\Omega_1)} + \frac{\beta}{2}\|u\|^2_{L_2(\Omega_2)} \\ \text{s.t} \quad & -\nabla^2 y = u \quad \text{in } \Omega, \\ & y = f \quad \text{on } \partial\Omega, \end{aligned} \tag{2.2}$$

where y is referred to as the *state* variable with $\bar{y}$ some known *desired state* and u as the *control variable*. Here Ω_1 and Ω_2 are subsets of the domain $\Omega \subset \mathbb{R}^d$, where $d \in \{2, 3\}$, on which the problem is defined with boundary $\partial\Omega$, and $\beta > 0$ is the (Tikhonov) regularization parameter. Note that we will limit ourselves to the cases $\Omega_2 = \Omega$ and $\Omega_2 = \partial\Omega$—the boundary control problem is addressed in the next section.

There are two common approaches for solving this optimization problem. One can consider the infinite-dimensional problem, write down the Lagrangian, and then discretize the first order conditions, which is referred to as the *optimize-then-discretize* approach, or one can first discretize the objective function and then build a discrete Lagrangian with corresponding first order conditions. The latter is the *discretize-then-optimize* approach. Recently, the paradigm that both approaches should coincide was used to derive discretization schemes for PDE-constrained optimization (see, for example, [25]).

The problem (2.2) represents a *steady* problem, i.e., $y = y(\mathbf{x})$, where $\mathbf{x}$ denotes the spatial variable. Using a Galerkin finite element method [12] and a discretize-then-optimize strategy, with the state y, control u, and *adjoint* state or Lagrange multiplier p all discretized using the same basis functions [40, 37], leads to the following first order system:

$$\begin{bmatrix} M_1 & 0 & K \\ 0 & \beta M & -M \\ K & -M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} M\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{c} \end{bmatrix}, \tag{2.3}$$

where $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{p}$ denote the vectors of coefficients in the finite element expansion in terms of the basis functions $\{\phi_j,\ j = 1, \ldots, n\}$ of y, u, and p, respectively, $\bar{\mathbf{y}}$ is the

vector corresponding to $\bar{y}$, and $\mathbf{c}$ corresponds to the Dirichlet boundary conditions imposed. Here, M denotes a finite element *mass matrix* over the domain Ω; similarly, M_1 is the finite element mass matrix for the domain Ω_1 and K a *stiffness matrix* over Ω. The matrices are of dimension $n \times n$ with n being the degrees of freedom of the finite element approximation. These are defined by

$$M = \{m_{ij},\ i,j = 1,\dots,n\}, \quad m_{ij} = \int_\Omega \phi_i \phi_j \ \mathrm{d}\Omega, \tag{2.4}$$
$$K = \{k_{ij},\ i,j = 1,\dots,n\}, \quad k_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \ \mathrm{d}\Omega.$$

Note that we often consider M to be a lumped mass matrix, that is,

$$M = \mathrm{diag}(m_{ii}), \quad m_{ii} = \sum_{j=1}^{n} \left| \int_\Omega \phi_i \phi_j \ \mathrm{d}\Omega \right|.$$

The matrix M_1 can be obtained analogously to the above by replacing Ω by Ω_1.

In literature such as [37, 48], solvers are designed which solve (2.3) in computational time independent of the mesh size h and any choice of regularization parameter β. The solver that we consider is based on the block diagonal preconditioner discussed in [37], in which the system (2.3) is written in classical saddle point form (2.1) with $A = \begin{bmatrix} M & 0 \\ 0 & \beta M \end{bmatrix}$ and $B = \begin{bmatrix} K & -M \end{bmatrix}$. The $(1,1)$-block is then approximated by the application of Chebyshev semi-iteration to each mass matrix for consistent mass matrices [58] or by simple inversion for lumped mass matrices, and the (negative) Schur complement

$$S = BA^{-1}B^T = KM^{-1}K + \frac{1}{\beta}M$$

is approximated by

$$\widehat{S} = \left(K + \frac{1}{\sqrt{\beta}}M\right) M^{-1} \left(K + \frac{1}{\sqrt{\beta}}M\right).$$

It is shown in [37] that $\lambda(\widehat{S}^{-1}S) \in [\frac{1}{2}, 1]$ for any choice of step-size h and regularization parameter β when this approximation is used. Using a multigrid process to approximate the inverse of the matrix $K + \frac{1}{\sqrt{\beta}}M$ gives a viable solution strategy.

In this paper, we attempt to extend this preconditioning framework to time-dependent analogues of the above problem. Specifically, we will consider the optimal control of the heat equation. This problem may be written as

$$\begin{aligned} &\min_{y,u} \ J(y,u) \\ &\text{s.t} \quad y_t - \nabla^2 y = u, \quad \text{for } (\mathbf{x},t) \in \Omega \times [0,T], \\ &\qquad\qquad y = f \quad \text{on } \partial\Omega, \\ &\qquad\qquad y = y_0 \quad \text{at } t = 0 \end{aligned} \tag{2.5}$$

for some functional $J(y,u)$, where f and y_0 may depend on $\mathbf{x}$ but not t. The functional that we consider here is a functional where we have observations (desired state) on the whole time-interval

$$J_1(y,u) = \frac{1}{2}\int_0^T \int_{\Omega_1} \left(y(\mathbf{x},t) - \bar{y}(\mathbf{x},t)\right)^2 \ \mathrm{d}\Omega_1 \mathrm{d}t + \frac{\beta}{2}\int_0^T \int_{\Omega_2} \left(u(\mathbf{x},t)\right)^2 \ \mathrm{d}\Omega_2 \mathrm{d}t. \tag{2.6}$$

Note that it is also possible to consider a problem where the desired state is only defined at a more limited set at times, for example, at only $t = T$, which would correspond to a functional of the form [36]

$$J_2(y,u) = \frac{1}{2}\int_{\Omega_1} \left(y(\mathbf{x},T) - \bar{y}(\mathbf{x})\right)^2 \, \mathrm{d}\Omega_1 + \frac{\beta}{2}\int_0^T \int_{\Omega_2} \left(u(\mathbf{x},t)\right)^2 \, \mathrm{d}\Omega_2 \mathrm{d}t.$$

We consider here only the problem relating to the functional $J_1(y,u)$, which we refer to as the "all-times case." Note that the state, control, and adjoint state are all now time-dependent functions. For now we again assume that $\Omega_2 = \Omega$.

As illustrated in [52], the matrix system arising from solving the problem (2.5) with $J(y,u) = J_1(y,u)$ varies according to whether a *discretize-then-optimize* or *optimize-then-discretize* strategy is applied. Applying the discretize-then-optimize approach, using the trapezoidal rule and the backward Euler scheme with N_t time steps of (constant) size τ to discretize the PDE in time, gives the matrix system [52]

$$\begin{bmatrix} \tau\mathcal{M}^{(1)}_{1/2} & 0 & \mathcal{K}^T \\ 0 & \beta\tau\mathcal{M}_{1/2} & -\tau\mathcal{M} \\ \mathcal{K} & -\tau\mathcal{M} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \tau\mathcal{M}^{(1)}_{1/2}\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{d} \end{bmatrix}, \tag{2.7}$$

where $\mathbf{y}$, $\mathbf{u}$, $\bar{\mathbf{y}}$, and $\mathbf{p}$ are vectors corresponding to the state, control, desired state, and adjoint at all time-steps $1, 2, \ldots, N_t$, and

$$\mathcal{M}_{1/2} = \begin{bmatrix} \frac{1}{2}M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & \frac{1}{2}M \end{bmatrix}, \quad \mathcal{M} = \begin{bmatrix} M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & M \end{bmatrix}, \tag{2.8}$$

$$\mathcal{M}^{(1)}_{1/2} = \begin{bmatrix} \frac{1}{2}M_1 & & & & \\ & M_1 & & & \\ & & \ddots & & \\ & & & M_1 & \\ & & & & \frac{1}{2}M_1 \end{bmatrix},$$

$$\mathcal{K} = \begin{bmatrix} M+\tau K & & & & \\ -M & M+\tau K & & & \\ & \ddots & \ddots & & \\ & & -M & M+\tau K & \\ & & & -M & M+\tau K \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} M\mathbf{y}_0 + \mathbf{c} \\ \mathbf{c} \\ \vdots \\ \mathbf{c} \\ \mathbf{c} \end{bmatrix}.$$

Note that if n is the number of degrees of freedom in the spatial representation only, then each of the matrices in (2.8) belongs to $\mathbb{R}^{nN_t \times nN_t}$ with blocks as indicated, where $M, M_1, K \in \mathbb{R}^{n\times n}$. The overall coefficient matrix in (2.7) is of dimension $3nN_t \times 3nN_t$.

If, alternatively, the optimize-then-discretize approach is used with $J(y,u) = J_1(y,u)$, the matrix system becomes [52]

$$\begin{bmatrix} \tau\mathcal{M}_0 & 0 & \mathcal{K}^T \\ 0 & \beta\tau\mathcal{M}_{1/2} & -\tau\mathcal{M} \\ \mathcal{K} & -\tau\mathcal{M} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \tau\mathcal{M}_0\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{d} \end{bmatrix}, \tag{2.9}$$

where

$$\mathcal{M}_0 = \begin{bmatrix} M_1 & & & & \\ & M_1 & & & \\ & & \ddots & & \\ & & & M_1 & \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{nN_t \times nN_t}.$$

The matrix systems (2.7) and (2.9) are the systems corresponding to the time-dependent distributed control problem. The efficient solution of these saddle point systems will be considered in this paper.

2.3. Neumann boundary control problems. Another important problem in the field of PDE-constrained optimization is the class of *Neumann boundary control problems*. Note that this problem corresponds to $\Omega_2 = \partial\Omega$ in (2.2). In practical applications, these are perhaps the most useful class of problems. We start once more by considering the boundary control of Poisson's equation written as

$$\begin{aligned} \min_{y,u} \ & \frac{1}{2} \|y - \bar{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2} \|u\|^2_{L_2(\partial\Omega)} \\ \text{s.t} \quad & -\nabla^2 y = f \quad \text{in } \Omega, \\ & \frac{\partial y}{\partial n} = u \quad \text{on } \partial\Omega, \end{aligned} \tag{2.10}$$

where f is the known source term, which may be zero, and the control, u, is applied in the form of a Neumann boundary condition. As for the distributed control case, we discretize y, u, and p using the same finite element basis functions.

The first order optimality conditions of a discretize-then-optimize approach yield the following matrix system:

$$\begin{bmatrix} M & 0 & K \\ 0 & \beta M_b & -N^T \\ K & -N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} M\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{f} \end{bmatrix}, \tag{2.11}$$

where M and K are as before (see (2.4)), M_b here denotes the boundary mass matrix over $\partial\Omega$, and N corresponds to entries arising from terms within the integral $\int_{\partial\Omega} u\mathrm{tr}(v)\mathrm{d}s$ (with u the boundary control and $\mathrm{tr}(v)$ denoting the trace function acting on a member of the Galerkin test space). The vector $\mathbf{f}$ corresponds to f, the source term of Poisson's equation. The matrix in (2.11) is essentially of dimension $(2n + n_b) \times (2n + n_b)$, where n is the number of degrees of freedom for y and n_b the number of degrees of freedom for the boundary control, u.

As well as this problem, we also investigate the time-dependent analogue, that is, the Neumann boundary control of the heat equation. We write the problem that we consider as

$$\begin{aligned} \min_{y,u} \ & \frac{1}{2} \int_0^T \int_\Omega \left(y(\mathbf{x},t) - \bar{y}(\mathbf{x},t)\right)^2 \ \mathrm{d}\Omega\mathrm{d}t + \frac{\beta}{2} \int_0^T \int_{\partial\Omega} \left(u(\mathbf{x},t)\right)^2 \ \mathrm{d}s\mathrm{d}t, \\ \text{s.t.} \quad & y_t - \nabla^2 y = f \quad \text{for } (\mathbf{x},t) \in \Omega \times [0,T], \\ & \frac{\partial y}{\partial n} = u \quad \text{on } \partial\Omega. \end{aligned} \tag{2.12}$$

Note that this is related to the distributed control problem (2.5) with $J(y,u) = J_1(y,u)$. Although we could seek to solve the optimize-then-discretize formulation of this problem in a similar way as for the distributed control problem, we focus our attention on the discretize-then-optimize formulation. In this case, applying the backward Euler scheme in time and the trapezoidal rule, we obtain the matrix system

$$\begin{bmatrix} \tau\mathcal{M}_{1/2} & 0 & \mathcal{K}^T \\ 0 & \beta\tau\mathcal{M}_{1/2,b} & -\tau\mathcal{N}^T \\ \mathcal{K} & -\tau\mathcal{N} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \tau\mathcal{M}_{1/2}\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{g} \end{bmatrix}, \tag{2.13}$$

where $\mathcal{M}$ and $\mathcal{K}$ are as defined in (2.8), and

$$\mathcal{M}_{1/2,b} = \begin{bmatrix} \frac{1}{2}M_b & & & & \\ & M_b & & & \\ & & \ddots & & \\ & & & M_b & \\ & & & & \frac{1}{2}M_b \end{bmatrix},$$

$$\mathcal{N} = \begin{bmatrix} N & & & & \\ & N & & & \\ & & \ddots & & \\ & & & N & \\ & & & & N \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} M\mathbf{y}_0 + \mathbf{f} \\ \mathbf{f} \\ \vdots \\ \mathbf{f} \\ \mathbf{f} \end{bmatrix}.$$

We will consider the iterative solution of the matrix systems (2.11) and (2.13), in addition to the distributed control problems previously stated, in section 3.

2.4. Possible extensions. In this section we wish to introduce some extensions of the above problems that in one form or another frequently appear in the field of optimization with PDE constraints. In many applications so-called box constraints for the state and/or the control have to be included. Here we highlight pointwise control constraints

$$u_a(x) \le u(x) \le u_b(x)$$

as well as pointwise state constraints

$$y_a(x) \le y(x) \le y_b(x).$$

These additional constraints can be handled very efficiently by so-called semismooth Newton methods [27, 23, 56, 28], whereas due to the reduced regularity of the Lagrange multiplier the state-constrained problem presents a more difficult problem [9]. It is also possible to include different or additional regularization terms in the objective function. A popular choice is the inclusion of a so-called sparsity term where the control u is given in the L_1-norm for which we write $\|\mathbf{u}\|_1$. This term can efficiently be treated as part of the semismooth Newton method (see [22]). Another possibility is to have differential operators acting on the control as part of the objective function, for which we write $\|Lu\|_2$. In this case efficient preconditioning depends on the nature of the operator L and how well it can be approximated. Recent examples for this can be found in [43, 4]. Combinations of all the above are of course possible and we address some possibilities in the next section.

3. Preconditioning. In this section, we motivate and discuss our proposed preconditioners for the matrix systems stated in section 2. These will be applied within the MINRES algorithm [35]. This section is structured as follows. In section 3.1.1, we propose a preconditioner for the matrix system (2.7) corresponding to a time-dependent distributed control problem, minimizing (2.6) and using a discretize-then-optimize formulation. We start with the case $\Omega_1 = \Omega$ and discuss the subdomain case next. In section 3.1.2, we motivate a preconditioner for (2.9), which is the same problem except with an optimize-then-discretize strategy employed. We then consider Neumann boundary control problems for the case $\Omega_1 = \Omega$; in section 3.2, we discuss the time-independent case corresponding to (2.11), and in section 3.3 we extend this theory to the time-dependent case, relating to (2.13). We only discuss the subdomain case $\Omega_1 \subset \Omega$ for the time-dependent problem in section 3.4. In section 4, we present numerical results to demonstrate that all our proposed solvers are effective in practice.

3.1. Time-dependent distributed control.

3.1.1. Minimizing J_1 with discretize-then-optimize. We start by considering the case $\Omega_1 = \Omega$, which gives $\mathcal{M}^{(1)}_{1/2} = \mathcal{M}_{1/2}$. Equation (2.7), which is the discretize-then-optimize formulation of (2.5) with $J(y,u) = J_1(y,u)$, can be written as a saddle point system with

$$A = \begin{bmatrix} \tau\mathcal{M}_{1/2} & 0 \\ 0 & \beta\tau\mathcal{M}_{1/2} \end{bmatrix}, \quad B = \begin{bmatrix} \mathcal{K} & -\tau\mathcal{M} \end{bmatrix},$$

in the notation of (2.1). The (negative) Schur complement of this system is therefore given by

$$S = \frac{1}{\tau}\mathcal{K}\mathcal{M}^{-1}_{1/2}\mathcal{K}^T + \frac{\tau}{\beta}\mathcal{M}\mathcal{M}^{-1}_{1/2}\mathcal{M}. \tag{3.1}$$

For this matrix system, we seek a (symmetric block diagonal) preconditioner of the form

$$\widehat{\mathcal{P}} = \begin{bmatrix} \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix} \tag{3.2}$$

to be used with MINRES.

For the approximation $\widehat{A}$, we apply a similar approach as for the Poisson control problem and take

$$\widehat{A} = \begin{bmatrix} \tau\widehat{\mathcal{M}}_{1/2} & 0 \\ 0 & \beta\tau\widehat{\mathcal{M}}_{1/2} \end{bmatrix}, \tag{3.3}$$

where $\widehat{\mathcal{M}}_{1/2}$ denotes the approximation of $\mathcal{M}_{1/2}$. Here a Chebyshev semi-iteration process is again taken to approximate consistent mass matrices or a simple inversion for lumped mass matrices.

We now wish to develop a result which enables us to find an accurate approximation to (3.1), as well as to approximate Schur complements that we will consider in section 3.1.2.

We start by noting that the matrix system (2.7) is of the form

$$\begin{bmatrix} \Phi_1 & 0 & \mathcal{K}^T \\ 0 & \beta\Phi_1 & -\Phi_2 \\ \mathcal{K} & -\Phi_2 & 0 \end{bmatrix}$$

with Schur complement given by

$$S = \mathcal{K}\Phi_1^{-1}\mathcal{K}^T + \frac{1}{\beta}\Phi_2\Phi_1^{-1}\Phi_2, \tag{3.4}$$

where Φ_1 and Φ_2 are symmetric positive definite, as they are block matrices solely consisting of mass matrices. (In section 3.1.2, we will consider approximations of Schur complements of the form (3.4), where Φ_1 and Φ_2 have the same such properties.)

We note that in all the cases we consider, the matrix $\Phi_1^{-1}\Phi_2$ simply involves scaled (positive) multiples of identity matrices. That is, all the relevant blocks are scalings of the same matrix $I \in \mathbb{R}^{n\times n}$. We may use the straightforward resulting observation that $\mathcal{M}\Phi_1^{-1}\Phi_2 = \Phi_1^{-1}\Phi_2\mathcal{M}$ with $\mathcal{M}$ defined as in (2.8) to demonstrate one further property that we will require in our analysis: that $\mathcal{K}\Phi_1^{-1}\Phi_2 + \Phi_1^{-1}\Phi_2\mathcal{K}^T$ is positive definite. We show this by applying Theorem 1 below with $\Delta = \Phi_1^{-1}\Phi_2$.

THEOREM 1. *The matrix $\mathcal{K}\Delta + \Delta\mathcal{K}^T$, where $\Delta = blkdiag(\alpha_1 I, \alpha_2 I, \ldots, \alpha_{N_t} I)$, $\alpha_1, \ldots, \alpha_{N_t} > 0$, $I \in \mathbb{R}^{n\times n}$, and $\mathcal{K}$ is as defined in (2.8), is positive definite.*

Proof. We show that $\mathbf{w}^T(\mathcal{K}\Delta+\Delta\mathcal{K}^T)\mathbf{w} > 0$ for all $\mathbf{w} := \begin{bmatrix} \mathbf{w}_1^T & \mathbf{w}_2^T & \cdots & \mathbf{w}_{N_t-1}^T & \mathbf{w}_{N_t}^T \end{bmatrix}^T$ with $\mathbf{w}_1, \ldots, \mathbf{w}_{N_t} \in \mathbb{R}^n$, and

$$\Delta = \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_{N_t} \end{bmatrix}, \quad \Delta_j \in \mathbb{R}^{n\times n}, \quad j = 1, \ldots, N_t,$$

with $\Delta_j = \alpha_j I$, $j = 1, \ldots, N_t$.

Using the symmetry of the mass and stiffness matrices M and K,

$$\mathcal{K}\Delta + \Delta\mathcal{K}^T = \begin{bmatrix} \Lambda_1 & -\Delta_1 M & & & & \\ -M\Delta_1 & \Lambda_2 & -\Delta_2 M & & & \\ & \ddots & \ddots & \ddots & & \\ & & -M\Delta_{N_t-2} & \Lambda_{N_t-1} & -\Delta_{N_t-1}M & \\ & & & -M\Delta_{N_t-1} & \Lambda_{N_t} & \end{bmatrix},$$

where $\Lambda_j = (M + \tau K)\Delta_j + \Delta_j(M + \tau K)$ for $j = 1, \ldots, N_t$ and therefore by straightforward manipulation that

$$\begin{aligned} \mathbf{w}^T(\mathcal{K}\Delta + \Delta\mathcal{K}^T)\mathbf{w} &= \sum_{j=1}^{N_t} \mathbf{w}_j^T[M\Delta_j + \Delta_j M + \tau K\Delta_j + \tau\Delta_j K]\mathbf{w}_j \\ &\quad - \sum_{j=1}^{N_t-1} \mathbf{w}_j^T(M\Delta_j)\mathbf{w}_{j+1} - \sum_{j=2}^{N_t} \mathbf{w}_j^T(\Delta_{j-1}M)\mathbf{w}_{j-1} \\ &= 2\tau\sum_{j=1}^{N_t} \mathbf{w}_j^T(K\Delta_j)\mathbf{w}_j + \sum_{j=1}^{N_t-1} (\mathbf{w}_j - \mathbf{w}_{j+1})^T(M\Delta_j)(\mathbf{w}_j - \mathbf{w}_{j+1}) \\ &\quad + \mathbf{w}_1^T(M\Delta_1)\mathbf{w}_1 + \mathbf{w}_{N_t}^T(M\Delta_{N_t})\mathbf{w}_{N_t}, \end{aligned} \tag{3.5}$$

where we have used the facts that $M\Delta_j = \Delta_j M$ and $K\Delta_j = \Delta_j K$ for $j = 1, \ldots, N_t$, which are clear by the definition of Δ.

As we now have that $\mathbf{w}^T(\mathcal{K}\Delta+\Delta\mathcal{K}^T)\mathbf{w}$ is a sum of positive multiples of (symmetric positive definite) mass and stiffness matrices, we deduce that $\mathbf{w}^T(\mathcal{K}\Delta+\Delta\mathcal{K}^T)\mathbf{w} > 0$ and hence that $\mathcal{K}\Delta + \Delta\mathcal{K}^T$ is positive definite. ∎

Having demonstrated the properties required, we are now in a position to prove a result bounding the eigenvalues of $\widehat{S}^{-1}S$, where

$$\widehat{S} = \left(\mathcal{K} + \frac{1}{\sqrt{\beta}}\Phi_2\right)\Phi_1^{-1}\left(\mathcal{K} + \frac{1}{\sqrt{\beta}}\Phi_2\right)^T \tag{3.6}$$

and S is given by (3.4). To do this, we consider the Rayleigh quotient $R := \frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}\mathbf{v}}$. This may be written as

$$R = \frac{\mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b} + \mathbf{a}^T\mathbf{b} + \mathbf{b}^T\mathbf{a}}, \tag{3.7}$$

where

$$\mathbf{a} = \Phi_1^{-1/2}\mathcal{K}^T\mathbf{v}, \quad \mathbf{b} = \frac{1}{\sqrt{\beta}}\Phi_1^{-1/2}\Phi_2\mathbf{v}.$$

Now, as $\mathbf{a}^T\mathbf{b} + \mathbf{b}^T\mathbf{a} = \frac{1}{\sqrt{\beta}}\mathbf{v}^T[\mathcal{K}\Phi_1^{-1}\Phi_2 + \Phi_2\Phi_1^{-1}\mathcal{K}^T]\mathbf{v} > 0$ due to Theorem 1 with $\Delta_j = \Phi_1^{-1}\Phi_2 = \Phi_2\Phi_1^{-1}$, it is clear from (3.7) that $R < 1$.

Further, showing that $R \geq \frac{1}{2}$ is a simple algebraic task, which requires only the fact that $\mathbf{b}^T\mathbf{b} > 0$ because of the positive definiteness of Φ_1 and Φ_2. (See [38] for further details.)

We have hence proved the next theorem.

THEOREM 2. *If S and $\widehat{S}$ are of the form stated in* (3.4) *and* (3.6) *respectively, with Φ_1, Φ_2 symmetric positive definite and $\Phi_1^{-1}\Phi_2 = blkdiag(\alpha_1 I, \alpha_2 I, \ldots, \alpha_{N_t} I)$, $\alpha_1, \ldots, \alpha_{N_t} > 0$, $I \in \mathbb{R}^{n\times n}$, then*

$$\lambda(\widehat{S}^{-1}S) \in \left[\frac{1}{2}, 1\right].$$

We note that Theorem 2 is an extension to a result discussed in [38] concerning convection-diffusion control.

We may now apply Theorem 2 with $\Phi_1 = \tau\mathcal{M}_{1/2}$ and $\Phi_2 = \tau\mathcal{M}$, as Φ_1 and Φ_2 defined in this way are clearly symmetric and positive definite and are such that $\Delta = \Phi_1^{-1}\Phi_2$ is symmetric positive definite and satisfies $\mathcal{M}\Delta = \Delta\mathcal{M}$. We therefore deduce that

$$\widehat{S} = \frac{1}{\tau}\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}\right)\mathcal{M}_{1/2}^{-1}\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}\right)^T \tag{3.8}$$

is an effective approximation to the Schur complement of the matrix system (2.7). We note that applying the inverses of the matrix $\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}$ and its transpose would not be feasible as this essentially means solving the PDE directly, which in itself is a computationally expensive task. Hence, for a practical algorithm we approximate $\widehat{S}$ using multigrid techniques for $\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}$ and its transpose, that is, we require a multigrid process for each of the diagonal blocks $M + \tau K + \frac{\tau}{\sqrt{\beta}}M \in \mathbb{R}^{n\times n}$. We apply a few cycles of such a multigrid process N_t times to approximate the inverse of $\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}$ and N_t times to approximate the inverse of $\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\mathcal{M}\right)^T$.

In conclusion, for an effective iterative method for solving (2.7), we recommend a MINRES method with a preconditioner of the form (3.2), with $\widehat{A}$ and $\widehat{S}$ as in (3.3) and (3.8). In section 4, we provide numerical results to demonstrate the effectiveness of our proposed preconditioner.

3.1.2. Minimizing J_1 with optimize-then-discretize. We now turn our attention to (2.9), the optimize-then-discretize formulation of (2.3) with $J(y,u) = J_1(y,u)$. Again, we may write this as a saddle point system of the form (2.1) with

$$A = \begin{bmatrix} \tau\mathcal{M}_0 & 0 \\ 0 & \beta\tau\mathcal{M}_{1/2} \end{bmatrix}, \quad B = \begin{bmatrix} \mathcal{K} & -\tau\mathcal{M} \end{bmatrix}.$$

We note that the $(1,1)$-block of this system, A, is not invertible, due to the rank-deficiency of $\mathcal{M}_0$, so when prescribing an approximation for a preconditioner, we recommend considering a perturbation of the matrix $\mathcal{M}_0$

$$\mathcal{M}_0^\gamma = \begin{bmatrix} M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & \gamma M \end{bmatrix}$$

for some constant γ such that $0 < \gamma \ll 1$, and taking as our approximation to A the following:

$$\widehat{A} = \begin{bmatrix} \tau\widehat{\mathcal{M}}_0 & 0 \\ 0 & \beta\tau\widehat{\mathcal{M}}_{1/2} \end{bmatrix}, \tag{3.9}$$

where $\widehat{\mathcal{M}}_0$ and $\widehat{\mathcal{M}}_{1/2}$ denote approximations to $\mathcal{M}_0^\gamma$ and $\mathcal{M}_{1/2}$, generated by using Chebyshev semi-iteration in the case of consistent mass matrices, or, in the case of lumped mass matrices, themselves.

Now, due to the noninvertibility of $\mathcal{M}_0$, the Schur complement of the matrix system (2.9) does not exist. Therefore it is less obvious what the $(2,2)$-block of our block diagonal preconditioner of the form (3.2) should be. The heuristic we use is to examine the perturbed saddle point system $\begin{bmatrix} \widehat{A} & B^T \\ B & 0 \end{bmatrix}$ and consider the Schur complement of this matrix system. This is given by the quantity

$$\tilde{S} := \frac{1}{\tau}\mathcal{K}\widehat{\mathcal{M}}_0^{-1}\mathcal{K}^T + \frac{\tau}{\beta}\mathcal{M}\mathcal{M}_{1/2}^{-1}\mathcal{M}.$$

Now, by simple manipulation, we observe that

$$\tilde{S} = \frac{1}{\tau}\mathcal{K}\widehat{\mathcal{M}}_0^{-1}\mathcal{K}^T + \frac{\tau}{\beta}\Gamma_1\widehat{\mathcal{M}}_0^{-1}\Gamma_1,$$

where

$$\Gamma_1 = \begin{bmatrix} \sqrt{2}M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & \sqrt{2\gamma}M \end{bmatrix}.$$

By applying Theorem 2 with $\Phi_1 = \tau\widehat{\mathcal{M}}_0$ and $\Phi_2 = \tau\Gamma_1$, we therefore deduce that

$$\widehat{S} = \frac{1}{\tau}\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\Gamma_1\right)\mathcal{M}_0^{-1}\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\Gamma_1\right)^T \tag{3.10}$$

satisfies

$$\lambda(\widehat{S}^{-1}\tilde{S}) \in \left[\frac{1}{2}, 1\right],$$

which tells us that $\widehat{S}$ is a good Schur complement approximation to the perturbed matrix system we have considered. As the matrix system (2.9) is very similar in structure to this perturbed system, it seems that this would also be a pragmatic choice for the $(2,2)$-block of our block diagonal preconditioner for this system.

Therefore, within the MINRES algorithm for solving (2.9), we again recommend a preconditioner of the form (3.2) with $\widehat{A}$ and $\widehat{S}$ as in (3.9) and (3.10). The numerical results of section 4 demonstrate that this is indeed an effective approach.

3.2. Time-independent Neumann boundary control. We now consider preconditioning the system (2.11), which arises when solving the time-independent Poisson boundary control problem. If we write the saddle point system in the form (2.1) with

$$A = \begin{bmatrix} M & 0 \\ 0 & \beta M_b \end{bmatrix}, \quad B = \begin{bmatrix} K & -N \end{bmatrix},$$

then constructing an approximation $\widehat{A}$ to the $(1,1)$-block A is relatively straightforward, as we treat both mass matrices M and M_b as before. However, an issue arises when we consider the effective approximation of the Schur complement of (2.11)

$$S = KM^{-1}K + \frac{1}{\beta}NM_b^{-1}N^T.$$

Because of the rank-deficiency of the $\frac{1}{\beta}NM_b^{-1}N^T$ term of the Schur complement, it is not as simple to find a clean and easy-to-invert approximation $\widehat{S}$ to S such that the eigenvalues of $\widehat{S}^{-1}S$ may be pinned down into an interval independent of both h and β, as for the distributed control case in section 2.2. We therefore seek an approximation which is robust for a range of h and β. We first wish to motivate our choices before analyzing them in more detail.

We assume now that all mass matrices are lumped. It is then easy to see that $NM_b^{-1}N^T$ is a diagonal matrix with nonzero entries on the diagonal for every boundary node. For simplicity we assume the degrees of freedom are ordered in such a way that the nodes located on the boundary can be found in the lower right corner of $NM_b^{-1}N^T$, i.e.,

$$NM_b^{-1}N^T = \begin{bmatrix} 0 & 0 \\ 0 & M_b \end{bmatrix}.$$

Now our task is to approximate the Schur complement S via

$$\widehat{S} = \left(K + \frac{1}{\sqrt{\beta}}\widehat{M}\right) M^{-1} \left(K + \frac{1}{\sqrt{\beta}}\widehat{M}\right)$$

for some matrix $\widehat{M}$ in such a way that the structure of the original Schur complement is maintained as much as possible. If we look at the last equation we see this gives

$$\widehat{S} = KM^{-1}K^T + \frac{1}{\beta}\widehat{M}M^{-1}\widehat{M} + \frac{1}{\sqrt{\beta}}\left(KM^{-1}\widehat{M} + \widehat{M}M^{-1}K\right).$$

We now look at the terms separately. The first one is part of the original Schur complement. The second one needs to be looked at more carefully. Hence

$$\begin{bmatrix} 0 & 0 \\ 0 & \alpha M_b \end{bmatrix} \begin{bmatrix} M_{y,i}^{-1} & 0 \\ 0 & M_{y,b}^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \alpha M_b \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha^2 M_b M_{y,b}^{-1} M_b \end{bmatrix}$$

with i and b denoting interior and boundary, respectively, and for some constant α. This tells us now that if

$$\alpha^2 M_b M_{y,b}^{-1} M_b \approx M_b,$$

we have found a good approximation to the Schur complement of the original matrix, which can be evaluated efficiently. A simplification will now motivate our choice of α as, if we approximate $M_b = hI_b$ (where I_b is the identity matrix of dimension equal to the number of boundary nodes) and $M_{y,b} = h^2 I$, we obtain that

$$\alpha^2 M_b M_{y,b}^{-1} M_b = M_b \iff \alpha^2 h h^{-2} h I = \alpha^2 I \approx hI, \tag{3.11}$$

and hence a good choice for α seems to be $\alpha = \sqrt{h}$. As a result, our recommended Schur complement approximation is now defined as

$$\widehat{S}_1 = \left(K + \sqrt{\frac{h}{\beta}}M_\Gamma\right) M^{-1} \left(K + \sqrt{\frac{h}{\beta}}M_\Gamma\right),$$

i.e., the matrix $\widehat{M}$ introduced earlier is given by $\sqrt{h}M_\Gamma$. We note that because of the diagonal nature of the mass matrices the matrix $M_\Gamma = NM_b^{-1}N^T$ is simple to evaluate. Another choice with a similar motivation is given by

$$\widehat{S}_2 = \left(K + \sqrt{\frac{h}{\beta}}M_\Gamma\right) (h\widehat{M}_\Gamma)^{-1} \left(K + \sqrt{\frac{h}{\beta}}M_\Gamma\right).$$

Here $\widehat{M}_\Gamma$ is given by the matrix M_b in the boundary components and a small scalar of order h for all nodes corresponding to the degrees of freedom on the interior, i.e.,

$$\widehat{M}_\Gamma = M_\Gamma + hI_i,$$

with I_i a diagonal matrix with ones on the diagonal for all interior degrees of freedom and zeros elsewhere. We now wish to analyze these two preconditioners in more detail by considering the eigenvalue distributions of $\widehat{S}_1^{-1}S$ and $\widehat{S}_2^{-1}S$. Our analysis is based on the two-dimensional problem, however it can be easily extended to the three-dimensional case.

Eigenvalues of $\widehat{S}_1^{-1}S$. Here we must consider the Rayleigh quotient

$$\begin{aligned}\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_1\mathbf{v}} &= \frac{\mathbf{v}^T KM^{-1}K\mathbf{v} + \frac{1}{\beta}\mathbf{v}^T M_\Gamma\mathbf{v}}{\mathbf{v}^T KM^{-1}K\mathbf{v} + \frac{h}{\beta}\mathbf{v}^T M_\Gamma M^{-1}M_\Gamma\mathbf{v} + \sqrt{\frac{h}{\beta}}\mathbf{v}^T\left[M_\Gamma M^{-1}K + KM^{-1}M_\Gamma\right]\mathbf{v}} \\ &= \frac{\mathbf{v}^T KM^{-1}K\mathbf{v} + \mathbf{v}^T\left(\frac{1}{\beta}M_\Gamma\right)\mathbf{v}}{\mathbf{v}^T KM^{-1}K\mathbf{v} + \frac{h}{\beta}\mathbf{v}^T M_\Gamma M^{-1}M_\Gamma\mathbf{v} + 2\sqrt{\frac{h}{\beta}}\mathbf{v}^T M_\Gamma M^{-1}K\mathbf{v}},\end{aligned}$$

which will provide us with the eigenvalues of $\widehat{S}_1^{-1}S$.

If $\mathbf{v} \in \text{null}(M_\Gamma)$, then $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1 \mathbf{v}} = 1$. If not, then we can write the above also as

$$\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1 \mathbf{v}} = \frac{1}{\frac{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{h}{\beta}\mathbf{v}^T M_\Gamma M^{-1} M_\Gamma \mathbf{v}}{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{1}{\beta}\mathbf{v}^T M_\Gamma \mathbf{v}} + \frac{2\sqrt{\frac{h}{\beta}}\mathbf{v}^T M_\Gamma M^{-1} K\mathbf{v}}{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{1}{\beta}\mathbf{v}^T M_\Gamma \mathbf{v}}}. \tag{3.12}$$

Using the fact that $M_\Gamma \left(\frac{1}{h}M\right)^{-1} M_\Gamma = h M_\Gamma M^{-1} M_\Gamma$ and M_Γ are spectrally equivalent, we can see that

$$0 < \frac{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{h}{\beta}\mathbf{v}^T M_\Gamma M^{-1} M_\Gamma \mathbf{v}}{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{1}{\beta}\mathbf{v}^T M_\Gamma \mathbf{v}} =: D_1 = \mathcal{O}(1),$$

where D_1 is a mesh and β-independent constant.

We now examine the term

$$\frac{2\sqrt{\frac{h}{\beta}}\mathbf{v}^T M_\Gamma M^{-1} K\mathbf{v}}{\mathbf{v}^T K M^{-1} K\mathbf{v} + \frac{1}{\beta}\mathbf{v}^T M_\Gamma \mathbf{v}} =: \frac{T_1}{T_2},$$

in particular its maximum and minimum values, more carefully. We assume now that $M \approx h^2 I$ and $M_\Gamma \approx hI$, ignoring all multiplicative constants. Furthermore, we note that the eigenvalues of K are within the interval $\left[c_K h^2, C_K\right]$, where c_K and C_K are constants independent of h and β (apart from a single zero eigenvalue with a corresponding eigenvector of ones—this corresponds to an arbitrary constant being a solution of the continuous Neumann problem for Poisson's equation).

As we work with lumped mass matrices throughout our work on Neumann boundary control, we observe that $T_1 \geq 0$, as it relates to a positive constant multiplied by the product of two matrices ($M_\Gamma M^{-1}$, which we have assumed to be approximately $h^{-1}I$, and K, which is symmetric positive definite). We also note that T_2 must be strictly positive.[1]

We now consider the maximum and minimum values of $\frac{T_1}{T_2}$. We consider the maximum such value by writing

$$\frac{T_1}{T_2} = \frac{\beta^{-1/2} h^{1/2} h^1 h^{-2} c}{h^{-2}c^2 + \beta^{-1} h} = \frac{\beta^{-1/2} h^{-1/2} c}{h^{-2}(c^2 + \beta^{-1} h^3)} = \frac{ac}{c^2 + a^2}$$

with $a = h^{3/2}\beta^{-1/2}$ and c corresponding to the relevant eigenvalue of K. Here, both a and c are positive. Therefore, in this case, $\frac{ac}{c^2+a^2} \leq \frac{1}{2}$ by straightforward algebraic manipulation. This means that the denominator in (3.12) will be bounded above by a constant independent of h, β, and τ, as both terms are of $\mathcal{O}(1)$. This gives us a lower bound for $\lambda_{\min}$.

As T_1 and T_2 are both nonnegative, we may write that $\frac{T_1}{T_2} \geq 0$ and hence that $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1 \mathbf{v}} \geq \frac{1}{D_1}$, giving us an upper bound for $\lambda_{\max}$.

Putting our analysis together, and reinstating multiplicative constants, we conclude that

$$\lambda_{\min}(\widehat{S}_1^{-1} S) = c_1, \quad \lambda_{\max}(\widehat{S}_1^{-1} S) = C_1,$$

where c_1 and C_1 are positive constants independent of h, β, and τ.

[1] This may be argued as follows. Both $\mathbf{v}^T K M^{-1} K\mathbf{v}$ and $\frac{1}{\beta}\mathbf{v}^T M_\Gamma \mathbf{v}$ are nonnegative terms. The former will be strictly positive unless $\mathbf{v}$ is the vector of ones, which corresponds to the zero eigenvalue of K. In this case, it is clear that the $\mathbf{v}^T M_\Gamma \mathbf{v}$ term will be strictly positive, as none of the entries of M_Γ are negative. So for each $\mathbf{v}$, at least one term will be strictly positive.

Eigenvalues of $\widehat{S}_2^{-1}S$. We may carry out a similar analysis for the approximation $\widehat{S}_2$ of S by considering the Rayleigh quotient

$$\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}} = \frac{\mathbf{v}^T K M^{-1}K\mathbf{v} + \mathbf{v}^T\left(\frac{1}{\beta}M_\Gamma\right)\mathbf{v}}{\mathbf{v}^T K(h\widehat{M}_\Gamma)^{-1}K\mathbf{v} + \frac{h}{\beta}\mathbf{v}^T M_\Gamma(h\widehat{M}_\Gamma)^{-1}M_\Gamma\mathbf{v} + 2\sqrt{\frac{h}{\beta}}\mathbf{v}^T M_\Gamma(h\widehat{M}_\Gamma)^{-1}K\mathbf{v}},$$

and writing that $M \approx h^2 I$, $\widehat{M}_\Gamma \approx hI$, and $M_\Gamma \approx \text{blkdiag}(0, hI_b)$.

Proceeding as we did for the analysis of $\widehat{S}_1$, we obtain that

$$\lambda_{\min}(\widehat{S}_2^{-1}S) = c_2, \quad \lambda_{\max}(\widehat{S}_2^{-1}S) = C_2,$$

where c_2 and C_2 are positive constants independent of h, β, and τ, provided we use lumped mass matrices.

We emphasize that due to the rank-deficient nature of the $\frac{1}{\beta}NM_b^{-1}N^T$ term of the Schur complement S, it is more difficult to obtain a complete picture of the eigenvalue distributions of $\widehat{S}_1^{-1}S$ and $\widehat{S}_2^{-1}S$ than for the preconditioned Schur complement in the distributed control case. Consequently, the bounding of $\lambda(\widehat{S}_1^{-1}S)$ and $\lambda(\widehat{S}_2^{-1}S)$ by constants of $\mathcal{O}(1)$ is less descriptive than the more specific bound outlined for distributed control in [37] and discussed in section 2.2.

However, the conclusion that the eigenvalues of $\widehat{S}_1^{-1}S$ and $\widehat{S}_2^{-1}S$ are certainly real and bounded above and below by constants of $\mathcal{O}(1)$, independently of h, β, and τ, indicates that either S_1 or S_2 should serve as an effective approximation of S—a hypothesis which is verified by the numerical results presented in section 4. We note that in the above analysis, we have assumed that lumped mass matrices are being used; however, numerical tests indicate that we still obtain a clean bound when using consistent mass matrices.

3.3. Time-dependent Neumann boundary control. In the case of the time-dependent boundary control problem, we are interested in approximating the Schur complement

$$S = \frac{1}{\tau}\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T + \frac{\tau}{\beta}\mathcal{N}\mathcal{M}_{1/2,b}^{-1}\mathcal{N}^T \tag{3.13}$$

of the saddle point matrix $\mathcal{A}$. We want to approximate the above by

$$\widehat{S}_3 = \tau^{-1}\left(\mathcal{K} + \frac{\tau}{\sqrt{\beta}}\widehat{\mathcal{M}}\right)\mathcal{M}_{1/2}^{-1}\left(\mathcal{K}^T + \frac{\tau}{\sqrt{\beta}}\widehat{\mathcal{M}}\right), \tag{3.14}$$

and for this to be a good approximation the choice of $\widehat{\mathcal{M}}$ is again crucial. We recall that we assumed $\mathcal{M}_{1/2,b}$ to be a block diagonal matrix of lumped boundary mass matrices and also that $\mathcal{M}_{1/2}$ consists of lumped mass matrices over the domain Ω. Hence the first term in (3.14) is given by $\tau^{-1}\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T$, which means that the first term in the Schur complement (3.13) is well represented in our approximation. We then obtain the next term from (3.14) as

$$\frac{\tau^{-1}\tau\tau}{\sqrt{\beta}\sqrt{\beta}}\widehat{\mathcal{M}}\mathcal{M}_{1/2}^{-1}\widehat{\mathcal{M}} = \frac{\tau}{\beta}\widehat{\mathcal{M}}\mathcal{M}_{1/2}^{-1}\widehat{\mathcal{M}}.$$

To understand how this approximates $\mathcal{N}\mathcal{M}_{1/2,b}^{-1}\mathcal{N}^T$, we need to study the structure of both matrix products more carefully. We recall that $\mathcal{M}_{1/2,b} = \text{blkdiag}(\frac{1}{2}M_b, M_b, \ldots,$

$M_b, \frac{1}{2}M_b)$ and that with some abuse of notation $\mathcal{N} = \text{blkdiag}_{rec}(N, \ldots, N)$, giving for the overall structure

$$\mathcal{N}\mathcal{M}_{1/2,b}^{-1}\mathcal{N}^T = \begin{bmatrix} 2NM_b^{-1}N^T & & & & \\ & NM_b^{-1}N^T & & & \\ & & \ddots & & \\ & & & NM_b^{-1}N^T & \\ & & & & 2NM_b^{-1}N^T \end{bmatrix}.$$

We see that as $\mathcal{M}_{1/2} = \text{blkdiag}\left(\frac{1}{2}M, M, \ldots, M, \frac{1}{2}M\right)$ and $\widehat{M} = \text{blkdiag}(\widehat{M}, \ldots, \widehat{M})$, the structure of the large problem looks as follows:

$$\widehat{\mathcal{M}}\mathcal{M}_{1/2}^{-1}\widehat{\mathcal{M}} = \begin{bmatrix} 2\widehat{M}M^{-1}\widehat{M} & & & & \\ & \widehat{M}M^{-1}\widehat{M} & & & \\ & & \ddots & & \\ & & & \widehat{M}M^{-1}\widehat{M} & \\ & & & & 2\widehat{M}M^{-1}\widehat{M} \end{bmatrix}.$$

This indicates that it is important for $\widehat{M}M^{-1}\widehat{M} \approx NM_b^{-1}N^T$, which we split up even further now. Consider an ordering of the degrees of freedom on the boundary and in the interior as before,

$$\widehat{M}M^{-1}\widehat{M} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha M_b \end{bmatrix} \begin{bmatrix} M_{y,i}^{-1} & 0 \\ 0 & M_{y,b}^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \alpha M_b \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha^2 M_b M_{y,b}^{-1} M_b \end{bmatrix},$$

and now note that

$$NM_b^{-1}N^T = \begin{bmatrix} 0 & 0 \\ 0 & M_b \end{bmatrix},$$

where $M_{y,i}$ and $M_{y,b}$ denote the splitting of the mass matrix M into its interior and boundary parts, respectively. Similar to before, we can show that $\alpha = \sqrt{h}$ is a good choice. A choice not very different from the above is given by the approximation

$$\widehat{S}_4 = \tau^{-1}\left(\mathcal{K} + \tau\sqrt{\frac{h}{\beta}}\widehat{\mathcal{M}}\right)(h\widehat{\mathcal{M}}_\Gamma)^{-1}\left(\mathcal{K} + \tau\sqrt{\frac{h}{\beta}}\widehat{\mathcal{M}}\right)^T, \tag{3.15}$$

where $\widehat{\mathcal{M}}_\Gamma$ consists of block diagonal matrices that have the boundary mass matrix for the boundary nodes and a suitably scaled identity matrix for the interior nodes. (See also the time-independent case.)

Eigenvalues of $\widehat{S}_4^{-1}S$. We now search for the eigenvalues of $\widehat{S}_4^{-1}S$, where

$$\widehat{S}_4 = \tau^{-1}\left(\mathcal{K} + \tau\sqrt{\frac{h}{\beta}}\widehat{\mathcal{M}}\right)(h\widehat{\mathcal{M}}_\Gamma)^{-1}\left(\mathcal{K} + \tau\sqrt{\frac{h}{\beta}}\widehat{\mathcal{M}}\right)^T, \tag{3.16}$$

by considering the Rayleigh quotient

$$\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_4\mathbf{v}} = \frac{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T\mathbf{v} + \tau\beta^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\mathcal{K}\mathbf{v} + \frac{\tau h}{\beta}\mathbf{v}^T\widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}}\mathbf{v} + 2\sqrt{\frac{h}{\beta}}\mathbf{v}^T\mathcal{K}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}}\mathbf{v}},$$

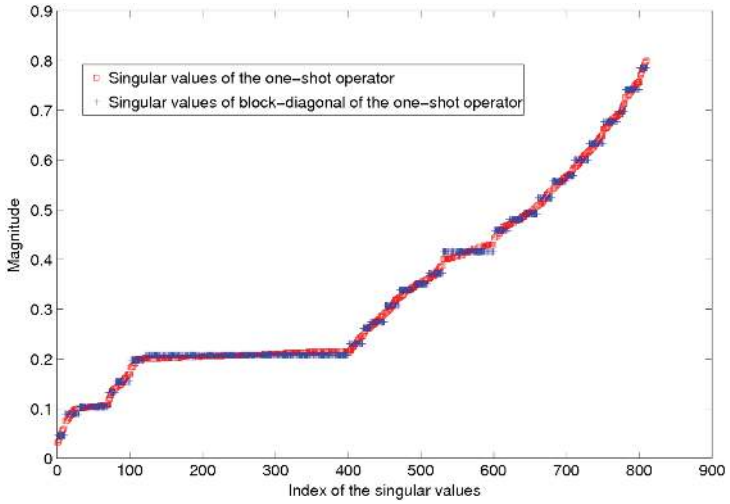


FIG. 3.1. *Singular values of $\mathcal{L}$ and $\mathcal{K}$ for a small example.*

using the fact that $\widehat{\mathcal{M}}_\Gamma = \mathcal{N}\mathcal{M}_{1/2,b}^{-1}\mathcal{N}^T$. Assuming that $\mathbf{v} \in \text{null}(\widehat{\mathcal{M}})$, we obtain that

$$\frac{\mathbf{v}^T S \mathbf{v}}{\mathbf{v}^T \widehat{S}_4 \mathbf{v}} = \mathcal{O}(1).$$

So we now consider the case where $\mathbf{v}$ is not in this nullspace; we then examine the term

$$\frac{1}{\frac{\tau^{-1}\mathbf{v}^T\mathcal{K}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\mathcal{K}\mathbf{v}+\frac{\tau h}{\beta}\mathbf{v}^T\widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}}\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T\mathbf{v}+\tau\beta^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\mathbf{v}} + \sqrt{\frac{h}{\beta}}\frac{\mathbf{v}^T\left(\mathcal{K}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}}+\widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\mathcal{K}^T\right)\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T\mathbf{v}+\tau\beta^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\mathbf{v}}}.$$

So if we now assume (neglecting constants for now) that $h\widehat{\mathcal{M}}_\Gamma \approx \mathcal{M}_{1/2} \approx h^2 I$ and $\widehat{\mathcal{M}} \approx \widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}} \approx hI$, we see that

$$\frac{\tau^{-1}\mathbf{v}^T\mathcal{K}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\mathcal{K}\mathbf{v} + \frac{\tau h}{\beta}\mathbf{v}^T\widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}}\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_{1/2}^{-1}\mathcal{K}^T\mathbf{v} + \tau\beta^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\mathbf{v}} = \mathcal{O}(1).$$

In order to simplify the analysis at this stage we simply assume that $\mathcal{K}$ is approximated by its block diagonal, i.e., $\mathcal{L} \approx \mathcal{K}$ (see Figure 3.1). We use this to approximate the above by

$$\sqrt{\frac{h}{\beta}}\frac{\mathbf{v}^T\left(\mathcal{L}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\widehat{\mathcal{M}} + \widehat{\mathcal{M}}(h\widehat{\mathcal{M}}_\Gamma)^{-1}\mathcal{L}\right)\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{L}\mathcal{M}_{1/2}^{-1}\mathcal{L}\mathbf{v} + \tau\beta^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\mathbf{v}} =: \frac{T_1}{T_2}.$$

We may proceed as in section 3.2 for the time-independent boundary control case to obtain (neglecting constants)

$$\frac{T_1}{T_2} = \frac{h^{1/2}\beta^{-1/2}h^{-1}c}{\tau^{-1}h^{-2}c^2 + \tau\beta^{-1}h} = \frac{\beta^{-1/2}h^{-1/2}c}{h^{-2}\tau^{-1}(c^2 + \tau^2\beta^{-1}h^3)} = \frac{\tau\beta^{-1/2}h^{3/2}c}{c^2 + \tau^2\beta^{-1}h^3} = \frac{ac}{c^2 + a^2} \leq \frac{1}{2}$$

with $a = \tau\beta^{-1/2}h^{3/2}$ and $c \in \left[c_K\tau h^2 + c_M h^2, C_k\tau + C_M h^2\right]$. This shows that the results for the time-independent case can be used here as well. For the minimum value of $\frac{T_1}{T_2}$, we may apply a similar analysis as in the case of $\widehat{S}_1^{-1}S$ and working once more with lumped mass matrices. We obtain that (reintroducing constants)

$$\lambda_{\min}(\widehat{S}_4^{-1}S) = c_4, \quad \lambda_{\max}(\widehat{S}_4^{-1}S) = C_4,$$

where c_4 and C_4 are positive constants independent of h, β and τ.

A similar analysis can be carried out for $\widehat{S}_3^{-1}S$. As for the time-independent case, it is more difficult to develop a complete picture of the eigenvalue distribution of the preconditioned Schur complement than for the distributed control case; however, it is useful to see that we may bound the eigenvalues by constants of $\mathcal{O}(1)$ independently of the parameters h, β, and τ. Indeed, the results shown in section 4 show that the performance for the preconditioners for the time-dependent and time-independent boundary control problems is quite similar, and we find that both approximations $\widehat{S}_3$ and $\widehat{S}_4$ are effective for this problem for a wide range of parameters.

3.4. The subdomain case. We now wish to address the case when the desired state is only given on a subdomain Ω_1 of Ω. The saddle point system is then defined by

$$A = \begin{bmatrix} \tau\mathcal{M}_{1/2}^{(1)} & 0 \\ 0 & \beta\tau\mathcal{M}_{1/2} \end{bmatrix}, \quad B = \begin{bmatrix} \mathcal{K} & -\tau\mathcal{M} \end{bmatrix},$$

and we note that the matrix A is only positive semidefinite as the matrix $\mathcal{M}_{1/2}^{(1)}$ is semidefinite. However, we wish to obtain an invertible approximation of A, as well as an effective Schur complement approximation, as in previous sections. For that purpose we introduce a parameter $\gamma \in \mathbb{R}$ such that the matrix $\mathcal{M}_{1/2}^{(1,\gamma)} = \text{blkdiag}\left(\frac{1}{2}M_1^\gamma, M_1^\gamma, \ldots, M_1^\gamma, \frac{1}{2}M_1^\gamma\right)$ with M_1^γ defined as

$$(M_1^\gamma)_{\Omega\setminus\Omega_1} = \gamma I \quad \text{or} \quad (M_1^\gamma)_{\Omega\setminus\Omega_1} = \gamma M_{\Omega\setminus\Omega_1}.$$

Note that we use the same notation for the small parameter, namely, γ, dealing with the zero parts of the $(1,1)$-block and believe it will be clear from the context what γ represents. The $(1,1)$-block of this perturbed problem may now be approximated by $\widehat{A} = \text{blkdiag}(\tau\widehat{\mathcal{M}}_{1/2}^{(1,\gamma)}, \beta\tau\widehat{\mathcal{M}}_{1/2})$, where $\widehat{\mathcal{M}}_{1/2}^{(1,\gamma)}$ now denotes the relevant approximation of mass matrices (Chebyshev semi-iteration or diagonal solves) within the matrix $\mathcal{M}_{1/2}^{(1,\gamma)}$. The Schur complement of this perturbed problem that we now wish to approximate is given by

$$\widetilde{S} = \frac{1}{\tau}\mathcal{K}(\mathcal{M}_{1/2}^{(1,\gamma)})^{-1}\mathcal{K}^T + \frac{\tau}{\beta}\mathcal{M}\mathcal{M}_{1/2}^{-1}\mathcal{M}.$$

Again our goal is to derive an approximation to the Schur complement that exhibits robustness with respect to the regularization parameter. For this we consider

$$\widehat{S} = \frac{1}{\tau}(\mathcal{K} + \widehat{\mathcal{M}})(\mathcal{M}_{1/2}^{(1,\gamma)})^{-1}(\mathcal{K} + \widehat{\mathcal{M}})^T,$$

where we have to define $\widehat{\mathcal{M}}$. Ideally, we have agreement between the terms $\frac{1}{\tau}\widehat{\mathcal{M}}(\mathcal{M}_{1/2}^{(1,\gamma)})^{-1}\widehat{\mathcal{M}} \approx \frac{\tau}{\beta}\mathcal{M}\mathcal{M}_{1/2}^{-1}\mathcal{M}$. Assuming now that all mass matrices are lumped we can give an elementwise description of what we wish to achieve, i.e.,

$$\widehat{m}_{ii}^2 = \frac{\tau^2}{\beta}\left(m^{(1,\gamma)}\right)_{ii} m_{ii},$$

that is,

$$\widehat{m}_{ii} = \frac{\tau}{\sqrt{\beta}} \sqrt{\left(m^{(1,\gamma)}\right)_{ii}} \sqrt{m_{ii}}. \tag{3.17}$$

We now have to distinguish between indices i that represent degrees of freedom within Ω_1 or in $\Omega \setminus \Omega_1$. In more detail,

$$\left(m^{(1,\gamma)}\right)_{ii} = \begin{cases} m_{ii} & \text{if } i \in \Omega_1, \\ \gamma & \text{otherwise.} \end{cases} \tag{3.18}$$

We have now established an expression for the elements of $\widehat{\mathcal{M}}$ in the case of the distributed control problem. We find that the resulting Schur complement approximation works well in practice—we demonstrate this once again with numerical results in section 4.

Choice of γ. We now explain how we select in practice the "perturbation parameter" γ that we have utilized in previous sections. We start by deriving the parameter γ for the case when optimize-then-discretize is used for the distributed control problem. We assume that we want both terms of the Schur complement

$$S = \mathcal{K}\widehat{\mathcal{M}}_0^{-1}\mathcal{K}^T + \tau\beta^{-1}\mathcal{M}_2$$

with $\widehat{\mathcal{M}}_0 = \text{blkdiag}(M, \ldots, M, \gamma M)$, $\mathcal{M}_2 = \text{blkdiag}(2M, M \ldots, M, 2M)$ to be "balanced" (see [4, 52]). We simplify this task by replacing $\mathcal{K}$ by its block diagonal $\mathcal{L} := \text{blkdiag}(L, \ldots, L)$, where $L = M + \tau K$. We now wish to balance the terms in this new approximation with a particular focus on the parameter γ, i.e.,

$$\widehat{S} = \mathcal{L}\widehat{\mathcal{M}}_0^{-1}\mathcal{L}^T + \tau\beta^{-1}\mathcal{M}_2.$$

Comparing the blocks in $\widehat{S}$ that involve γ, we obtain

$$\gamma^{-1}h^{-2}L^2 \approx \tau\beta^{-1}h^2 I, \tag{3.19}$$

using the approximation $M = h^2 I$ for a two-dimensional problem. In this heuristic, we want to balance the smallest eigenvalues of both terms; for $L^2 = \tau^2 K^2 + \tau KM + \tau MK + M^2$ these will be of the order $\tau^2 h^4$ (neglecting constants). In order for γ to balance both terms in (3.19), we get

$$\gamma^{-1}h^{-2}\tau^2 h^4 \approx \tau\beta^{-1}h^2$$

and therefore that

$$\tau\beta \approx \gamma. \tag{3.20}$$

Note that the above heuristic holds for the two-dimensional case. In complete analogy, we can derive that

$$\tau\beta \approx \gamma \tag{3.21}$$

is also a good choice for problems in three dimensions. If one wants to balance the largest eigenvalues in both terms the parameter γ might not be small, depending on the choice of τ and β. In a very similar way we can derive a heuristic for the parameter

γ in the subdomain case. To solve the distributed control problem we replace the zero entries by γ to give

$$\gamma^{-1}\tau^2 h^4 \approx \tau\beta^{-1}h^2 \Rightarrow \gamma = \tau\beta h^2. \tag{3.22}$$

Rees and Greif [41] also introduce a similar parameter γ that is part of a perturbation of the $(1,1)$-block of a saddle point problem coming from the treatment of a quadratic program using interior point methods. They construct a preconditioner with an augmented $(1,1)$-block, i.e., $A+\gamma^{-1}BB^T$, using the classical saddle point notation, where their parameter γ is chosen to balance the two summands A and BB^T, similar to our heuristic above.

4. Numerical results. The results presented in this section are based on an implementation of the above described algorithms within the deal.II [2] framework using $Q1$ finite elements. For the AMG preconditioner, we used the Trilinos ML package [15] that implements a smoothed aggregation AMG. Within the AMG we typically used 10 steps of a Chebyshev smoother in combination with the application of two V-cycles. Our implementation of MINRES was taken from [12] and was stopped with a tolerance of 10^{-4} for the relative pseudoresidual. Our experiments are performed for $T = 1$ and $\tau = 0.05$, i.e., 20 time-steps. We consider homogeneous Dirichlet conditions for distributed control problems, though we are of course not limited to them, and also a zero forcing term $f = 0$ for Neumann boundary control problems. We carried out the examples on the domain $\Omega = [0,1]^3$. Whenever we show the degrees of freedom these are only the degrees of freedom for one grid point in time (i.e., for a single time-step). Implicitly, we are solving a linear system of dimension three times the number of time-steps (N_t) times the degrees of freedom of the spatial discretization (n). For example, a spatial discretization with 274,625 unknowns and 20 time-steps corresponds to an overall linear system of dimension 16,477,500. All results are performed on a Centos Linux machine with Intel Xeon CPU X5650 at 2.67 GHz CPUs and 48 GB of RAM.

4.1. Distributed control. We start by giving results for the distributed control examples presented earlier. For the distributed control problems we impose a zero Dirichlet condition. This results in the computed state not matching the desired state quite as well very close to the boundary. Another option would be to impose a Dirichlet condition where the state corresponds to the desired state on $\partial\Omega$.

4.1.1. The all-times case—whole domain. The example we consider for the distributed control problem is given by the all-times case, where the functional $J(y,u)$ contains observations for all time-steps. We have the choice of using the trapezoidal rule (which corresponds to the discretize-then-optimize formulation) or the rectangular rule (which corresponds to the optimize-then-discretize formulation) for the discretization of the state integral. We will show results for both cases that desired to drive the state close to the desired state given by

$$\bar{\mathbf{y}} = 64t\sin\left(2\pi\left((x_0-0.5)^2+(x_1-0.5)^2+(x_2-0.5)^2\right)\right)$$

with a zero initial value. An illustration of the desired state, the computed state, and the control is shown in Figure 4.1 for one particular point in time, i.e., one particular time-step. The results with the Schur complement approximation as presented in section 3.1.1 (trapezoidal rule) are shown in Table 4.1 and the results for the approach presented in section 3.1.2 (rectangular rule) are shown in Table 4.2. We can see that the number of iterations remains constant with varying mesh size and regularization parameter β.

4.1.2. The all-times case—subdomain problem. We now show results for the subdomain problem when the desired state is again defined by

$$\bar{\mathbf{y}} = 64t \sin\left(2\pi\left((x_0 - 0.5)^2 + (x_1 - 0.5)^2 + (x_2 - 0.5)^2\right)\right)$$

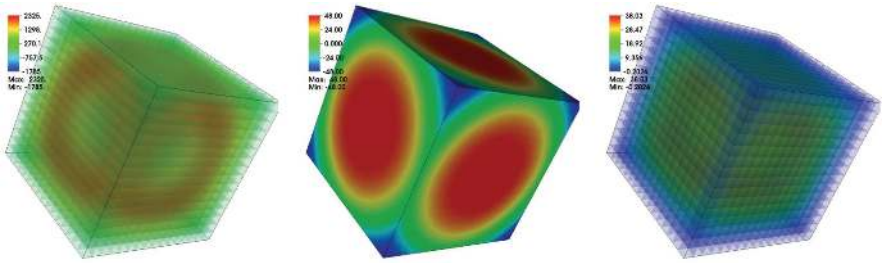
and the domain Ω_1 is defined by

$$\Omega_1 = \left\{x \in [0,1]^3 : 0.4 \leq x_1, x_2 \leq 0.7\right\}.$$

Results for this case are shown in Table 4.3, where we can again see that the iteration numbers are small and robust with respect to the mesh parameter and the regularization parameter. The timings are slightly higher than in the case for the whole domain. This is because in our experience the AMG approximation sometimes deteriorates for small parameters and as we now include γ in our approximation we decided to use four V-cycles instead of two.

4.2. Boundary control. We now present results for the time-independent and time-dependent Neumann boundary control problems discussed earlier.

4.2.1. Time-independent boundary control. The time-independent boundary control problem example that we present starts from initial value zero, matching



(a) Control (b) Desired state (c) Computed state

FIG. 4.1. *Control, desired state, and state for distributed control with* $\beta = 1e-4$ *at grid point* 15 *in time.*

TABLE 4.1
Results for discretize-then-optimize approach via trapezoidal rule.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	10(2)	12(2)	12(2)
35937	10(14)	12(17)	12(18)
274625	10(148)	12(171)	12(170)

TABLE 4.2
Results for optimize-then-discretize approach via rectangular rule.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	12(3)	10(2)	8(1)
35937	12(16)	10(14)	10(14)
274625	14(196)	10(152)	10(147)

TABLE 4.3
Results for discretize-then-optimize approach via trapezoidal rule for a subdomain problem.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	12(5)	13(5)	15(5)
35937	12(28)	15(35)	17(38)
274625	12(332)	15(386)	19(495)

TABLE 4.4
Results obtained with Schur complement approximation $\widehat{S}_1$.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	26(1)	28(1)	22(1)
35937	32(2)	38(2)	30(2)
274625	34(22)	48(31)	46(29)
2146689	38(211)	60(289)	64(314)

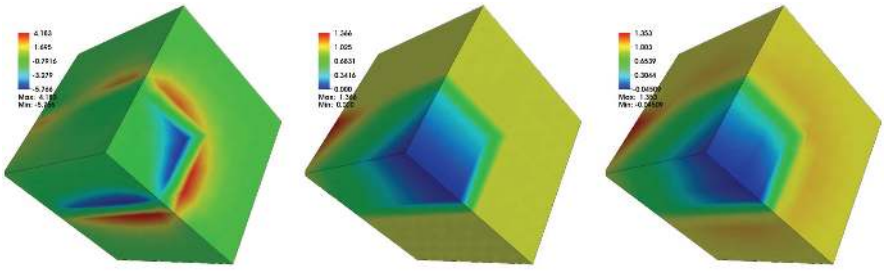
the desired state given by

$$\bar{\mathbf{y}} = \begin{cases} \sin(x_1) + x_2 x_0 & \text{if } x_0 > 0.5 \text{ and } x_1 < 0.5, \\ 1 & \text{otherwise.} \end{cases}$$

The desired state, computed state, and control are shown in Figure 4.2. The CPU times and iteration numbers for the MINRES algorithm with varying mesh size and regularization parameter are shown in Table 4.4 for the Schur complement approximation $\widehat{S}_1$ and in Table 4.5 for $\widehat{S}_2$. We see that $\widehat{S}_1$ performs better in all cases, although the results for $\widehat{S}_2$ are not dramatically different. We see for both approaches a slow growth in the iteration numbers, which is expected when dealing with a pure Neumann problem (see [5]). We observe some rather small growth with decreasing β, especially for small meshes, but with the iteration numbers still reasonably small. We also observe improved preformance when h^3 and β are further apart. The results we experience matched our expectations based on the theory detailed in section 3.2.

4.2.2. Time-dependent boundary control. The setup for the example time-dependent boundary control problem we present again starts with an initial value of zero and the following time-dependent desired state:

$$\bar{\mathbf{y}} = \begin{cases} \sin(t) + x_0 x_1 x_2 & \text{if } x_0 > 0.5 \text{ and } x_1 < 0.5, \\ 1 & \text{otherwise.} \end{cases}$$



(a) Control (b) Desired state (c) Computed state

FIG. 4.2. *Control, desired state, and state for boundary control with $\beta = 1e-4$.*

TABLE 4.5
Results obtained with Schur complement approximation $\widehat{S}_2$.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	38(1)	38(1)	30(1)
35937	44(3)	54(3)	44(3)
274625	48(31)	74(48)	70(44)
2146689	54(263)	98(466)	108(513)

TABLE 4.6
Results obtained with Schur complement approximation $\widehat{S}_3$.

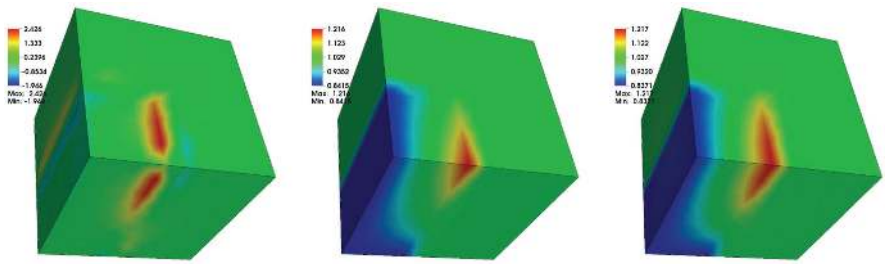
DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	34(7)	38(7)	28(6)
35937	38(49)	48(62)	38(48)
274625	48(620)	62(800)	58(725)

TABLE 4.7
Results obtained with Schur complement approximation $\widehat{S}_4$.

DoF	MINRES(T)	MINRES(T)	MINRES(T)
	$\beta = 1e-2$	$\beta = 1e-4$	$\beta = 1e-6$
4913	40(8)	42(8)	36(7)
35937	50(65)	59(73)	42(54)
274625	62(808)	80(1002)	68(855)

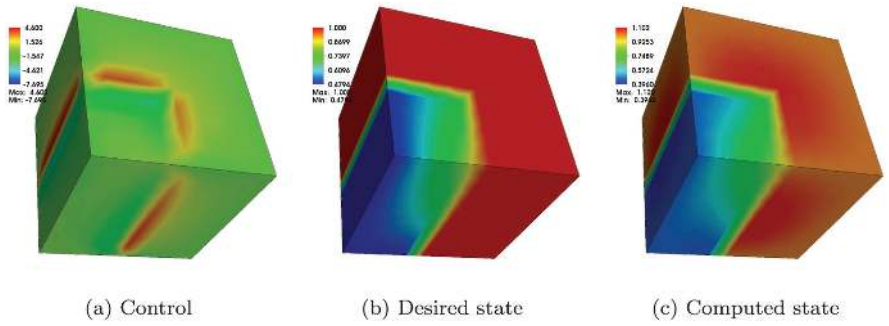
The desired state as well as the computed state and control are depicted for grid point 20 in time (i.e., the 20th time step) in Figure 4.3 and for grid point 10 (the 10th time step) in Figure 4.4. We again computed results for both Schur complement approximations presented earlier; the results are given in Table 4.6 for $\widehat{S}_3$ and in Table 4.7 for $\widehat{S}_4$. We again see higher iteration numbers for the second approximation $\widehat{S}_4$ and benign growth with respect to the mesh size, but again with improved results if h^3 and β are far apart. The results here reflect the results for the time-independent case, which we expect due to our theoretical study presented in section 3.3.

5. Concluding remarks and outlook. We have presented various setups for the optimal control of the heat equation. We derived the discretized first order conditions for the distributed and boundary control cases and showed that both problems



(a) Control (b) Desired state (c) Computed state

FIG. 4.3. *Control, desired state, and state for boundary control with* $\beta = 1e-6$ *at grid point* 20 *in time.*



(a) Control (b) Desired state (c) Computed state

FIG. 4.4. *Control, desired state, and state for boundary control with* $\beta = 1e-6$ *at grid point* 10 *in time.*

lead to a linear system with saddle point structure. We then extended the analysis for a regularization-robust preconditioner from the time-independent distributed control case to the time-dependent distributed control case. We also provided some bounds for the case of Neumann boundary control for the time-dependent and time-independent setups. We then gave an extensive numerical study of the preconditioners derived earlier and showed that the dependence with respect to the mesh size, regularization parameter, and time-step could be removed for the distributed control case. The numerical results for the pure Neumann control problem illustrated a benign dependence on the mesh size (similar to the forward problem) and very little dependence with respect to the regularization parameter β. These results have already been used in a different work on time-periodic parabolic problems with control constraints (see [51]), where good numerical results were obtained. The work presented in this paper also serves as a framework for the consideration of other time-dependent optimal control problems. The techniques presented could be adapted for the case where the control is only applied in a subdomain, or examples with additional constraints such as box constraints being imposed on the state or control.

A possible future extension of this work would be to develop robust preconditioners for more complicated PDEs with respect to all parameters involved. As well, one could generate solvers for the subdomain case, as discussed in this manuscript, or the boundary control setting. Furthermore, one drawback of the procedure described in this paper, which could be tackled in future work, is the necessary storage requirement for the vectors corresponding to the control, state, and adjoint. Although it is possible to condense the system by, for example, eliminating the control, more research would be required here. Various approaches have been applied to time-dependent PDE-constrained optimization in the past. For instance, checkpointing [17], a method which involves storing only certain checkpoints of the state and computing the adjoint state based on these, has been investigated. We note that our one-shot approach is not ideally suited for this method but rather could be treated using ideas based on instantaneous control [10, 24, 21], multiple shooting [21], and parareal schemes [31, 32, 33]. Possibly the simplest idea of all is to split up the interval into subintervals and use our approach to solve the relevant subproblems, for which all the analysis presented here carries over. However, we note that the solution obtained using this approach is suboptimal [21]. Parareal and shooting methods maintain the splitting into subintervals but ensure agreement of the control and state where the time-slices meet each other. We believe that the techniques presented here can be

used within multiple shooting methods such as that presented in [21]—this is another area of further work which will be investigated.

Acknowledgement. The authors would like to thank an anonymous referee for a careful reading of the manuscript and helpful comments.

REFERENCES

[1] U. Ascher and E. Haber, *A multigrid method for distributed parameter estimation problems.*, Electron. Trans. Numer. Anal., 15 (2003), pp. 1–17.
[2] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.* II—*a general-purpose object-oriented finite element library*, ACM Trans. Math. Software, 33 (2007), pp. Art. 24, 27.
[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer, 14 (2005), pp. 1–137.
[4] M. Benzi, E. Haber, and L. Taralli, *A preconditioning technique for a class of PDE-constrained optimization problems*, Adv. Comput. Math., 35 (2011), pp. 149–173.
[5] P. Bochev and R. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Rev., 47 (2005), pp. 50–66.
[6] A. Borzì, *Multigrid methods for parabolic distributed optimal control problems.*, J. Comput. Appl. Math., 157 (2003), pp. 365–382.
[7] A. Borzì and V. Schulz, *Multigrid methods for PDE optimization.*, SIAM Rev., 51 (2009), pp. 361–395.
[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] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.
[10] H. Choi, M. Hinze, and K. Kunisch, *Instantaneous control of backward-facing step flows*, Appl. Numer. Math., 31 (1999), pp. 133–158.
[11] S. Dollar, *Iterative Linear Algebra for Constrained Optimization*, Ph.D. thesis, University of Oxford, 2005; also available online from http://web.comlab.ox.ac.uk/oucl/research/na/thesis/thesisdollar.pdf.
[12] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2005.
[13] R. Falgout, *An Introduction to Algebraic Multigrid*, Comput. Sci. Engrg., 8 (2006), pp. 24–33.
[14] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Ser. Adv. Numer. Math., John Wiley & Sons, Chichester, UK, 1996.
[15] M. Gee, C. Siefert, J. Hu, R. Tuminaro, and M. Sala, *ML* 5.0 *Smoothed Aggregation User's Guide*, Tech. rep. SAND2006-2649, Sandia National Laboratories, 2006.
[16] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput, 23 (2001), pp. 1376–1395.
[17] A. Griewank and A. Walther, *Algorithm* 799*: revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation*, ACM Trans. Math. Software, 26 (2000), pp. 19–45.
[18] E. Haber, *A parallel method for large scale time domain electromagnetic inverse problems.*, Appl. Numer. Math., 58 (2008), pp. 422–434.
[19] E. Haber, U. M. Ascher, and D. Oldenburg, *On optimization techniques for solving nonlinear inverse problems.*, Inverse Problems, 16 (2000), pp. 1263–1280.
[20] W. Hackbusch, *Multigrid methods and applications*, Springer Ser. Comput. Math. 4, Springer-Verlag, Berlin, 1985.
[21] M. Heinkenschloss, *A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems.*, J. Comput. Appl. Math., 173 (2005), pp. 169–198.
[22] R. Herzog and E. W. Sachs, *Preconditioned conjugate gradient method for optimal control problems with control and state constraints*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2291–2317.
[23] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.
[24] M. Hinze, *Optimal and Instantaneous Control of the Instationary Navier-Stokes Equations*, Habilitation, TU Berlin, 2000.

[25] M. Hinze, M. Köster, and S. Turek, *A Hierarchical Space-Time Solver for Distributed Control of the Stokes Equation*, Priority Programme 1253, Tech. rep. SPP1253-16-01, 2008.
[26] M. Hinze, M. Köster, and S. Turek, *A Space-Time Multigrid Solver for Distributed Control of the Time-Dependent Navier-Stokes System*, Priority Programme 1253, Tech. rep. SPP1253-16-02, 2008.
[27] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, in Mathematical Modelling: Theory and Applications, Springer-Verlag, New York, 2009.
[28] K. Ito and K. Kunisch, *Semi-smooth Newton methods for state-constrained optimal control problems*, Systems Control Lett., 50 (2003), pp. 221–228.
[29] K. Ito and K. Kunisch, *Lagrange multiplier approach to variational problems and applications*, Adv. Des. Control 15, SIAM, Philadelphia, 2008.
[30] C. Keller, N. Gould, and A. Wathen, *Constraint Preconditioning for Indefinite Linear Systems*, SIAM J. Matrix Anal. Appl, 21 (2000), pp. 1300–1317.
[31] J. Lions, Y. Maday, and G. Turinici, *A "parareal" in time discretization of PDE's*, C. R. Math. Acad. Sci. Ser. Paris, 332 (2001), pp. 661–668.
[32] Y. Maday and G. Turinici, *A parareal in time procedure for the control of partial differential equations*, C. R. Math., 335 (2002), pp. 387–392.
[33] T. P. Mathew, M. Sarkis, and C. E. Schaerer, *Analysis of block parareal preconditioners for parabolic optimal control problems.*, SIAM J. Sci. Comput., 32 (2010), pp. 1180–1200.
[34] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput, 21 (2000), pp. 1969–1972.
[35] C. C. Paige and M. A. Saunders, *Solutions of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal, 12 (1975), pp. 617–629.
[36] J. W. Pearson, M. Stoll, and A. J. Wathen, *Robust Iterative Solution of a Class of Time-Dependent Optimal Control Problems*, submitted.
[37] J. W. Pearson and A. J. Wathen, *A New Approximation of the Schur Complement in Preconditioners for PDE Constrained Optimization*, Numer. Linear Algebra Appl., (2011), DOI 10.1002/nla.814.
[38] J. W. Pearson and A. J. Wathen, *Fast Iterative Solvers for Convection-Diffusion Control Problems*, submitted.
[39] A. Potschka, M. Mommer, J. Schlöder, and H. Bock, *A Newton-Picard Approach for Efficient Numerical Solution of Time-Periodic Parabolic PDE Constrained Optimization Problems*, Interdisciplinary Center for Scientific Computing, Heidelberg University, 2010.
[40] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 271–298.
[41] T. Rees and C. Greif, *A preconditioner for linear systems arising from interior point optimization methods.*, SIAM J. Sci. Comput., 29 (2007), pp. 1992–2007.
[42] T. Rees and M. Stoll, *Block-triangular preconditioners for PDE-constrained optimization*, Numeri. Linear Algebra Appl., 17 (2010), pp. 977–996.
[43] T. Rees and M. Stoll, *A fast solver for an H_1 regularized optimal control problem*, submitted.
[44] T. Rees, M. Stoll, and A. Wathen, *All-at-once preconditioners for PDE-constrained optimization*, Kybernetika, 46 (2010), pp. 341–360.
[45] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, Frontiers Appl. Math. 3, SIAM, Philadelphia, 1987, pp. 73–130.
[46] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.
[47] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput, 7 (1986), pp. 856–869.
[48] J. Schöberl and W. 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.
[49] V. Simoncini, *Block triangular preconditioners for symmetric saddle-point problems*, Appl. Numer. Math, 49 (2004), pp. 63–80.
[50] V. Simoncini, *Reduced order solution of structured linear systems arising in certain PDE-constrained optimization problems*, Comput. Optim. Appl., to appear.
[51] M. Stoll, *All-at-once solution of a time-dependent time-periodic PDE-constrained optimization problems*, submitted.
[52] M. Stoll and A. Wathen, *All-at-once solution of time-dependent PDE-constrained optimization problems*, Technical Report 1017, Mathematical Institute, University of Oxford, 2010.
[53] M. Stoll and A. Wathen, *All-at-once solution of time-dependent Stokes control*, J. Comput. Phys., to appear.
[54] S. Takacs and W. Zulehner, *Convergence analysis of multigrid methods with collective point smoothers for optimal control problems*, Comput. Vis. Sci., 14 (2011), pp.131–141.

[55] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, AMS, Providence, RI, 2010.
[56] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems*, SIAM Philadelphia, 2011.
[57] H. A. van der Vorst, *Iterative Krylov methods for large linear systems*, Cambridge Monogr. Appl. Comput. Math. 13, Cambridge University Press, Cambridge, UK, 2003.
[58] A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electron. Trans. Numer. Anal., 34 (2008), pp. 125–135.
[59] W. Zulehner, *Analysis of iterative methods for saddle point problems: a unified approach*, Math. Comp., 71 (2002), pp. 479–505.