Fast Interior Point Solution of Quadratic Programming Problems Arising from PDE-Constrained Optimization

Noname manuscript No.
(will be inserted by the editor)

Fast Interior Point Solution of Quadratic Programming Problems Arising from PDE-Constrained Optimization

John W. Pearson · Jacek Gondzio

Received: date / Accepted: date

Abstract Interior point methods provide an attractive class of approaches for solving linear, quadratic and nonlinear programming problems, due to their excellent efficiency and wide applicability. In this paper, we consider PDE-constrained optimization problems with bound constraints on the state and control variables, and their representation on the discrete level as quadratic programming problems. To tackle complex problems and achieve high accuracy in the solution, one is required to solve matrix systems of huge scale resulting from Newton iteration, and hence fast and robust methods for these systems are required. We present preconditioned iterative techniques for solving a number of these problems using Krylov subspace methods, considering in what circumstances one may predict rapid convergence of the solvers in theory, as well as the solutions observed from practical computations.

Keywords Interior point methods · PDE-constrained optimization · Krylov subspace methods · Preconditioning · Schur complement

PACS 65F08 · 65F10 · 65F50 · 76D05 · 76D55 · 93C20

1 Introduction

We are concerned with optimization problems which involve partial differential equations. Problems of this type appear for example in numerous applications of optimal control, where one wishes state variables to be close to a certain desired form and hopes to achieve it by an appropriate choice of control variables. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded open domain with sufficiently smooth boundary $\partial\Omega$. An optimal control problem with constraints may be written as:

$$\min_{y \in Y,\ u \in U} \quad \mathcal{J}(y, u) \quad \text{s.t.} \quad c(y, u) = 0, \tag{1}$$

where the state y and control u belong to appropriate function spaces Y and U, respectively. The objective $\mathcal{J} : Y \times U \mapsto \mathbb{R}$ and the constraints $c : Y \times U \mapsto \Lambda$, where Λ is another function space, are assumed to satisfy certain smoothness conditions to guarantee the existence and uniqueness of the solution. Many real-life problems may be modelled as optimal control problems (1). There exists rich literature on the subject which addresses specific applications and provides theoretical background to such problems. The rigorous analysis of optimal control problems requires the use of nontrivial function spaces and involves sophisticated techniques from functional analysis. We refer the interested reader to excellent books on

John W. Pearson
School of Mathematics, Statistics and Actuarial Science, University of Kent, Cornwallis Building (East), Canterbury, CT2 7NF, UK
E-mail: j.w.pearson@kent.ac.uk

Jacek Gondzio
School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK
E-mail: j.gondzio@ed.ac.uk

the subject [22,24,45], while for simplicity in this paper we assume that Y, U and Λ are all equal to $L_2(\Omega)$.

The objective function $\mathcal{J}$ may take many different forms but it is often given as:

$$\mathcal{J}(y,u) = \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L_2(\Omega)}, \tag{2}$$

which corresponds to balancing between two goals: keeping the state y close to a certain desired form $\widehat{y}$, and minimizing the "energy" of the applied control u. The constraints c in (1) involve some PDE operator(s), and restrict y and u to Ω and its boundary $\partial\Omega$. Additionally they may include simple bounds on y and u. In Section 3 we will introduce two particular classes of optimal control problems: time-invariant and time-dependent PDE-constrained problems.

Computational techniques for PDE-constrained optimal control problems involve a discretization of the underlying PDE. There are two options for doing this, and the typical paradigm in PDE-constrained optimization literature is for both approaches to solve the problem in a similar manner. The first is to apply an *optimize-then-discretize* method, involving constructing continuous optimality conditions, and then discretizing these. However we find that this approach is inconvenient when considering the resulting discrete systems for the problems considered in this paper, specifically with regard to the reduction of the dimension of the system, as well as symmetry of the matrix involved. The alternative method, which we apply in this paper, is the *discretize-then-optimize* approach: here a discrete cost functional is constructed and discretized constraints are formulated. Then optimality conditions are derived for such (possibly huge) problems. Our motivation for using this approach originates from an observation that for a particular (quadratic) cost functional (2) the discretized PDE-constrained problem takes the form of a quadratic optimization problem for linear PDEs. The use of fine discretization leads to a substantial size of the resulting optimization problem. Therefore we will apply an interior point algorithm to solve it.

Interior point methods (IPMs) are very well-suited to solving quadratic optimization problems and they excel when sizes of problems grow large [17,52], which makes them perfect candidates for discretized PDE-constrained optimal control problems. The use of IPMs in PDE-constrained optimization is not new. There have been several developments which address theoretical aspects, including the functional analysis viewpoint, and study the convergence properties of an interior point algorithm [46,49,51], and many others which focus on the practical (computational) aspects. IPMs belong to a broad class of methods which rely on the use of Newton methods to compute optimizing directions. There have been several successful attempts to use Newton-based approaches in the PDE-constrained optimization context [4,5,25,28]. The main computational challenge in these approaches is the solution of the linear system which determines the Newton direction. For fine PDE discretizations such systems quickly get very large. Additionally, when IPMs are applied, the added interior point diagonal scaling matrices degrade the conditioning of such systems [17] and make them numerically challenging. Direct methods for sparse linear algebra [10] can handle the ill-conditioning well but struggle with excessive memory requirements when problems get larger. Inexact interior point methods [16,18,50] overcome this difficulty by employing iterative methods to solve the Newton equations.

Because of the unavoidable ill-conditioning of these equations the success of any iterative scheme for their solution depends on the ability to design efficient *preconditioners* which can improve spectral properties of linear systems. The development of such preconditioners is a very active research area. Preconditioners for IPMs in PDE-constrained optimization exploit the vast experience gathered for saddle point systems [2], but face an extra difficulty originating from the presence of IPM scaling. There have already been several successful attempts to design preconditioners for such systems, see [1,3,18] and the references therein.

In this paper, we propose a general methodology to design efficient preconditioners for such systems. Our approach is derived from the *matching strategy* originally developed for a particular Poisson control problem [37]. We adapt it to much more challenging circumstances of saddle point systems arising in IPMs applied to solve the PDE-constrained optimal control problems. We briefly comment on the enjoyable spectral properties of the preconditioned system, and provide computational results to demonstrate that they work well in practice.

This paper is structured as follows. In Section 2 we briefly recall a few basic facts about interior point methods for quadratic programming. In Section 3 we demonstrate how IPMs can be applied to PDE-constrained optimization problems. In Section 4 we introduce the preconditioners proposed for problems

originating from optimal control. We consider separately two different cases of time-independent and time-dependent problems. In Section 5 we illustrate our findings with computational results and, finally, in Section 6 we give our conclusions.

2 Interior point methods for quadratic programming

Within this paper, we are interested in the solution of *quadratic programming* (QP) problems. In their most basic form, such problems may be written as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{3}$$

We consider the case where $A \in \mathbb{R}^{m\times n}$ $(m \leq n)$ has full row rank, $Q \in \mathbb{R}^{n\times n}$ is positive semidefinite, $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. This formulation is frequently considered alongside its *dual problem*

$$\begin{aligned} \max_{\mathbf{y}} \quad & \mathbf{b}^\top \mathbf{y} - \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} \\ \text{s.t.} \quad & A^\top \mathbf{y} + \mathbf{z} - Q\mathbf{x} = \mathbf{c}, \\ & \mathbf{y} \ \text{ free}, \quad \mathbf{z} \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{z} \in \mathbb{R}^n$, and $\mathbf{y} \in \mathbb{R}^m$. We note that a subset of this setup is that of linear programming (LP) problems, where $Q = 0$.

In this manuscript, we consider the solution of quadratic programming problems using interior point methods [17]. The nonnegativity constraints $\mathbf{x} \geq \mathbf{0}$ are "replaced" with the logarithmic barrier penalty function, and the Lagrangian associated with the barrier subproblem is formed:

$$\mathcal{L}_\mu(\mathbf{x}, \mathbf{y}) = \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} + \mathbf{y}^\top (A\mathbf{x} - \mathbf{b}) - \mu \sum_j \log(x_j).$$

Differentiating $\mathcal{L}_\mu$ with respect to $\mathbf{x}$ and $\mathbf{y}$ and defining $z_j = \mu/x_j, \ \forall j$, gives the *first order optimality conditions* (or *Karush-Kuhn-Tucker conditions*):

$$\begin{aligned} A\mathbf{x} &= \mathbf{b}, \\ A^\top \mathbf{y} + \mathbf{z} - Q\mathbf{x} &= \mathbf{c}, \\ x_j z_j &= \mu, \quad j = 1, 2, \ldots, n, \\ (\mathbf{x}, \mathbf{z}) &\geq 0, \end{aligned} \tag{4}$$

in which the standard complementarity condition for (3), that is $x_j z_j = 0, \ \forall j$, is replaced with the perturbed complementarity condition $x_j z_j = \mu, \ \forall j$. IPMs drive the barrier term μ to zero and gradually reveal the activity of the primal variables x_j and dual slacks z_j. This is achieved by applying Newton's method to the system of (mildly) nonlinear equations (4)

$$\begin{bmatrix} -Q & A^\top & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{x} \\ \boldsymbol{\delta}\mathbf{y} \\ \boldsymbol{\delta}\mathbf{z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}_d \\ \boldsymbol{\xi}_p \\ \boldsymbol{\xi}_c \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\delta}\mathbf{x}$, $\boldsymbol{\delta}\mathbf{y}$ and $\boldsymbol{\delta}\mathbf{z}$ denote Newton directions, $\boldsymbol{\xi}_p$, $\boldsymbol{\xi}_d$ and $\boldsymbol{\xi}_c$ denote primal and dual infeasibilities and the violation of complementarity conditions. X and Z denote diagonal matrices with elements of $\mathbf{x}$ and $\mathbf{z}$ spread on the diagonals, respectively. By eliminating $\boldsymbol{\delta}\mathbf{z}$, the Newton system (5) is further reduced to a saddle point form

$$\begin{bmatrix} -Q - X^{-1}Z & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{x} \\ \boldsymbol{\delta}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}_d - X^{-1}\boldsymbol{\xi}_c \\ \boldsymbol{\xi}_p \end{bmatrix}. \tag{6}$$

Since for any $j = 1, 2, \ldots, n$ at least one of the variables x_j and z_j reaches zero at optimality, the elements of the diagonal scaling matrix $X^{-1}Z$ added to the $(1,1)$-block may significantly differ in magnitude: some

of them go to zero while the others go to infinity. This feature of IPMs [17] is a challenge for any linear equation solver applied to (6). We skip further details about IPMs and refer the interested reader to [17, 52]. We also highlight that $\mathbf{y}$ in this description relates to a dual variable, whereas for PDE-constrained optimization the function y corresponds to a primal variable – we elect to use the standard notation within the respective fields.

However, before moving on to PDE-constrained optimization, it is worth drawing the reader's attention to the fact that, although in (3) we assume only the one-sided bound $\mathbf{x} \geq \mathbf{0}$, IPMs can also be easily applied to variables with two-sided bounds:

$$\mathbf{x}_a \leq \mathbf{x} \leq \mathbf{x}_b.$$

This requires introducing two nonnegative Lagrange multipliers associated with two inequalities. Later on we will denote them as $\mathbf{z}_a$ and $\mathbf{z}_b$, respectively.

3 PDE-constrained optimization

We now wish to demonstrate how interior point methods may be applied to PDE-constrained optimization problems. These are a crucial class of problems which may be used to model a range of applications in science and industry, for example fluid flow, chemical and biological processes, shape optimization, imaging problems, and mathematical finance, to name but a few. However the problems are often of complex structure, and sophisticated techniques are frequently required to achieve accurate solutions for the models being considered. We recommend the works [22,45], which provide an excellent introduction to the field.

Let us first consider a time-independent linear PDE-constrained optimization problem with additional bound constraints:

$$\begin{aligned} \min_{y,u} \quad & \frac{1}{2}\left\|y-\widehat{y}\right\|^2_{L_2(\Omega)} + \frac{\beta}{2}\left\|u\right\|^2_{L_2(\Omega)} \\ \text{s.t.} \quad & \mathcal{L}y = u, \quad \text{in } \Omega, \\ & y = f, \quad \text{on } \partial\Omega, \\ & y_a \leq y \leq y_b, \quad \text{a.e. in } \Omega, \\ & u_a \leq u \leq u_b, \quad \text{a.e. in } \Omega. \end{aligned} \tag{7}$$

Here y, $\widehat{y}$, u denote the *state*, *desired state* and *control variables*, with $\mathcal{L}$ some PDE operator, and β a positive *regularization parameter*. The problem is solved on domain Ω (with boundary $\partial\Omega$), for given functions f, y_a, y_b, u_a, u_b.

We will now apply the discretize-then-optimize approach to (7), commencing with the construction of a Lagrangian on the discrete space. The alternative optimize-then-discretize method will guarantee an accurate solution of the continuous first order optimality conditions, however when applied in conjunction with interior point methods the resulting matrix systems are not necessarily symmetric, nor can they be reduced to such low dimensions for these problems as the matrix systems illustrated later in this section. For these reasons, we find it is advantageous to apply the discretize-then-optimize approach for the interior point solution of PDE-constrained optimization problems – we highlight that this follows the approach used in important literature on the field such as [5,28]. Provided reasonable choices are made for the discretization of the problem, it is frequently observed that both methods lead to very similar behaviour in the solutions, and indeed this paradigm has recently been used to derive discretization schemes for PDE-constrained optimization (see [20], for instance).

We wish to construct a finite element discretization of the cost functional in (7): for the problems considered in this paper it is beneficial to use equal order finite elements for state and control variables, and observe that a discretized approximation of the cost functional is

$$\frac{1}{2}\left\|y-\widehat{y}\right\|^2_{L_2(\Omega)} + \frac{\beta}{2}\left\|u\right\|^2_{L_2(\Omega)} \approx \frac{1}{2}\mathbf{y}^\top M\mathbf{y} - \mathbf{y}_d^\top\mathbf{y} + \underbrace{\frac{1}{2}\int_\Omega \widehat{y}^2 \,\mathrm{d}\Omega}_{\text{constant}} + \frac{\beta}{2}\mathbf{u}^\top M\mathbf{u},$$

where $\mathbf{y}$, $\mathbf{u}$ are the discretized versions of y, u. The (symmetric) finite element *mass matrix* M contains entries of the form $[M]_{ij} = \int_\Omega \phi_i \phi_j \, \mathrm{d}\Omega$, where $\{\phi_i\}$ are the finite element basis functions used, and $\mathbf{y}_d$ contains entries of the form $\int_\Omega \widehat{y}\phi_i \, \mathrm{d}\Omega$.

We therefore write (7) on the discrete level as

$$\begin{aligned} \min_{\mathbf{y},\mathbf{u}} \quad & \frac{1}{2}\mathbf{y}^\top M\mathbf{y} - \mathbf{y}_d^\top \mathbf{y} + \frac{\beta}{2}\mathbf{u}^\top M\mathbf{u} \\ \text{s.t.} \quad & K\mathbf{y} - M\mathbf{u} = \mathbf{f}, \\ & \mathbf{y}_a \leq \mathbf{y} \leq \mathbf{y}_b, \\ & \mathbf{u}_a \leq \mathbf{u} \leq \mathbf{u}_b, \end{aligned} \tag{8}$$

with $\mathbf{f}$, $\mathbf{y}_a$, $\mathbf{y}_b$, $\mathbf{u}_a$, $\mathbf{u}_b$ the discrete versions of f, y_a, y_b, u_a, u_b. The matrix K depends on the PDE operator $\mathcal{L}$ considered: for example when a Poisson control problem (with $\mathcal{L} = -\nabla^2$) is examined, K denotes a finite element *stiffness matrix* with entries $[K]_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, \mathrm{d}\Omega$. Alternatively for convection-diffusion control problems (with $\mathcal{L} = -\nu\nabla^2 + (\vec{w}\cdot\nabla)$, and without stabilization applied within the solution method), K contains a sum of diffusion and convection terms with $[K]_{ij} = \int_\Omega \left(\nu\nabla\phi_i \cdot \nabla\phi_j + (\vec{w}\cdot\nabla\phi_j)\phi_i\right) \mathrm{d}\Omega$.

We observe that, using our equal order finite element method, the matrices $M, K \in \mathbb{R}^{N\times N}$, where N denotes the number of finite element nodes used, and furthermore that $\mathbf{y}, \mathbf{u} \in \mathbb{R}^N$.

It can be easily seen that the problem statement (8) is in the form of the quadratic programming problem (3), with

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix}, \quad Q = \begin{bmatrix} M & 0 \\ 0 & \beta M \end{bmatrix}, \quad A = \begin{bmatrix} K & -M \end{bmatrix},$$
$$\mathbf{c} = \begin{bmatrix} -\mathbf{y}_d \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{x}_a = \begin{bmatrix} \mathbf{y}_a \\ \mathbf{u}_a \end{bmatrix}, \quad \mathbf{x}_b = \begin{bmatrix} \mathbf{y}_b \\ \mathbf{u}_b \end{bmatrix}.$$

It should be highlighted that, as there has been relatively little previous research on interior point methods for PDE-constrained optimization, there are a number of theoretical considerations that one should account for. As discussed in the paper [46], the majority of the theory available for primal-dual interior point methods is based on finite-dimensional mathematical programming, as opposed to the function space setting of optimal control problems. The authors then proceed to carry out a global and local convergence analysis in the L^∞ and L^q (for $q < \infty$) settings. It is also important to note that the regularity properties of the optimal state and control are different, which as highlighted in [5] is a crucial feature of the continuous (infinite dimensional) problem which tends to be overlooked when moving to a discretized setting. It is essential to recognise the differences between the continuous formulations involving control constraints and state constraints [5,46], in particular the greater scope for a rigorous analysis of the control constrained problem, as well as the possibility of generating provably mesh-independent algorithms (including interior point methods) for problems with control constraints, in constrast to problems with state constraints [5]. As the main objective of this paper is to demonstrate the possibility of solving large scale linear systems that arise from interior point methods, we focus for the most part on the challenges faced on the discrete level, however it is crucial to also be aware of the issues present when examining the associated infinite dimensional problem, and in particular the implications of the discretization strategy employed.

In the next section we consider interior point methods for solving problems of structure (8), for a range of operators $\mathcal{L}$ and all $\beta > 0$. Although there has at this point been relatively little research into such strategies, we highlight that the paper [46] considers the numerical solution of problems of this type with control constraints only, and [1] derives effective preconditioners for large values of β and $\mathcal{L}y = -\nabla^2 y + y$. We also point to the development of solvers of different forms to those presented in this paper: in [18] reduced-space preconditioners are considered for optimal control problems, and in [9] multigrid methods are discussed for a class of control problems.

3.1 Newton iteration

We now wish to derive the equations arising from a Newton iteration applied to the (nonlinear) problem (7). Let us define

$$\mathcal{J}(\mathbf{y},\mathbf{u}) = \frac{1}{2}\mathbf{y}^\top M\mathbf{y} - \mathbf{y}_d^\top \mathbf{y} + \frac{\beta}{2}\mathbf{u}^\top M\mathbf{u}$$

to be the discrete functional which we wish to minimize. Applying the discretized version of the PDE constraint, alongside a barrier function for the bound constraints as in the previous section, leads to the Lagrangian

$$\begin{aligned}\mathcal{L}_\mu(\mathbf{y},\mathbf{u},\boldsymbol{\lambda}) = \mathcal{J}(\mathbf{y},\mathbf{u}) + \boldsymbol{\lambda}^\top(K\mathbf{y} - M\mathbf{u} - \mathbf{f}) \\ - \mu\sum_j \log\left(y_j - y_{a,j}\right) - \mu\sum_j \log\left(y_{b,j} - y_j\right) \\ - \mu\sum_j \log\left(u_j - u_{a,j}\right) - \mu\sum_j \log\left(u_{b,j} - u_j\right),\end{aligned}$$

of which we wish to find the stationary point(s). Here $\boldsymbol{\lambda}$ denotes the discretized *adjoint variable* (or *Lagrange multiplier*), y_j, $y_{a,j}$, $y_{b,j}$, u_j, $u_{a,j}$, $u_{b,j}$ denote the j-th entries of $\mathbf{y}$, $\mathbf{y}_a$, $\mathbf{y}_b$, $\mathbf{u}$, $\mathbf{u}_a$, $\mathbf{u}_b$, and μ is the *barrier parameter* used.

Differentiating $\mathcal{L}_\mu$ with respect to $\mathbf{y}$, $\mathbf{u}$ and $\boldsymbol{\lambda}$ gives the *first order optimality conditions* (or *Karush-Kuhn-Tucker conditions*):

$$M\mathbf{y} - \mathbf{y}_d + K^\top\boldsymbol{\lambda} - \mathbf{z}_{y,a} + \mathbf{z}_{y,b} = \mathbf{0}, \tag{9}$$
$$\beta M\mathbf{u} - M\boldsymbol{\lambda} - \mathbf{z}_{u,a} + \mathbf{z}_{u,b} = \mathbf{0}, \tag{10}$$
$$K\mathbf{y} - M\mathbf{u} - \mathbf{f} = \mathbf{0}, \tag{11}$$

where the j-th entries of $\mathbf{z}_{y,a}$, $\mathbf{z}_{y,b}$, $\mathbf{z}_{u,a}$, $\mathbf{z}_{u,b}$ are defined as follows

$$(\mathbf{z}_{y,a})_j = \frac{\mu}{y_j - y_{a,j}}, \quad (\mathbf{z}_{y,b})_j = \frac{\mu}{y_{b,j} - y_j}, \quad (\mathbf{z}_{u,a})_j = \frac{\mu}{u_j - u_{a,j}}, \quad (\mathbf{z}_{u,b})_j = \frac{\mu}{u_{b,j} - u_j}. \tag{12}$$

Note that, by construction, the following bound constraints apply for the Lagrange multipliers enforcing the constraints on y and u:

$$\mathbf{z}_{y,a} \geq \mathbf{0}, \qquad \mathbf{z}_{y,b} \geq \mathbf{0}, \qquad \mathbf{z}_{u,a} \geq \mathbf{0}, \qquad \mathbf{z}_{u,b} \geq \mathbf{0}.$$

Applying a Newton iteration to (9)–(12) gives, at each Newton step,

$$M\boldsymbol{\delta}\mathbf{y} + K^\top\boldsymbol{\delta}\boldsymbol{\lambda} - \boldsymbol{\delta}\mathbf{z}_{y,a} + \boldsymbol{\delta}\mathbf{z}_{y,b} = \mathbf{y}_d - M\mathbf{y}^* - K^\top\boldsymbol{\lambda}^* + \mathbf{z}^*_{y,a} - \mathbf{z}^*_{y,b}, \tag{13}$$
$$\beta M\boldsymbol{\delta}\mathbf{u} - M\boldsymbol{\delta}\boldsymbol{\lambda} - \boldsymbol{\delta}\mathbf{z}_{u,a} + \boldsymbol{\delta}\mathbf{z}_{u,b} = -\beta M\mathbf{u}^* + M\boldsymbol{\lambda}^* + \mathbf{z}^*_{u,a} - \mathbf{z}^*_{u,b}, \tag{14}$$
$$K\boldsymbol{\delta}\mathbf{y} - M\boldsymbol{\delta}\mathbf{u} = \mathbf{f} - K\mathbf{y}^* + M\mathbf{u}^*, \tag{15}$$
$$(\mathbf{y}^* - \mathbf{y}_a)\circ\boldsymbol{\delta}\mathbf{z}_{y,a} + \mathbf{z}^*_{y,a}\circ\boldsymbol{\delta}\mathbf{y} = \mu\mathbf{e} - (\mathbf{y}^* - \mathbf{y}_a)\circ\mathbf{z}^*_{y,a}, \tag{16}$$
$$(\mathbf{y}_b - \mathbf{y}^*)\circ\boldsymbol{\delta}\mathbf{z}_{y,b} - \mathbf{z}^*_{y,b}\circ\boldsymbol{\delta}\mathbf{y} = \mu\mathbf{e} - (\mathbf{y}_b - \mathbf{y}^*)\circ\mathbf{z}^*_{y,b}, \tag{17}$$
$$(\mathbf{u}^* - \mathbf{u}_a)\circ\boldsymbol{\delta}\mathbf{z}_{u,a} + \mathbf{z}^*_{u,a}\circ\boldsymbol{\delta}\mathbf{u} = \mu\mathbf{e} - (\mathbf{u}^* - \mathbf{u}_a)\circ\mathbf{z}^*_{u,a}, \tag{18}$$
$$(\mathbf{u}_b - \mathbf{u}^*)\circ\boldsymbol{\delta}\mathbf{z}_{u,b} - \mathbf{z}^*_{u,b}\circ\boldsymbol{\delta}\mathbf{u} = \mu\mathbf{e} - (\mathbf{u}_b - \mathbf{u}^*)\circ\mathbf{z}^*_{u,b}. \tag{19}$$

Here, $\mathbf{y}^*$, $\mathbf{u}^*$, $\boldsymbol{\lambda}^*$, $\mathbf{z}^*_{y,a}$, $\mathbf{z}^*_{y,b}$, $\mathbf{z}^*_{u,a}$, $\mathbf{z}^*_{u,b}$ denote the most recent Newton iterates for $\mathbf{y}$, $\mathbf{u}$, $\boldsymbol{\lambda}$, $\mathbf{z}_{y,a}$, $\mathbf{z}_{y,b}$, $\mathbf{z}_{u,a}$, $\mathbf{z}_{u,b}$, with $\boldsymbol{\delta}\mathbf{y}$, $\boldsymbol{\delta}\mathbf{u}$, $\boldsymbol{\delta}\boldsymbol{\lambda}$, $\boldsymbol{\delta}\mathbf{z}_{y,a}$, $\boldsymbol{\delta}\mathbf{z}_{y,b}$, $\boldsymbol{\delta}\mathbf{z}_{u,a}$, $\boldsymbol{\delta}\mathbf{z}_{u,b}$ the Newton updates, $\mathbf{e}$ defines the vector of ones of appropriate dimension, and $\circ$ relates to the multiplication componentwise of two vectors.

In matrix form, (13)–(19) read

$$\begin{bmatrix} M & 0 & K^\top & -I & I & 0 & 0 \\ 0 & \beta M & -M & 0 & 0 & -I & I \\ K & -M & 0 & 0 & 0 & 0 & 0 \\ Z_{y,a} & 0 & 0 & Y - Y_a & 0 & 0 & 0 \\ -Z_{y,b} & 0 & 0 & 0 & Y_b - Y & 0 & 0 \\ 0 & Z_{u,a} & 0 & 0 & 0 & U - U_a & 0 \\ 0 & -Z_{u,b} & 0 & 0 & 0 & 0 & U_b - U \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{y} \\ \boldsymbol{\delta}\mathbf{u} \\ \boldsymbol{\delta}\boldsymbol{\lambda} \\ \boldsymbol{\delta}\mathbf{z}_{y,a} \\ \boldsymbol{\delta}\mathbf{z}_{y,b} \\ \boldsymbol{\delta}\mathbf{z}_{u,a} \\ \boldsymbol{\delta}\mathbf{z}_{u,b} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_d - M\mathbf{y}^* - K^\top\boldsymbol{\lambda}^* + \mathbf{z}^*_{y,a} - \mathbf{z}^*_{y,b} \\ -\beta M\mathbf{u}^* + M\boldsymbol{\lambda}^* + \mathbf{z}^*_{u,a} - \mathbf{z}^*_{u,b} \\ \mathbf{f} - K\mathbf{y}^* + M\mathbf{u}^* \\ \mu\mathbf{e} - (\mathbf{y}^* - \mathbf{y}_a)\circ\mathbf{z}^*_{y,a} \\ \mu\mathbf{e} - (\mathbf{y}_b - \mathbf{y}^*)\circ\mathbf{z}^*_{y,b} \\ \mu\mathbf{e} - (\mathbf{u}^* - \mathbf{u}_a)\circ\mathbf{z}^*_{u,a} \\ \mu\mathbf{e} - (\mathbf{u}_b - \mathbf{u}^*)\circ\mathbf{z}^*_{u,b} \end{bmatrix},$$

where Y, U, $Z_{y,a}$, $Z_{y,b}$, $Z_{u,a}$, $Z_{u,b}$ are diagonal matrices, with the most recent iterates for $\mathbf{y}$, $\mathbf{u}$, $\mathbf{z}_{y,a}$, $\mathbf{z}_{y,b}$, $\mathbf{z}_{u,a}$, $\mathbf{z}_{u,b}$ appearing on the diagonal entries. Similarly, the matrices Y_a, Y_b, U_a, U_b are diagonal matrices corresponding to $\mathbf{y}_a$, $\mathbf{y}_b$, $\mathbf{u}_a$, $\mathbf{u}_b$.

Now, we may write that fourth, fifth, sixth and seventh rows lead to

$$\boldsymbol{\delta}\mathbf{z}_{y,a} = -(Y - Y_a)^{-1} Z_{y,a} \boldsymbol{\delta}\mathbf{y} - Z_{y,a} + \mu (Y - Y_a)^{-1} \mathbf{e}, \tag{20}$$

$$\boldsymbol{\delta}\mathbf{z}_{y,b} = (Y_b - Y)^{-1} Z_{y,b} \boldsymbol{\delta}\mathbf{y} - Z_{y,b} + \mu (Y_b - Y)^{-1} \mathbf{e}, \tag{21}$$

$$\boldsymbol{\delta}\mathbf{z}_{u,a} = -(U - U_a)^{-1} Z_{u,a} \boldsymbol{\delta}\mathbf{u} - Z_{u,a} + \mu (U - U_a)^{-1} \mathbf{e}, \tag{22}$$

$$\boldsymbol{\delta}\mathbf{z}_{u,b} = (U_b - U)^{-1} Z_{u,b} \boldsymbol{\delta}\mathbf{u} - Z_{u,b} + \mu (U_b - U)^{-1} \mathbf{e}, \tag{23}$$

whereupon we may consider instead the solution of the reduced system

$$\begin{bmatrix} M + D_y & 0 & K^\top \\ 0 & \beta M + D_u & -M \\ K & -M & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{y} \\ \boldsymbol{\delta}\mathbf{u} \\ \boldsymbol{\delta}\boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mu (Y - Y_a)^{-1} \mathbf{e} - \mu (Y_b - Y)^{-1} \mathbf{e} + \mathbf{y}_d - M\mathbf{y}^* - K^\top \boldsymbol{\lambda}^* \\ \mu (U - U_a)^{-1} \mathbf{e} - \mu (U_b - U)^{-1} \mathbf{e} - \beta M \mathbf{u}^* + M \boldsymbol{\lambda}^* \\ \mathbf{f} - K\mathbf{y}^* + M\mathbf{u}^* \end{bmatrix}, \tag{24}$$

where

$$D_y = (Y - Y_a)^{-1} Z_{y,a} + (Y_b - Y)^{-1} Z_{y,b}, \tag{25}$$

$$D_u = (U - U_a)^{-1} Z_{u,a} + (U_b - U)^{-1} Z_{u,b}. \tag{26}$$

The conditions written in (24) are applied, alongside the imposition of (20)–(23), at each Newton iteration.

Note that, due to the fact that state and control bounds are enforced as strict inequalities at each Newton step, the diagonal matrices D_y and D_u are positive definite.

Of course, it is perfectly natural to consider a problem with only state constraints or only control constraints (or indeed only lower or upper bound constraints). For such cases we may follow exactly the same working to obtain a matrix system of the form (24), removing individual matrices corresponding to constraints that we do not apply.

3.2 Algorithm

We now present the structure of the interior point algorithm, adapted from the paper [17], that we apply to the problems considered in this paper. The essence of the method is to traverse the interior of the feasible region where solutions may arise – we do this by applying a relaxed Newton iteration, reducing the barrier parameter by a factor σ at each Newton step. Having computed the Newton updates $\boldsymbol{\delta}\mathbf{y}$, $\boldsymbol{\delta}\mathbf{u}$, $\boldsymbol{\delta}\boldsymbol{\lambda}$, $\boldsymbol{\delta}\mathbf{z}_{y,a}$, $\boldsymbol{\delta}\mathbf{z}_{y,b}$, $\boldsymbol{\delta}\mathbf{z}_{u,a}$, $\boldsymbol{\delta}\mathbf{z}_{u,b}$, we make a step in this direction that also guarantees that the strict bounds are enforced at each iteration. Upon convergence the iterates approach the true solution of the optimization problem, with the additional state and control constraints automatically satisfied.

Let us now consider appropriate stopping criteria for the method. Two natural requirements are for the norms of the primal and dual infeasibilities (at the k-th iteration)

$$\boldsymbol{\xi}_p^k = \mathbf{f} - K\mathbf{y}^k + M\mathbf{u}^k, \qquad \boldsymbol{\xi}_d^k = \begin{bmatrix} \mathbf{y}_d - M\mathbf{y}^k - K^\top \boldsymbol{\lambda}^k + \mathbf{z}_{y,a}^k - \mathbf{z}_{y,b}^k \\ -\beta M \mathbf{u}^k + M\boldsymbol{\lambda}^k + \mathbf{z}_{u,a}^k - \mathbf{z}_{u,b}^k \end{bmatrix},$$

to be lower than some prescribed tolerances ϵ_p, ϵ_d, respectively. Additionally, we require the error in the complementarity products

$$\boldsymbol{\xi}_c^k = \begin{bmatrix} \mu\mathbf{e} - (\mathbf{y}^k - \mathbf{y}_a) \circ \mathbf{z}_{y,a}^k \\ \mu\mathbf{e} - (\mathbf{y}_b - \mathbf{y}^k) \circ \mathbf{z}_{y,b}^k \\ \mu\mathbf{e} - (\mathbf{u}^k - \mathbf{u}_a) \circ \mathbf{z}_{u,a}^k \\ \mu\mathbf{e} - (\mathbf{u}_b - \mathbf{u}^k) \circ \mathbf{z}_{u,b}^k \end{bmatrix}, \tag{27}$$

to fall below some specified tolerance ϵ_c.

We present the algorithm that we apply – its structure is similar to the algorithm outlined in [17, Section 2].

Interior Point Method for Quadratic Programming

Parameters

$\alpha_0 = 0.995$, step-size factor to boundary

$\sigma \in (0,1)$, barrier reduction parameter

$\epsilon_p,\ \epsilon_d,\ \epsilon_c$, stopping tolerances,

Interior point method stops when $\left\|\boldsymbol{\xi}_p^k\right\| \le \epsilon_p,\ \left\|\boldsymbol{\xi}_d^k\right\| \le \epsilon_d,\ \left\|\boldsymbol{\xi}_c^k\right\| \le \epsilon_c$

Initialize IPM

Initial guesses for $\mathbf{y}^0,\ \mathbf{u}^0,\ \boldsymbol{\lambda}^0,\ \mathbf{z}_{y,a}^0,\ \mathbf{z}_{y,b}^0,\ \mathbf{z}_{u,a}^0,\ \mathbf{z}_{u,b}^0$

Barrier parameter μ_0

Primal infeasibility $\boldsymbol{\xi}_p^0 = \mathbf{f} - K\mathbf{y}^0 + M\mathbf{u}^0$

Dual infeasibility $\boldsymbol{\xi}_d^0 = \begin{bmatrix} \mathbf{y}_d - M\mathbf{y}^0 - K^\top\boldsymbol{\lambda}^0 + \mathbf{z}_{y,a}^0 - \mathbf{z}_{y,b}^0 \\ -\beta M\mathbf{u}^0 + M\boldsymbol{\lambda}^0 + \mathbf{z}_{u,a}^0 - \mathbf{z}_{u,b}^0 \end{bmatrix}$

Complementarity products $\boldsymbol{\xi}_c^0$, as in (27) with $k = 0$

Interior Point Method

while $\left(\left\|\boldsymbol{\xi}_p^k\right\| > \epsilon_p \ \text{ or } \ \left\|\boldsymbol{\xi}_d^k\right\| > \epsilon_d \ \text{ or } \ \left\|\boldsymbol{\xi}_c^k\right\| > \epsilon_c\right)$

Reduce barrier parameter $\mu_{k+1} = \sigma\mu_k$

Solve Newton system (24) for primal-dual Newton direction $\boldsymbol{\delta}\mathbf{y},\ \boldsymbol{\delta}\mathbf{u},\ \boldsymbol{\delta}\boldsymbol{\lambda}$

Use (20)–(23) to find $\boldsymbol{\delta}\mathbf{z}_{y,a},\ \boldsymbol{\delta}\mathbf{z}_{y,b},\ \boldsymbol{\delta}\mathbf{z}_{u,a},\ \boldsymbol{\delta}\mathbf{z}_{u,b}$

Find $\alpha_P,\ \alpha_D$ s.t. bound constraints on primal and dual variables hold

Set $\alpha_P = \alpha_0\alpha_P,\ \alpha_D = \alpha_0\alpha_D$

Make step: $\mathbf{y}^{k+1} = \mathbf{y}^k + \alpha_P\boldsymbol{\delta}\mathbf{y},\ \mathbf{u}^{k+1} = \mathbf{u}^k + \alpha_P\boldsymbol{\delta}\mathbf{u},\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha_D\boldsymbol{\delta}\boldsymbol{\lambda}$

$\mathbf{z}_{y,a}^{k+1} = \mathbf{z}_{y,a}^k + \alpha_D\boldsymbol{\delta}\mathbf{z}_{y,a},\ \mathbf{z}_{y,b}^{k+1} = \mathbf{z}_{y,b}^k + \alpha_D\boldsymbol{\delta}\mathbf{z}_{y,b}$

$\mathbf{z}_{u,a}^{k+1} = \mathbf{z}_{u,a}^k + \alpha_D\boldsymbol{\delta}\mathbf{z}_{u,a},\ \mathbf{z}_{u,b}^{k+1} = \mathbf{z}_{u,b}^k + \alpha_D\boldsymbol{\delta}\mathbf{z}_{u,b}$

Update infeasibilities:

$\boldsymbol{\xi}_p^{k+1} = \mathbf{f} - K\mathbf{y}^{k+1} + M\mathbf{u}^{k+1}$,

$\boldsymbol{\xi}_d^{k+1} = \begin{bmatrix} \mathbf{y}_d - M\mathbf{y}^{k+1} - K^\top\boldsymbol{\lambda}^{k+1} + \mathbf{z}_{y,a}^{k+1} - \mathbf{z}_{y,b}^{k+1} \\ -\beta M\mathbf{u}^{k+1} + M\boldsymbol{\lambda}^{k+1} + \mathbf{z}_{u,a}^{k+1} - \mathbf{z}_{u,b}^{k+1} \end{bmatrix}$

Compute error of complementarity products as in (27)

Set iteration number $k = k + 1$

end

It is clear from the presentation of this method that the dominant computational work arises from the solution of Newton system (24). It is therefore crucial to construct fast and robust solvers for this system, and this is what we focus on in Section 4.

3.3 Time-dependent problems

It is also important to be able to handle time-dependent problems using this methodology, due to the complexity and practical utility of such setups. To provide a brief illustration of how this may be

accomplished, let us consider the time-dependent problem:

$$\begin{aligned}\min_{y,u} \quad & \frac{1}{2}\int_0^T \int_\Omega (y-\widehat{y})^2 \, \mathrm{d}\Omega \mathrm{d}t + \frac{\beta}{2}\int_0^T \int_\Omega u^2 \, \mathrm{d}\Omega \mathrm{d}t \\ \text{s.t.} \quad & y_t + \mathcal{L}y = u, \quad \text{in } \Omega \times (0,T], \\ & y = f, \quad \text{on } \partial\Omega \times (0,T], \\ & y = y_0, \quad \text{at } t = 0, \\ & y_a \leq y \leq y_b, \quad \text{a.e. in } \Omega \times (0,T], \\ & u_a \leq u \leq u_b, \quad \text{a.e. in } \Omega \times (0,T].\end{aligned}$$

The state, control and adjoint variables are now solved in a space-time domain $\Omega \times (0,T]$, with $\mathcal{L}$ the time-independent component of the PDE operator.

As in [35,44] for heat equation control problems, we may apply a discretize-then-optimize approach, using the trapezoidal rule to approximate the integrals within the cost functional, and the backward Euler method to account for the time derivative. We thus rewrite the problem in the discrete setting as follows:

$$\begin{aligned}\min_{y,u} \quad & \frac{\tau}{2}\mathbf{y}^\top \mathcal{M}_{1/2}\mathbf{y} - \tau \mathbf{y}_{d,T}^\top \mathbf{y} + \frac{\beta\tau}{2}\mathbf{u}^\top \mathcal{M}_{1/2}\mathbf{u} \\ \text{s.t.} \quad & \mathcal{K}\mathbf{y} - \tau\mathcal{M}\mathbf{u} = \mathbf{f}_T, \\ & \mathbf{y}_a \leq \mathbf{y} \leq \mathbf{y}_b, \\ & \mathbf{u}_a \leq \mathbf{u} \leq \mathbf{u}_b.\end{aligned}$$

Here the matrix $\mathcal{M}_{1/2} = \text{blkdiag}(\frac{1}{2}M, M, ..., M, \frac{1}{2}M)$, $\mathcal{M} = \text{blkdiag}(M, ..., M)$, and

$$\mathcal{K} = \begin{bmatrix} M+\tau K & & & & \\ -M & M+\tau K & & & \\ & \ddots & \ddots & & \\ & & -M & M+\tau K & \\ & & & -M & M+\tau K \end{bmatrix}, \quad \mathbf{y}_{d,T} = \begin{bmatrix} \frac{1}{2}\mathbf{y}_{d,1} \\ \mathbf{y}_{d,2} \\ \vdots \\ \mathbf{y}_{d,N_t-1} \\ \frac{1}{2}\mathbf{y}_{d,N_t} \end{bmatrix}, \quad \mathbf{f}_T = \begin{bmatrix} M\mathbf{y}_0 + \mathbf{f} \\ \mathbf{f} \\ \vdots \\ \mathbf{f} \\ \mathbf{f} \end{bmatrix},$$

where K corresponds to the time-independent part of the PDE operator, and τ denotes the (constant) time-step taken. The vectors $\mathbf{y}_{d,i}$ relate to the values of $\widehat{y}$ at the i-th time-step, and $\mathbf{y}_0$ is the vector representation of y_0. We denote by $N_t := \frac{T}{\tau}$ the number of time-steps taken.

We apply Newton iteration to the discrete optimality conditions, in an analogous way to the time-independent problem. This yields the matrix system

$$\begin{bmatrix} \tau\mathcal{M}_{1/2} & 0 & \mathcal{K}^\top & -I & I & 0 & 0 \\ 0 & \beta\tau\mathcal{M}_{1/2} & -\tau\mathcal{M} & 0 & 0 & -I & I \\ \mathcal{K} & -\tau\mathcal{M} & 0 & 0 & 0 & 0 & 0 \\ Z_{y,a} & 0 & 0 & Y - Y_a & 0 & 0 & 0 \\ -Z_{y,b} & 0 & 0 & 0 & Y_b - Y & 0 & 0 \\ 0 & Z_{u,a} & 0 & 0 & 0 & U - U_a & 0 \\ 0 & -Z_{u,b} & 0 & 0 & 0 & 0 & U_b - U \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{y} \\ \boldsymbol{\delta}\mathbf{u} \\ \boldsymbol{\delta}\boldsymbol{\lambda} \\ \boldsymbol{\delta}\mathbf{z}_{y,a} \\ \boldsymbol{\delta}\mathbf{z}_{y,b} \\ \boldsymbol{\delta}\mathbf{z}_{u,a} \\ \boldsymbol{\delta}\mathbf{z}_{u,b} \end{bmatrix} \tag{28}$$

$$= \begin{bmatrix} \tau\mathbf{y}_{d,T} - \tau\mathcal{M}_{1/2}\mathbf{y}^* - \mathcal{K}^\top\boldsymbol{\lambda}^* + \mathbf{z}_{y,a}^* - \mathbf{z}_{y,b}^* \\ -\beta\tau\mathcal{M}_{1/2}\mathbf{u}^* + \tau\mathcal{M}\boldsymbol{\lambda}^* + \mathbf{z}_{u,a}^* - \mathbf{z}_{u,b}^* \\ \mathbf{f}_T - \mathcal{K}\mathbf{y}^* + \tau\mathcal{M}\mathbf{u}^* \\ \mu\mathbf{e} - (\mathbf{y}^* - \mathbf{y}_a) \circ \mathbf{z}_{y,a}^* \\ \mu\mathbf{e} - (\mathbf{y}_b - \mathbf{y}^*) \circ \mathbf{z}_{y,b}^* \\ \mu\mathbf{e} - (\mathbf{u}^* - \mathbf{u}_a) \circ \mathbf{z}_{u,a}^* \\ \mu\mathbf{e} - (\mathbf{u}_b - \mathbf{u}^*) \circ \mathbf{z}_{u,b}^* \end{bmatrix},$$

with $\mathbf{z}_{y_a}$, $\mathbf{z}_{y_b}$, $\mathbf{z}_{u_a}$, $\mathbf{z}_{u_b}$ the same as for the time-independent setting, except now measured over all points in space and time.

Reducing (28) as for the time-independent case gives a block matrix system

$$\begin{bmatrix} \tau\mathcal{M}_{1/2}+\mathcal{D}_y & 0 & \mathcal{K}^\top \\ 0 & \beta\tau\mathcal{M}_{1/2}+\mathcal{D}_u & -\tau\mathcal{M} \\ \mathcal{K} & -\tau\mathcal{M} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta y} \\ \boldsymbol{\delta u} \\ \boldsymbol{\delta\lambda} \end{bmatrix} \tag{29}$$
$$= \begin{bmatrix} \mu(Y-Y_a)^{-1}\mathbf{e}-\mu(Y_b-Y)^{-1}\mathbf{e}+\tau\mathbf{y}_{d,T}-\tau\mathcal{M}_{1/2}\mathbf{y}^*-\mathcal{K}^\top\boldsymbol{\lambda}^* \\ \mu(U-U_a)^{-1}\mathbf{e}-\mu(U_b-U)^{-1}\mathbf{e}-\beta\tau\mathcal{M}_{1/2}\mathbf{u}^*+\tau\mathcal{M}\boldsymbol{\lambda}^* \\ \mathbf{f}_T-\mathcal{K}\mathbf{y}^*+\tau\mathcal{M}\mathbf{u}^* \end{bmatrix},$$

with $\mathcal{D}_y$, $\mathcal{D}_u$ analogous to D_y, D_u, as defined in (25), (26), except with the quantities measured within the entire space-time domain.

4 Preconditioning for the Newton system

For the matrix systems considered in this paper, particularly those arising from time-dependent problems, great care must be taken when seeking an appropriate scheme for obtaining an accurate solution. The dimensions of these systems mean that a direct method is often infeasible, so we find that the natural approach is to develop preconditioned Krylov subspace solvers.

When seeking preconditioners for such methods, we exploit the fact that the matrix systems for the PDE-constrained optimization problems are of *saddle point form*:

$$\underbrace{\begin{bmatrix} \Phi & \Psi^\top \\ \Psi & \Theta \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \tag{30}$$

Here $\Phi \in \mathbb{R}^{n\times n}$, $\Psi \in \mathbb{R}^{m\times n}$ and $\Theta \in \mathbb{R}^{m\times m}$ (with $m \le n$, as in Section 2). Further Φ and Θ are symmetric matrices, meaning that $\mathcal{A}$ is itself symmetric, and all of the matrices are sparse for the finite element method used. We recommend [2] for a thorough overview of saddle point systems and their numerical properties.

The study of preconditioners for systems of this form is a well-established subject area: indeed it is known that two 'ideal' preconditioners are given by

$$\mathcal{P}_D = \begin{bmatrix} \Phi & 0 \\ 0 & S \end{bmatrix}, \qquad \mathcal{P}_T = \begin{bmatrix} \Phi & 0 \\ \Psi & -S \end{bmatrix},$$

where $S := -\Theta + \Psi\Phi^{-1}\Psi^T$ defines the (negative) *Schur complement* of $\mathcal{A}$. It can be shown [23,26,29] that the eigenvalues of the preconditioned systems are given by

$$\lambda\left(\mathcal{P}_D^{-1}\mathcal{A}\right) \in \left\{1, \frac{1}{2}(1\pm\sqrt{5})\right\}, \qquad \text{if } \Theta = 0,$$
$$\lambda\left(\mathcal{P}_T^{-1}\mathcal{A}\right) \in \{1\}, \qquad \text{generally,}$$

provided that these systems are invertible.

In practice, of course, one would not wish to invert Φ and S exactly within a preconditioner, so the main challenge is to devise effective approximations $\widehat{\Phi}$ and $\widehat{S}$ which can be applied within a block diagonal or block triangular preconditioner of the form

$$\mathcal{P} = \begin{bmatrix} \widehat{\Phi} & 0 \\ 0 & \widehat{S} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \widehat{\Phi} & 0 \\ \Psi & -\widehat{S} \end{bmatrix}. \tag{31}$$

Such preconditioners are very often found to be extremely potent in practice, and in many cases one can prove their effectiveness as well (we discuss this further in Section 4.1).

A major objective within the remainder of this paper is to develop effective representations of the $(1,1)$-block Φ and Schur complement S for matrix systems arising from interior point solvers.

4.1 Time-independent problems

We now wish to apply saddle point theory to matrix systems arising from time-independent problems. So consider the matrix system (24), for instance in the case where the matrix K arises from a Laplacian operator (considered for Poisson control) or convection-diffusion operator. This system is of saddle point form (30), with

$$\mathit{\Phi} = \begin{bmatrix} M + D_y & 0 \\ 0 & \beta M + D_u \end{bmatrix}, \qquad \mathit{\Psi} = \begin{bmatrix} K & -M \end{bmatrix}, \qquad \mathit{\Theta} = \begin{bmatrix} 0 \end{bmatrix}.$$

Let us consider approximating the $(1,1)$-block and Schur complement of this matrix system. For this problem M is a positive definite matrix, with positive diagonal entries, and the same applies to K in the case of Poisson control problems.

We now highlight that mass matrices may in fact be well approximated by their diagonal: for instance, in the case of $Q1$ mass matrices on a uniform two dimensional domain, the eigenvalues of $[\text{diag}(M)]^{-1}M$ are all contained within the interval $[\frac{1}{4}, \frac{9}{4}]$ (see [47]). As D_y and D_u are diagonal and positive definite, one option for approximating $\mathit{\Phi}$ is hence to take

$$\widehat{\mathit{\Phi}} = \begin{bmatrix} \text{diag}\,(M + D_y) & 0 \\ 0 & \text{diag}\,(\beta M + D_u) \end{bmatrix}.$$

The effectiveness of the approximation may be measured in some sense by the eigenvalues of $\widehat{\mathit{\Phi}}^{-1}\mathit{\Phi}$, which may themselves be determined by the Rayleigh quotient

$$\begin{aligned} \frac{\mathbf{v}^\top \mathit{\Phi}\mathbf{v}}{\mathbf{v}^\top \widehat{\mathit{\Phi}}\mathbf{v}} &= \frac{\mathbf{v}_1^\top (M + D_y)\mathbf{v}_1 + \mathbf{v}_2^\top (\beta M + D_u)\mathbf{v}_2}{\mathbf{v}_1^\top \left[\text{diag}(M + D_y)\right]\mathbf{v}_1 + \mathbf{v}_2^\top \left[\text{diag}(\beta M + D_u)\right]\mathbf{v}_2} \\ &= \frac{\mathbf{v}_1^\top M\mathbf{v}_1 + \beta\mathbf{v}_2^\top M\mathbf{v}_2 + \mathbf{v}_1^\top D_y\mathbf{v}_1 + \mathbf{v}_2^\top D_u\mathbf{v}_2}{\mathbf{v}_1^\top \left[\text{diag}(M)\right]\mathbf{v}_1 + \beta\mathbf{v}_2^\top \left[\text{diag}(M)\right]\mathbf{v}_2 + \mathbf{v}_1^\top D_y\mathbf{v}_1 + \mathbf{v}_2^\top D_u\mathbf{v}_2} \\ &\in \left[\min\left\{\frac{\mathbf{v}_1^\top M\mathbf{v}_1 + \beta\mathbf{v}_2^\top M\mathbf{v}_2}{\mathbf{v}_1^\top \left[\text{diag}(M)\right]\mathbf{v}_1 + \beta\mathbf{v}_2^\top \left[\text{diag}(M)\right]\mathbf{v}_2}, 1\right\},\right. \\ &\qquad \left.\max\left\{\frac{\mathbf{v}_1^\top M\mathbf{v}_1 + \beta\mathbf{v}_2^\top M\mathbf{v}_2}{\mathbf{v}_1^\top \left[\text{diag}(M)\right]\mathbf{v}_1 + \beta\mathbf{v}_2^\top \left[\text{diag}(M)\right]\mathbf{v}_2}, 1\right\}\right] \\ &\in \left[\min\left\{\lambda_{\min}\left(\left[\text{diag}(M)\right]^{-1}M\right), 1\right\}, \max\left\{\lambda_{\max}\left(\left[\text{diag}(M)\right]^{-1}M\right), 1\right\}\right], \end{aligned} \tag{32}$$

where (32) follows from the fact that $\mathbf{v}_1^\top D_y\mathbf{v}_1 + \mathbf{v}_2^\top D_u\mathbf{v}_2$ is non-negative. Here $\mathbf{v} = \begin{bmatrix}\mathbf{v}_1^\top, \ \mathbf{v}_2^\top\end{bmatrix}^\top \neq \mathbf{0}$, with $\mathbf{v}_1$, $\mathbf{v}_2$ vectors of appropriate length, and $\lambda_{\min}$, $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix. We therefore see that if $[\text{diag}(M)]^{-1}M$ is well-conditioned, then the same is true of $\widehat{\mathit{\Phi}}^{-1}\mathit{\Phi}$.

As an alternative for our approximation $\widehat{\mathit{\Phi}}$, one may apply a Chebyshev semi-iteration method [14, 15, 48] to approximate the inverses of $M + D_y$ and $\beta M + D_u$. This is a slightly more expensive process to approximate this component of the entire system (in general the matrices with the most complex structure are K and $K^\top$), however due to the tight clustering of the eigenvalues of $[\text{diag}(\mathit{\Phi})]^{-1}\mathit{\Phi}$ we find greater accuracy in the results obtained.

The main task at this stage is to approximate the Schur complement

$$S = K(M + D_y)^{-1}K^\top + M(\beta M + D_u)^{-1}M. \tag{33}$$

The aim is to build an approximation such that the eigenvalues of the preconditioned Schur complement are tightly clustered. We motivate our approximation based on a 'matching' strategy originally derived in [37] for the Poisson control problem without bound constraints: for this particular problem, K is the finite element stiffness matrix, and the matrices $D_y = D_u = 0$. It was shown that by 'capturing' both terms ($KM^{-1}K$ and $\frac{1}{\beta}M$) of the Schur complement, one obtains the result

$$\lambda\left(\left[\left(K + \frac{1}{\sqrt{\beta}}M\right)M^{-1}\left(K + \frac{1}{\sqrt{\beta}}M\right)\right]^{-1}\left[KM^{-1}K + \frac{1}{\beta}M\right]\right) \in \left[\frac{1}{2}, 1\right], \tag{34}$$

independently of problem size, as well as the value of β.

Furthermore, it is possible to prove a lower bound of the preconditioned Schur complement for a very general matrix form, as demonstrated below.

Theorem 1 *Let S_G and $\widehat{S}_G$ be the general matrices*

$$S_G = \bar{X}\bar{X}^\top + \bar{Y}\bar{Y}^\top, \qquad \widehat{S}_G = (\bar{X}+\bar{Y})(\bar{X}+\bar{Y})^\top,$$

which we assume to be invertible, and with real $\bar{X}$, $\bar{Y}$. Then the eigenvalues of $\widehat{S}_G^{-1}S_G$ are real, and satisfy $\lambda \geq \frac{1}{2}$.

Proof As S_G and $\widehat{S}_G$ are invertible, they are symmetric positive definite by constuction. To examine the spectrum of $\widehat{S}_G^{-1}S_G$ we therefore consider the Rayleigh quotient (for real $\mathbf{v} \neq \mathbf{0}$):

$$R := \frac{\mathbf{v}^\top S_G \mathbf{v}}{\mathbf{v}^\top \widehat{S}_G \mathbf{v}} = \frac{\boldsymbol{\chi}^\top\boldsymbol{\chi} + \boldsymbol{\omega}^\top\boldsymbol{\omega}}{(\boldsymbol{\chi}+\boldsymbol{\omega})^\top(\boldsymbol{\chi}+\boldsymbol{\omega})}, \qquad \boldsymbol{\chi} = \bar{X}^\top\mathbf{v}, \quad \boldsymbol{\omega} = \bar{Y}^\top\mathbf{v},$$

which is itself clearly real. By the invertibility of S_G and $\widehat{S}_G$, both numerator and denominator are positive. Therefore

$$\frac{1}{2}(\boldsymbol{\chi}-\boldsymbol{\omega})^\top(\boldsymbol{\chi}-\boldsymbol{\omega}) \geq 0 \quad \Leftrightarrow \quad \boldsymbol{\chi}^\top\boldsymbol{\chi} + \boldsymbol{\omega}^\top\boldsymbol{\omega} \geq \frac{1}{2}(\boldsymbol{\chi}+\boldsymbol{\omega})^\top(\boldsymbol{\chi}+\boldsymbol{\omega}) \quad \Leftrightarrow \quad R \geq \frac{1}{2},$$

which gives the result. □

For the Schur complement given by (33), the matrices $\bar{X} = K(M+D_y)^{-1/2}$ and $\bar{Y} = M(\beta M + D_u)^{-1/2}$, which we use below to derive our approximation. Note that to demonstrate an upper bound for this problem, one would write

$$\begin{aligned}
R &= 1 - \frac{2\boldsymbol{\omega}^\top\boldsymbol{\chi}}{(\boldsymbol{\chi}+\boldsymbol{\omega})^\top(\boldsymbol{\chi}+\boldsymbol{\omega})} \\
&= 1 - \frac{2\mathbf{v}^\top M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\mathbf{v}}{\mathbf{v}^\top K(M+D_y)^{-1}K^\top\mathbf{v} + \mathbf{v}^\top M(\beta M+D_u)^{-1}M\mathbf{v} + 2\mathbf{v}^\top M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\mathbf{v}} \\
&\leq 1 - \min_{\mathbf{v}\neq\mathbf{0}} \left\{ \frac{2\mathbf{v}^\top M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\mathbf{v}}{\mathbf{v}^\top\left[K(M+D_y)^{-1}K^\top + M(\beta M+D_u)^{-1}M + 2M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\right]\mathbf{v}} \right\} \\
&= 1 - \min_{\mathbf{v}\neq\mathbf{0}} \left\{ \left(1 + \frac{\mathbf{v}^\top\left[K(M+D_y)^{-1}K^\top + M(\beta M+D_u)^{-1}M\right]\mathbf{v}}{2\mathbf{v}^\top M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\mathbf{v}} \right)^{-1} \right\},
\end{aligned} \tag{35}$$

provided $\mathbf{v} \notin \ker(K^\top)$. We may therefore draw the following conclusions:

- The Rayleigh quotient R is certainly finite, as the case $\boldsymbol{\chi}+\boldsymbol{\omega} = \mathbf{0}$ is disallowed by the assumption of invertibility of $\widehat{S}_G$.
- Furthermore, depending on the (typically unknown) entries of D_y, the term $\mathbf{v}^\top K(M+D_y)^{-1}K^\top\mathbf{v}$ should be large compared with the term $\mathbf{v}^\top M(\beta M+D_u)^{-1/2}(M+D_y)^{-1/2}K^\top\mathbf{v}$ arising in the denominator above, due to the fact that K has larger eigenvalues than M in general. The term being minimized in (35) will therefore not take a large negative value in general, and hence R will not become excessively large.
- However, it is generally not possible to demonstrate a concrete upper bound unless $\bar{X}$ and $\bar{Y}$ have structures which can be exploited. The reason for this is that the diagonal matrices D_y and D_u that determine the distribution of the eigenvalues can take any positive value (including arbitrarily small or infinitely large values, in finite precision), depending on the behaviour of the Newton iterates, which is impossible to control. In practice, we find it is rare for the largest eigenvalues of the preconditioned Schur complement to exceed values of roughly $5-10$.

– However, using the methodology of Theorem 1, results of this form have been demonstrated for problems such as convection-diffusion control [36] and heat equation control [35] (without additional bound constraints). We also highlight that, in [39,42], preconditioners for problems with bound constraints[1], solved with active set Newton methods, are derived. In [39], parameter-independent bounds are derived for a preconditioned Schur complement, however the additional requirement is imposed that M is a *lumped* (i.e. diagonal) mass matrix. As we do not assume that the mass matrices are lumped in this work, we may not exploit this method to obtain an upper eigenvalue bound.
– In general, the eigenvalues of $\widehat{S}_G^{-1}S_G$ are better clustered if the term $\bar{X}\bar{Y}^\top + \bar{Y}\bar{X}^\top$ is positive semi-definite, or 'nearly' positive semi-definite. The worst case would arise in the setting where $\boldsymbol{\chi} \approx -\boldsymbol{\omega}$, however for our problem the matrices $\bar{X}$ and $\bar{Y}$ do not relate closely to each other as the activities in the state and control variables do not share many common features.

We now provide an indicative result for the situation which corresponds to the limiting case when the barrier parameter $\mu \to 0$ and all state and control bounds are satisfied as strict inequalities, i.e. all bounds remain inactive at the optimum. In such a case all Lagrange multipliers $\mathbf{z}_{y,a}$, $\mathbf{z}_{y,b}$, $\mathbf{z}_{u,a}$ and $\mathbf{z}_{u,b}$ would take small values of order μ and so would the diagonal matrices D_y and D_u defined by (25) and (26), respectively. In the limit we would observe $D_y = 0$ and $D_u = 0$.

Lemma 1 *If $D_y = D_u = 0$, and the matrix $K + K^\top$ is positive semi-definite[2], then the eigenvalues of $\widehat{S}_G^{-1}S_G$ satisfy $\lambda \leq 1$.*

Proof From the above working, we have that

$$R = 1 - \frac{2\boldsymbol{\omega}^\top\boldsymbol{\chi}}{(\boldsymbol{\chi}+\boldsymbol{\omega})^\top(\boldsymbol{\chi}+\boldsymbol{\omega})} = 1 - \frac{\frac{1}{\sqrt{\beta}}\mathbf{v}^\top(K+K^\top)\mathbf{v}}{\mathbf{v}^\top K M^{-1}K^\top\mathbf{v} + \frac{1}{\beta}\mathbf{v}^\top M\mathbf{v} + \frac{1}{\sqrt{\beta}}\mathbf{v}^\top(K+K^\top)\mathbf{v}},$$

using the assumption that $D_y = D_u = 0$. The denominator of the quotient above is clearly positive, due to the positive definiteness of M, and the numerator is non-negative by the assumption of positive semi-definiteness of $K + K^\top$. This automatically leads to the statement $R \leq 1$, and hence the result. □

The 'matching strategy' presented here guarantees a lower bound for the preconditioned Schur complement of matrices of this form, provided some very weak assumptions hold[3], and often results in the largest eigenvalue being of moderate magnitude. We therefore wish to make use of this matching approach to generate effective Schur complement approximations for the very general class of matrix systems considered in this manuscript. In particular, we consider matrices K of general form (as opposed to the stiffness matrix as in (34)), as well as diagonal matrices D_y and D_u which can be extremely ill-conditioned. Motivated by Theorem 1, we may therefore consider a matching strategy for the Schur complement (33), by writing

$$\widehat{S}_1 := \big(K + \widehat{M}_1\big)(M + D_y)^{-1}\big(K + \widehat{M}_1\big)^\top, \tag{36}$$

where $\widehat{M}_1$ is chosen such that the matrix $\widehat{M}_1(M + D_y)^{-1}\widehat{M}_1^\top$ captures the second term of the exact Schur complement (33). That is,

$$\widehat{M}_1(M + D_y)^{-1}\widehat{M}_1^\top \approx M(\beta M + D_u)^{-1}M.$$

This leads to the following requirement when selecting $\widehat{M}_1$:

$$\widehat{M}_1 \approx M(\beta M + D_u)^{-1/2}(M + D_y)^{1/2}.$$

We take diagonal approximations where appropriate, in order to avoid having to construct square roots of matrices, which would be extremely expensive computationally. That is, we take

$$\widehat{M}_1 = M\big[\mathrm{diag}(\beta M + D_u)\big]^{-1/2}\big[\mathrm{diag}(M + D_y)\big]^{1/2}. \tag{37}$$

We now present a result concerning this choice for $\widehat{M}_1$.

[1] For the problems considered in [39], bounds for $\alpha_y y + \alpha_u u$ are specified, where α_y and α_u are given constants.

[2] This assumption holds for both Poisson control and convection-diffusion control problems, for instance.

[3] The main assumption made is that $\widehat{S}_G$ is invertible. This certainly holds unless $(\bar{X} + \bar{Y})^\top\mathbf{v} = \mathbf{0}$ for some $\mathbf{v}$, which in our setting implies that $M^{-1}(\beta M + D_u)^{1/2}(M + D_y)^{-1/2}K^\top$ has an eigenvalue exactly equal to -1. As the matrices M, D_y, D_u and K are unlikely to interact closely at any Newton step, this is extremely unlikely to occur and our assumption is therefore reasonable.

Lemma 2 *When the Schur complement* (33) *is approximated by* $\widehat{S}_1$*, and with* $\widehat{M}_1$ *given by* (37)*, then, provided* $K + \widehat{M}_1$ *is invertible, the eigenvalues of* $\widehat{S}_1^{-1}S$ *satisfy*

$$\lambda \geq \frac{1}{2} \cdot \frac{\min\left\{\lambda_{\min}\left(\left[\mathrm{diag}(M)\right]^{-1}M\right), 1\right\}}{\max\left\{\lambda_{\max}\left(\left[\mathrm{diag}(M)\right]^{-1}M\right), 1\right\}}.$$

In other words the eigenvalues are bounded below by a fixed constant, depending only on the finite element discretization used.

Proof Selecting $\widehat{M}_1$ as in (37) gives that the eigenvalues of $\widehat{S}_1^{-1}S$ are determined by the Rayleigh quotient

$$R := \frac{\mathbf{v}^\top S\mathbf{v}}{\mathbf{v}^\top \widehat{S}_1\mathbf{v}} = \frac{\boldsymbol{\chi}^\top\boldsymbol{\chi} + \boldsymbol{\omega}^\top\boldsymbol{\omega}}{(\boldsymbol{\chi}+\boldsymbol{\gamma})^\top(\boldsymbol{\chi}+\boldsymbol{\gamma})} = \frac{\boldsymbol{\chi}^\top\boldsymbol{\chi} + \dfrac{\boldsymbol{\omega}^\top\boldsymbol{\omega}}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}\,\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}{(\boldsymbol{\chi}+\boldsymbol{\gamma})^\top(\boldsymbol{\chi}+\boldsymbol{\gamma})},$$

where for this problem the vectors of interest are $\boldsymbol{\chi} = (M+D_y)^{-1/2}K^\top\mathbf{v}$, $\boldsymbol{\omega} = (\beta M + D_u)^{-1/2}M\mathbf{v}$ and $\boldsymbol{\gamma} = (M+D_y)^{-1/2}\left[\mathrm{diag}(M+D_y)\right]^{1/2}\left[\mathrm{diag}(\beta M+D_u)\right]^{-1/2}M\mathbf{v}$. As the numerator and denominator both consist of positive quantities, using the assumption that $K+\widehat{M}_1$ is invertible, with the possible exception of $\boldsymbol{\chi}^\top\boldsymbol{\chi}$ which may be zero, we can state that

$$R \geq \min\left\{\frac{\boldsymbol{\omega}^\top\boldsymbol{\omega}}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}, 1\right\} \cdot \frac{\boldsymbol{\chi}^\top\boldsymbol{\chi} + \boldsymbol{\gamma}^\top\boldsymbol{\gamma}}{(\boldsymbol{\chi}+\boldsymbol{\gamma})^\top(\boldsymbol{\chi}+\boldsymbol{\gamma})} \geq \frac{1}{2} \cdot \min\left\{\frac{\boldsymbol{\omega}^\top\boldsymbol{\omega}}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}, 1\right\},$$

by setting $\bar{X} = K(M+D_y)^{-1/2}$ and $\bar{Y} = M\left[\mathrm{diag}(\beta M+D_u)\right]^{-1/2}\left[\mathrm{diag}(M+D_y)\right]^{1/2}(M+D_y)^{-1/2}$ within Theorem 1.

We then observe that the quotient $\frac{\boldsymbol{\omega}^\top\boldsymbol{\omega}}{\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}$ can be decomposed as

$$\begin{aligned}
&\frac{\mathbf{w}_1^\top(\beta M+D_u)^{-1}\mathbf{w}_1}{\mathbf{w}_1^\top\left[\mathrm{diag}(\beta M+D_u)\right]^{-1/2}\left[\mathrm{diag}(M+D_y)\right]^{1/2}(M+D_y)^{-1}\left[\mathrm{diag}(M+D_y)\right]^{1/2}\left[\mathrm{diag}(\beta M+D_u)\right]^{-1/2}\mathbf{w}_1} \\
&= \frac{\mathbf{w}_1^\top(\beta M+D_u)^{-1}\mathbf{w}_1}{\mathbf{w}_1^\top\left[\mathrm{diag}(\beta M+D_u)\right]^{-1}\mathbf{w}_1} \cdot \frac{\mathbf{w}_2^\top\left[\mathrm{diag}(M+D_y)\right]^{-1}\mathbf{w}_2}{\mathbf{w}_2^\top(M+D_y)^{-1}\mathbf{w}_2},
\end{aligned}$$

where $\mathbf{w}_1 = M\mathbf{v} \neq \mathbf{0}$ and $\mathbf{w}_2 = \left[\mathrm{diag}(M+D_y)\right]^{1/2}\left[\mathrm{diag}(\beta M+D_u)\right]^{-1/2}\mathbf{w}_1 \neq \mathbf{0}$.

Now, it may be easily shown that

$$\begin{aligned}
\frac{\mathbf{w}_1^\top(\beta M+D_u)^{-1}\mathbf{w}_1}{\mathbf{w}_1^\top\left[\mathrm{diag}(\beta M+D_u)\right]^{-1}\mathbf{w}_1} &\geq \left[\max\left\{\lambda_{\max}\left(\left[\mathrm{diag}(M)\right]^{-1}M\right), 1\right\}\right]^{-1}, \\
\frac{\mathbf{w}_2^\top\left[\mathrm{diag}(M+D_y)\right]^{-1}\mathbf{w}_2}{\mathbf{w}_2^\top(M+D_y)^{-1}\mathbf{w}_2} &\geq \min\left\{\lambda_{\min}\left(\left[\mathrm{diag}(M)\right]^{-1}M\right), 1\right\},
\end{aligned}$$

using the working earlier in this section. Combining these bounds gives the desired result. □

Clearly, it is valuable to have this insight that using our approximation $\widehat{M}_1$ retains the parameter independence of the lower bound for the eigenvalues of $\widehat{S}_1^{-1}S$. We note that this can potentially be a weak bound, as the large diagonal entries in D_y and D_u are likely to dominate the behaviour of $M+D_y$ and $\beta M + D_u$, thus driving the eigenvalues of the preconditioned Schur complement closer to 1.

We highlight that, in practice, one may also approximate the inverses of $K+\widehat{M}_1$ and its transpose effectively using a multigrid process. We apply the Aggregation-based Algebraic Multigrid (AGMG) software [30–33] for this purpose within our iterative solvers.

Combining our approximations of Φ and S, we propose the following block diagonal preconditioner of the form (31):

$$\mathcal{P}_1 = \begin{bmatrix} (M+D_y)_{\mathrm{approx}} & 0 & 0 \\ 0 & (\beta M+D_u)_{\mathrm{approx}} & 0 \\ 0 & 0 & \widehat{S}_1 \end{bmatrix},$$

where $(M + D_y)_{\text{approx}}$, $(\beta M + D_u)_{\text{approx}}$ indicate our choice of approximations for $M + D_y$, $\beta M + D_u$ (i.e. diagonal approximation, or Chebyshev semi-iteration method), and $\widehat{S}_1$ is given by (36). This preconditioner is symmetric positive definite, and may thus be applied within a symmetric solver such as MINRES [34].

It is useful to consider the distribution of eigenvalues of the preconditioned system, as this will control the convergence properties of the MINRES method. The fundamental result we use for our analysis of saddle point matrices (30) is stated below [40, Lemma 2.1].

Theorem 2 *If Φ is symmetric positive definite, Ψ is full rank, and $\Theta = 0$, the eigenvalues of $\mathcal{A}$ are contained within the following intervals:*

$$\lambda(\mathcal{A}) \in \left[\frac{1}{2}\left(\mu_{\min} - \sqrt{\mu_{\min}^2 + 4\sigma_{\max}^2}\right), \frac{1}{2}\left(\mu_{\max} - \sqrt{\mu_{\max}^2 + 4\sigma_{\min}^2}\right)\right] \cup \left[\mu_{\min}, \frac{1}{2}\left(\mu_{\max} + \sqrt{\mu_{\max}^2 + 4\sigma_{\max}^2}\right)\right],$$

where $\mu_{\max}$, $\mu_{\min}$ denote the largest and smallest eigenvalues of Φ, with $\sigma_{\max}$, $\sigma_{\min}$ the largest and smallest singular values of Ψ.

We now wish to apply a result of this form to the preconditioned system. The preconditioned matrix, when a general block diagonal preconditioner of the form (31) is used, is given by

$$\mathcal{P}^{-1}\mathcal{A} = \begin{bmatrix} \widehat{\Phi} & 0 \\ 0 & \widehat{S} \end{bmatrix}^{-1} \begin{bmatrix} \Phi & \Psi^\top \\ \Psi & 0 \end{bmatrix} = \begin{bmatrix} \widehat{\Phi}^{-1}\Phi & \widehat{\Phi}^{-1}\Psi^\top \\ \widehat{S}^{-1}\Psi & 0 \end{bmatrix}.$$

Now, to analyse the properties of this system, let

$$\lambda(\widehat{\Phi}^{-1}\Phi) \in [\phi_{\min}, \phi_{\max}], \qquad \lambda(\widehat{S}^{-1}S) \in [s_{\min}, s_{\max}],$$

where $\phi_{\min}, s_{\min} > 0$. The analysis of this section gives us information about these values.

By the similarity property of matrix systems (using that for our problem $\widehat{\Phi}$ and $\widehat{S}$ are positive definite) the eigenvalues will be the same as those of

$$\begin{aligned} \mathcal{P}^{-1/2}\mathcal{A}\mathcal{P}^{-1/2} &= \begin{bmatrix} \widehat{\Phi}^{-1/2} & 0 \\ 0 & \widehat{S}^{-1/2} \end{bmatrix} \begin{bmatrix} \Phi & \Psi^\top \\ \Psi & 0 \end{bmatrix} \begin{bmatrix} \widehat{\Phi}^{-1/2} & 0 \\ 0 & \widehat{S}^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} \widehat{\Phi}^{-1/2}\Phi\widehat{\Phi}^{-1/2} & \widehat{\Phi}^{-1/2}\Psi^\top\widehat{S}^{-1/2} \\ \widehat{S}^{-1/2}\Psi\widehat{\Phi}^{-1/2} & 0 \end{bmatrix}. \end{aligned}$$

The eigenvalues of the $(1,1)$-block of this matrix, $\widehat{\Phi}^{-1/2}\Phi\widehat{\Phi}^{-1/2}$, are the same as those of $\widehat{\Phi}^{-1}\Phi$ by similarity, and so are contained in $[\phi_{\min}, \phi_{\max}]$. The singular values of the $(2,1)$-block are given by the square roots of the eigenvalues of $\widehat{S}^{-1/2}\Psi\widehat{\Phi}^{-1}\Psi^\top\widehat{S}^{-1/2}$, i.e. the square roots of the eigenvalues of $\widehat{S}^{-1}(\Psi\widehat{\Phi}^{-1}\Psi^\top)$ by similarity. Writing the Rayleigh quotient (for $\mathbf{v} \neq \mathbf{0}$),

$$\frac{\mathbf{v}^\top\Psi\widehat{\Phi}^{-1}\Psi^\top\mathbf{v}}{\mathbf{v}^\top\widehat{S}\mathbf{v}} = \frac{\mathbf{v}^\top\Psi\widehat{\Phi}^{-1}\Psi^\top\mathbf{v}}{\mathbf{v}^\top\Psi\Phi^{-1}\Psi^\top\mathbf{v}} \cdot \frac{\mathbf{v}^\top\Psi\Phi^{-1}\Psi^\top\mathbf{v}}{\mathbf{v}^\top\widehat{S}\mathbf{v}} = \underbrace{\frac{\bar{\mathbf{v}}^\top\widehat{\Phi}^{-1}\bar{\mathbf{v}}}{\bar{\mathbf{v}}^\top\Phi^{-1}\bar{\mathbf{v}}}}_{\in[\phi_{\min},\phi_{\max}]} \cdot \underbrace{\frac{\mathbf{v}^\top\Psi\Phi^{-1}\Psi^\top\mathbf{v}}{\mathbf{v}^\top\widehat{S}\mathbf{v}}}_{\in[s_{\min},s_{\max}]},$$

where $\bar{\mathbf{v}} = \Psi^\top\mathbf{v}$, enables us to pin the singular values of the $(2,1)$-block within $\left[\sqrt{\phi_{\min}s_{\min}}, \sqrt{\phi_{\max}s_{\max}}\right]$.

So, using Theorem 2, the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are contained within the interval stated below.

Lemma 3 *If Φ and S are symmetric positive definite, and the above bounds on $\lambda(\widehat{\Phi}^{-1}\Phi)$ and $\lambda(\widehat{S}^{-1}S)$ hold, then the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ satisfy*

$$\lambda(\mathcal{P}^{-1}\mathcal{A}) \in \left[\frac{1}{2}\left(\phi_{\min} - \sqrt{\phi_{\min}^2 + 4\phi_{\max}s_{\max}}\right), \frac{1}{2}\left(\phi_{\max} - \sqrt{\phi_{\max}^2 + 4\phi_{\min}s_{\min}}\right)\right] \cup \left[\phi_{\min}, \frac{1}{2}\left(\phi_{\max} + \sqrt{\phi_{\max}^2 + 4\phi_{\max}s_{\max}}\right)\right].$$

It is therefore clear that, for our problem, a good approximation of the Schur complement will guarantee clustered eigenvalues of the preconditioned system, and therefore rapid convergence of the MINRES method. As we have observed for our problem, the quantities of interest are therefore the largest eigenvalues of $\widehat{S}^{-1}S$, which can vary at every step of a Newton method.

We now present a straightforward result concerning the eigenvectors of a preconditioned saddle point system of the form under consideration.

Proposition 1 *Consider an eigenvalue λ that satisfies*

$$\begin{bmatrix} \Phi & \Psi^\top \\ \Psi & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \lambda \begin{bmatrix} \widehat{\Phi} & 0 \\ 0 & \widehat{S} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}, \tag{38}$$

with Φ, $S = \Psi\Phi^{-1}\Psi^\top$, $\widehat{\Phi}$, $\widehat{S}$ symmetric positive definite. Then either λ is an eigenvalue of $\widehat{\Phi}^{-1}\Phi$, or λ, $\mathbf{v}_1$ and $\mathbf{v}_2$ satisfy

$$\left(\lambda\widehat{\Phi} - \Phi - \frac{1}{\lambda}\Psi^\top\widehat{S}^{-1}\Psi\right)\mathbf{v}_1 = \mathbf{0}, \qquad \mathbf{v}_2 = \frac{1}{\lambda}\widehat{S}^{-1}\Psi\mathbf{v}_1.$$

Proof Equation (38) is equivalent to

$$\Psi^\top\mathbf{v}_2 = \big(\lambda\widehat{\Phi} - \Phi\big)\mathbf{v}_1, \tag{39}$$

$$\Psi\mathbf{v}_1 = \lambda\widehat{S}\mathbf{v}_2. \tag{40}$$

Let us first consider the case where $\Psi\mathbf{v}_1 = \mathbf{0}$ (there are at most $n-m$ such linearly independent vectors that correspond to eigenvectors). Then (40) tells us that $\mathbf{v}_2 = \mathbf{0}$, from which we conclude from (39) that $(\lambda\widehat{\Phi} - \Phi)\mathbf{v}_1 = \mathbf{0}$. Therefore, in this case, the eigenvalues are given by eigenvalues of $\widehat{\Phi}^{-1}\Phi$, with eigenvectors of the form $\begin{bmatrix}\mathbf{v}_1^\top, & \mathbf{0}^\top\end{bmatrix}^\top$ – there are at most $n-m$ such solutions.

If $\Psi\mathbf{v}_1 \neq \mathbf{0}$, we may rearrange (40) to obtain

$$\mathbf{v}_2 = \frac{1}{\lambda}\widehat{S}^{-1}\Psi\mathbf{v}_1 \quad \Rightarrow \quad \Psi^\top\mathbf{v}_2 = \frac{1}{\lambda}\Psi^\top\widehat{S}^{-1}\Psi\mathbf{v}_1,$$

which we may substitute into (39) to obtain

$$\frac{1}{\lambda}\Psi^\top\widehat{S}^{-1}\Psi\mathbf{v}_1 = \big(\lambda\widehat{\Phi} - \Phi\big)\mathbf{v}_1.$$

This may be trivially rearranged to obtain the required result. □

We observe that the eigenvalues and eigenvectors of the $(1,1)$-block and Schur complement (along with their approximations) interact strongly with each other. This decreases the likelihood of many extreme eigenvalues of $\widehat{S}^{-1}S$ arising in practice, as this would have implications on the numerical properties of Φ and Ψ (which for our problems do not interact at all strongly). However the working provided here shows that this is very difficult to prove rigorously, due to the wide generality of the saddle point systems being examined – we must also rely on the physical properties of the PDE operators within the optimization framework. Our numerical experiments of Section 5 indicate that the eigenvalues of $\widehat{S}^{-1}S$, and therefore the preconditioned system, are tightly clustered, matching some of the observations made in this section.

As an alternative to the block diagonal preconditioner $\mathcal{P}_1$, we may take account of information on the block lower triangular parts of the matrix system, and apply the block triangular preconditioner

$$\mathcal{P}_2 = \begin{bmatrix} (M + D_y)_{\text{approx}} & 0 & 0 \\ 0 & (\beta M + D_u)_{\text{approx}} & 0 \\ K & -M & -\widehat{S}_1 \end{bmatrix},$$

within a non-symmetric solver such as GMRES [41].

It is possible to carry out eigenvalue analysis for the block triangular preconditioner $\mathcal{P}_2$ in the same way as for the block diagonal preconditioner $\mathcal{P}_1$. However it is well known that the convergence of non-symmetric solvers such as GMRES does not solely depend on the eigenvalues of the preconditioned system, and therefore such an analysis would be less useful in practice.

We now consider a completely different strategy for preconditioning the matrix system. We may first rearrange (24) to the form

$$\begin{bmatrix} \beta M + D_u & -M & 0 \\ -M & 0 & K \\ 0 & K^\top & M + D_y \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta u} \\ \boldsymbol{\delta \lambda} \\ \boldsymbol{\delta y} \end{bmatrix} = \begin{bmatrix} \mu(U - U_a)^{-1}\mathbf{e} - \mu(U_b - U)^{-1}\mathbf{e} - \beta M \mathbf{u}^* + M\boldsymbol{\lambda}^* \\ \mathbf{f} - K\mathbf{y}^* + M\mathbf{u}^* \\ \mu(Y - Y_a)^{-1}\mathbf{e} - \mu(Y_b - Y)^{-1}\mathbf{e} + \mathbf{y}_d - M\mathbf{y}^* - K^\top \boldsymbol{\lambda}^* \end{bmatrix}. \quad (41)$$

The matrix within (41) is a saddle point system of the form (30), with

$$\Phi = \begin{bmatrix} \beta M + D_u & -M \\ -M & 0 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} 0 & K^\top \end{bmatrix}, \qquad \Theta = \begin{bmatrix} M + D_y \end{bmatrix}.$$

This approach also has the desirable feature that the $(1,1)$-block Φ can be inverted almost precisely, as all that is required is a method for approximating the inverse of a mass matrix (to be applied twice). Once again, a very cheap and accurate method is Chebyshev semi-iteration [14,15,48], so we apply this strategy within our preconditioner.

Once again, the main challenge is to approximate the Schur complement:

$$\begin{aligned} S = & -(M + D_y) + \begin{bmatrix} 0 & K^\top \end{bmatrix} \begin{bmatrix} \beta M + D_u & -M \\ -M & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ K \end{bmatrix} \\ = & -(M + D_y) + \begin{bmatrix} 0 & K^\top \end{bmatrix} \begin{bmatrix} 0 & -M^{-1} \\ -M^{-1} & -M^{-1}(\beta M + D_u)M^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ K \end{bmatrix} \\ = & -\Big[K^\top M^{-1}(\beta M + D_u)M^{-1}K + (M + D_y)\Big]. \end{aligned}$$

Let us consider a 'matching' strategy once again, and write for our approximation:

$$\widehat{S}_2 := -\big(K^\top + \widehat{M}_2\big)M^{-1}(\beta M + D_u)M^{-1}\big(K + \widehat{M}_2^\top\big),$$

where $\widehat{M}_2$ is selected to incorporate the second term of S, i.e.

$$\widehat{M}_2 M^{-1}(\beta M + D_u)M^{-1}\widehat{M}_2^\top \approx M + D_y,$$

which may be achieved if

$$\widehat{M}_2 \approx (M + D_y)^{1/2}(\beta M + D_u)^{-1/2}M.$$

For a practical preconditioner, we in fact select

$$\widehat{M}_2 = \big[\mathrm{diag}(M + D_y)\big]^{1/2}\big[\mathrm{diag}(\beta M + D_u)\big]^{-1/2}M.$$

To approximate $K^\top + \widehat{M}_2$ and $K + \widehat{M}_2^\top$ in practice, we again make use of the AGMG software to apply a multigrid process to the relevant matrices within $\widehat{S}_2$.

One may therefore build a block triangular preconditioner for the permuted system (41), of the form $\mathcal{P}_T$ in (31). Rearranging the matrix system (and hence the preconditioner) to the form (24), we are therefore able to construct the following preconditioner for (24):

$$\mathcal{P}_3 = \begin{bmatrix} -\widehat{S}_2 & 0 & K^\top \\ 0 & \beta M + D_u & -M_{\mathrm{cheb}} \\ 0 & -M_{\mathrm{cheb}} & 0 \end{bmatrix},$$

where M_{cheb} relates to a Chebyshev semi-iteration process for the mass matrix M. We notice that this relates to observations made on nullspace preconditioners for saddle point systems in [38].

It is clear that to apply the preconditioner $\mathcal{P}_3$, we require a non-symmetric solver such as GMRES, as it is not possible to construct a positive definite preconditioner with this rearrangement of the matrix system. Within such a solver, a key positive property of this strategy is that we may approximate Φ almost perfectly (and cheaply), and may apply $K^\top$ exactly within $\mathcal{P}_3$ without a matrix inversion. An associated disadvantage is that our approximation of S is more expensive to apply than the approximation $\widehat{S}_1$ used within the preconditioners $\mathcal{P}_1$ and $\mathcal{P}_2$ – whereas Theorem 1 may again be applied[4] to verify a lower bound for the eigenvalues of the preconditioned Schur complement, the values of the largest eigenvalues are frequently found to be higher than for the Schur complement approximation $\widehat{S}_1$ described earlier.

[4] In the notation of Theorem 1, the matrices involved are $\bar{X} = K^\top M^{-1}(\beta M + D_u)^{1/2}$ and $\bar{Y} = (M + D_y)^{1/2}$.

4.2 Time-dependent problems

Due to the huge dimensions of the matrix systems arising from time-dependent PDE-constrained optimization problems, it is very important to consider preconditioners for the resulting systems, which are of the form (29). These are again of saddle point type (30), with

$$\Phi = \begin{bmatrix} \tau\mathcal{M}_{1/2} + \mathcal{D}_y & 0 \\ 0 & \beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u \end{bmatrix}, \qquad \Psi = \begin{bmatrix} \mathcal{K} & -\tau\mathcal{M} \end{bmatrix}, \qquad \Theta = \begin{bmatrix} 0 \end{bmatrix}.$$

As for the time-independent case we may approximate Φ using diagonal solves or the Chebyshev semi-iteration method applied to the matrices from each time-step.

To approximate the Schur complement of (29),

$$\mathcal{S} = \mathcal{K}\big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)^{-1}\mathcal{K}^\top + \tau^2\mathcal{M}\big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)^{-1}\mathcal{M},$$

we again apply a matching strategy to obtain

$$\widehat{\mathcal{S}}_{1,T} := \big(\mathcal{K} + \widehat{\mathcal{M}}_{1,T}\big)\big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)^{-1}\big(\mathcal{K} + \widehat{\mathcal{M}}_{1,T}\big)^\top,$$

where

$$\widehat{\mathcal{M}}_{1,T}\big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)^{-1}\widehat{\mathcal{M}}_{1,T}^\top \approx \tau^2\mathcal{M}\big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)^{-1}\mathcal{M}.$$

This in turn motivates the choice

$$\widehat{\mathcal{M}}_{1,T} = \tau\mathcal{M}\left[\operatorname{diag}\big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)\right]^{-1/2}\left[\operatorname{diag}\big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)\right]^{1/2},$$

and we require two multigrid processes per time-step to apply $\widehat{\mathcal{S}}_{1,T}^{-1}$ efficiently.

Combining our approximations of $(1,1)$-block and Schur complement, we may apply

$$\mathcal{P}_{1,T} = \begin{bmatrix} \big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)_{\text{approx}} & 0 & 0 \\ 0 & \big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)_{\text{approx}} & 0 \\ 0 & 0 & \widehat{\mathcal{S}}_{1,T} \end{bmatrix}$$

within MINRES, for example, or

$$\mathcal{P}_{2,T} = \begin{bmatrix} \big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)_{\text{approx}} & 0 & 0 \\ 0 & \big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)_{\text{approx}} & 0 \\ \mathcal{K} & -\tau\mathcal{M} & -\widehat{\mathcal{S}}_{1,T} \end{bmatrix},$$

within a nonsymmetric solver such as GMRES.

Alternatively, in complete analogy to the time-independent setting, one could rearrange the matrix system such that the $(1,1)$-block may be approximated accurately, and select the preconditioner

$$\mathcal{P}_{3,T} = \begin{bmatrix} -\widehat{\mathcal{S}}_{2,T} & 0 & \mathcal{K}^\top \\ 0 & \beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u & -\tau\mathcal{M}_{\text{cheb}} \\ 0 & -\tau\mathcal{M}_{\text{cheb}} & 0 \end{bmatrix}.$$

Inverting $\mathcal{M}_{\text{cheb}}$ requires the application of Chebyshev semi-iteration to N_t mass matrices M, and the Schur complement approximation is given by

$$\widehat{\mathcal{S}}_{2,T} := -\frac{1}{\tau^2}\big(\mathcal{K}^\top + \widehat{\mathcal{M}}_{2,T}\big)\mathcal{M}^{-1}\big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)\mathcal{M}^{-1}\big(\mathcal{K} + \widehat{\mathcal{M}}_{2,T}^\top\big),$$

with

$$\widehat{\mathcal{M}}_{2,T} = \tau\left[\operatorname{diag}\big(\tau\mathcal{M}_{1/2} + \mathcal{D}_y\big)\right]^{1/2}\left[\operatorname{diag}\big(\beta\tau\mathcal{M}_{1/2} + \mathcal{D}_u\big)\right]^{-1/2}\mathcal{M}.$$

Similar eigenvalue results can be shown for the Schur complement approximation $\widehat{\mathcal{S}}_{1,T}$ as for the approximations used in the time-independent case.

Remark 1 We highlight that a class of methods which is frequently utilized when solving PDE-constrained optimization problems, aside from the iterative methods discussed in this paper, is that of multigrid. We recommend [8] for an overview of such methods for PDE-constrained optimization, [7] for a convergence analysis of multigrid applied to these problems, [20,21] for schemes derived for solving flow control problems, and [6] for a method tailored to problems with additional bound constraints. These solvers require the careful construction of prolongation/restriction operators, as well as smoothing methods, tailored to the precise problem at hand. Applying multigrid to the entire coupled matrix systems resulting from the problems considered in this paper, as opposed to employing this technology to solve sub-blocks of the system within an iterative method, also has the potential to be a powerful approach for solving problems with bound constraints. Similar multigrid methods have previously been applied to the interior point solution of PDE-constrained optimization problems in one article [9], and we believe that a carefully tailored scheme could be a viable alternative when solving at least some of the numerical examples considered in Section 5.

Alternative problem formulations

We have sought to illustrate our interior point solvers, and in particular the preconditioned iterative methods for the solution of the associated Newton systems, using quadratic tracking functionals with a quadratic cost for the control, as in (2). We now wish to briefly outline some of the possible extensions to this problem that we believe we could apply our method to, as below:

- *Boundary control problems.* Our methodology could be readily extended to problems where the control (or state) variable is regularized on the boundary only within the cost functional, for instance where
$$\mathcal{J}(y,u) = \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L_2(\partial\Omega)}.$$
For such problems, we would need to take account of boundary mass matrices within the saddle point system that arises, however preconditioners have previously been designed for such problems that take into account these features (see [35], for instance).
- *Control variable regularized on a subdomain.* Analogously, problems may be considered using our preconditioning approach where the cost functional is of the form
$$\mathcal{J}(y,u) = \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L_2(\Omega_1)},$$
where $\Omega_1 \subset \Omega$. The matching strategy of Section 4.1 may be modified to account for the matrices of differing structures.
- *Alternative regularizations.* A further possibility is for the control (or state) variable to be regularized using a different term, for instance an H^1 regularization term of the following form:
$$\mathcal{J}(y,u) = \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{H^1(\Omega)} = \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|\nabla u\|^2_{L_2(\Omega)}.$$
Upon discretization, stiffness matrices arise within the $(1,1)$-block in addition to mass matrices, however the preconditioning method introduced in this paper may still be applied, by accounting for the new matrices within the matching strategy for the Schur complement.
- *Time-dependent problems.* Finally, we highlight that modifications to the cost functional considered for time-dependent problems in Section 3.3 may be made. For instance, one may measure the control (or state) variables at the final time only, that is
$$\mathcal{J}(y,u) = \frac{1}{2}\int_0^T \int_\Omega \big(y(\mathbf{x},t)-\widehat{y}(\mathbf{x},t)\big)^2 \,\mathrm{d}\Omega\mathrm{d}t + \frac{\beta}{2}\int_\Omega u(\mathbf{x},T)^2 \,\mathrm{d}\Omega.$$
On the discrete level, this will lead to mass matrices being removed from portions of the $(1,1)$-block, and this information may be built into new preconditioners [35,44].

We emphasize that there are some examples of cost functional, for instance a functional where a curl function is applied to state or control, or one which includes terms of the form $\int \max\{0, \det(\nabla y)\}$ (see [19]), where the preconditioning approach presented here would not be directly applicable. As PDE-constrained optimization problems are widespread and varied in type, much useful further work could be carried out on extending the method presented in this paper to more diverse classes of optimization problems.

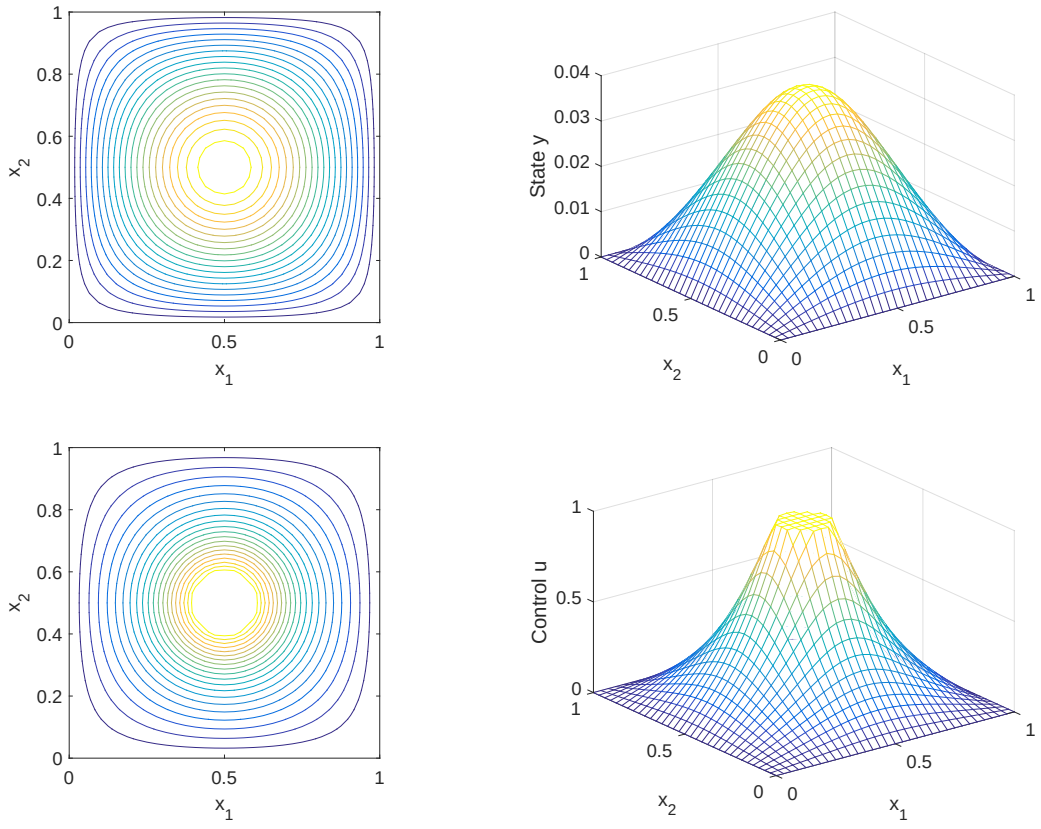

Fig. 1 Contour and mesh plots of the solution to the Poisson control example with control constraints, for state variable y (top) and control variable u (bottom), with $\beta = 10^{-2}$.

5 Numerical experiments

Having motivated our numerical methods for the solution of the problems considered, we now wish to test our solvers on a range of examples. These test problems are of both time-independent and time-dependent form, and are solved on a desktop with a quad-core 3.2GHz processor. For each test problem, we discretize the state, control and adjoint variables using $Q1$ finite elements. Within the interior point method, the value of the barrier reduction parameter σ is set to be 0.1, with $\alpha_0 = 0.995$, and $\epsilon_p = \epsilon_d = \epsilon_c = 10^{-6}$. To solve the Newton systems arising from the interior point method, we use the IFISS software package [11,43] to construct the relevant finite element matrices. When the symmetric block diagonal preconditioner $\mathcal{P}_1$ is used, we solve the Newton systems using the MINRES algorithm to a relative preconditioned residual norm tolerance of 10^{-8}, and the Chebyshev semi-iteration method to approximate the inverse of the $(1,1)$-block (apart from within one experiment where we use a diagonal approximation), as well as the AGMG method to approximate the inverse Schur complement. Where the block triangular preconditioners $\mathcal{P}_2$ and $\mathcal{P}_3$ are applied, we solve the Newton systems with the preconditioned GMRES method to a tolerance of 10^{-8}; we apply 20 steps of Chebyshev semi-iteration to approximate the $(1,1)$-block, and once again utilize AGMG for the Schur complement approximations. We highlight that it would be feasible to relax the tolerances for MINRES and GMRES in order to lower the overall CPU time for the interior point scheme [16], however we elect to solve the matrix systems relatively accurately in order to fully demonstrate the potency of our preconditioned iterative methods. All results are computed using MATLAB R2015a.

Control constrained problems. The first experiments we carry out involve a Poisson control problem, with $\mathcal{L} = -\nabla^2$ applied on $\Omega := [0,1]^2$, $y = 0$ on the boundary of Ω, and the desired state given by $\widehat{y} = e^{-64\left((x_1-0.5)^2+(x_2-0.5)^2\right)}$, where the spatial coordinates $\vec{x} = [x_1,\ x_2]^\top$. We solve this problem using the MINRES algorithm with preconditioner $\mathcal{P}_1$, using both the Chebyshev semi-iteration method

Table 1 Results for the Poisson control example with control constraints, for a range of values of h and β, and preconditioner $\mathcal{P}_1$. Presented are the number of interior point (Newton) iterations required to achieve convergence (blue, left), and average number of MINRES steps per interior point iteration before a relative preconditioned residual norm of 10^{-8} is achieved (black, right). Results are given with a Chebyshev semi-iteration method applied to the $(1,1)$-block (top), and with a diagonal approximation (bottom).

$\mathcal{P}_1$ Chebyshev		$\beta = 1$ $u \geq 0$ $u \leq 0.01$		$\beta = 10^{-1}$ $u \geq 0$ $u \leq 0.1$		$\beta = 10^{-2}$ $u \geq 0$ $u \leq 1$		$\beta = 10^{-3}$ $u \geq 0$ $u \leq 3$		$\beta = 10^{-4}$ $u \geq 0$ $u \leq 20$		$\beta = 10^{-5}$ $u \geq 0$ $u \leq 100$		$\beta = 10^{-6}$ $u \geq 0$ $u \leq 300$	
h	2^{-2}	10	5.6	11	6.3	13	6.2	15	6.6	18	7.5	19	7.2	20	7.4
	2^{-3}	10	5.7	13	6.1	14	6.3	16	7.8	19	8.3	20	8.7	21	9.3
	2^{-4}	10	5.6	13	6.1	15	6.5	19	7.4	22	8.6	22	8.5	21	8.8
	2^{-5}	11	5.4	16	5.8	18	6.3	21	7.0	23	8.8	25	8.9	24	9.4
	2^{-6}	11	5.5	16	5.8	20	6.2	22	6.8	26	15.5	24	8.9	30	9.4
	2^{-7}	12	5.2	18	5.5	20	6.2	20	7.1	27	8.4	25	8.6	31	9.2

$\mathcal{P}_1$ Diagonal		$\beta = 1$ $u \geq 0$ $u \leq 0.01$		$\beta = 10^{-1}$ $u \geq 0$ $u \leq 0.1$		$\beta = 10^{-2}$ $u \geq 0$ $u \leq 1$		$\beta = 10^{-3}$ $u \geq 0$ $u \leq 3$		$\beta = 10^{-4}$ $u \geq 0$ $u \leq 20$		$\beta = 10^{-5}$ $u \geq 0$ $u \leq 100$		$\beta = 10^{-6}$ $u \geq 0$ $u \leq 300$	
h	2^{-2}	9	9.4	11	10.4	13	9.5	15	9.2	18	10.1	19	9.4	20	17.6
	2^{-3}	10	15.1	12	16.7	14	16.9	16	18.4	19	17.5	20	18.5	21	19.5
	2^{-4}	10	15.5	15	18.6	16	19.9	19	22.7	22	21.6	22	23.4	21	24.3
	2^{-5}	11	16.3	16	16.2	19	19.5	21	21.1	23	24.7	25	25.7	24	25.8
	2^{-6}	11	15.5	16	20.2	20	16.9	22	18.9	26	32.1	24	18.9	31	26.7
	2^{-7}	12	14.3	18	15.7	21	16.1	20	18.5	28	28.8	25	19.3	31	23.4

Table 2 Results for the Poisson control example with state constraints, for a range of values of h and β. Presented are the number of interior point iterations required to achieve convergence (blue, left), and average number of GMRES steps needed (black, right). Results are given when the preconditioners $\mathcal{P}_2$ (top) and $\mathcal{P}_3$ (bottom) are used.

$\mathcal{P}_2$		$\beta = 1$ $-0.1 \leq y \leq 0.002$		$\beta = 10^{-2}$ $-0.1 \leq y \leq 0.175$		$\beta = 10^{-4}$ $-0.1 \leq y \leq 0.9$		$\beta = 10^{-6}$ $-0.1 \leq y \leq 1$	
h	2^{-2}	11	5.3	8	5.0	9	5.0	10	5.0
	2^{-3}	12	9.9	9	10.2	10	13.3	10	10.9
	2^{-4}	13	11.4	10	12.9	11	16.8	11	13.5
	2^{-5}	14	12.1	11	13.3	13	27.4	12	15.0
	2^{-6}	16	12.5	12	13.6	14	17.8	13	15.7
	2^{-7}	17	12.7	13	14.6	16	16.9	14	16.3

$\mathcal{P}_3$		$\beta = 1$ $-0.1 \leq y \leq 0.002$		$\beta = 10^{-2}$ $-0.1 \leq y \leq 0.175$		$\beta = 10^{-4}$ $-0.1 \leq y \leq 0.9$		$\beta = 10^{-6}$ $-0.1 \leq y \leq 1$	
h	2^{-2}	11	5.0	8	5.1	9	5.0	10	5.0
	2^{-3}	12	9.6	9	9.1	10	10.5	10	10.5
	2^{-4}	13	11.2	10	10.3	11	12.3	11	12.4
	2^{-5}	14	12.1	11	10.8	13	12.9	12	13.5
	2^{-6}	16	12.6	12	11.4	14	13.3	13	13.9
	2^{-7}	17	13.1	13	13.0	16	13.5	14	14.5

and the matrix diagonal to approximate the $(1,1)$-block within the preconditioner. The results obtained are shown in Table 1, for a range of mesh-sizes h and regularization parameters β. A solution plot for $\beta = 10^{-2}$ is also shown in Figure 1. We select box constraints for the control variable only, based on the value of β used and the behaviour of the optimal control problem when no bound constraints are imposed – we are careful to make sure that the constraints are sensible physically, but also challenging for our interior point solver. The constraints taken for each value of β are stated in Table 1. It is worth pointing out that increasing the accuracy of discretization (decreasing h by a factor of 2) typically adds about one extra interior point iteration, which once again demonstrates that interior point methods are not very sensitive to the problem dimension (as discussed in [17], for instance). We find that both the number of iterations of the interior point method, and the average number of MINRES iterations per interior point (Newton) step, are very reasonable for the problem considered. Whereas we observe an increase in iterative steps for the more challenging case of smaller β, all numbers are low, in particular the very

Table 3 Number of interior point (Newton) iterations, average number of iterations of the Krylov subspace method per interior point step, and CPU time required to solve the Poisson control example with state constraints, when the preconditioners $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ are used. Results are presented for a range of h, and fixed $\beta = 10^{-2}$.

$\beta = 10^{-2}$		$\mathcal{P}_1$			$\mathcal{P}_2$			$\mathcal{P}_3$		
		IPM	Krylov	CPU	IPM	Krylov	CPU	IPM	Krylov	CPU
h	2^{-2}	8	8.0	**0.13**	8	5.0	**0.20**	8	5.1	**0.22**
	2^{-3}	9	11.8	**0.23**	9	10.2	**0.35**	9	9.1	**0.34**
	2^{-4}	10	14.5	**0.46**	10	12.9	**0.63**	10	10.3	**0.57**
	2^{-5}	11	14.1	**1.8**	11	13.3	**2.6**	11	10.8	**2.4**
	2^{-6}	13	14.8	**9.1**	12	13.6	**11.4**	12	11.4	**10.1**
	2^{-7}	14	14.9	**37.4**	13	14.6	**54.4**	13	13.0	**53.8**

Table 4 Results for the Helmholtz problem with state constraints, for a range of values of h and β, as well as values of k. Presented are the number of interior point iterations required to achieve convergence (blue, left), and average number of GMRES steps needed (black, right). Results are given when the preconditioners $\mathcal{P}_2$ (top) and $\mathcal{P}_3$ (bottom) are used.

$\mathcal{P}_2$		$k = 20$						$k = 50$			
		$\beta = 10^{-2}$		$\beta = 10^{-4}$		$\beta = 10^{-6}$		$\beta = 10^{-2}$		$\beta = 10^{-4}$	
		$-0.0005 \le y \le 0.0005$		$-0.05 \le y \le 0.05$		$-0.6 \le y \le 0.6$		$-10^{-5} \le y \le 10^{-5}$		$-0.001 \le y \le 0.001$	
h	2^{-2}	7	4.3	10	5.3	10	4.7	5	4.0	8	4.7
	2^{-3}	8	9.2	10	11.6	11	12.4	5	6.8	8	12.3
	2^{-4}	8	10.6	11	17.7	12	30.8	6	10.4	8	17.9
	2^{-5}	9	11.2	12	18.8	12	19.6	6	6.1	9	20.5
	2^{-6}	9	10.4	12	15.9	13	22.7	7	10.3	10	23.1
	2^{-7}	10	10.2	13	15.6	14	15.5	8	10.6	10	20.1

$\mathcal{P}_3$		$k = 20$						$k = 50$			
		$\beta = 10^{-2}$		$\beta = 10^{-4}$		$\beta = 10^{-6}$		$\beta = 10^{-2}$		$\beta = 10^{-4}$	
		$-0.0005 \le y \le 0.0005$		$-0.05 \le y \le 0.05$		$-0.6 \le y \le 0.6$		$-10^{-5} \le y \le 10^{-5}$		$-0.001 \le y \le 0.001$	
h	2^{-2}	7	4.2	10	5.2	10	3.9	5	3.7	8	5.1
	2^{-3}	8	9.3	10	11.8	11	8.2	5	6.2	8	10.4
	2^{-4}	8	9.9	11	13.9	12	10.0	6	9.3	8	16.8
	2^{-5}	9	10.9	12	15.2	12	10.2	6	5.1	9	19.1
	2^{-6}	9	10.2	12	15.1	13	10.4	7	9.5	10	21.8
	2^{-7}	10	10.3	13	14.7	14	10.2	8	10.0	10	18.5

encouraging iteration counts for moderate regularization parameters. We also find that, as one might expect, the computational cheapness of a diagonal approximation of the $(1,1)$-block is counteracted by the higher MINRES iteration numbers that result.

Problems with state constraints. We next examine a Poisson control problem involving state constraints, where $\widehat{y} = \sin(\pi x_1)\sin(\pi x_2)$, and $y = \widehat{y}$ on the boundary of Ω. We apply the preconditioners $\mathcal{P}_2$ (with Chebyshev semi-iteration used to approximate the $(1,1)$-block) and $\mathcal{P}_3$, and solve using GMRES to a tolerance of 10^{-8} for a range of h and β. Again the results, which are presented in Table 2, are very promising when either preconditioner is used, and a large degree of robustness is achieved despite the very general matrix systems which can arise at each interior point iteration. We highlight that the iteration counts are likely to vary depending on how severe the box constraints that we impose are, as the structure of the matrices can change drastically. In Table 3 we present results for this problem (for $\beta = 10^{-2}$) with preconditioners $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ – we observe in particular that the CPU times scale in an approximately linear fashion with the dimension of the matrix systems being solved.

In order to illustrate the potential of our solvers to handle PDE constraints of varying forms, in Table 4 we present results where the PDE constraint is an indefinite Helmholtz equation, that is $\mathcal{L}y = -\nabla^2 y - k^2 y$ for a given (positive) parameter k. We highlight that the forward Helmholtz equation itself is a notoriously difficult problem to solve numerically [12], and a great deal of research has been undertaken concerning the preconditioning of such systems (we recommend [13] for a discussion of shifted Laplacian preconditioners for these problems). We therefore emphasize that, given the challenges involved and the inherent indefiniteness of the problem, it is extremely difficult to obtain completely robust solvers, and much future research could be undertaken in this area. However the results obtained indicate that, at

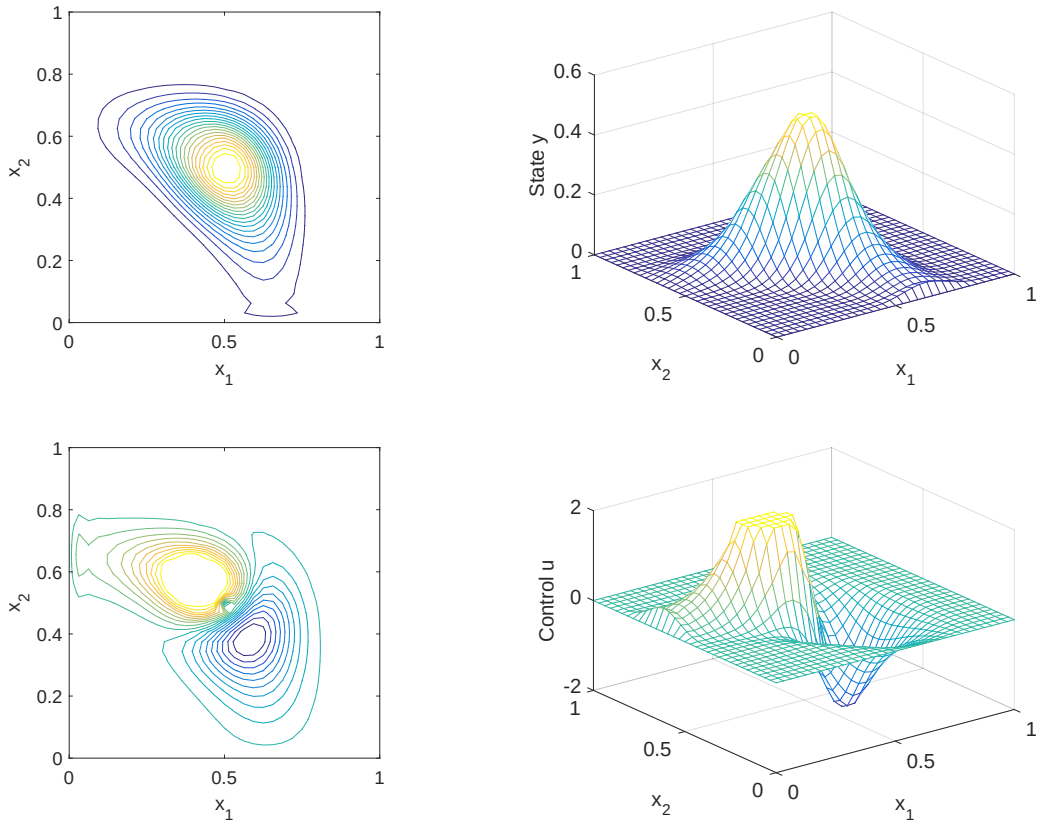

Fig. 2 Contour and mesh plots of the solution to the convection-diffusion control example with state and control constraints, for state variable y (top) and control variable u (bottom), with $\beta = 10^{-2}$.

least for some test problems, the interior point method presented can be applied for a range of parameter setups.

Both state and control constraints. In Table 5 we investigate a problem of convection-diffusion control type, with $\mathcal{L} = -0.01\nabla^2 + \left[-\frac{1}{\sqrt{2}},\ \frac{1}{\sqrt{2}}\right]^\top \cdot \nabla$, and $\widehat{y} = e^{-64\left((x_1-0.5)^2+(x_2-0.5)^2\right)}$. We now impose both state and control constraints (as specified for each value of β), and test the preconditioners $\mathcal{P}_2$ and $\mathcal{P}_3$ using GMRES. We also present a solution plot for $\beta = 10^{-2}$ in Figure 2. For convection-diffusion control problems such as this, we find there is a great advantage in applying the preconditioner $\mathcal{P}_3$ over the preconditioner $\mathcal{P}_2$, due in part to the accurate approximation of the $(1,1)$-block within it. Indeed this is demonstrated by the numbers of GMRES iterations required, which are much lower when using the preconditioner $\mathcal{P}_3$, especially for the final interior point iterations when convergence is close to being achieved. The GMRES solver with $\mathcal{P}_3$ demonstrates excellent robustness considering the complexity of the problem.

3D test problems. It is also important to emphasize that the methodology presented in this work can be readily applied to three dimensional test problems – indeed these are problems for which it is generally accepted that preconditioned iterative methods are essential, as the huge computer storage requirements associated with such problems ensure that direct methods are out of reach. We therefore experiment using a Poisson control problem applied on the domain $\Omega := [0,1]^3$, with desired state $\widehat{y} = e^{-64\left((x_1-0.5)^2+(x_2-0.5)^2+(x_3-0.5)^2\right)}$ and spatial coordinates $\vec{x} = [x_1,\ x_2,\ x_3]^\top$. We present numerical results in Table 6, demonstrating that, as for two dimensional problems, rapid convergence is achieved with robustness in problem size and regularization parameter.

Time-dependent PDE constraints. To demonstrate that our solvers are also able to handle matrix systems of vast dimension arising from time-dependent PDE-constrained optimization problems, we present results in Table 7 for a heat equation control problem, with the PDE constraint given by

Table 5 Results for the convection-diffusion control example with state and control constraints, for a range of values of h and β. Presented are the number of interior point iterations required to achieve convergence (blue, left), and average number of GMRES steps needed (black, right). Results are given when the preconditioners $\mathcal{P}_3$ (top) and $\mathcal{P}_2$ (bottom) are used.

$\mathcal{P}_3$		$\beta = 10^{-1}$ $0 \le y \le 0.2$ $-0.75 \le u \le 0.75$		$\beta = 10^{-2}$ $0 \le y \le 0.5$ $-2 \le u \le 2$		$\beta = 10^{-3}$ $0 \le y \le 0.5$ $-3 \le u \le 3$		$\beta = 10^{-4}$ $0 \le y \le 0.75$ $-5 \le u \le 5$		$\beta = 10^{-5}$ $0 \le y \le 0.75$ $-6 \le u \le 6$	
h	2^{-2}	13	8.9	14	9.1	15	9.5	14	8.7	14	8.8
	2^{-3}	14	11.3	15	10.9	15	12.1	15	12.2	15	11.8
	2^{-4}	15	13.1	15	11.8	16	13.4	16	13.3	16	14.1
	2^{-5}	17	13.9	17	13.3	16	14.7	19	13.7	19	14.9
	2^{-6}	19	14.6	19	14.5	17	17.9	22	16.6	23	16.3
	2^{-7}	21	23.0	21	14.9	22	17.3	26	17.7	27	18.4

$\mathcal{P}_2$		$\beta = 10^{-1}$ $0 \le y \le 0.2$ $-0.75 \le u \le 0.75$		$\beta = 10^{-2}$ $0 \le y \le 0.5$ $-2 \le u \le 2$		$\beta = 10^{-3}$ $0 \le y \le 0.5$ $-3 \le u \le 3$		$\beta = 10^{-4}$ $0 \le y \le 0.75$ $-5 \le u \le 5$		$\beta = 10^{-5}$ $0 \le y \le 0.75$ $-6 \le u \le 6$	
h	2^{-2}	13	10.1	14	11.3	14	11.3	14	11.1	14	11.4
	2^{-3}	14	20.9	15	19.8	15	24.8	15	21.8	15	22.4
	2^{-4}	16	35.1	15	20.6	16	42.6	16	37.6	17	53.0
	2^{-5}	17	44.1	17	40.4	16	45.6	19	64.3	19	69.3
	2^{-6}	19	52.6	19	48.4	17	47.3	22	66.7	23	73.6
	2^{-7}	21	50.0	21	48.2	22	63.5	26	75.0	27	81.7

Table 6 Results for the three dimensional Poisson control example with control constraints, for a range of values of h and β, and preconditioner $\mathcal{P}_1$. Presented are the number of interior point (Newton) iterations required to achieve convergence (blue, left), and average number of MINRES steps per interior point iteration before a relative preconditioned residual norm of 10^{-8} is achieved (black, right).

$\mathcal{P}_1$ Chebyshev		$\beta = 1$ $u \ge 0$ $u \le 0.01$		$\beta = 10^{-1}$ $u \ge 0$ $u \le 0.1$		$\beta = 10^{-2}$ $u \ge 0$ $u \le 1$		$\beta = 10^{-3}$ $u \ge 0$ $u \le 3$		$\beta = 10^{-4}$ $u \ge 0$ $u \le 20$		$\beta = 10^{-5}$ $u \ge 0$ $u \le 100$		$\beta = 10^{-6}$ $u \ge 0$ $u \le 300$	
h	2^{-2}	7	10.8	8	10.7	9	11.1	10	11.2	11	11.2	12	11.1	12	11.4
	2^{-3}	8	10.7	9	10.8	11	10.4	11	11.0	12	11.0	13	11.1	12	11.1
	2^{-4}	9	10.9	10	10.8	11	10.8	12	10.9	12	11.2	13	11.1	13	11.1
	2^{-5}	11	14.1	12	14.0	12	13.8	13	13.8	13	13.6	14	13.5	14	13.8

$y_t - \nabla^2 y = u$ (for $t \in (0, 1]$), and with additional control constraints imposed. The number of interior point iterations, and average MINRES iteration count when $\mathcal{P}_{1,T}$ is applied, are provided for a range of h and β. As mentioned earlier, the backward Euler method is used for the time discretization, and values of $\tau = 0.04$, 0.02 and 0.01 are tested for the time-step (in other words with 25, 50 and 100 time intervals). In Table 8, we present results obtained for the same problem using block triangular preconditioner $\mathcal{P}_{2,T}$ with GMRES. We once again observe a high degree of robustness in problem size (whether increased by refining the mesh in the spatial coordinates, or by decreasing the time-step) and regularization parameter.

Our final investigation involves the optimal control of the wave equation, which is the same problem as above, except with the PDE operator $y_{tt} - \nabla^2 y = u$ and with an initial condition imposed on y_t (which we set to be zero). The recent work [27] derives an implicit scheme for this problem, which involves averaging the Laplacian term in the PDE operator. Within the matrix $\mathcal{K}$, this leads to discrete approximations of the operator $I - \frac{\tau^2}{2}\nabla^2$ on the block diagonal entries, as well as additional entries on the two blocks below the diagonal (corresponding to the operators $-2I$ and $I - \frac{\tau^2}{2}\nabla^2$, respectively). The method is designed to be unconditionally convergent, while also removing the requirement of a Courant–Friedrichs–Lewy (CFL) condition of the form $\tau \le h$ [27]. We investigate the potency of our preconditioners for this matrix system. In Table 9, we present the average number of MINRES or GMRES iterations required to solve the systems arising from the interior point method. Although there is a larger variation in the number of steps required, due to the additional terms within the matrix system, the performance of the method is very encouraging considering the high complexity of the problem. We emphasize once again that the performance of the method is dependent somewhat on the severity of

Table 7 Results for the heat equation control example with control constraints, for a range of values of h, τ, and β, and preconditioner $\mathcal{P}_{1,T}$. Presented are the number of interior point iterations required to achieve convergence (blue, left), and average number of MINRES steps needed (black, right).

$\mathcal{P}_{1,T}$ $\tau = 0.04$		$\beta = 10^{-1}$ $0 \le u \le 0.1$		$\beta = 10^{-2}$ $0 \le u \le 1$		$\beta = 10^{-3}$ $0 \le u \le 3$		$\beta = 10^{-4}$ $0 \le u \le 30$	
h	2^{-2}	13	13.1	15	16.5	16	19.7	21	31.3
	2^{-3}	15	13.7	16	16.6	18	20.5	24	30.6
	2^{-4}	16	14.0	18	17.1	20	20.8	24	28.2
	2^{-5}	16	14.0	19	17.5	21	21.0	25	27.5
	2^{-6}	18	14.5	19	17.5	22	21.1	27	27.2

$\mathcal{P}_{1,T}$ $\tau = 0.02$		$\beta = 10^{-1}$ $0 \le u \le 0.1$		$\beta = 10^{-2}$ $0 \le u \le 1$		$\beta = 10^{-3}$ $0 \le u \le 3$		$\beta = 10^{-4}$ $0 \le u \le 30$	
h	2^{-2}	14	13.0	16	15.9	17	19.8	23	31.9
	2^{-3}	15	13.4	17	15.6	19	20.5	25	30.9
	2^{-4}	16	13.7	18	16.0	21	20.9	25	27.6
	2^{-5}	17	14.0	19	16.4	22	21.1	28	28.3
	2^{-6}	15	13.4	19	16.2	22	20.8	27	27.6

$\mathcal{P}_{1,T}$ $\tau = 0.01$		$\beta = 10^{-1}$ $0 \le u \le 0.1$		$\beta = 10^{-2}$ $0 \le u \le 1$		$\beta = 10^{-3}$ $0 \le u \le 3$		$\beta = 10^{-4}$ $0 \le u \le 30$	
h	2^{-2}	14	12.2	16	15.4	18	19.6	24	31.0
	2^{-3}	15	12.4	18	15.7	19	19.9	28	30.9
	2^{-4}	16	12.8	18	15.7	21	20.2	27	28.2
	2^{-5}	16	12.8	18	15.7	22	20.5	30	28.3
	2^{-6}	17	13.0	19	15.8	22	20.4	29	28.5

Table 8 Results for the heat equation control example with control constraints, for a range of values of h, τ, and β, and preconditioner $\mathcal{P}_{2,T}$. Presented are the number of interior point iterations required to achieve convergence (blue, left), and average number of GMRES steps needed (black, right).

$\mathcal{P}_{2,T}$		$\tau = 0.04$								$\tau = 0.02$							
		$\beta = 10^{-1}$ $0 \le u \le 0.1$		$\beta = 10^{-2}$ $0 \le u \le 1$		$\beta = 10^{-3}$ $0 \le u \le 3$		$\beta = 10^{-4}$ $0 \le u \le 30$		$\beta = 10^{-1}$ $0 \le u \le 0.1$		$\beta = 10^{-2}$ $0 \le u \le 1$		$\beta = 10^{-3}$ $0 \le u \le 3$		$\beta = 10^{-4}$ $0 \le u \le 30$	
h	2^{-2}	13	8.1	15	9.9	16	11.7	21	18.1	14	7.9	16	9.6	17	11.8	23	18.5
	2^{-3}	15	8.4	16	9.9	18	11.8	24	17.2	15	8.2	17	9.6	19	12.1	25	17.7
	2^{-4}	16	8.5	18	10.3	20	12.1	24	16.3	16	8.4	18	9.8	21	12.4	25	16.2
	2^{-5}	16	8.5	19	10.4	21	12.2	25	16.1	17	8.5	19	10.0	22	12.3	28	16.8
	2^{-6}	18	8.8	19	10.4	22	12.7	27	16.3	15	8.2	19	9.9	22	12.6	27	16.5

the box constraints imposed, however the numerical results obtained for a range of time-independent and time-dependent PDE-constrained optimization problems demonstrate the potency of the solvers presented in this manuscript.

Table 9 Results for the wave equation example with control constraints, for a range of values of h, τ, and β. Presented are the average number of MINRES (with preconditioner $\mathcal{P}_{1,T}$) and GMRES (with preconditioner $\mathcal{P}_{2,T}$) iterations required to solve the Newton systems obtained.

		$\mathcal{P}_{1,T}$, $h = 2^{-4}$			$\mathcal{P}_{1,T}$, $h = 2^{-5}$			$\mathcal{P}_{2,T}$, $h = 2^{-4}$			$\mathcal{P}_{2,T}$, $h = 2^{-5}$		
		β			β			β			β		
		10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
τ	0.04	13.7	17.7	13.3	14.7	18.0	13.4	10.1	12.1	9.9	11.1	12.7	10.0
	0.02	12.5	12.5	13.1	11.7	12.7	13.1	8.9	9.2	9.9	8.6	9.1	9.8
	0.01	14.7	10.9	10.6	31.6	50.8	10.9	10.9	13.5	8.1	22.8	23.7	8.5
	0.005	26.9	29.2	30.7	37.3	42.9	35.7	21.5	22.2	24.0	21.8	22.5	22.1

6 Concluding remarks

In this paper we have presented a practical method for the interior point solution of a number of PDE-constrained optimization problems with state and control constraints, by reformulating the minimization of the discretized system as a quadratic programming problem. Having outlined the structure of the algorithm for solving these problems, we derived fast and feasible preconditioned iterative methods for solving the resulting Newton systems, which is the dominant portion of the algorithm in terms of computational work. Encouraging numerical results indicate the effectiveness and utility of our approach.

The problems we considered involved Poisson control, heat equation control, and both steady and time-dependent convection-diffusion control. A natural extension of this work would be to consider the control of systems of PDEs, for instance Stokes control and other problems in fluid flow, as well as the control of nonlinear PDEs, which arises in a wide range of practical scientific applications. The latter task would be accomplished by reformulating the discretization as a nonlinear programming problem – the robust solution of such formulations is a substantial challenge within the optimization community, but would represent significant progress in tackling real-world optimal control problems.

Acknowledgements. The authors are grateful to two anonymous referees for their careful reading of the manuscript and helpful comments. The first author was partially funded for this research by the Engineering and Physical Sciences Research Council (EPSRC) Fellowship EP/M018857/1.

References

1. A. Battermann and M. Heinkenschloss, *Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems*. In: W. Desch, F. Kappel, and K. Kunisch (eds), *Control and Estimation of Distributed Parameter Systems*, pp.15–32, 1998.
2. M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14, pp.1–137, 2005.
3. M. Benzi, E. Haber, and L. Taralli, *Multilevel algorithms for large-scale interior point methods*, SIAM Journal on Scientific Computing, 31, pp.4152–4175, 2009.
4. L. Bergamaschi and A. Martinez, *RMCP: Relaxed mixed constraint preconditioners for saddle point linear systems arising in geomechanics*, Computer Methods in Applied Mechanics and Engineering, 221–222, pp.54–62, 2012.
5. M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of a Moreau-Yosida based active set strategy and interior point methods for constrained optimal control problems*, SIAM Journal on Optimization, 11, pp.495–521, 2000.
6. A. Borzì and K. Kunisch, *A multigrid scheme for elliptic constrained optimal control problems*, Computational Optimization and Applications, 31(3), pp.309–333, 2005.
7. A. Borzì, K. Kunisch, and D. Y. Kwak, *Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system*, SIAM Journal on Control and Optimization, 41(5), pp.1477–1497, 2003.
8. A. Borzì and V. Schulz, *Multigrid methods for PDE optimization*, SIAM Review, 51(2), pp.361–395, 2009.
9. A. Drăgănescu and C. Petra, *Multigrid preconditioning of linear systems for interior point methods applied to a class of box-constrained optimal control problems*, SIAM Journal on Numerical Analysis, 50(1), pp.328–353, 2012.
10. I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford University Press, New York, 1987.
11. H. C. Elman, A. Ramage, and D. J. Silvester, *IFISS: a computational laboratory for investigating incompressible flow problems*, SIAM Review, 56, pp.261–273, 2014.
12. O. G. Ernst, and M. J. Gander, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, Numerical Analysis of Multiscale Problems, Volume 83 of Lecture Notes in Computational Science and Engineering, pp.325–363, 2011.
13. M. J. Gander, I. G. Graham, and E. A. Spence, *Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed?*, Numerische Mathematik, 31(3), pp.567–614, 2015.
14. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. I*, Numerische Mathematik, 3, pp.147–156, 1961.
15. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. II*, Numerische Mathematik, 3, pp.157–168, 1961.
16. J. Gondzio, *Convergence analysis of an inexact feasible interior point method for convex quadratic programming*, SIAM Journal on Optimization, 23, pp.1510–1527, 2013.
17. J. Gondzio, *Interior point methods 25 years later*, European Journal of Operational Research, 218, pp.587-601, 2012.
18. M. J. Grotte, J. Huber, D. Kourounis, and O. Schenk, *Inexact interior-point method for PDE-constrained nonlinear optimization*, SIAM Journal on Scientific Computing, 36, pp.A1251–A1276, 2014.
19. R. Herzog and K. Kunisch, *Algorithms for PDE-constrained optimization*, GAMM-Mitteilungen, 33(2), pp.163–176, 2010.
20. M. Hinze, M. Köster, and S. Turek, *A hierarchical space-time solver for distributed control of the Stokes equation*, Priority Programme 1253, Preprint Number SPP1253-16-01, 2008.

21. 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, Preprint Number SPP1253-16-02, 2008.
22. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, Springer-Verlag, New York, 2009.
23. I. C. F. Ipsen, *A note on preconditioning non-symmetric matrices*, SIAM Journal on Scientific Computing, 23(3), pp.1050–1051, 2001.
24. K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, Vol. 15 of Advances in Design and Control, Society for Industrial and Applied Mathematics, 2008.
25. C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM Journal on Scientific Computing, 15, pp.645–667, 1994.
26. Y. A. Kuznetsov, *Efficient iterative solvers for elliptic finite element problems on nonmatching grids*, Russian Journal of Numerical Analysis and Mathematical Modelling, 10, pp.187–211, 1995.
27. B. Li, J. Liu, and M. Xiao, *A fast and stable preconditioned iterative method for optimal control problem of wave equations*, SIAM Journal on Scientific Computing, 37(6), pp.A2508–A2534, 2015.
28. H. D. Mittelmann and H. Maurer, *Solving elliptic control problems with interior point and SQP methods: control and state constraints*, Journal of Computational and Applied Mathematics, 120, pp.175–195, 2000.
29. M. F. Murphy, G. H. Golub, and A. J. Wathen, A note on preconditioning for indefinite linear systems, SIAM Journal on Scientific Computing, 21, pp.1969–1972, 2000.
30. A. Napov and Y. Notay, *An algebraic multigrid method with guaranteed convergence rate*, SIAM Journal on Scientific Computing, 34(2), pp.A1079–A1109, 2012.
31. Y. Notay, *An aggregation-based algebraic multigrid method*, Electronic Transactions on Numerical Analysis, 37, pp.123–146, 2010.
32. Y. Notay, *Aggregation-based algebraic multigrid for convection-diffusion equations*, SIAM Journal on Scientific Computing, 34(4), pp.A2288–A2316, 2012.
33. Y. Notay, *AGMG software and documentation*; see `http://homepages.ulb.ac.be/~ynotay/AGMG`.
34. C. C. Paige and M. A. Saunders, *Solutions of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12(4), pp.617–629, 1975.
35. J. W. Pearson, M. Stoll, and A. J. Wathen, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, SIAM Journal on Matrix Analysis and Applications, 33(4), pp.1126–1152, 2012.
36. J. W. Pearson and A. J. Wathen, *Fast iterative solvers for convection-diffusion control problems*, Electronic Transactions on Numerical Analysis, 40, pp.294–310, 2013.
37. J. W. Pearson and A. J. Wathen, *A new approximation of the Schur complement in preconditioners for PDE-constrained optimization*, Numerical Linear Algebra with Applications, 19, pp.816–829, 2012.
38. J. Pestana and T. Rees, *Null-space preconditioners for saddle point systems*, Rutherford Appleton Laboratory Technical Report RAL-TR-2015-003, 2015.
39. M. Porcelli, V. Simoncini, and M. Tani, *Preconditioning of active-set Newton methods for PDE-constrained optimal control problems*, SIAM Journal on Scientific Computing, 37(5), pp.S472–S502, 2016.
40. T. Rusten and R. Winther, *A preconditioned iterative method for saddle point problems*, SIAM Journal on Matrix Analysis and Applications, 13, pp.887–904, 1992.
41. Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific Computing, 7(3), pp.856–869, 1986.
42. A. Schiela and S. Ulbrich, *Operator preconditioning for a class of inequality constrained optimal control problems*, SIAM Journal on Optimization, 24(1), pp.435–466, 2014.
43. D. Silvester, H. Elman, and A. Ramage, *Incompressible Flow and Iterative Solver Software (IFISS), Version 3.3*, `http://www.manchester.ac.uk/ifiss`, 2014.
44. M. Stoll and A. Wathen, *All-at-once solution of time-dependent PDE-constrained optimization problems*, Oxford Centre for Collaborative Applied Mathematics Technical Report 10/47, 2010.
45. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, American Mathematical Society, Providence, Rhode Island, 2010.
46. M. Ulbrich and S. Ulbrich, *Primal-dual interior-point methods for PDE-constrained optimization*, Mathematical Programming, 117(1–2), pp.435–485, 2009.
47. A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA Journal of Numerical Analysis, 7(4), pp.449–457, 1987.
48. A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electronic Transactions on Numerical Analysis, 34, pp.125–135, 2009.
49. M. Weiser, *Interior point methods in function space*, SIAM Journal on Control and Optimization, 44, pp.1766–1786, 2005.
50. M. Weiser and P. Deuflhard, *Inexact central path following algorithms for optimal control problems*, SIAM Journal on Control and Optimization, 46, pp.792–815, 2007.
51. M. Weiser, T. Gänzler, and A. Schiela, *A control reduced primal interior point method for a class of control constrained optimal control problems*, Computational Optimization and Applications, 41, pp.127–145, 2008.
52. S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.