MAX-PLANCK-GESELLSCHAFT

John W. Pearson, Martin Stoll

Fast Iterative Solution of Reaction-Diffusion Control Problems Arising from Chemical Processes

MAX–PLANCK–INSTITUT
FÜR DYNAMIK KOMPLEXER
TECHNISCHER SYSTEME
MAGDEBURG

Max Planck Institute Magdeburg
Preprints

MPIMD/12-16 November 14, 2012

Impressum:

Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg

Publisher:
Max Planck Institute for Dynamics of Complex Technical Systems

Address:
Max Planck Institute for Dynamics of
Complex Technical Systems
Sandtorstr. 1
39106 Magdeburg

www.mpi-magdeburg.mpg.de/preprints

FAST ITERATIVE SOLUTION OF REACTION-DIFFUSION CONTROL PROBLEMS ARISING FROM CHEMICAL PROCESSES

JOHN W. PEARSON* AND MARTIN STOLL†

Abstract. PDE-constrained optimization problems, and the development of preconditioned iterative methods for the efficient solution of the arising matrix system, is a field of numerical analysis that has recently been attracting much attention. In this paper, we analyze and develop preconditioners for matrix systems that arise from the optimal control of reaction-diffusion equations, which themselves result from chemical processes. Important aspects in our solvers are saddle point theory, mass matrix representation and effective Schur complement approximation, as well as the outer (Newton) iteration to take account of the nonlinearity of the underlying PDEs.

Key words. PDE-constrained optimization, reaction-diffusion, chemical processes, Newton iteration, preconditioning, Schur complement.

AMS subject classifications. 65F08, 65F10, 65F50, 92E20, 93C20

1. Introduction.

Optimal control problems, including PDE-constrained optimization problems, have a number of applications in mathematical and physical problems [4]. One such field in which problems can be posed in this way is that of chemical processes [4, 18, 19, 20, 21]. In this case the underlying PDEs are reaction-diffusion equations, and therefore the PDE constraints in our formulation are nonlinear PDEs.

When solving such reaction-diffusion control problems using a finite element method, and employing a Lagrange-Newton iteration to take account of the nonlinearity involved in the PDEs, the resulting matrix system upon each Newton iteration will be large, sparse and of saddle point structure. It is therefore desirable to devise preconditioned iterative methods to solve these systems efficiently, and in such a way that the structure of the matrix is exploited. Work in devising preconditioners for PDE-constrained optimization problems has been considered for simpler problems previously, for instance Poisson control [43, 45, 49], Stokes control [48, 50, 35] and heat equation control [42].

In this paper, we will consider an optimal control formulation of a reaction-diffusion control problem, which generates a symmetric matrix system upon each Newton iteration (such an iteration is required to take account of the nonlinear terms within the underlying PDEs). The solvers for the matrix systems that we will discuss are Minres [41] and Bicg[13] – the choice of the appropriate method depends on whether the preconditioner we use is positive definite. We will search for block diagonal, symmetric positive definite preconditioners for the matrix systems we examine. In order to do this, we will need to approximate the $(1,1)$-block by accurately representing the inverse of mass matrices, as well as devise an effective approximation of the Schur complement of the matrix system. We aim to provide heuristic guidance as to the effectiveness of the Schur complement approximations we advocate, and also demonstrate with numerical tests why it these are sensible choices for a number of practical problems.

This paper is structured as follows. In Section 2, we discuss the underlying

[*] Numerical Analysis Group, Mathematical Institute, University of Oxford, 24–29 St Giles', Oxford, OX1 3LB, UK (john.pearson@worc.ox.ac.uk),

[†] Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtotstr. 1, 39106 Magdeburg Germany (stollm@mpi-magdeburg.ac.uk)

chemical problem (detailing the time-independent and time-dependent variants of the problem), and represent it in terms of a matrix problem. In Section 3, we introduce some basic saddle point theory and use this to devise effective preconditioners for the matrices which arise. In Section 4, we present numerical results to demonstrate the performance of our iterative solver in practice. Finally, in Section 5, we make some concluding remarks.

2. Problem formulation and discretization. Throughout this paper we consider an optimal control problem based on that considered in [4]. The objective function that has to be minimized is given by

$$\begin{aligned} J(u,v,c) = \ & \frac{\alpha_u}{2}\left\|u-u_Q\right\|^2_{L_2(Q)}+\frac{\alpha_v}{2}\left\|v-v_Q\right\|^2_{L_2(Q)} \\ & +\frac{\alpha_{TU}}{2}\left\|u(T)-u_\Omega\right\|^2_{L_2(\Omega)}+\frac{\alpha_{TV}}{2}\left\|v(T)-v_\Omega\right\|^2_{L_2(\Omega)}+\frac{\alpha_c}{2}\left\|c\right\|^2_{L_2(\Sigma)}, \end{aligned} \tag{2.1}$$

where u and v refer to concentrations of reactants (which in this problem are *state variables*), and c is the *control variable*, which also influences the underlying reaction. The domains of interest are given by a spatial domain $\Omega \subset \mathbb{R}^d$ with $d \in \{2,3\}$. The time domain is here given by the interval $t \in [0,T]$ and we have then the space-time domain Q given as $Q := \Omega \times [0,T]$ as well as the space-time boundary given by $\Sigma = \partial\Omega \times (0,T)$. The goal of the optimization is to compute the quantities u, v, and c, in such a way that they are close in the L_2 sense to what is often referred to as desired states (u_Ω,u_Q,v_Ω,v_Q). Note that we have 4 desired states in this problem – 2 which are defined at all time periods, and 2 which are solely defined at the final time at which the problem is being solved. These are known quantities and typically can come from measurements and observations. In order for the objective function to resemble a physical or chemical process the variables need to satisfy the physics of the process of interest, which are typically modeled using one or more PDEs plus additional constraints. In our case the constraint of the objective $J(u,v,c)$ is given by the reaction-diffusion equations, i.e.

$$u_t - D_1\Delta u + k_1 u = -\gamma_1 uv, \quad \text{in } Q, \tag{2.2}$$

$$v_t - D_2\Delta v + k_2 v = -\gamma_2 uv, \quad \text{in } Q, \tag{2.3}$$

$$D_1\partial_\nu u + b(x,t,u) = c, \quad \text{on } \Sigma, \tag{2.4}$$

$$D_2\partial_\nu v + \widetilde{\epsilon} v = 0, \quad \text{on } \Sigma, \tag{2.5}$$

$$u(x,0) = u_0(x), \quad \text{in } \Omega, \tag{2.6}$$

$$v(x,0) = v_0(x), \quad \text{in } \Omega, \tag{2.7}$$

$$c \in C_{ad} = \{c \in L_\infty(\Sigma) : c_a \le c \le c_b \text{ a.e. on } \Sigma\}. \tag{2.8}$$

The constants k_1, k_2, α_u, α_v, α_{TU}, α_{TV}, α_c, $\widetilde{\epsilon}$, γ_1 and γ_2 are non-negative constants. The function c describing the boundary condition (2.4) is the so-called control variable. Equations (2.6) and (2.7) defines the initial conditions for both concentrations. Additionally, we can impose so-called box constraints on the control as stated in Equation (2.8). In [21] Griesse and Volkwein also consider an integral constraint on c, which we will not consider here. In some cases it might also be sensible to include state constraints for the concentrations u and v, which would be described by

$$u_a \le u \le u_b, \quad v_a \le v \le v_b.$$

State constraints typically bring additional difficulties to optimal control problems (see [9, 32]) and are not considered further in this present paper. For the remainder of this paper we will follow the assumptions of $b(x,t,u)=0$ and $\widetilde{\epsilon}=0$ as studied in [21]. There are typically two ways of how to proceed from the above problem. The first is the so-called discretize-then-optimize approach, where we discretize the objective function and constraint to build a discrete Lagrangian, and then impose the optimality conditions in the discrete setting. The second is the so-called optimize-then-discretize approach, where we instead build a Lagrangian for the infinite dimensional problem and then discretize the first-order conditions. There is no preferred approach and we refer to [29] for a discussion of the two cases. We note that recently it has become a paradigm to create discretization schemes such that both approaches lead to the same discrete first order system. We also need to deal with the nonlinearity of the PDE constraint. We here apply a simple Sequential Quadratic Programming (SQP) or Lagrange-Newton method. Before we proceed to the derivation of optimality conditions and discretization we split the problem into two stages.

Derivation of the Newton system without control constraints. In this section we wish to further describe how the above problem can be treated and in particular focus on the treatment of the nonlinearity. We proceed by formally building the Lagrangian subject to the reaction-diffusion system, i.e.

$$\begin{aligned}
u_t - D_1\Delta u + k_1 u = & -\gamma_1 uv, \quad \text{in } Q,\\
v_t - D_2\Delta v + k_2 v = & -\gamma_2 uv, \quad \text{in } Q,\\
D_1\partial_\nu u = & \, c, \quad \text{on } \Sigma,\\
D_2\partial_\nu v = & \, 0, \quad \text{on } \Sigma,\\
u(x,0) = & \, u_0(x), \quad \text{in } \Omega,\\
v(x,0) = & \, v_0(x), \quad \text{in } \Omega,
\end{aligned}$$

to give

$$\begin{aligned}
\mathcal{L}(u,v,c,p,q) = J(v,u,c) &+ \int_Q p(u_t - D_1\Delta u + k_1 u + \gamma_1 uv)\\
&+ \int_Q q(v_t - D_2\Delta v + k_2 v + \gamma_2 uv)\\
&+ \int_\Sigma p_\Sigma(D_1\partial_\nu u - c) + \int_\Sigma q_\Sigma(D_2\partial_\nu v).
\end{aligned}$$

Here we have split up p and q into interior and boundary parts (p & p_Σ, and q & q_Σ). Note that we only included the PDE part without boundary and initial conditions, which of course would need to be done as well but for reasons of exposition we feel that the derivation below is more accessible; we refer to [4, 21] for more rigorous discussions. We now take the Fréchet derivative, and consider for brevity of presentation the case where $\alpha_{TU}=\alpha_{TV}=0$ of $\mathcal{L}$ to obtain the following[1]:

$$\begin{aligned}\partial\mathcal{L}=\frac{\partial}{\partial\varepsilon}\Bigg(&\frac{\alpha_u}{2}\int_Q(u+\varepsilon du-u_Q)^2+\int_Q(v+\varepsilon dv-v_Q)^2+\int_\Sigma(c+\varepsilon dc)^2\\&+\int_Q(p+\varepsilon dp)((u+\varepsilon du)_t-D_1\Delta(u+\varepsilon du)+k_1(u+\varepsilon du)+\gamma_1(u+\varepsilon du)(v+\varepsilon dv))\\&+\int_Q(q+\varepsilon dq)((v+\varepsilon dv)_t-D_2\Delta(v+\varepsilon dv)+k_2(v+\varepsilon dv)+\gamma_2(u+\varepsilon du)(v+\varepsilon dv))\\&+\int_\Sigma(p_\Sigma+\varepsilon dp_\Sigma)(D_1\partial_\nu(u+\varepsilon du)-c-\varepsilon dc)+\int_\Sigma(q_\Sigma+\varepsilon dq_\Sigma)(D_2\partial_\nu(v+\varepsilon dv))\Bigg)|_{\varepsilon=0}.\end{aligned}$$

We now wish to simplify all the expressions involved – we commence by considering the terms coming from the functional

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\left(\frac{\alpha_u}{2}\int_Q(u+\varepsilon du-u_Q)^2+\frac{\alpha_v}{2}\int_Q(v+\varepsilon dv-v_Q)^2+\frac{\alpha_c}{2}\int_Q(c+\varepsilon dc)^2\right)|_{\varepsilon=0}\\&\quad=\alpha_u\int_Q(u-u_Q)du+\alpha_v\int_Q(v-v_Q)dv+\alpha_c\int_\Sigma cdc.\end{aligned}$$

Further, we obtain that

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\left(\int_Q(p+\varepsilon dp)((u+\varepsilon du)_t-D_1\Delta(u+\varepsilon du)+k_1(u+\varepsilon du)+\gamma_1(u+\varepsilon du)(v+\varepsilon dv))\right)|_{\varepsilon=0}\\&\quad=\int_Q dp(u_t-D_1\Delta u+k_1u+uv)+\int_Q p(du_t-D_1\Delta du+k_1du+\gamma_1(vdu+udv)),\end{aligned}$$

where we now rewrite the second term to obtain

$$\int_Q(p_tdu+k_1pdu-D_1\Delta pdu+\gamma_1pvdu+\gamma_1pudv)-\int_\Omega(pdu)|_0^T-\int_\Sigma p\partial_\nu du+du\partial_\nu p,$$

using Green's first identity twice. Similarly, we simplify

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\left(\int_Q(q+\varepsilon dq)((v+\varepsilon dv)_t-D_2\Delta(v+\varepsilon dv)+k_2(v+\varepsilon dv)+\gamma_2(u+\varepsilon du)(v+\varepsilon dv))\right)|_{\varepsilon=0}\\&\quad=\int_Q(v_t-D_2\Delta v+k_2v+\gamma_2uv)dq+\int_Q q((dv)_t-D_2\Delta dv+k_2dv+\gamma_2(vdu+udv)),\end{aligned}$$

and examine the second part of this expression to obtain

$$\int_Q(q_t-D_2\Delta q+k_2q+\gamma_2qu)dv+\gamma_2qvdu-\int_\Omega(qdv)|_0^T-\int_\Sigma q\partial_\nu dv+dq\partial_\nu v.$$

Finally, we simplify

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\int_\Sigma(p_\Sigma+\varepsilon dp_\Sigma)(D_1\partial_\nu(u+\varepsilon du)-c-\varepsilon dc)|_{\varepsilon=0}+\frac{\partial}{\partial\varepsilon}\int_\Sigma(q_\Sigma+\varepsilon dq_\Sigma)(D_2\partial_\nu(v+\varepsilon dv))|_{\varepsilon=0}=\\&\int_\Sigma dp_\Sigma(D_1\partial_\nu u-c)+\int_\Sigma p_\Sigma(D_1\partial_\nu du-dc)+\int_\Sigma dq_\Sigma(D_2\partial_\nu q)+\int_\Sigma q_\Sigma(D_2\partial_\nu dv).\end{aligned}$$

[1] We note that very similar results can be obtained in the case where $\alpha_{TU}=\alpha_{TV}=0$ is not assumed.

We now use all of the above working to write down the first order or Karush-Kuhn-Tucker (KKT) conditions. This means that $\partial\mathcal{L}$ has to vanish for all du, dv, dp, dq, dc. Considering this gives straightforwardly that $\int_\Sigma p\partial_\nu du = \int_\Sigma du\partial_\nu p = 0$. Now, using the Fundamental Lemma of the Calculus of Variations, we obtain the following optimality system:

$$\begin{aligned} -p_t - D_1\Delta p + k_1 p + \gamma_1 pv + \gamma_2 qv + \alpha_u(u-u_Q) &= 0, && \text{in } Q,\\ -q_t - D_2\Delta q + k_2 q + \gamma_2 qu + \gamma_1 pu + \alpha_v(v-v_Q) &= 0, && \text{in } Q,\\ \alpha_c c - D_1^{-1} p &= 0, && \text{in } \Sigma,\\ u_t - D_1\Delta u + k_1 u + \gamma_1 uv &= 0, && \text{in } Q,\\ v_t - D_2\Delta v + k_2 v + \gamma_2 uv &= 0, && \text{in } Q,\\ \partial_\nu u - D_1^{-1} c &= 0, && \text{in } \Sigma,\\ \partial_\nu p = \partial_\nu q = \partial_\nu v &= 0, && \text{in } \Sigma, \end{aligned}$$

This is now a set of nonlinear equations describing the first order conditions and we can abbreviate this using the notation $\Phi(x) = 0$. We can now use Newton's method to solve this problem via the relation $\Phi'(x_k)s_k = -\Phi(x_k)$. Note we did not explicitly state the initial conditions for forward and adjoint PDEs as these carry through the above process. We now have to construct the Fréchet derivative of Φ, which we evaluate component-wise

$$\begin{aligned} &\frac{\partial}{\partial\varepsilon}\left(-(q+\varepsilon s_q)_t - D_2\Delta(q+\varepsilon s_q) + k_2(q+\varepsilon s_q)\right. \\ &\quad \left.+\gamma_2(q+\varepsilon s_q)(u+\varepsilon s_u) + \gamma_1(p+\varepsilon s_p)(u+\varepsilon s_u) + \alpha_u((u+\varepsilon s_u)-u_Q)\right)|_{\varepsilon=0} = b_1, \end{aligned} \tag{2.9}$$

which then gives

$$-(s_q)_t - D_2\Delta s_q + k_2 s_q + \gamma_2(qs_u + s_q u) + \gamma_1(ps_u + s_p u) + \alpha_u s_u = b_1.$$

We next look at

$$\begin{aligned} &\frac{\partial}{\partial\varepsilon}\left(-(p+\varepsilon s_p)_t - D_1\Delta(p+\varepsilon s_p) + k_1(p+\varepsilon s_p) + \gamma_1(p+\varepsilon s_p)(v+\varepsilon s_v)\right. \\ &\quad \left.+\gamma_2(q+\varepsilon s_q)(v+\varepsilon s_v) + \alpha_v((v+\varepsilon s_v)-v_Q)\right)|_{\varepsilon=0} = b_2, \end{aligned} \tag{2.10}$$

and obtain that

$$-(s_p)_t - D_1\Delta s_p + k_1 s_p + \gamma_1(ps_v + s_p v) + \gamma_2(qs_v + s_q v) + \alpha_v s_v = b_2.$$

Next we consider the gradient equation

$$\frac{\partial}{\partial\varepsilon}\left(\alpha_c c + \alpha_c\varepsilon s_c + p + \varepsilon s_p\right)|_{\varepsilon=0} = b_3 \tag{2.11}$$

and its easily seen that this simplifies to

$$\alpha_c s_c + s_p = b_3.$$

We are left with looking at the Newton equations for the forward reaction-diffusion PDEs

$$\frac{\partial}{\partial\varepsilon}\left((u+\varepsilon s_u)_t - D_1\Delta(u+\varepsilon s_u) + k_1(u+\varepsilon s_u) + \gamma_1(u+\varepsilon s_u)(v+\varepsilon s_v)\right)|_{\varepsilon=0} = b_4 \tag{2.12}$$

which becomes

$$(s_u)_t - D_1\Delta(s_u) + k_1(s_u) + \gamma_1(s_u v + u s_v) = b_4,$$

and the second equation

$$\frac{\partial}{\partial\varepsilon}\left((v+\varepsilon s_v)_t - D_2\Delta(v+\varepsilon s_v) + k_2(v+\varepsilon s_v) + \gamma_2(u+\varepsilon s_u)(v+\varepsilon s_v)\right)|_{\varepsilon=0} = b_5 \tag{2.13}$$

is now given by

$$(s_v)_t - D_2\Delta s_v + k_2 s_v + \gamma_2(u s_v + s_u v) = b_5.$$

Here we denoted with $\mathbf{b} = [b_1, b_2, b_3, b_4, b_5] := -\Phi(x_k)$ the right hand side of the Newton system. Note that we did not write down the boundary conditions but they of course carry through to the Newton system. If we now write everything together into an infinite dimensional system the system matrix describing the Newton process is given by

$$\begin{bmatrix} \alpha_u Id & \gamma_1 p + \gamma_2 q & 0 & \mathcal{L}'_u & \gamma_2 v \\ \gamma_2 q + \gamma_1 p & \alpha_v Id & 0 & \gamma_1 u & \mathcal{L}'_v \\ 0 & 0 & \alpha_c Id & -D_1^{-1} Id & 0 \\ \mathcal{L}_u & \gamma_1 u & -D_1^{-1} Id & 0 & 0 \\ \gamma_2 v & \mathcal{L}_v & 0 & 0 & 0 \end{bmatrix}, \tag{2.14}$$

where

$$\begin{aligned} \mathcal{L}_u &= \frac{\partial}{\partial t} - D_1\Delta + k_1 Id + \gamma_1 v, \quad \mathcal{L}'_u = -\frac{\partial}{\partial t} - D_1\Delta + k_1 Id + \gamma_1 v, \\ \mathcal{L}_v &= \frac{\partial}{\partial t} - D_2\Delta + k_2 Id + \gamma_2 u, \quad \mathcal{L}'_v = -\frac{\partial}{\partial t} - D_2\Delta + k_2 Id + \gamma_2 u, \end{aligned}$$

and Id denotes the identity operator.

In order to numerically solve the above problem we need to discretize the system (2.14) and the right hand side $-\Phi(x_k)$.

We first note that the system (2.14) is in saddle point form (as defined in Section 3) and its discrete counterpart is given by

$$\begin{bmatrix} \tau\mathcal{M}_1 & 0 & \mathcal{K}^T \\ 0 & \alpha_c\tau\mathcal{M}_c & -\tau D_1^{-1}\mathcal{N}^T \\ \mathcal{K} & -\tau D_1^{-1}\mathcal{N} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \\ \mathbf{p} \end{bmatrix} =: \mathbf{b}, \tag{2.15}$$

with

$$\begin{aligned} \mathcal{M}_1 &= \text{blkdiag}\left(M_1^{(1)}, M_1^{(2)}, \ldots, M_1^{(N_t-1)}, M_1^{(N_t)}\right), \\ \mathcal{M}_c &= \text{blkdiag}(M_c, M_c, \ldots, M_c, M_c), \\ \mathcal{N} &= \begin{bmatrix} N & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & N & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & N & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & N \end{bmatrix}, \end{aligned}$$

where

$$M_1^{(i)} = \begin{bmatrix} \alpha_u M & \gamma_1 M_{p_{(i)}} + \gamma_2 M_{q_{(i)}} \\ \gamma_1 M_{p_{(i)}} + \gamma_2 M_{q_{(i)}} & \alpha_v M \end{bmatrix},$$

M denotes a standard finite element mass matrix, M_c is a boundary mass matrix, the matrix N consists of evaluations of inner products from the term $\int_{\partial\Omega} w\mathrm{tr}(v)$ with w a function on the boundary $\partial\Omega$, v a test function for the domain Ω and tr the trace operator. The matrices $M_{p_{(i)}}$ and $M_{q_{(i)}}$ are mass-like matrices whose entries are terms of the form $\int_\Omega \bar{p}\phi_i\phi_j$ and $\int_\Omega \bar{q}\phi_i\phi_j$ respectively (where $\bar{p}$ and $\bar{q}$ represent the previous Newton iterates of the adjoint variables – or Lagrange multipliers – p and q), and the vectors $\mathbf{y}$ and $\mathbf{p}$ correspond to the discretized state $(\mathbf{u}, \mathbf{v})$ and adjoint $(\mathbf{p}, \mathbf{q})$ variables respectively. The quantity N_t denotes the number of time-steps used.

Finally the matrix $\mathcal{K}$ represents the discretized PDE, and can be written as

$$\mathcal{K} = \begin{bmatrix} L^{(1)} & & & & \\ -M_d & L^{(2)} & & & \\ & \ddots & \ddots & & \\ & & -M_d & L^{(N_t-1)} & \\ & & & -M_d & L^{(N_t)} \end{bmatrix}$$

where

$$M_d = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix},$$

and

$$L^{(i)} = \begin{bmatrix} M + \tau(D_1 K + k_1 M + \gamma_1 M_{v_{(i)}}) & \tau\gamma_1 M_{u_{(i)}} \\ \tau\gamma_2 M_{v_{(i)}} & M + \tau(D_2 K + k_2 M + \gamma_2 M_{u_{(i)}}) \end{bmatrix},$$

with K the standard finite element stiffness matrix, and $M_{u_{(i)}}$ and $M_{v_{(i)}}$ mass-like matrices with terms of the form $\int_\Omega \bar{u}\phi_i\phi_j$ and $\int_\Omega \bar{v}\phi_i\phi_j$, where $\bar{u}$ and $\bar{v}$ corresponding to the previous Newton iterates of the state variables u and v.

Note that if we write

$$\underbrace{\begin{bmatrix} \alpha_u Id & \gamma_1 p + \gamma_2 q & 0 & \mathcal{L}'_u & \gamma_2 v \\ \gamma_2 q + \gamma_1 p & \alpha_v Id & 0 & \gamma_1 u & \mathcal{L}'_v \\ 0 & 0 & \alpha_c Id & -D_1^{-1} Id & 0 \\ \mathcal{L}_u & \gamma_1 u & -D_1^{-1} Id & 0 & 0 \\ \gamma_2 v & \mathcal{L}_v & 0 & 0 & 0 \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} u^{(k+1)} - u^{(k)} \\ v^{(k+1)} - v^{(k)} \\ c^{(k+1)} - c^{(k)} \\ p^{(k+1)} - p^{(k)} \\ q^{(k+1)} - q^{(k)} \end{bmatrix} = \widetilde{\mathbf{b}}.$$

as

$$\mathcal{A} \begin{bmatrix} u^{(k+1)} \\ v^{(k+1)} \\ c^{(k+1)} \\ p^{(k+1)} \\ q^{(k+1)} \end{bmatrix} = \mathcal{A} \begin{bmatrix} u^{(k)} \\ v^{(k)} \\ c^{(k)} \\ p^{(k)} \\ q^{(k)} \end{bmatrix} + \widetilde{\mathbf{b}} = \begin{bmatrix} \alpha_u u_Q + (\gamma_1 p^{(k)} + \gamma_2 q^{(k)}) v^{(k)} \\ \alpha_v v_Q + (\gamma_2 q^{(k)} + \gamma_1 p^{(k)}) u^{(k)} \\ 0 \\ \gamma_1 u^{(k)} v^{(k)} \\ \gamma_2 v^{(k)} u^{(k)} \end{bmatrix}.$$

we can solve for the updated states, control and adjoints directly.

Problem with control constraints. The problem we have discussed so far did not include any additional constraints on the control c. We now wish to briefly discuss how point-wise constraints on the control, i.e.

$$\underline{c}(x,t) \leq c(x,t) \leq \bar{c}(x,t).$$

The treatment of control constraints can typically be dealt with using a semi-smooth Newton method introduced in [6] and for further information we refer to [26, 29, 53]. For the special case of the reaction-diffusion system we refer to [4, 20, 21, 18, 19]. In general the gradient equation of the Lagrangian becomes a variational inequality, which is in turn solved using the semi-smooth Newton method or equivalently [26] a Primal-Dual Active Set method. In contrast to [6] we employ a penalty technique that has been used very successfully for state-constraint optimization problems called the Moreau-Yosida penalty function [24, 31, 36] and has also been used to control constrained problems [51]. There the constraints

$$\underline{c}(x,t) \leq c(x,t) \leq \bar{c}(x,t).$$

are incorporated into the objective function via a penalization term, i.e. we now wish to minimize

$$J(y,u,c) + \frac{1}{2\bar{\varepsilon}} \left\|\max\left\{0, \bar{c} - c\right\}\right\|^2_{L_2(\Sigma)} + \frac{1}{2\bar{\varepsilon}} \left\|\min\left\{0, \underline{c} - c\right\}\right\|^2_{L_2(\Sigma)}$$

subject to the above mentioned state equation. We can now proceed using the semi-smooth Newton approach solving the linear systems of the form

$$\begin{bmatrix} \tau\mathcal{M}_1 & 0 & \mathcal{K}^T \\ 0 & \alpha_c\tau\mathcal{L}_c & -\tau D_1^{-1}\mathcal{N}^T \\ \mathcal{K} & -\tau D_1^{-1}\mathcal{N} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \\ \mathbf{p} \end{bmatrix} =: \widetilde{\mathbf{b}}, \tag{2.16}$$

where

$$\mathcal{L}_c = \begin{bmatrix} M_c + \bar{\varepsilon}^{-1} G_{\mathcal{A}^{(1)}} M_c G_{\mathcal{A}^{(1)}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_c + \bar{\varepsilon}^{-1} G_{\mathcal{A}^{(N_t)}} M_c G_{\mathcal{A}^{(N_t)}} \end{bmatrix}.$$

Here $\mathcal{A}^{(i)} = \mathcal{A}_+^{(i)} \cup \mathcal{A}_-^{(i)}$ defines the active sets for every time-step of the discretized problem, i.e.

$$\mathcal{A}_+^{(i)} = \{j \in \{1, 2, \ldots, N\} : (\mathbf{c}_i)_j > (\bar{c}_i)_j\} \tag{2.17}$$

$$\mathcal{A}_-^{(i)} = \{j \in \{1, 2, \ldots, N\} : (\mathbf{c}_i)_j < (\underline{c}_i)_j\} \tag{2.18}$$

using the control $\mathbf{c}$ from the previous iteration. This method is schematically shown in Algorithm 1, where we assume here that the problem is already discretized. Here

3. Solving the linear systems.

Krylov solvers. We now discuss how to efficiently solve the linear system that arises at the heart of the Lagrange-Newton method we have discussed in the previous section. We here decide to employ Krylov subspace methods which have proven to be very efficient for optimal control problems subject to PDE constraints [49, 45, 47, 46]. In our case, as the system matrix is symmetric and indefinite we could

```
Choose initial values for $\mathbf{c}^{(0)}$, $\mathbf{p}^{(0)}$, $\mathbf{q}^{(0)}$, $\mathbf{u}^{(0)}$, $\mathbf{v}^{(0)}$
Set the active sets $\mathcal{A}_+^{(0)}$, $\mathcal{A}_-^{(0)}$ and $\mathcal{A}_I^{(0)}$ by using $\mathbf{c}^{(0)}$ in (2.17), (2.18)
for $k = 1, 2, \ldots$ do
    Solve (2.16) (a system on the free variables from the previous iteration ( $\mathcal{A}_I^{(k-1)}$))
    Set the active sets $\mathcal{A}_+^{(k)}$, $\mathcal{A}_-^{(k)}$ and $\mathcal{A}_I^{(k)}$ by using $\mathbf{c}^{(k)}$ as given in (2.17), (2.18)
    if $\mathcal{A}_+^{(k)} = \mathcal{A}_+^{(k-1)}$, $\mathcal{A}_-^{(k)} = \mathcal{A}_-^{(k-1)}$, and $\mathcal{A}_I^{(k)} = \mathcal{A}_I^{(k-1)}$ then
        STOP (Algorithm converged)
    end if
end for
```

Algorithm 1: Active Set algorithm

employ the Minres [41] method introduced by Paige and Saunders. As a short term recurrence method [11] it only uses a minimal amount of storage and one matrix-vector-multiplication per iteration. Minres minimizes the 2-norm of the residual $\mathbf{r}_k = \mathbf{b} - \mathcal{A}\mathbf{x}_k$ over the current Krylov subspace where $\mathbf{x}_k$ is the approximation at step k of this procedure. Of course, any Krylov method will only be effective if a preconditioner $\mathcal{P}$ is introduced such that the properties of the left-preconditioned system

$$\mathcal{P}^{-1}\mathcal{A}\mathbf{x} = \mathcal{P}^{-1}\mathbf{b},$$

where $\mathcal{P}$ is constructed in order to resemble well the matrix $\mathcal{A}$, and also be easy to invert. For excellent introductions to the topic of constructing preconditioners for saddle point problems we refer to [5, 10] and the references mentioned therein. As a guideline for constructing good preconditioners we use a result that was presented in [38, 37], where it is shown that if the saddle point matrix

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

is invertible, then the (ideal) block preconditioner

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

where A is the unchanged $(1,1)$-block of the saddle point matrix and $S = BA^{-1}B^T$ is the (negative) *Schur complement* of $\mathcal{A}$ satisfies $\lambda(\mathcal{P}^{-1}\mathcal{A}) \left\{1, \frac{1\pm\sqrt{5}}{2}\right\}$. Therefore $\mathcal{P}$ is an extremely effective preconditioner for $\mathcal{A}$. Of course, in practice we would not wish to explicitly invert A and S to apply the ideal preconditioner – however if we construct good approximations to the $(1,1)$-block and the Schur complement of the system (2.15) an appropriate iterative solver is should converge rapidly when used with this preconditioner. As we already pointed out earlier the $(1,1)$-block of the preconditioner might be indefinite and in this case we cannot employ a symmetric Krylov subspace solver. Now faced with the decision of choosing a nonsymmetric Krylov method, we wish to point out that it is not straightforward to pick the "best method" (see [39]) and even the convergence of the Krylov subspace solver might not be adequately described by the matrix eigenvalues [17]. Nevertheless, in practice a good clustering of the eigenvalues often leads to fast convergence of the iterative scheme and it can be seen that for a good preconditioner many methods behave in a similar way.

It is also possible to employ multigrid approaches to such saddle point problems. This class of methods has previously been shown to demonstrate good performance when applied to solve a number of PDE-constrained optimization problems, subject to both steady and transient PDEs [27, 28, 52, 7, 8, 2, 23, 22, 1].

We emphasize again that the matrix systems we seek to solve fit into this saddle point framework, with

$$A = \begin{bmatrix} \tau\mathcal{M}_1 & 0 \\ 0 & \alpha_c\tau\mathcal{M}_c \end{bmatrix}, \quad B = \begin{bmatrix} \mathcal{K} & -\tau D_1^{-1}\mathcal{N} \end{bmatrix}.$$

Approximating the $(1,1)$-block. In the case of a PDE-constrained optimization problem with a linear PDE as the constraint, the $(1,1)$-block of the resulting matrix system is a block diagonal matrix containing mass matrices (see e.g. [49, 45, 47]), which can be handled very efficiently. In our case we have to take into account that the $(1,1)$-block now contains blocks of the form

$$\begin{bmatrix} \alpha_u M & \gamma_1 M_{p_i} + \gamma_2 M_{q_i} \\ \gamma_1 M_{p_i} + \gamma_2 M_{q_i} & \alpha_v M \end{bmatrix},$$

which demonstrates one of the major complexities encountered when attempting to solve such nonlinear problems numerically. When we seek to approximate these blocks, we use the saddle point theory as stated above to take as our approximation

$$A_0^{(i)} = \begin{bmatrix} \alpha_u M - \alpha_v^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right) M^{-1} \alpha_v^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right) & 0 \\ 0 & \alpha_v M \end{bmatrix}.$$

Note that these complicated looking matrices are actually straightforward to handle as we assume that the mass matrices are lumped here[2]. The block $\mathcal{M}_c$, which also forms part of the $(1,1)$-block of our matrix systems, may be approximated using Chebyshev semi-iteration [15, 16, 54] for consistent mass matrices, or by simply inversion for lumped mass matrices.

Approximating the Schur complement. We now focus on the approximation of the Schur complement, which is given by

$$S = \tau^{-1}\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T + \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T.$$

One approach that proved successful for moderate values of the parameter α_c is to use the approximation

$$\widehat{S} = \tau^{-1}\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T \tag{3.1}$$

(see [45]) for smaller values this approximation did not provide satisfying results. Hence, approximations that provide robustness with respect to the crucial problem parameters have been investigated (see [49, 55, 34, 44, 42]). The idea presented in [44] uses an approximation

$$\widehat{S} = \tau^{-1}(\mathcal{K} + \widehat{\mathcal{M}})\mathcal{M}_1^{-1}(\mathcal{K} + \widehat{\mathcal{M}})^T$$

[2] In the case where mass matrices are not lumped, we believe that we may take a similar approximation, but replace the mass matrices by their diagonals within the preconditioner.

where $\widehat{\mathcal{M}}$ is chosen to accommodate a better approximation of the term that was initially dropped from S. Before we go into the details we wish to state that we use the approximation for the $(1,1)$-block within the Schur complement approximation, i.e.

$$S \approx \tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T.$$

We will use this as the basis for our Schur complement approximation. Starting with the assumption that we wish our approximation to look like

$$\widehat{S} = \tau^{-1}(\mathcal{K}+\widehat{\mathcal{M}})\widehat{\mathcal{M}}_1^{-1}(\mathcal{K}+\widehat{\mathcal{M}})^T$$

where we have to determine $\widehat{\mathcal{M}}$. Studying $\widehat{S}$ more closely reveals that

$$\widehat{S} = \tau^{-1}\left(\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} + \widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}\right)$$

and we wish the first two terms in $\widehat{S}$ to match the Schur complement as closely as possible. Therefore, we wish that

$$\tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T \approx \tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}.$$

We now recall the block structure (assuming $\widehat{\mathcal{M}}$ to be block diagonal) of the matrices to see that the last equation gives

$$\begin{bmatrix} \tau\alpha_c^{-1}D_1^{-2}NM_c^{-1}N^T & 0 \\ 0 & 0 \end{bmatrix} \approx \begin{bmatrix} \tau^{-1}\widehat{M}_1 A_0^{-(i)}\widehat{M}_1 & 0 \\ 0 & \tau^{-1}\alpha_v^{-1}\widehat{M}_2 M^{-1}\widehat{M}_2 \end{bmatrix}.$$

where $A_0^{-(i)} := \left(A_0^{(i)}\right)^{-1}$. We will set $\widehat{M}_2$ to zero and choose the entries of $\widehat{M}_1$ such that $\tau\alpha_c^{-1}NM_c^{-1}N^T = \tau^{-1}\widehat{M}_1 A_0^{-(i)}\widehat{M}_1$, i.e. the diagonal elements of $\widehat{M}_1$ assuming for now that M_c is a mass matrix for the whole domain to give

$$\tau\alpha_c^{-1}D_1^{-2}m_{jj} = \tau^{-1}\frac{\widehat{m}_{jj}^2}{a_{0i,jj}}$$

or equivalently

$$\widehat{m}_{jj}^2 = \tau^2\alpha_c^{-1}D_1^{-2}a_{0i,jj}m_{jj} \Rightarrow \widehat{m}_{jj} = \tau\sqrt{\alpha_c^{-1}D_1^{-2}}\sqrt{a_{0i,jj}m_{jj}}. \qquad (3.2)$$

Note that so far we have ignored that M_c is a boundary mass matrix that scales differently to a mass matrix on the whole domain by an order of h. We illustrate how to deal with this on a simple example when we wish the following to hold

$$\widehat{M}M^{-1}\widehat{M} = NM_b^{-1}N^T,$$

where M is the mass matrix on the whole domain and M_b on the boundary. Using the approximations $M \approx h^2 I$ and $M \approx hI$ we get

$$h^{-2}\widehat{M}^2 \approx \widehat{M}M^{-1}\widehat{M} = NM_b^{-1}N^T \approx h^{-1}NN^T.$$

As all matrices in the last expression are diagonal (N is a rectangular matrix with entries only when boundary degree of freedom is paired with boundary degree of freedom) we get

$$\widehat{m}_{jj}^2 = hm_{jj}^2 \Rightarrow \widehat{m}_{jj} = \sqrt{h}m_{jj}.$$

We can now incorporate the different orders of scaling into (3.2) using $\sqrt{h}$ to give

$$\widehat{m}_{jj} = \tau\sqrt{\frac{ha_{0i,jj}m_{jj}}{D_1^2\alpha_c}}.$$

The approach we presented so far was relying on the fact that the matrix

$$A_0^{(i)} = \begin{bmatrix} \alpha_u M - \alpha_v^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right) M^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right) & 0 \\ 0 & \alpha_v M \end{bmatrix}$$

is positive definite, which is in general not satisfied. In this case even though the system matrix is symmetric we could chose a nonsymmetric solver using a nonsymmetric preconditioner, which contains blocks of the form $A_0^{(i)}$ that are indefinite. We again only formulate our results for a block diagonal preconditioner even though a block triangular preconditioner may also be appropriate. It is straightforward to use our techniques for block triangular preconditioners and we will show results using them in Section 4. We can therefore approximate the Schur complement using

$$\widehat{S} = \tau^{-1}(\mathcal{K} + \widehat{\mathcal{M}}_1)\mathcal{M}_1^{-1}(\mathcal{K}^T + \widehat{\mathcal{M}}_2)$$

where in general $\widehat{\mathcal{M}}_1$ is not equal to $\widehat{\mathcal{M}}_2$. In a similar fashion to the above we get that

$$\tau\alpha_c^{-1}D_1^{-2}m_{jj} = \tau^{-1}\frac{\widehat{m}_{1,jj}\widehat{m}_{2,jj}}{a_{0i,jj}},$$

which leads to

$$\tau^2 a_{0i,jj}\alpha_c^{-1}D_1^{-2}m_{jj} = \widehat{m}_{1,jj}\widehat{m}_{2,jj}.$$

If we also incorporate the difference in scalings for the boundary mass matrix we have

$$h\tau^2 a_{0i,jj}\alpha_c^{-1}D_1^{-2}m_{jj} = \widehat{m}_{1,jj}\widehat{m}_{2,jj}$$

and we can now choose $\widehat{m}_{1,jj}$ and $\widehat{m}_{2,jj}$ in a balanced fashion to give

$$\widehat{m}_{1,jj} = \tau\sqrt{\frac{h}{\alpha_c}}D_1^{-1}a_{0i,jj} \text{ and } \widehat{m}_{2,jj} = \tau\sqrt{\frac{h}{\alpha_c}}D_1^{-1}m_{jj} \tag{3.3}$$

or the following, which we found to work very well in practice:

$$\widehat{m}_{1,jj} = \tau\sqrt{\frac{h}{\alpha_c}}D_1^{-1}\sqrt{m_{jj}}\sqrt{|a_{0i,jj}|} \text{ and } \widehat{m}_{2,jj} = \tau\sqrt{\frac{h}{\alpha_c}}D_1^{-1}\sqrt{m_{jj}}\sqrt{|a_{0i,jj}|}. \tag{3.4}$$

Preconditioning for control constraints. The system (2.16) again needs to be preconditioned effectively. The $(1,1)$-block now contains the matrix $\alpha_c\tau\mathcal{L}_c$, which is a simple block-diagonal matrix that can be treated in the same way as the $(1,1)$-block of the problem without control constraints. Approximating the Schur complement

$$S = \tau^{-1}\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T + \frac{\tau}{\alpha_c D_1^2}\mathcal{N}\mathcal{L}_c^{-1}\mathcal{N}^T$$

is again the more challenging task. We now wish to use the technique employed earlier, i.e.,

$$\widehat{S} = \tau^{-1}(\mathcal{K} + \widehat{\mathcal{M}})\widehat{\mathcal{M}}_1^{-1}(\mathcal{K} + \widehat{\mathcal{M}})^T$$

where $\widehat{\mathcal{M}}_1$ approximates the $(1,1)$-block. We again wish that

$$\tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} = \frac{\tau}{\alpha_c D_1^2}\mathcal{N}\mathcal{L}_c^{-1}\mathcal{N}^T.$$

Note that $\frac{\tau}{\alpha_c D_1^2}\mathcal{N}\mathcal{L}_c^{-1}\mathcal{N}^T$ is a block-diagonal matrix with a block of the form

$$\frac{\tau}{\alpha_c D_1^2} N\left(M_c + \bar{\varepsilon}^{-1} G_{\mathcal{A}^{(i)}} M_c G_{\mathcal{A}^{(i)}}\right)^{-1} N^T$$

alternating with a zero block. Hence, we see that $\widehat{M}$ has also an alternating block-diagonal structure. Using the notation $l_{ci,jj}$ for the diagonal entries of $\left(M_c + \bar{\varepsilon}^{-1} G_{\mathcal{A}^{(i)}} M_c G_{\mathcal{A}^{(i)}}\right)$, the special structure of N (N only has entries on the diagonal belonging to boundary degrees of freedom) and that all mass matrices are lumped, It holds for all nonzero entries in $\widehat{\mathcal{M}}$, as only entries corresponding to the boundary degrees of freedom will be nonzero,

$$\tau^{-1}\widehat{m}_{i,jj} a_{0i,jj}^{-1} \widehat{m}_{i,jj} = \frac{\tau}{\alpha_c D_1^2} m_{c,jj} l_{ci,jj}^{-1} m_{c,jj},$$

where the index i always refers to the entries corresponding to the i-th grid point in time. This gives

$$\widehat{m}_{i,jj} = \sqrt{\frac{\tau^2}{\alpha_c D_1^2} a_{0i,jj} l_{ci,jj}^{-1} m_{c,jj}^2}$$

with

$$\widehat{m}_{1i,jj} = \left(\alpha_u M - \alpha_v^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right) M^{-1}\left(\gamma_1 M_{p_i} + \gamma_2 M_{q_i}\right)\right)_{jj}$$

and

$$l_{ci,jj} = \left(M_c + \bar{\varepsilon}^{-1} G_{\mathcal{A}^{(i)}} M_c G_{\mathcal{A}^{(i)}}\right)_{jj}.$$

3.1. Eigenvalue analysis. In this section, we aim to provide guidance as to how to establish the effectiveness of our Schur complement approximations stated above, by analyzing what we expect the behaviour of the eigenvalues of $\widehat{S}_1^{-1}S$ and $\widehat{S}_2^{-1}S$ to be. Due to the complexity of the underlying problems and the linear algebraic issues involved, we make a few simplifying assumptions for our analysis – the resulting eigenvalue estimates should thus be regarded as heuristic guidance, rather than rigorous proof.

In this section, we consider the preconditioned Schur complements in the case without further control constraints (we find that in practice, the control-constrained case results in very similar eigenvalue spectra). One simplifying assumption that we make throughout in our analysis is to ignore the single zero eigenvalue of K, that is to consider the matrix K corresponding to a Dirichlet problem rather than a Neumann problem – we find that this makes little difference in practice, but the presence of a zero eigenvalue of K would make our analysis much harder to proceed with. Finally, we consider only the case where the $(1,1)$-block of the matrix system is positive definite as we believe this is the only case for which analyzing this problem in detail would be feasible using our methodology – we find that our solvers may work well if this is not the case, but not as effectively as in the positive definite case.

When analyzing this problem, we consider only the more physically complex 3D problem; results for the 2D problem are similar.

To begin our analysis, we make use of Theorem 2 of [12], in which Feingold and Varga establish Gershgorin-type theorems for block matrices:

Theorem 1. *For the partitioned matrix*

$$\Lambda = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & & \vdots & \\ A_{N1} & A_{N2} & \dots & A_{NN} \end{bmatrix},$$

where A_{jj}, $j = 1, ..., N$ are square, each eigenvalue λ of Λ satisfies

$$\left(\left\|(A_{jj} - \lambda I_j)^{-1}\right\|\right)^{-1} \leq \sum_{k=1, k\neq j}^{n} \|A_{jk}\|$$

for at least one j, $1 \leq j \leq N$.

We may use this to state the following Lemma (for the 3D problem)[3]:

Lemma 1. *The eigenvalues of $L^{(j)}$ (and hence $(L^{(j)})^T$), $j = 1, ..., N_t$, are within sets of the type (excluding multiplicative constants of $\mathcal{O}(1)$)*

$$\left\{\left|\lambda - \left(h^3 + \tau(D_1\eta + k_1h^3 + \gamma_1h^3\bar{v})\right)\right| \leq \tau\gamma_1h^3\bar{u}\right\} \cup \\ \left\{\left|\lambda - \left(h^3 + \tau(D_2\eta + k_2h^3 + \gamma_2h^3\bar{u})\right)\right| \leq \tau\gamma_2h^3\bar{v}\right\},$$

where h denotes the mesh-size used, and η denotes an eigenvalue of K (which is in the range $[h^3, h]$, using our simplifying assumption). Therefore, for each eigenvalue λ of $L^{(j)}$:

$$\begin{aligned} & h^3 + \tau\left(D_1\eta + k_1h^3 + \gamma_1h^3(\bar{v} - \bar{u})\right) \leq \mathrm{Re}(\lambda) \leq h^3 + \tau\left(D_1\eta + k_1h^3 + \gamma_1h^3(\bar{u} + \bar{v})\right) \\ \text{or} \quad & h^3 + \tau\left(D_2\eta + k_2h^3 + \gamma_2h^3(\bar{u} - \bar{v})\right) \leq \mathrm{Re}(\lambda) \leq h^3 + \tau\left(D_2\eta + k_2h^3 + \gamma_2h^3(\bar{u} + \bar{v})\right). \end{aligned}$$

Furthermore, we may prove a similar result for the matrix $\mathcal{M}_1$, as follows:

Lemma 2. *The eigenvalues of the j-th block of $\mathcal{M}_1$, that is (the symmetric matrix) $\begin{bmatrix} \alpha_uM & \gamma_1M_{p_{(j)}} + \gamma_2M_{q_{(j)}} \\ \gamma_1M_{p_{(j)}} + \gamma_2M_{q_{(j)}} & \alpha_vM \end{bmatrix}$, for $j = 1, ..., N_t$, are within sets of the type (excluding multiplicative constants of $\mathcal{O}(1)$)*

$$\left\{\left|\lambda - \alpha_uh^3\right| \leq \gamma_1h^3\bar{p} + \gamma_2h^3\bar{q}\right\} \cup \left\{\left|\lambda - \alpha_vh^3\right| \leq \gamma_1h^3\bar{p} + \gamma_2h^3\bar{q}\right\}.$$

Therefore, for each eigenvalue λ of $\begin{bmatrix} \alpha_uM & \gamma_1M_{p_{(j)}} + \gamma_2M_{q_{(j)}} \\ \gamma_1M_{p_{(j)}} + \gamma_2M_{q_{(j)}} & \alpha_vM \end{bmatrix}$ (which are all real due to symmetry of the relevant matrices):

$$\begin{aligned} & (\alpha_u - \gamma_1\bar{p} - \gamma_2\bar{q})h^3 \leq \lambda \leq (\alpha_u + \gamma_1\bar{p} + \gamma_2\bar{q})h^3 \\ \text{or} \quad & (\alpha_v - \gamma_1\bar{p} - \gamma_2\bar{q})h^3 \leq \lambda \leq (\alpha_v + \gamma_1\bar{p} + \gamma_2\bar{q})h^3, \end{aligned}$$

[3] We use here that the eigenvalues of M are given by h^3 up to constants of $\mathcal{O}(1)$, and that the eigenvalues of $M_{u_{(j)}}$ and $M_{v_{(j)}}$ are of the form $\bar{u}h^3$ and $\bar{v}h^3$ up to constants of $\mathcal{O}(1)$, with $\bar{u}$ and $\bar{v}$ representing the values of the most recent Newton iterates of u and v.

and so if $\mathcal{M}_1$ is positive definite, its largest eigenvalue is

$$\lambda_{\max} = (\max\{\alpha_u, \alpha_v\} + \gamma_1 \bar{p} + \gamma_2 \bar{q})h^3.$$

Eigenvalues of $\widehat{S}_1^{-1}S$. Lemmas 1 and 2 lead to the following statement about the eigenvalues of $\widehat{S}_1^{-1}S$, where

$$S = \frac{1}{\tau}\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T + \frac{\tau}{\alpha_c D_1^2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T, \quad \widehat{S}_1 = \frac{1}{\tau}\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T,$$

in the case where $\mathcal{M}_1$ is positive definite.

LEMMA 3. *Suppose $\mathcal{M}_1$ is positive definite. Then the eigenvalues of $\widehat{S}_1^{-1}S$ are contained within the following interval:*

$$\lambda(\widehat{S}_1^{-1}S) \in [1, 1+\mu],$$

where

$$\mu = \frac{C_1 h^5(\max\{\alpha_u, \alpha_v\} + \gamma_1 \bar{p} + \gamma_2 \bar{q})}{\alpha_c D_1^2 \min\left\{(D_1\eta + k_1 h^3 + \gamma_1 h^3(\bar{v} - \bar{u})), (D_2\eta + k_2 h^3 + \gamma_2 h^3(\bar{u} - \bar{v}))\right\}^2},$$

and C_1 is a constant of $\mathcal{O}(1)$.

Proof. As both S and $\widehat{S}_1$ are symmetric matrices, we may prove the result using a Rayleigh quotient argument. We write that

$$\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1\mathbf{v}} = \frac{\frac{1}{\tau}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v} + \frac{\tau}{\alpha_c D_1^2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}{\frac{1}{\tau}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}} = 1 + \frac{\tau^2}{\alpha_c D_1^2}\frac{\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}.$$

This quantity is clearly bounded below by 1, as the numerator of the second term is non-negative and the denominator positive.[4] For the upper bound, we need to consider the upper bound of $\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}$ and the lower bound of $\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}$.

In the (frequently occurring) case $\mathbf{v} \in \text{null}(N^T)$, the term $\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}$ is equal to zero, and we are done – $\widehat{S}_1$ is the exact Schur complement in this case. In the case $\mathbf{v} \notin \text{null}(N^T)$, the term $\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}$ is bounded above by h^2 in the 3D case, excluding multiplicative constants of $\mathcal{O}(1)$.

The lower bound of $\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}$ will correspond to the smallest real eigenvalue(s) of $\mathcal{K}$ (and $\mathcal{K}^T$) and the largest eigenvalue(s) of $\mathcal{M}_1$. For this, we may use Lemmas 1 and 2 to give that

$$\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v} \geq \frac{\left(\tau \min\left\{\left(D_1\eta + k_1 h^3 + \gamma_1 h^3(\bar{v} - \bar{u})\right), \left(D_2\eta + k_2 h^3 + \gamma_2 h^3(\bar{u} - \bar{v})\right)\right\}\right)^2}{(\max\{\alpha_u, \alpha_v\} + \gamma_1 \bar{p} + \gamma_2 \bar{q})h^3}.$$

Therefore, inserting our bounds for $\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}$ and $\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}$ into our expression for $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1\mathbf{v}}$ gives (reintroducing our multiplicative constant of $\mathcal{O}(1)$)

$$\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T \widehat{S}_1\mathbf{v}} \leq 1 + \frac{h^5(\max\{\alpha_u, \alpha_v\} + \gamma_1 \bar{p} + \gamma_2 \bar{q})}{\alpha_c D_1^2 \min\left\{(D_1\eta + k_1 h^3 + \gamma_1 h^3(\bar{v} - \bar{u})), (D_2\eta + k_2 h^3 + \gamma_2 h^3(\bar{u} - \bar{v}))\right\}^2},$$

[4] Once again, we do not take account of the zero eigenvalue of K; to deal with this issue (and ensure that $\widehat{S}_1$ invertible) we recommend introducing an artificial Dirichlet boundary condition within the preconditioner.

and so the result is proved. □

We are able to bound the minimum of $\lambda(\widehat{S}^{-1}S)$ robustly in this case. We are naturally interested in the limits of the upper bound as $h \to 0$ and $\tau \to 0$, as we wish the preconditioner to behave as robustly as possible as the matrix system grows in size. Clearly, the upper bound is independent of the parameter τ. As $h \to 0$, the upper bound may tend to a term proportional to h^{-1} in the worst case (corresponding to a value of η of $\mathcal{O}(h^3)$). So a Schur complement approximation $\widehat{S}_1^{-1}S$ will not generate a totally robust solver for h (and certainly not for many of the other parameters), but as the value of η can vary within the interval $[h^3, h]$, we observe that dependence on h is reasonably mild in practice. However, we believe that the choice $\widehat{S}_1$ of Schur complement approximation is not the optimal one.

Eigenvalues of $\widehat{S}_2^{-1}S$. We now wish to demonstrate a bound for $\widehat{S}_2^{-1}S$, where:

$$\widehat{S}_2 = \frac{1}{\tau}(\mathcal{K} + \widehat{\mathcal{M}})\widehat{\mathcal{M}}_1^{-1}(\mathcal{K} + \widehat{\mathcal{M}})^T,$$

and $\widehat{\mathcal{M}}$ is such that $\tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} = \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T$, in the case where $\mathcal{M}_1$ and $\widehat{\mathcal{M}}_1$ are positive definite. In order to do this, we make use of the following Lemma:

LEMMA 4. *The smallest eigenvalues of the matrix*

$$\mathcal{K}\begin{bmatrix} \omega_1\bar{I} & & & & \\ & \omega_2\bar{I} & & & \\ & & \ddots & & \\ & & & \omega_{N_t-1}\bar{I} & \\ & & & & \omega_{N_t}\bar{I} \end{bmatrix} + \begin{bmatrix} \omega_1\bar{I} & & & & \\ & \omega_2\bar{I} & & & \\ & & \ddots & & \\ & & & \omega_{N_t-1}\bar{I} & \\ & & & & \omega_{N_t}\bar{I} \end{bmatrix}\mathcal{K}^T,$$

where $\bar{I} = blkdiag(I, 0)$ and $\omega_j > 0$, $j = 1, ..., N_t - 1$, in the case where K has no zero eigenvalue, are bounded below by values of the following order:

$$-\tau h^3 \max_{j=1,\dots,N_t}\{\omega_j\gamma_2\bar{v}_j\},$$

where $\bar{v}_j$ corresponds to the most recent Newton iterates of v at the j-th time-step

Proof. Expanding out the above matrix gives the following:

$$\underbrace{\begin{bmatrix} \omega_1\Pi_1 & -\omega_1 M_d & & & \\ -\omega_1 M_d & \omega_2\Pi_2 & -\omega_2 M_d & & \\ & -\omega_2 M_d & \ddots & \ddots & \\ & & \ddots & \ddots & -\omega_{N_t-1}M_d \\ & & & -\omega_{N_t-1}M_d & \omega_{N_t}\Pi_{N_t} \end{bmatrix}}_{\bar{\mathcal{K}}},$$

where

$$\Pi_j = \begin{bmatrix} 2M + \tau N_{(j)} + \tau N_{(j)}^T & \tau\gamma_2 M_{v_{(j)}} \\ \tau\gamma_2 M_{v_{(j)}} & 0 \end{bmatrix}$$

and $N_{(j)} = D_1 K + k_1 M + \gamma_1 M_{v_{(j)}}$ for $j = 1, ..., N_t$.

We may now work with the Rayleigh quotient $\mathbf{w}^T\bar{\mathcal{K}}\mathbf{w}$, where we denote that $\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_{N_t} \end{bmatrix}^T$, with $\mathbf{w}_j$ vectors of the same dimension as the matrices

Π_j and M_d. By expanding out the terms, similarly to the working carried out to prove Theorem 1 in [42], we then obtain

$$\begin{aligned}
\mathbf{w}^T\bar{\mathcal{K}}\mathbf{w} = {} & \sum_{j=1}^{N_t} 2\omega_j \mathbf{w}_j^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \mathbf{w}_j + \sum_{j=1}^{N_t} \omega_j \mathbf{w}_j^T \Pi_j \mathbf{w}_j \\
& \quad - \sum_{j=1}^{N_t-1} \omega_j \mathbf{w}_j^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \mathbf{w}_{j+1} - \sum_{j=2}^{N_t} \omega_{j-1} \mathbf{w}_j^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \mathbf{w}_{j-1} \\
= {} & \sum_{j=1}^{N_t-1} \omega_j (\mathbf{w}_j - \mathbf{w}_{j+1})^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} (\mathbf{w}_j - \mathbf{w}_{j+1}) \\
& \quad + \omega_1 \mathbf{w}_1^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \mathbf{w}_1 + \omega_{N_t} \mathbf{w}_{N_t}^T \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \mathbf{w}_{N_t} \\
& \quad + \sum_{j=1}^{N_t} \omega_j \mathbf{w}_j^T \Pi_j \mathbf{w}_j \\
\geq {} & \mathbf{w}^T \mathrm{blkdiag}\left(\omega_1 \Pi_1, ..., \omega_{N_t} \Pi_{N_t}\right) \mathbf{w}
\end{aligned}$$

using the positive semi-definiteness of the matrix M_d. Therefore, the minimum possible value of $\mathbf{w}^T\bar{\mathcal{K}}\mathbf{w}$ is certainly bounded above by the largest negative value of the block diagonal matrix written above. This will be given by an eigenvalue of the form (using Theorem 1)

$$-\omega_j \tau \gamma_2 \bar{v}_j h^3,$$

for some $j = 1, ..., N_t$, where $\bar{v}_j$ is as defined above. The result follows directly from this bound. □

We will use this result to demonstrate a heuristic bound for the eigenvalues of $\widehat{S}_2^{-1} S$. As we are again dealing with symmetric matrices, we seek to use a Rayleigh quotient argument to do this. We consider the quantity

$$\begin{aligned}
\frac{\mathbf{v}^T S \mathbf{v}}{\mathbf{v}^T \widehat{S}_2 \mathbf{v}} &= \frac{\tau^{-1}\mathbf{v}^T \mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v} + \tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T \mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v} + \tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v} + 2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}\mathbf{v}} \\
&= \frac{1}{\frac{\tau^{-1}\mathbf{v}^T \mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v} + \tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T \mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v} + \tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}} + \frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T \mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v} + \tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}}.
\end{aligned}$$

We observe that the term $\frac{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}$ is likely to be an important one, so we seek to analyses it in more depth. We see that the term corresponds to the spectrum of $(\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T)^{-1}(\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T)$, and therefore, by a simple Rayleigh quotient argument, the spectrum of $\widehat{\mathcal{M}}_1^{-1}\mathcal{M}_1$. Now, by the way $\widehat{\mathcal{M}}_1$ is constructed, we know that $\lambda(\widehat{\mathcal{M}}_1^{-1}\mathcal{M}_1) \in \left\{1, \frac{1\pm\sqrt{5}}{2}\right\}$. Now, as $\mathcal{M}_1$ and $\widehat{\mathcal{M}}_1$ are positive definite (by assumption), we may apply Sylvester's law of inertia (see [30]) to deduce that, in this case, $\lambda(\widehat{\mathcal{M}}_1^{-1}\mathcal{M}_1) \in \left\{1, \frac{1+\sqrt{5}}{2}\right\}$, and therefore that $\frac{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}} \in \left[1, \frac{1+\sqrt{5}}{2}\right]$.

We may now demonstrate a lower bound for $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}}$ as follows:

$$
\begin{aligned}
\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}} &\geq \frac{1}{\frac{1+\sqrt{5}}{2}+\max\left(\frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}\right)} \\
&= \frac{1}{\frac{1+\sqrt{5}}{2}+\max\left(\frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\frac{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}\right)} \\
&\geq \frac{1}{\frac{1+\sqrt{5}}{2}+\max\left(\frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\min\left(\frac{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}\right)\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}\right)}.
\end{aligned}
$$

At this point we note that $\frac{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}$ can be bounded below by $\frac{2}{1+\sqrt{5}}$ (by previous working), so we may write

$$
\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}} \geq \frac{1}{\frac{1+\sqrt{5}}{2}+\frac{1+\sqrt{5}}{2}\max\left(\frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}\right)}.
$$

The quantity $\frac{2\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}$ may be written as $\frac{2\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}+\mathbf{b}^T\mathbf{b}}$, where $\mathbf{a}=\tau^{-1/2}\widehat{\mathcal{M}}_1^{-1/2}\mathcal{K}^T\mathbf{v}$ and $\mathbf{b}=\tau^{1/2}\alpha_c^{-1/2}D_1^{-1}\mathcal{M}_c^{-1/2}\mathcal{N}^T\mathbf{v}$. As $\mathbf{a}^T\mathbf{a}>0$ (by assumption of positive definiteness of $\widehat{\mathcal{M}}_1$, and again ignoring the zero eigenvalue of K), we may use the same trick as in our previous work (see [43, 42]) to bound this above by 1. Therefore, putting all the pieces together we may write that

$$
\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}} \geq \frac{1}{1+\sqrt{5}},
$$

excluding multiplicative constants of $\mathcal{O}(1)$.

For the upper bound of $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}}$, we may now write

$$
\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\widehat{S}_2\mathbf{v}} \leq \frac{1}{1+\min\left(\frac{\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}\mathbf{v}+\tau^{-1}\mathbf{v}^T\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}+\tau\alpha_c^{-1}D_1^{-2}\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v}}\right)}, \tag{3.5}
$$

for which we need to establish the smallest eigenvalue (that is to say largest negative eigenvalue) of

$$
\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T+\tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\right)^{-1}\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}+\tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\right).
$$

We may do this by considering the Rayleigh quotient

$$
\frac{\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}+\tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\right)\mathbf{v}}{\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T+\tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\right)\mathbf{v}}.
$$

Now, using the fact that $\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} = \mathrm{blkdiag}(M_{pq,1}, 0, ..., M_{pq,2}, 0, ..., M_{pq,N_t}, 0)$, where $M_{pq,j} := \alpha_u I - \alpha_v^{-1}(\gamma_1 M_{p_j} + \gamma_2 M_{q_j})^{-1} M (\gamma_1 M_{p_j} + \gamma_2 M_{q_j})^{-1} M$, we may assume (by arguing based on the assumption that the $(1,1)$-block is positive definite and hence that the $\alpha_u I$ term of $\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}}$ is dominant) that this is of the form $\mathrm{blkdiag}(\omega_1 \bar{I}, ..., \omega_{N_t}\bar{I})$ of Lemma 4, with $\omega_j = \alpha_u - \alpha_v(\gamma_1 \bar{p}_j + \gamma_2 \bar{q}_j)^{-2}$, (and $\omega_j > 0$ by the assumption of positive definiteness of the $(1,1)$-block. We may therefore apply Lemma 4 to write that the $\mathbf{v}^T \left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} + \tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\right)\mathbf{v}$ term is bounded below by

$$-h^3 \max_{j=1,\dots,N_t} \{\omega_j \gamma_2 \bar{v}_j\} = -h^3 \max_{j=1,\dots,N_t} \left\{\left(\alpha_u - \alpha_v(\gamma_1 \bar{p}_j + \gamma_2 \bar{q}_j)^{-2}\right)\gamma_2 \bar{v}_j\right\}.$$

We now aim to find the smallest (positive) value of the denominator of the Rayleigh quotient, that is the $\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\right)\mathbf{v}$ term. This is of course at least $\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}$, which we observe

$$\begin{aligned}\tau^{-1}\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v} &= \tau^{-1}\frac{\mathbf{v}^T\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\mathbf{v}}{\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}\\ &\geq \tau^{-1}\lambda_{\min}\left(\widehat{\mathcal{M}}_1^{-1}\mathcal{M}_1\right)\cdot\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}\\ &= \tau^{-1}\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v},\end{aligned}$$

where we now have considered a Rayleigh quotient relating to the eigenvalues of $\widehat{\mathcal{M}}_1^{-1}\mathcal{M}_1$. We may use that in Lemma 3 we showed

$$\begin{aligned}&\mathbf{v}^T\mathcal{K}\mathcal{M}_1^{-1}\mathcal{K}^T\mathbf{v}\\ &\quad\geq \frac{\left(h^3 + \tau\min\left\{\left(D_1\eta + k_1h^3 + \gamma_1h^3(\bar{v}-\bar{u})\right), \left(D_2\eta + k_2h^3 + \gamma_2h^3(\bar{u}-\bar{v})\right)\right\}\right)^2}{(\max\{\alpha_u,\alpha_v\} + \gamma_1\bar{p} + \gamma_2\bar{q})h^3},\end{aligned}$$

and hence we can write that

$$\begin{aligned}&\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\right)\mathbf{v}\\ &\quad\geq \frac{\left(h^3 + \tau\min\left\{\left(D_1\eta + k_1h^3 + \gamma_1h^3(\bar{v}-\bar{u})\right), \left(D_2\eta + k_2h^3 + \gamma_2h^3(\bar{u}-\bar{v})\right)\right\}\right)^2}{\tau(\max\{\alpha_u,\alpha_v\} + \gamma_1\bar{p} + \gamma_2\bar{q})h^3}.\end{aligned}$$

where we have used that $\mathbf{v}^T\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\mathbf{v} \geq 0$.

Hence:

$$\begin{aligned}&\frac{\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\widehat{\mathcal{M}} + \tau^{-1}\widehat{\mathcal{M}}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T\right)\mathbf{v}}{\mathbf{v}^T\left(\tau^{-1}\mathcal{K}\widehat{\mathcal{M}}_1^{-1}\mathcal{K}^T + \tau\alpha_c^{-1}D_1^{-2}\mathcal{N}\mathcal{M}_c^{-1}\mathcal{N}^T\right)\mathbf{v}}\\ &\quad\geq -\frac{\tau h^6\gamma_2\max_{j=1,\dots,N_t}\{\omega_j\bar{v}_j(\max\{\alpha_u,\alpha_v\} + \gamma_1\bar{p}_j + \gamma_2\bar{q}_j)\}}{\left(h^3 + \tau\min\left\{\left(D_1\eta + k_1h^3 + \gamma_1h^3(\bar{v}_j-\bar{u}_j)\right), \left(D_2\eta + k_2h^3 + \gamma_2h^3(\bar{u}_j-\bar{v}_j)\right)\right\}\right)^2}.\end{aligned}$$

Examining this expression shows that the Rayleigh quotient is bounded above by a parameter which should behave uniformly as h decreases, and will not increase in magnitude as τ decreases. Therefore, the size of this Rayleigh quotient will not worsen as the problem size becomes larger, either from step-size or time-step decreasing.

Putting all our working together, we have demonstrated heuristically that

$$\lambda(\widehat{S}_2^{-1}S) \in \left[c_{1,1}, \frac{1}{1-C_{1,\tau}}\right],$$

where $c_{1,1}$ is a positive constant which is robust with respect to h and τ, and $C_{1,\tau} \leq 1$ is a parameter[5] which is robust as $h \to 0$ and gets smaller as $\tau \to 0$.

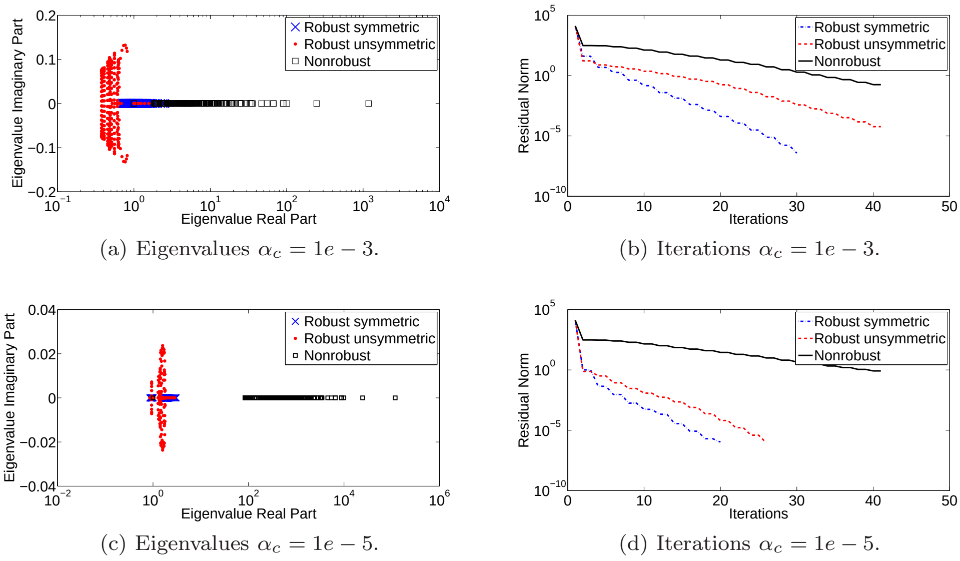
The above result indicates that as the problem size gets larger (as either h or τ gets smaller), the eigenvalue spectrum of our approximation $\widehat{S}_2$ to the (positive definite) Schur complement should not worsen – this is the property we desire.

We note that proving similar results as those above for the case where $\mathcal{M}_1$ is indefinite would be a much harder task. However, we observe that in practice, our strategy of "matching" the first and last terms of the Schur complement by choosing a suitable approximation of the form $(\mathcal{K}+\widehat{\mathcal{M}})\widehat{\mathcal{M}}_1^{-1}(\mathcal{K}+\widehat{\mathcal{M}})^T$ often "captures" the dominant terms as h and τ decreases (the first term) and α_c and D_1 decreases (the second term).

To motivate our choices for the next section even further we wish to illustrate the eigenvalue distribution coming from our Schur complement approximation. In Figures 3.1 and 3.2 we show the eigenvalue distribution of $\hat{S}^{-1}S$ (left) for two values of the regularization parameter α_c and for each Schur complement approximation. The right plots in Figures 3.1 and 3.2 show the performance of GMRES using these approximations for the preconditioner. The term robust symmetric refers to the approximation (3.4), robust unsymmetric to (3.3) and non-robust to (3.1). It can be seen from the results that the approximation (3.4) provides the best eigenvalue distribution and lowest number of iterations. This reflects our experience for larger test problems such as the ones provided in the next section. Thus, we will use this approximation for the remainder of the paper.

4. Numerical Experiments. In this section we present numerical results for the above mentioned algorithm with Schur complement approximation $\hat{S}_2$. We have implemented this methodology using the finite element package deal.II [3] with $Q1$ finite elements. For the AMG preconditioner, we used the Trilinos ML package [14] that implements a smoothed aggregation AMG. Within the algebraic multigrid we typically used a Chebyshev smoother (10 steps) in combination with the application of 6 of V-cycles. Our implementation is currently a proof-of-concept implementation as at present we reinitialize the AMG preconditioner upon each application. Another option would be to store various preconditioners, which is prohibitive from a computer memory point-of-view. The development of an efficient technique using multigrid or a fixed number of a simple iterative solver such as a Gauss-Seidel or Jacobi method should be investigated in the future. We therefore wish to emphasize that the timings presented here are not as rapid as they would be were the recomputation of the preconditioner at each application not required. We expect that if the varying preconditioners are handled efficiently the timings being reduced drastically. This could also lead to the now relatively larger number of V-cycles can be reduced – we choose to use this number of V-cycles as from our experience the performance of the AMG can be sensitive to parameter changes, which we wished to avoid here. Our implementation

[5] It is clear from practical considerations that the parameter $C_{1,\tau} \leq 1$, as otherwise $\widehat{S}_2^{-1}S$ would have an infinite eigenvalue, which is not the case as $\widehat{S}_2^{-1}S$ is clearly invertible. Our working indicates also that $C_{1,\tau}$ is not too close to 1 either, and numerical tests validate this hypothesis.



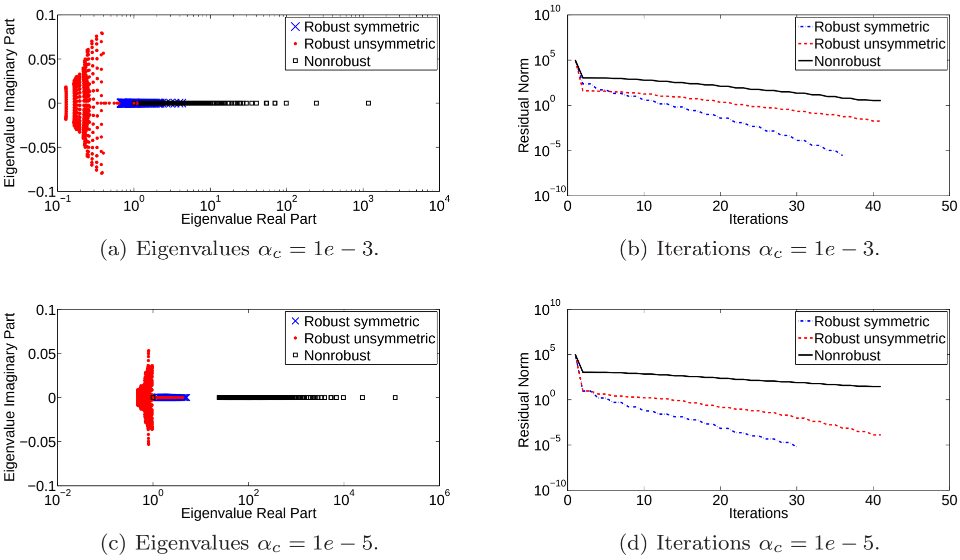
(a) Eigenvalues $\alpha_c = 1e-3$.

(b) Iterations $\alpha_c = 1e-3$.

(c) Eigenvalues $\alpha_c = 1e-5$.

(d) Iterations $\alpha_c = 1e-5$.

Fig. 3.1: Small problem ($\dim(S) = 1080$): Eigenvalues of $Sv = \lambda \hat{S} v$ for various approximations of the Schur complement (left). GMRES iterations for the saddle point problem using the three different Schur complement approximations (right).



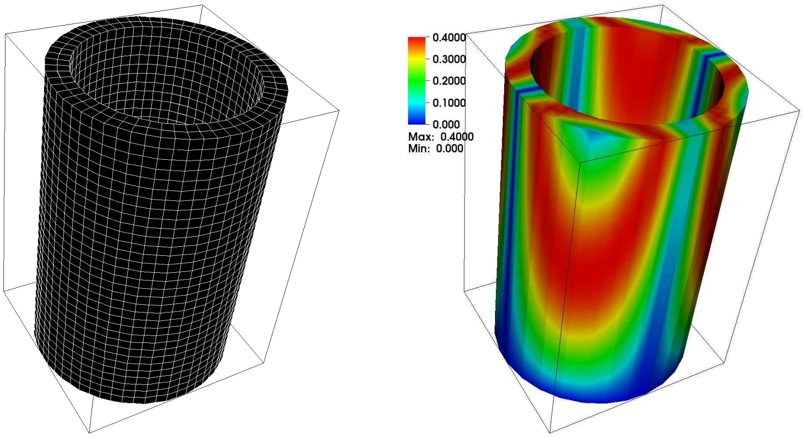
(a) Eigenvalues $\alpha_c = 1e-3$.

(b) Iterations $\alpha_c = 1e-3$.

(c) Eigenvalues $\alpha_c = 1e-5$.

(d) Iterations $\alpha_c = 1e-5$.

Fig. 3.2: Slightly larger problem ($\dim(S) = 5000$): Eigenvalues of $Sv = \lambda \hat{S} v$ for various approximations of the Schur complement (left). GMRES iterations for the saddle point problem using the three different Schur complement approximations (right).



(a) Domain (b) Desired state for first reactant

Fig. 4.1: Cylindrical Shell domain for computations and desired state for the first reactant.

of Bicg was stopped with a tolerance of 10^{-4} for the relative residual. Additionally, we stopped the SQP method whenever the relative change between two consecutive solutions was smaller than a given tolerance, which we will specify in our examples. More sophisticated techniques [40] for this could be employed in the future. We feel that as our purpose is to illustrate the performance of our preconditioner the choice made here is appropriate. Our experiments are performed for $T = 1$ and $\tau = 0.05$, i.e. 20 time-steps. We have taken the parameters $\alpha_{TU} = \alpha_{TV} = 0$ in all our numerical experiments, though we find it makes little difference computationally if this is not the case. We only consider three-dimensional examples here and will specify $\Omega \subset \mathbb{R}^3$ for each example. All results are performed on a Centos Linux machine with Intel(R) Xeon(R) CPU X5650 @ 2.67GHz CPUs and 48GB of RAM.

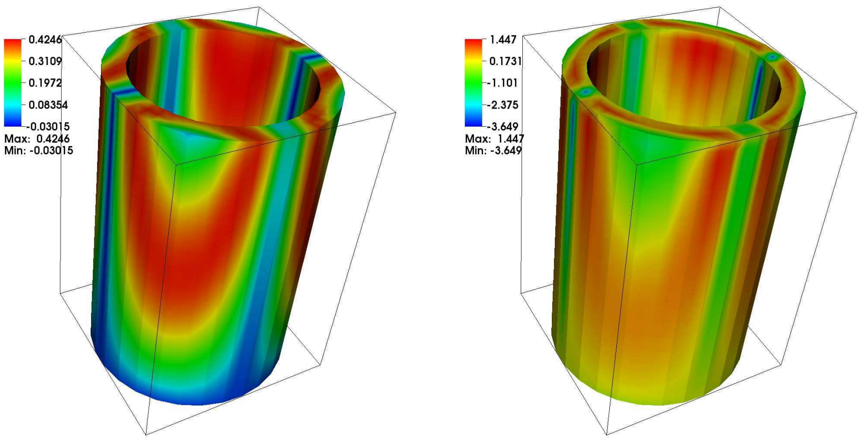
No Control Constraints.

Example 1. The first example we show is a cylindrical shell domain shown in Figure 4.1a with inner radius 0.8, outer radius 1.0 and height 3.0. The parameter setup for this problem is as follows: the desired state for the first reactant is shown in Figure 4.1b and is given by

$$u_Q = t\,|\sin(2x_0x_1x_2)|\,,$$

and the desired state for the second reactant is given by $v_Q = 0$. Additionally, we have $k_1 = k_2 = D_1 = D_2 = 1$ and $\gamma_1 = \gamma_2 = 0.15$.

In Table 4.1 we show the iteration numbers for the SQP method as well as the number of Bicg iterations needed for one SQP step. CPU timings are also provided. The results indicate that there is some mesh-dependence of the preconditioner, which for our experience can often be observed for boundary control problems. We also see a very benign growth with respect to the regularization parameter. This illustrates the robustness of our approach with respect to the regularization parameter for the control term.



(a) Computed state for first reactant at time step 7

(b) Computed control at time step 7

Fig. 4.2: Computed control and state for the first reactant at time step 7 for $\alpha_c = 1e{-}5$ and $\alpha_u = \alpha_v = 1.0$.

DoF	time		BiCG	time		BiCG
		$\alpha_c = 1e-5$			$\alpha_c = 1e-3$	
538 240	1175	step 1	46	1145	step 1	44
		step 2	46		step 2	44
3 331 520	17701	step 1	96	14910	step 1	80
		step 2	98		step 2	82

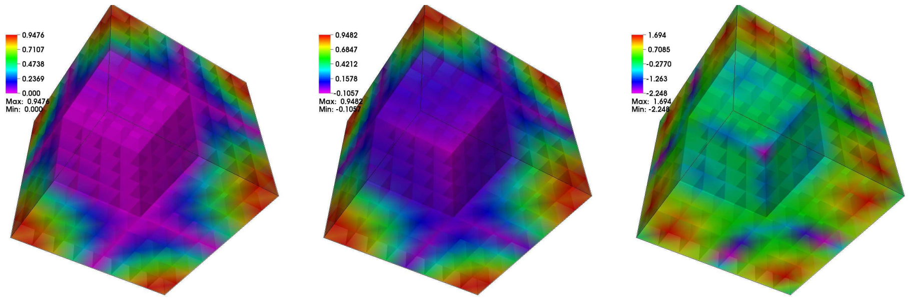
Table 4.1: Results for the cylindrical shell for varying mesh-size and regularization parameter α_c.

Example 2. The setup used for the second example is similar to the one we presented previously. We again use the desired states

$$u_Q = t\left|\sin(2x_0x_1x_2)\right|, v_Q = 0$$

with the parameters $k_1 = k_2 = D_1 = D_2 = 1$ and $\gamma_1 = \gamma_2 = 0.15$. In contrast to the last example we now solve the optimization problem on a Hyper L consisting of the cube on $[-1,1]^3$ with the cube $(0,1]^3$ removed (see Figure 4.3). Again, we wish to vary the control regularization parameter α_c and the mesh-parameter. Table 4.2 shows the results for the setup presented here including timings and iteration numbers. We can again observe a mild growth in iteration numbers with varying mesh-size and also a growth for very small values of α_c, however we find all iteration numbers are very reasonable considering the complexity of the matrix system being solved.

We consider the same problem as before but now wish to vary some values that have been previously been assumed to be fixed. The default setup is again $k_1 = k_2 = D_1 = D_2 = 1$, and $\gamma_1 = \gamma_2 = 0.15$. In the remainder of this section we will vary one



(a) Computed state first reactant

(b) Desired state

(c) Computed control

Fig. 4.3: Desired state, computed control and state for the first reactant at time step 18 for $\alpha_c = 1e-5$ and $\alpha_u = \alpha_v = 1.0$.

DoF	time		Bicg	time		Bicg
		$\alpha_c = 1e-5$			$\alpha_c = 1e-3$	
382 840	1523	step 1 step 2	60 66	1131	step 1 step 2	40 48
2 670 200	19511	step 1 step 2	116 119	18615	step 1 step 2	75 75

Table 4.2: Results for the Hyper L for varying mesh-size and regularization parameter α_c.

of these parameters and keep the other ones fixed. Obviously, this does not cover all the relevant choices that might be possible but this should indicate the effectiveness of our approach for a large range of parameter regimes. All computations are carried out on a fixed mesh that leads to a saddle point system of dimension 382840. We also note that each of these problems represents a completely different setup of the PDE and the optimization problem. The sole purpose of presenting the results in Table 4.3 is to show that the iteration numbers for all of these scenarios are reasonable, or sometimes very low. Clearly, there are some specific parameter regimes for which this approach will not be as effective as for the cases presented,[6] but for a wide range of parameters (h, τ, α_u, α_v, α_c, D_1, D_2, k_1, k_2, γ_1, γ_2) we find that our approach works very well.

[6] Experimental evidence indicates that the main case where the method is less effective occurs when γ_1 and γ_2 are large (i.e. when the $(1,1)$-block of the matrix system may have large negative eigenvalues) and α_c is small (i.e. when the term of the Schur complement corresponding to the indefinite part of the $(1,1)$-block does not dominate the positive semi-definite second term). From a linear algebra perspective, this is a difficult regime, as it involves approximating an indefinite $(1,1)$-block and Schur complement with little specific structure which can be exploited in our preconditioners. We also note that such a parameter regime may lead to an indefinite Hessian within the SQP step being carried out – in this case more sophisticated SQP schemes incorporating line-search or trust region approaches may be needed and could be explored in future work.

parameter	time		Bicg	time		Bicg
	$\alpha_c = 1e-5$			$\alpha_c = 1e-3$		
$D_1 = D_2 = 0.1$	3397	step 1	40	2866	step 1	36
		step 2	43		step 2	36
		step 3	45			
$D_1 = D_2 = 100$	2119	step 1	20	4195	step 1	12
		step 2	26		step 2	13
$\gamma_1 = \gamma_2 = 0.05$	1285	step 1	60	732	step 1	44
		step 2	64		step 2	48
$\gamma_1 = \gamma_2 = 0.75$	3382	step 1	60	2532	step 1	44
		step 2	71		step 2	53

Table 4.3: Results for varying parameters on the Hyper L domain for fixed dimension 382840 and varying regularization parameter α_c.

Control Constraints. We now present results for the case when control constraints are present. The domain of interest is again the the Hyper L presented earlier, with the desired states given by

$$u_Q = t\left|\sin(2x_0x_1x_2)\cos(2x_0x_1x_2)\right|, v_Q = 0,$$

and $k_1 = k_2 = D_1 = D_2 = 1$, $\gamma_1 = \gamma_2 = 0.15$. We only work with an upper bound on the control given by

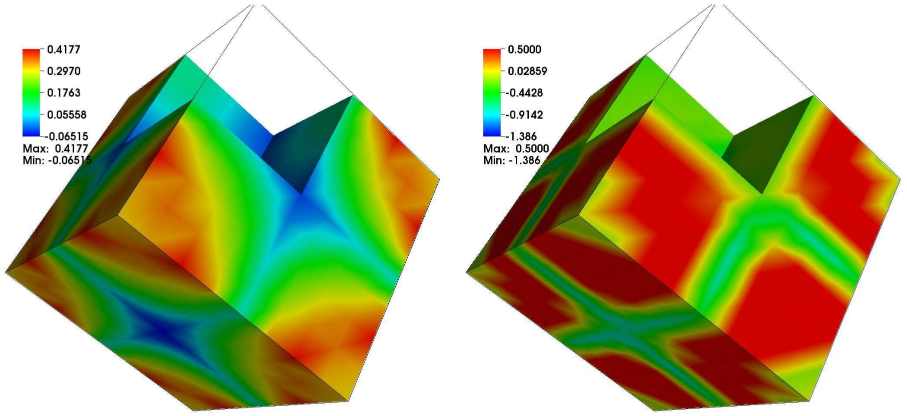
$$\overline{c} = 0.5.$$

The results for varying α_c and different mesh-parameters are shown in Table 4.4. We note that the convergence of the outer Newton method was observed to depend on the tolerance we used for the solution of the linear system (see [33]). The smaller value for α_c shown in Table 4.4 required the tolerance for the iterative solver to be reduced as otherwise we could not observe convergence of the Newton method. Our stopping criterion for the Newton method is based on the coincidence of subsequent active sets but a more sophisticated stopping criterion might be able to avoid the convergence issue of the Newton method [25, 40]. Table 4.4 shows the number of SQP steps, the number of semi-smooth Newton steps for the control constraints and the average number of Bicg iterations for one SQP step. We see that in both cases there is a benign growth with respect to the mesh-size. The difference between the two different values of α_c is probably due to the fact that as we change α_c the values for the control c change, which means that more variables will be active than in the case for the larger value of α_c.

In addition, we wish to illustrate robustness with respect to the penalty parameter $\bar{\varepsilon}$. We here keep the mesh and the regularization parameter ($\alpha_c = 1e-3$) fixed and consider different values of $\bar{\varepsilon}$. Table 4.5 illustrates that again the resulting iteration numbers are rather low. We also observed that the performance of the Newton method depended on the tolerance with which the linear systems were solved. For the rather low tolerance of $1e-9$ we found the Newton scheme and the SQP-scheme to converge with very few iterations. We observe that for rather small values of the penalty

DoF	time		NM/∅ BiCG	time		NM/∅ BiCG
		$\alpha_c = 1e-5$			$\alpha_c = 1e-3$	
60 920	3014	step 1	6/78.2	2070	step 1	5/35.6
		step 2	6/93.2		step 2	5/40.6
382 840	14502	step 1	6/92	12087	step 1	5/42.4
		step 2	6/107.6		step 2	5/48.4

Table 4.4: Results for the Hyper L for varying mesh-size and regularization parameter α_c.



(a) Computed state first reactant

(b) Computed control

Fig. 4.4: Desired state, computed control and state for the first reactant at time step 18 for $\alpha_c = 1e-5$ and $\alpha_u = \alpha_v = 1.0$.

parameter the convergence of the outer SQP method was slower than for larger values. This might be caused by the use of our simple SQP scheme and as we mentioned before more sophisticated schemes might be able to avoid this.

DoF		NM/∅ BiCG		NM/∅ BiCG		NM/∅ BiCG
		$\bar{\varepsilon} = 1e-2$		$\bar{\varepsilon} = 1e-4$		$\bar{\varepsilon} = 1e-6$
60 920	step 1	5/41.2	step 1	5/33.6	step 1	5/24.8
	step 2	5/44.2	step 2	5/34.8	step 2	5/25.2
					step 3	5/25.2
					step 4	5/25.2
					step 5	5/25.2

Table 4.5: Results on Hyper L domain for varying penalty parameter $\bar{\varepsilon}$.

5. Conclusions. In this paper we have established a Lagrange-Newton system for a reaction-diffusion optimization problem typically used in modeling chemical processes. At the heart of the nonlinear solvers lies the solution of large-scale linear systems in saddle point form that we have shown can be solved using efficient preconditioning techniques for a wide range of cases. We have introduced a preconditioner that efficiently approximates the $(1,1)$-block of the saddle point system and additionally derived robust approximations to the Schur complement. We have provided guidance on the eigenvalues of the preconditioned Schur complement and our numerical results illustrate that for a variety of problem setups (including box constraints on the control) our method produces low iteration numbers. The method presented here not only enables the accurate solution of chemical process models but also provides fast techniques to do this.

Acknowledgements. The first author was supported for this work by the Engineering and Physical Sciences Research Council (UK), Grant EP/P505216/1. The authors also gratefully acknowledge the award of a European Science Foundation (ESF) Exchange Grant under the OPTPDE programme. Both authors would like to thank Andy Wathen for fruitful discussions regarding the presented work.

REFERENCES

[1] S. S. Adavani and G. Biros, *Multigrid algorithms for inverse problems with linear parabolic PDE constraints.*, SIAM J. Sci. Comput., 31 (2008), pp. 369–397.

[2] U. Ascher and E. Haber, *A multigrid method for distributed parameter estimation problems.*, ETNA, Electron. Trans. Numer. Anal., 15 (2003), pp. 1–17,.

[3] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II—a general-purpose object-oriented finite element library*, ACM Trans. Math. Software, 33 (2007), pp. Art. 24, 27.

[4] W. Barthel, C. John, and F. Tröltzsch, *Optimal boundary control of a system of reaction diffusion equations.*, ZAMM, Z. Angew. Math. Mech., 90 (2010), pp. 966–982.

[5] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer, 14 (2005), pp. 1–137.

[6] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[7] A. Borzì, *Multigrid methods for parabolic distributed optimal control problems.*, J. Comput. Appl. Math., 157 (2003), pp. 365–382.

[8] A. Borzì and V. Schulz, *Multigrid methods for PDE optimization.*, SIAM Rev., 51 (2009), pp. 361–395.

[9] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[10] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[11] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal, 21 (1984), pp. 352–362.

[12] D. G. Feingold and R. S. Varga, *Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem*, Pacific J. Math., 12 (1962), pp. 1241–1250.

[13] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical analysis (Proc 6th Biennial Dundee Conf., Univ. Dundee, Dundee, 1975), Springer, Berlin, 1976, pp. 73–89. Lecture Notes in Math., Vol. 506.

[14] M. Gee, C. Siefert, J. Hu, R. Tuminaro, and M. Sala, *ML 5.0 smoothed aggregation user's guide*, Tech. Rep. SAND2006-2649, Sandia National Laboratories, 2006.

[15] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. I*, Numer. Math., 3 (1961), pp. 147–156.

[16] ———, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. II*, Numer. Math., 3 (1961), pp. 157–168.

[17] A. Greenbaum, V. Pták, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), p. 465.

[18] R. Griesse, *Parametric sensitivity analysis in optimal control of a reaction diffusion system. I: Solution differentiability.*, Numer. Funct. Anal. Optimization, 25 (2004), pp. 93–117.

[19] ———, *Parametric sensitivity analysis in optimal control of a reaction-diffusion system. II: Practical methods and examples.*, Optim. Methods Softw., 19 (2004), pp. 217–242.

[20] R. Griesse and S. Volkwein, *A primal-dual active set strategy for optimal boundary control of a nonlinear reaction-diffusion system.*, SIAM J. Control Optimization, 44 (2005), pp. 467–494.

[21] R. Griesse and S. Volkwein, *Parametric sensitivity analysis for optimal boundary control of a 3D reaction-diffusion system.*, New York, NY: Springer, 2006.

[22] E. Haber, *A parallel method for large scale time domain electromagnetic inverse problems.*, Appl. Numer. Math., 58 (2008), pp. 422–434.

[23] E. Haber, U. M. Ascher, and D. Oldenburg, *On optimization techniques for solving nonlinear inverse problems.*, Inverse Probl., 16 (2000), pp. 1263–1280.

[24] R. Herzog and E. W. Sachs, *Preconditioned conjugate gradient method for optimal control problems with control and state constraints*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2291–2317.

[25] M. Hintermüller, M. Hinze, and M. Tber, *An adaptive finite-element Moreau-Yosida-based solver for a non-smooth Cahn-Hilliard problem.*, Optim. Methods Softw., 26 (2011), pp. 777–811.

[26] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.

[27] M. Hinze, M. Köster, and S. Turek, *A Hierarchical Space-Time Solver for Distributed Control of the Stokes Equation*, tech. rep., SPP1253-16-01, 2008.

[28] ———, *A Space-Time Multigrid Solver for Distributed Control of the Time-Dependent Navier-Stokes System*, tech. rep., SPP1253-16-02, 2008.

[29] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, Springer-Verlag, New York, 2009.

[30] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1990.

[31] K. Ito and K. Kunisch, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM J. Contr. and Opt., 43 (2004), pp. 357–376.

[32] 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 (SIAM), Philadelphia, PA, 2008.

[33] C. Kanzow, *Inexact semismooth newton methods for large-scale complementarity problems*, Optimization Methods and Software, 19 (2004), pp. 309–325.

[34] M. Kollmann and M. Kolmbauer, *A Preconditioned MinRes Solver for Time-Periodic Parabolic Optimal Control Problems*, Sumitted,Numa-Report 2011-06, (August 2011).

[35] M. Kollmann and W. Zulehner, *A Robust Preconditioner for Distributed Optimal Control for Stokes Flow with Control Constraints*, Submitted, (2012).

[36] K. Krumbiegel, I. Neitzel, and A. Rösch, *Sufficient optimality conditions for the Moreau-Yosida-type regularization concept applied to the semilinear elliptic optmimal control problems with pointwise state constraints*, Tech. Rep. 1503/2010, WIAS, 2010.

[37] Y. A. Kuznetsov, *Efficient iterative solvers for elliptic finite element problems on nonmatching grids*, Russian Journal of Numerical Analysis and Mathematical Modelling, 10 (1995), pp. 187–211.

[38] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput, 21 (2000), pp. 1969–1972.

[39] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl, 13 (1992), pp. 778–795.

[40] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research, Springer-Verlag, New York, 1999.

[41] C. C. Paige and M. A. Saunders, *Solutions of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal, 12 (1975), pp. 617–629.

[42] J. W. Pearson, M. Stoll, and A. J. Wathen, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, To appear SIAM J. Matrix Anal.Appl, (2012).

[43] J. W. Pearson and A. J. Wathen, *Fast Iterative Solvers for Convection-Diffusion Control Problems*, in preparation, (2011).

[44] J. W. Pearson and A. J. Wathen, *A new approximation of the schur complement in precon-*

ditioners for pde-constrained optimization, Numerical Linear Algebra with Applications, (2011), p. Online first.

[45] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM Journal on Scientific Computing, 32 (2010), pp. 271–298.

[46] T. Rees and M. Stoll, *Block-triangular preconditioners for pde-constrained optimization*, Numerical Linear Algebra with Applications, 17 (2010), pp. 977–996.

[47] T. Rees, M. Stoll, and A. Wathen, *All-at-once preconditioners for PDE-constrained optimization*, Kybernetika, 46 (2010), pp. 341–360.

[48] T. Rees and A. Wathen, *Preconditioning iterative methods for the optimal control of the Stokes equation*, SIAM Journal on Scientific Computing, 33 (2011), pp. 2903–2926.

[49] J. Schöberl and W. Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to pde-constrained optimization problems*, SIAM J. Matrix Anal. Appl, 29 (2007), pp. 752–773.

[50] M. Stoll and A. Wathen, *All-at-once solution of time-dependent Stokes control*, Accepted Journal of Computational Physics, (2012).

[51] ———, *Preconditioning for partial differential equation constrained optimization with control constraints*, Numer. Lin. Alg. Appl., 19 (2012), pp. 53–71.

[52] S. Takacs and W. Zulehner, *Convergence analysis of multigrid methods with collective point smoothers for optimal control problems*, Computing and Visualization in Science, 14 (2011), pp. 131– 141.

[53] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems*, SIAM Philadelphia, 2011.

[54] A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electronic Transactions in Numerical Analysis, 34 (2008), pp. 125–135.

[55] W. Zulehner, *Non-standard norms and robust estimates for saddle point problems*, SIAM J. Matrix Analysis Applications, 32 (2011), pp. 536–560.