Operator preconditioning for PDE-constrained optimisation and multiscale problems

Magne Nordaas

Thesis submitted for the degree of Ph.D.

Department of Informatics
University of Oslo

February 2017

Acknowledgements

First of all I am grateful to my supervisors, Kent-André Mardal and Bjørn Fredrik Nielsen for their guidance and the many discussions we had.

I want to thank all my colleagues at Simula and BioComp. You made my three years at Simula so much more enjoyable. In particular I should mention Simon Funke, who shared an office with me for most of this time.

Finally, I am grateful to my family for their endless support.

Contents

Introduction

1 Motivation

Many physical processes studied in science and engineering are modelled with partial differential equations (PDEs). The PDE model describes the global evolution of the process in terms of a (usually) local equation of its partial derivatives. Examples include the diffusion of heat through a material, elastic and electromagnetic wave propagation, and the flow of viscous fluids. These are only a few examples, as PDEs are ubiquitous in science and engineering.

In many cases, it is not the physical process itself that we are interested in, but instead some quantity derived from it. Mathematically, we describe this quantity in terms of a functional. Often, it is the extreme values of this functional that is of interest. That is, we want to identify the process that maximise or minimise the functional value, over a range of input parameters. For example, we may want to find a heat sink that optimise the cooling, a material composition that reproduces observed wave scattering, or the optimal placement of turbines in a tidal energy farm. These are examples of PDE-constrained optimisation problems.

Solving a PDE-constrained optimisation problem means identifying the best PDE solution over a range of parameters. Generally, this is involves more than just solving the PDE, which can be a challenging task in itself. The optimisation side of the problem also have to be solved, so a technique for solving the PDE must be combined with a technique for solving the optimisation problem. The many ways this can be done, and a vast range of applications, results in a rich theory of PDE-constrained optimisation.

The three works included in this thesis cover different aspects of PDE-constrained optimisation. The first paper concerns theory of PDE-constrained optimisation, more specifically parameter-robust solution techniques for a PDE-constrained optimisation problem. The second paper is about coupling PDEs on domains of different topological dimension. Although the problem considered is not a PDE-constrained optimisation problem, it does have a similar structure and can be analysed with similar techniques. The third paper is more application oriented, and concerns the use of PDE-constrained optimisation techniques to reconstruct cerebral blood flow from medical images.

The remainder of this introduction is structured as follows. Section 2 provides an introduction to PDE-constrained optimisation, and provides a common context to the three works included in this thesis. Sections 3 to 5 introduces the individual papers.

2 Methods

Most PDEs do not have closed form solutions and can only be solved approximately with numerical methods. In order to do so, it is necessary to formulate a solution algorithm and a discretisation scheme for the PDE in question. Similarly, an algorithm must be formulated for computing the optimal solution.

This section is intended as a very gentle introduction to PDE-constrained optimisation, and introduces notation and nomenclature used in the remainder of the text. It also serves as context for the papers included in this thesis, providing a common framing of the problems and the methods considered therein.

In this part of the thesis, technical details are kept to a minimum. In particular, it is assumed that all necessary smoothness conditions are satisfied.

2.1 An example: Inverse Poisson

The Poisson equation is commonly used as the prototypical example of a partial differential equations. Correspondingly, we use the inverse Poisson problem as a prototype example for PDE-constrained optimisation.

The forward problem

The Poisson equation has many applications. For instance, it can model the electric potential generated by distribution of electric charge, or a stationary distribution of heat in a conductive medium. The simplest Poisson problem reads

$$-\Delta u = f \quad \text{in } \Omega \tag{1a}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{1b}$$

Here, f is a known variable, and u is the unknown in the equation, and Ω is the domain of the equations. The second equation (1b) is a boundary condition, needed to close the problem. For simplicity, here u vanishes on the boundary $\partial\Omega$, but many other boundary conditions for u are possible.

The inverse problem

For the inverse Poisson problem we reverse the problem formulation. Given u, can we recover f such that (1) holds? Clearly, for a smooth u, we could directly compute $f = -\Delta u$. In practice however, a function u obtained from measurement will always contain some amount of noise, for example we might have $u = u_{\text{signal}} + u_{\text{noise}}$. A direct computation would then result in $f = -\Delta u_{\text{signal}} - \Delta u_{\text{noise}}$. Since the noisy component typically does not vary smoothly, the computed f would be dominated by the inverted noise, even if the relative amplitude of the noise is small. In mathematical terms, we say that the Laplacian is unbounded, and that the inverse problem is ill-posed since a small change in the input produces a large change in output.

A remedy for the ill-posedness of the inverse problem is to reformulate it as a regularised minimisation problem. That is, instead of trying to match the data exactly, we seek a solution that minimises the misfit with the data without being “too wild". There are many ways to formulate this minimisation problem, corresponding to different choices of metrics. The simplest of these is the least-squares approach, and that is what will be presented here.

Formal problem definition

Suppose that we have some given (noisy) measurement data d. Then the functional J_α, measuring misfit is defined as

$$J_\alpha(f, u) = \frac{1}{2}\int_\Omega (u - d)^2\,dx + \frac{\alpha}{2}\int_\Omega f^2\,dx. \tag{2}$$

The first term of J_α is the squared distance, and the second term, where $\alpha > 0$, is an example of Tikhonov regularisation, which is discussed below in section 2.6. The minimisation problem then reads

$$\begin{aligned} &\text{minimise:} && J_\alpha(f, u) \\ &\text{subject to:} && \begin{cases} -\Delta u = f & \text{in } \Omega \\ \quad u = 0 & \text{on } \partial\Omega \end{cases} \end{aligned} \tag{3}$$

Problem (3) is a formulation of the inverse Poisson problem that can be solved using standard methods, as we will see in the next section. It belongs to a class of problems called constrained minimisation problems.

2.2 Constrained optimisation

In this subsection we will discuss some solution methods for (3) and similar problems that are relevant to this thesis. For future reference, it is practical to adopt a more more abstract formulation of (3), as follows.

$$\begin{aligned} &\underset{(f,u)\in F\times V}{\text{minimise:}} && J(f, u) \\ &\text{subject to:} && e(f, u) = 0. \end{aligned} \tag{4}$$

Here the functional J is assumed to be convex. We also introduce the standard nomenclature: We call f the control variable and u the state variable. The functional J is called the objective, and $e(u, f) = 0$ is called the state equation. Specifying the function spaces for the control and state variables is an essential part of the problem formulation. Here we have denoted the control space as F and the state space as V. Moreover we assume state equation hold in Z^*, i.e. with $e : F \times V \to Z^*$, with F, V and Z being reflexive Banach spaces.

The method of Lagrange multipliers

The standard technique for minimising a convex function on the real line, is to identify the point where its derivative vanishes. This technique could be applied to (4) if it were not for the presence of the constraint. Therefore, methods for solving constrained problems commonly reformulate or modify the problem so that the constraint can be eliminated.

The method of Lagrange multipliers essentially replaces the constraint with a new variable, which we will call the adjoint state. A new functional, the Lagrangian, combining the objective functional and the state equation, is defined

$$\mathcal{L}(f, u, z) = J(f, u) + ze(f, u),$$

with $z \in Z$ (and size Z is reflexive, z can be identified with a functional mapping Z^* into $\mathbb{R}$). Let us perform a heuristic computation to see how the Lagrangian $\mathcal{L}$ relates to the optimisation problem (4).

$$\begin{aligned} \inf_{f,u} \{J(f,u) \mid e(f,u) = 0\} &= \inf_{f,u} \left\{ J(f,u) + \begin{cases} 0 & e(f,u) = 0 \\ \infty & e(f,u) \neq 0 \end{cases} \right\} \\ &= \inf_{f,u} \left\{ J(f,u) + \sup_z ze(f,u) \right\} \\ &= \inf_{f,u} \sup_z \mathcal{L}(f,u,z). \end{aligned} \tag{5}$$

In (5) we see that original constrained minimisation condition on J is transformed to a saddle point condition on the Lagrangian $\mathcal{L}$. Since we know that the derivative vanishes in a saddle point (and we assume here that $\mathcal{L}$ is differentiable), we have a necessary optimality condition,

$$D\mathcal{L} = 0,$$

where D denotes the (Fréchet) derivative operator. Writing out the individual components of $D\mathcal{L}$, we obtain a system of three equations,

$$D_f J(f,u) + zD_f e(f,u) = D_f \mathcal{L} = 0, \tag{6a}$$

$$D_u J(f,u) + zD_u e(f,u) = D_u \mathcal{L} = 0, \tag{6b}$$

$$e(f,u) = D_z \mathcal{L} = 0. \tag{6c}$$

In the literature the three equations are commonly given names. The first equation (6a) is sometimes called the design equation, the second equation is known as the adjoint equation, and the third we recognise as the state equation. We refer to the system of equations (6) as the optimality conditions.

The reduced problem

Another method for eliminating the constraint in (4) is to consider the state u as function of the control f. In general, some form of the implicit function theorem is needed to make the dependence of u on f explicit, but for the simplistic example (3) this is already clear.

The dependence on the state u is eliminated from the so-called *reduced functional* $\hat{J}$, which we define

$$\hat{J}(f) = J(f, u(f))$$

Since $\hat{J}$ is a function in f only, with the constraint eliminated, we can apply unconstrained solution techniques to minimise $\hat{J}$. In particular, we have the optimality condition

$$D\hat{J}(f) = 0. \tag{7}$$

For the reduced problem (7), we need a way to compute the total derivative of the reduced functional $\hat{J}$. Applying the chain rule, we have

$$D\hat{J}(f) = D_f J(f, u(f)) + D_u J(f, u(f)) D_f u(f).$$

Now we could compute $D_f u(f)$, known as the *tangent linear model*, and then evaluate $D\hat{f}$, but there is another approach that will appear more natural. By differentiating the state equation and again using the chain rule, we get

$$D_f e(f, u(f)) + D_u e(f, u(f)) D_f u(f) = 0,$$

hence

$$\begin{aligned} D\hat{J} &= D_f J(f(u(f)) + D_u J(f, u(f)) D_f u(f) \\ &= D_f J(f(u(f)) - D_u J(f, u(f)) \big[D_u e(f, u(f))\big]^{-1} D_f e(f, u(f)) \\ &= D_f J(f, u(f)) + z D_f e(f, u(f)), \end{aligned} \tag{8}$$

Where z solves an *adjoint equation* similar to (6b),

$$z D_u e(f, u(f)) = -D_u J(f, u(f)).$$

After computing the derivative, an updated value for f can be computed with a descent step, and the process can be repeated until convergence.

2.3 Variational forms and discretisation

A solution to the minimisation problem (4) is found by solving the equations provided by conditions (6) or (7). For this, we need numerical methods, both for the PDE and for the optimisation method.

Variational formulation

The PDE formulation in (1) is usually rewritten using the integration by parts formula,

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \qquad \forall v \in H_0^1(\Omega). \tag{9}$$

Equation (9) is said to be a *variational* or *weak* form of the Poisson problem (1). The term variational here refers to that the function v in (9) is allowed to vary, and the equality should hold for any choice of v for which the integration by parts formula makes sense. The set all such v comprise a linear space denoted $H_0^1(\Omega)$.

It is convenient to introduce a shorthand notation for (9):

$$\langle Au, v\rangle = \langle Mf, v\rangle\, dx \qquad \forall v \in H_0^1(\Omega).$$

where A and M are linear operators defined

$$\begin{aligned}\langle Au, v\rangle &= \int_\Omega \nabla u \cdot \nabla v\, dx,\\ \langle Mu, v\rangle &= \int_\Omega fv\, dx.\end{aligned}$$

With this notation, the inverse Poisson problem fits in the framework of (4), with

$$\begin{aligned}J(f,u) &= \frac{1}{2}\langle M(u-d), (u-d)\rangle + \frac{\alpha}{2}\langle Mf, f\rangle,\\ ze(f,u) &= \langle Au - Mf, z\rangle.\end{aligned}$$

An elementary computation now lets us express the optimality conditions (6) in terms of the operators A and M,

$$\begin{aligned}\alpha Mf - M^*z &= 0,\\ Mu + A^*z &= Md,\\ Au - Mf &= 0,\end{aligned}$$

where * denotes the dual operator defined $\langle A^*z, u\rangle = \langle Au, z\rangle$. The optimality conditions form a linear system

$$\begin{bmatrix}\alpha M & & -M^*\\ & M & A^*\\ -M & A & \end{bmatrix}\begin{bmatrix}f\\ u\\ z\end{bmatrix} = \begin{bmatrix}0\\ Md\\ 0\end{bmatrix} \tag{10}$$

We write this more compactly

$$\mathcal{A}_\alpha x = b$$

where $\mathcal{A}_\alpha : X \to X^*$ is the coefficient matrix in (10), $X = V \times F \times Z$.

The finite element method

Solving equation (10) gives us a solution to the inverse Poisson problem (3). But before can solve this equation numerically, we need to discretise it. That is, we need to derive a finite-dimensional approximation of the problem that can be solved by a computer.

Finite element methods (FEM) comprise a class of discretisation schemes commonly used for PDEs. For the equation (10) however, we only need the simplest of finite element methods, the “Lagrange triangle", described below.

Assume that $\mathcal{T}_h$ is a triangularisation of the domain Ω. That is, $\mathcal{T}_t$ is a collection of triangles that cover Ω, in the sense that $\cup_{T\in\mathcal{T}_h} T = \overline{\Omega}$, such that the intersection of any two triangles is either empty, a shared side, or a shared vertex. A finite-dimensional subspace of

$H_0^1(\Omega)$ is then defined

$$V_h = \{\phi \in H_0^1(\Omega) \mid \phi|_T \in \mathbb{P}_1(T),\, T \in \mathcal{T}_h\},$$

where $\mathcal{P}_k(T)$ denotes the linear space of polynomials of degree at most k on T. A basis $\{\phi_i\}_{i=1}^N$ spanning V_h can then be uniquely determined from the condition

$$\phi_i(v_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad v_j \in \mathcal{V}(\mathcal{T}_h), \quad i, j \in \{1, \ldots, N\},$$

where $\mathcal{V}(\mathcal{T}_h)$ denotes the set of vertices in the triangularisation. Function $u_h, f_h \in V_h$ have unique representation $\mathbf{u}_h, \mathbf{f}_h \in \mathbb{R}^N$ with respect to this basis, i.e. $u_h = \sum_{j=1}^N (\mathbf{u}_h)_j \phi_j$ and a similar identity for f_h. Taking $u = u_h$, $f = h_h$ and $v = \phi_i$, $i = 1, \ldots, N$ in the variational form of the Poisson problem (9), we obtain a system of scalar equations,

$$\begin{aligned} \int_\Omega \nabla u_h \cdot \nabla \phi_i \, dx &= \int_\Omega \sum_{j=1}^N (\mathbf{u}_h)_j \nabla \phi_i \cdot \nabla \phi_j \, dx \\ &= \int_\Omega \sum_{j=1}^N (\mathbf{f}_h)_j \phi_j \phi_i \, dx = \int_\Omega f_h \phi_i \, dx. \end{aligned}$$

This linear system of N equations is more compactly expressed in matrix notation

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{M}_h \mathbf{f}_h,$$

where $\mathbf{A}_h, \mathbf{M}_h \in \mathbb{R}^{N \times N}$ are matrix representations of the restrictions of A and M to $V_h \subset H_0^1(\Omega)$, defined

$$\begin{aligned} (\mathbf{A}_h)_{i,j} &= \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, dx \\ (\mathbf{M}_h)_{i,j} &= \int_\Omega \phi_i \phi_j \, dx. \end{aligned}$$

Returning to the optimality conditions (10), we can replace the operators A and M with their discrete counterparts. This results in linear system that can be solved on a computer:

$$\begin{bmatrix} \alpha \mathbf{M}_h & & -\mathbf{M}_h^T \\ & \mathbf{M}_h & \mathbf{A}_h^T \\ -\mathbf{M}_h & \mathbf{A}_h & \end{bmatrix} \begin{bmatrix} \mathbf{f}_h \\ \mathbf{u}_h \\ \mathbf{z}_h \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{M}_h \mathbf{d}_h \\ 0 \end{bmatrix} \tag{11}$$

Here $\mathbf{z}_h \in \mathbb{R}^N$ denotes the representation of $z_h \in V_h$ with respect to the chosen basis, and the coefficient matrix in (11) is $3N$ by $3N$.

2.4 Solving the optimisation problem

A number of techniques exists for solving the discrete problem (11). In this section we outline two iterative methods that are relevant to papers included in this thesis. While the system (11) can be solved directly (with Gauss-elimination or Cholesky factorisation), we are interested in techniques that can also be applied to the more general, usually nonlinear and infinite-dimensional system (6).

The minimum residual method

The saddle-point system (11) can be solved with a suitable Krylov subspace method, for example the minimum residual method (minres). The advantage of Krylov solvers is that under certain conditions, they are optimal in the in the sense that computational cost of solving a linear system to a given tolerance scales linearly with the number of unknowns. This will require that the system is sparse, and usually also requires a good preconditioning technique as described below in section 2.5.

A disadvantage of using Krylov solvers for the system (11) is that all three equations contributes to the residual used to determine convergence. This important because the residuals of the three equations have different interpretation. A residual in the first equation simply means that the solution is not optimal. On the other hand, a residual in the third equation comes from the PDE constraint not being satisfied, meaning that the solution could be potentially be non-physical. If we terminate the iterative procedure while the residual is not negligible, we have to be certain that the PDE-constraint is satisfied well enough before we can conclude that the iterate is an approximate solution to the constrained optimisation problem.

Gradient descent methods

The reduced problem (7) can be solved with an iterative procedure. Assuming a give current iterate f^k, the derivative $D\hat{J}(f^k)$ is computed according to (8). The next iterate is then obtained by descent step,

$$f^{k+1} = f^k - \omega_k R_k^{-1} D\hat{J}(f^k), \qquad k = 0, 1, \dots, \tag{12}$$

where R_k^{-1} is a symmetric positive definite operator mapping the linear functional $D\hat{J}(f^k)$ to an element in the same space as f^k, and ω_k is a step length, typically determined by a line search method to ensure convergence. For example taking to R_k to be the Riesz operator mapping F^* to F results in a steepest descent method, and taking R_k to be the inverse Hessian $\left(D^2\hat{J}(f_k)\right)^{-1}$ results in a Newton method.

The discretised method can be derived as in the previous section. We omit the details, and write out the method in matrix notation,

$$\begin{bmatrix} \mathbf{A}_h & & \\ \mathbf{M}_h & \mathbf{A}_h^T & \\ & -\mathbf{R}_k & \omega_k^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{u}_h^{k+1} \\ \mathbf{z}_h^{k+1} \\ \mathbf{f}_h^{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_h \mathbf{f}_h^k \\ \mathbf{M}_h \mathbf{d}_h \\ (\omega_k^{-1} - \alpha \mathbf{R}_k)\mathbf{f}_h^k \end{bmatrix} \qquad k = 0, 1, \dots \tag{13}$$

where $\mathbf{R}_{h,k}$ is matrix representation of a discretisation of $\mathcal{R}_k$. The iterative scheme provides a new control vector $\mathbf{f}_h^{k+1}$ (as well as a state $\mathbf{u}_h^{k+1}$ and adjoint $\mathbf{z}_h^{k+1}$) from the previous control, by solving the equations in (13) sequentially from top to bottom, starting from some initial guess $\mathbf{f}_h^0$.

2.5 Analysis of saddle-point systems

The linear system (11) is an example of a *saddle-point system.* Given reflexive Banach spaces V and Q, such a system can be written in the standard form

$$\mathcal{A}x = \begin{bmatrix} A & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} = b \tag{14}$$

where $A : V \to V^*$ and $B : V \to Q^*$ are bounded linear operators and $(f, g) \in V^* \times Q^* = X^*$. Other examples of saddle-point systems include the Stokes equations and the equations of linear elasticity, and mixed formulations of the Poisson problem and the biharmonic problem.

Stability

It is well-known that system (14) has a solution $(u, p) \in V \times Q = X$ if the Ladyzhenskaya-Babuška-Brezzi (LBB) conditions are satisfied. The theorem and its proof can be found in [12, chapter 2]. Here, a weaker form of the theorem will suffice.

Theorem 2.1. *Assume A is bounded, self-adjoint and positive, and assume B is bounded. In addition assume that there exists constants $c_A, c_B > 0$, such that*

$$\langle Av, v\rangle \geq c_A^2 \|v\|_V^2 \qquad \forall v \in \ker B \subset V \tag{15a}$$

$$\sup_{v\in V} \frac{\langle Bv, q\rangle}{\|v\|_V} \geq c_B \|q\|_Q \qquad \forall q \in Q \tag{15b}$$

Then the saddle-point system (14) *has a unique solution $(u, p) \in V \times Q$ for every $(f, g) \in V^* \times Q^*$, and moreover*

$$\|(u, p)\|_{V\times Q} \leq c\|(f, g)\|_{V^*\times Q^*}$$

for some constant c independent of u and p.

Its important to note that the estimates (15) do not imply that corresponding estimates hold when U and Q are replaced with subspaces $V_h \subset V$ and $Q_h \subset Q$.

Preconditioning

Recall that the operator $\mathcal{A}$ in (14) maps X into its dual space X^*. This property commonly holds for linear operators that arises from variational problems. To formulate the minimum residual method and other Krylov subspace methods, it is then necessary introduce a preconditioner $\mathcal{R}^{-1} : X^* \to X$ in order to construct the Krylov subspace

$$\mathcal{K}_k = \operatorname{span}\{\mathcal{R}^{-1}b, (\mathcal{R}^{-1}\mathcal{A})\mathcal{R}^{-1}b, \ldots, (\mathcal{R}^{-1}\mathcal{A})^{k-1}\mathcal{R}^{-1}b\}.$$

For the minimum residual method, the preconditioner has to be a symmetric ($\mathcal{R}^* = \mathcal{R}$) and positive definite isomorphism, which means that the choice of preconditioner is equivalent to a choice of inner product on X, in accordance with the Riesz-Fréchet theorem.

The performance of the minimum residual method depends on the spectrum of the preconditioned coefficient matrix. A crude bound for the convergence rate can be determined from the estimate

$$\langle r^k, \mathcal{R}^{-1}r^k\rangle \leq 2\left(\frac{\kappa(\mathcal{R}^{-1}\mathcal{A})^2 - 1}{\kappa(\mathcal{R}^{-1}\mathcal{A})^2 + 1}\right)^k \langle r_0, \mathcal{R}^{-1}r_0\rangle,$$

where $r_k = b - \mathcal{A}x_k$ is the residual after k iterations, and κ is the condition number

$$\kappa(\mathcal{R}^{-1}\mathcal{A}) = \frac{\sup_{\lambda\in\sigma_p(\mathcal{R}^{-1}\mathcal{A})}|\lambda|}{\inf_{\lambda\in\sigma_p(\mathcal{R}^{-1}\mathcal{A})}|\lambda|}$$

where $\sigma_p(\mathcal{R}^{-1}\mathcal{A})$ denotes the point spectrum (eigenvalues) of $\mathcal{R}^{-1}\mathcal{A}$. For a more detailed discussion of minres convergence rates, see e.g. [36, 32, 48] and the references therein. In particular, the choice of preconditioner is important to ensure good convergence of the method independently of discretisation parameters.

Assuming the conditions of theorem 2.1 hold, we know that there are constants $c_1, c_2 > 0$ such that

$$c_1^{-1}\|x\|_{\mathcal{R}} \leq \|\mathcal{A}x\|_{\mathcal{R}^{-1}} \leq c_2\|x\|_{\mathcal{R}},$$

where we use the notation $\|\cdot\|_{\mathcal{R}}$ and $\|\cdot\|_{\mathcal{R}^{-1}}$ to emphasise that the norms on X and X^* are induced by $\mathcal{R}$ and $\mathcal{R}^{-1}$, respectively. Elementary computations, which we omit here, reveals that this estimate can be used to determine an upper bound for the condition number of the preconditioned system (and hence an upper bound for the minres convergence rate) from

$$\kappa(\mathcal{R}^{-1}\mathcal{A}) = \|\mathcal{A}\|\|\mathcal{A}^{-1}\| \leq c_1c_2.$$

In particular, if the estimates (15) can be shown to hold with constants independent of problem parameters, it follows that the upper bound for the minres convergence rate will also be bounded independently of the parameters.

Practical preconditioning

When solving discretised problems, $\mathcal{A}$ and $\mathcal{R}$ are replaced by discrete counterparts $\mathcal{A}_h$ and $\mathcal{R}_h$, which will typically have sparse matrix representations in the finite element basis. However, the preconditioned minres method calls for the inverse of $\mathcal{R}_h$ which will usually have a dense matrix representation in the same basis. In order to avoid the computationally expensive computation $\mathcal{R}_h^{-1}$, we replace $\mathcal{R}_h^{-1}$ with an approximation $\hat{\mathcal{R}}_h^{-1}$. The ideal approximation $\hat{\mathcal{R}}_h^{-1}$ induces a norm equivalent to that of $\mathcal{R}_h^{-1}$ (uniformly in h), and allows for inexpensive matrix-vector operations. Such approximations can be constructed with multilevel methods, see e.g. [8, 51, 10].

2.6 Regularisation

Remark. The functional in (2) can be written

$$J_\alpha = J_0 + \frac{\alpha}{2} \int_\Omega f^2 \, dx. \tag{16}$$

Here, J_0 measures the misfit, which we can consider the “real" quantity we seek to minimise, with second term being artificial penalty term. A naive approach would be solve the inverse Poisson problem without the second term in the functional, i.e. to solve problem (3) sense with $\alpha = 0$. Unfortunately, this does not work.

In order to obtain meaningful solutions to a problem, we want the problem to be well-posed as defined by Jacques Hadamard[33]: A problem is well-posed if a solutions exists, the solution is unique and depend continuously on the data.

Inverse problems are typically *not* well-posed, and the last term of the functional in (2) has been added to recover a well-posedness. This technique is an example of Tikhonov regularisation (see e.g. [69, chapter 1]), and the parameter $\alpha > 0$ determines the “strength" of the regularisation: As $\alpha \to 0$, the problem becomes harder to solve, reflecting that it is ill-posed in the limit $\alpha = 0$. In practical terms, as α decreases, the condition number of the optimality system increases, and the convergence rate of the minres iterations deteriorates.

2.7 Generalization

The example presented here is very simple because of the quadratic functional and linear constraint. Generally, PDE-constrained optimisation problems occurring in practice are complicated by non-smooth functionals, nonlinear PDE constraints, and additional inequality constraint on the control or state variable. Consequently, the theory of PDE-constrained optimisation is far richer than the elementary exposition presented here may indicate.

The mentioned complicating features typically results parametrised perturbations to the problem and nested iterative schemes. For example, if the optimality conditions (6) are nonlinear, they will have to be solved iteratively, for example with a Newton method. Within each Newton iteration, a linear saddle point system similar to (10) have to be solved. Although the exposition above treats a simplistic PDE-constrained optimisation problem, the methods are still applicable to more complex problems, as building blocks in more complex solution methods.

3 Parameter-robust preconditioning for a class of optimal control problems

Consider the optimal control problem where we minimise the functional

$$J(f,u) = \frac{1}{2}\|Tu - d\|_H^2 + \frac{\alpha}{2}\|f\|_F^2 \tag{17a}$$

over f and u in some Hilbert spaces F and V, and where $T : V \to H$ is an observation operator, so that the first term in (17a) measures misfit with some data d. The minimisation is subject to the constraint

$$Au + Bf = 0, \tag{17b}$$

where $A : V \to W^*$ and $B : F \to W^*$ are bounded linear operator, for example those arising from variational formulation of a linear PDE. The problem (17) can be solved with the method of Lagrange multipliers, as described in section 2.2. The saddle point of the Lagrangian

$$\mathcal{L}(f,u,w) = J(f,u) + (Au + Bf)w.$$

is the solution of the linear optimality system

$$\mathcal{A}x = \begin{bmatrix} \alpha R_F & & B \\ & T^*T & A^* \\ B & A & \end{bmatrix} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ T^*d \\ 0 \end{bmatrix} = b, \tag{18}$$

where $R_F : F \to F^*$ is the operator defining the inner product on F, i.e. $(f,g)_F = Rfg$ for all $f, g \in F$.

3.1 Distributed control and observation

The most studied examples of (17) are variants of the “prototype problem” (3), with a tracking-type functional constrained by an elliptic PDE. As an example, consider the problem

$$\text{minimise:} \quad J(f,u) = \frac{1}{2}\|u - d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|f\|_{L^2(\Omega)}^2 \tag{19a}$$

$$\text{subject to:} \quad \begin{cases} u - \Delta u + f = 0 & \text{in } \Omega \\ \nabla u \cdot n = 0 & \text{on } \partial\Omega \end{cases} \tag{19b}$$

Note that form PDE (19b) is chosen to simplify the presentation here, and the analysis can be generalised to similar elliptic PDE constraints. The more important structure of (19), as we will see, is that the control, state, and observational data are all distributed in over the same domain Ω.

Assuming the same Galerkin discretisation $V_h \subset H^1(\Omega)$ is employed for all three fields, the discretised optimality system reads

$$\mathcal{A}x = \begin{bmatrix} \alpha R_0 & & R_0 \\ & R_0 & R_1 \\ R_0 & R_1 & \end{bmatrix} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ R_0 d \\ 0 \end{bmatrix} = b, \tag{20}$$

where R_0 and R_1 are operators defining the $L^2(\Omega)$ and $H^1(\Omega)$ inner products V_h. Problem (20) is a special case of the general problem (18), with $T^*T = B = R_F = R_0$, equipping it with a lot more structure to exploit thant the general problem.

Now the first equation of (20) allows for eliminating the control variable (see e.g. [14, 38] for the case with additional control constraints), and this would reduce the computational costs of solving the system. Here we choose to retain the full system to preserve the standard saddle point form of the system so that it can be analysed with the techniques of sections 2.5 and 2.5, noting that analysis for the full system readily carries over to the system with the control eliminated.

Since the optimisation problem is not well-posed in the absence of regularisation, the condition number of the coefficient matrix in (18) blows up as $\alpha \to 0$. As a consequence, the iteration numbers for a minres solver applied to (18) tend to increase as α decreases.

Considerable effort has gone into developing preconditioners for optimality systems (18) that are robust with respect to the regularisation parameter α, in the sense that the condition number for the preconditioned system is bounded independently of $\alpha \in (0, 1)$ as well as of the mesh size.

For example, in [61] a robust preconditioner for (20) is developed. This block-diagonal preconditioner reads

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha R_0 & & \\ & R_0 & \\ & & \hat{S} \end{bmatrix} \tag{21}$$

and the robustness of the preconditioner follows from the Schur complement approximation

$$S = \frac{1}{\alpha} R_0 + R_1 R_0^{-1} R_1 \approx \left(R_1 + \frac{1}{\sqrt{\alpha}} R_0 \right) R_0^{-1} \left(R_1 + \frac{1}{\sqrt{\alpha}} R_0 \right) = \hat{S},$$

which is valid for small values of α. A practical preconditioner is obtained by replacing the inverse action R_0 and R_1 with less expensive approximations, e.g. Gauss-Seidel iterations for R_0 and multigrid cycles for R_1. This method has been adapted to optimal control problems with other PDE constraints, e.g. parabolic PDEs[53], convection diffusion[55], Stokes[66] and Navier-Stokes[52]. Adaptions to problems with additional constraints can be found in [65] and [54].

In [63] a robust preconditioner for (18) is derived using an operator preconditioning[3, 48] approach with nonstandard norms. The preconditioner, which differs from that in (21), is defined

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha \mathcal{R}_0 & & \\ & \mathcal{R}_0 + \sqrt{\alpha} \mathcal{R}_1 & \\ & & \frac{1}{\alpha} \mathcal{R}_0 + \frac{1}{\sqrt{\alpha}} \mathcal{R}_1 \end{bmatrix} \tag{22}$$

This technique has has been adapted to Stokes control in [74], time-periodic parabolic optimal control in [40].

3.2 Generalizations

The base problem (17) is naturally generalised by letting observation and control be restricted to parts of the domain of the state equation. For example let $\Omega_1 \subset \Omega$ and $\Omega_2 \subset \Omega$ be open subsets, and consider the problem

$$\begin{aligned}
&\text{minimise:} \quad J(f,u) = \frac{1}{2}\|u-d\|^2_{L^2(\Omega_1)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega_2)},\\
&\text{subject to:} \quad \begin{cases} u - \Delta u + f = 0 & \text{in } \Omega_2 \\ u - \Delta u = 0 & \text{in } \Omega \setminus \Omega_2 \\ \nabla u \cdot n = 0 & \text{on } \partial\Omega. \end{cases}
\end{aligned}$$

This problem was investigated in a more general setting, including additional constraints, in [62]. However, when Ω_1 and Ω_2 are proper subsets, it is not clear how to obtain a practical preconditioner, and the problem is even more challenging if the two regions do not overlap.

Other interesting problems arise when the controlled or observed region is restricted to the boundary or another set of lower topological dimension. Some classes of boundary control problems with observation across the domain were studied in [53] and robust preconditioners were derived by constructing an approximate Schur complement similar to (21). In [20], a robust preconditioning technique for a similar problem where the control is regularised in the weaker norm of the dual space $H^1(\Omega)^*$.

The model problem for boundary observation reads

$$\text{minimise:} \quad J(f,u) = \frac{1}{2}\|u-d\|^2_{L^2(\partial\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega)}, \tag{24a}$$

$$\text{subject to:} \quad \begin{cases} u - \Delta u + f = 0 & \text{in } \Omega \\ \nabla u \cdot n = 0 & \text{on } \partial\Omega. \end{cases} \tag{24b}$$

Note that the only change compared to (20) is that only boundary values are fitted to the data. The corresponding optimality system reads

$$\mathcal{A}x = \begin{bmatrix} \alpha R_0 & & R_0 \\ & R_{0,\partial\Omega} & R_1 \\ R_0 & R_1 & \end{bmatrix} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ R_{0,\partial\Omega} d \\ 0 \end{bmatrix} = b, \tag{25}$$

which is similar to (20) but the operator $R_{0,\partial\Omega}$, associated with the $L^2(\partial\Omega)$ inner product, has a nontrivial kernel, consisting of all the functions that vanish on the boundary. This significantly changes the structure of the problem compared (20), in particular the Schur complement used to derive (21) cannot be formed and the coercivity estimates used to derive (22) breaks down.

3.3 Paper I

In this paper we derive a robust preconditioning technique for the optimality system problem (25). The method is based on a reformulation of the problem, assuming higher regularity of the state variable. This approach allows establishing parameter-dependent norms in which the operator of (25) defines an isomorphism.

It is well-known that the solution of elliptic PDEs may have better than H^1-regularity, depending on the domain and boundary conditions, as well as the smoothness of the source term and any coefficients in the PDE. For the simple PDE under consideration, the following holds.

Theorem 3.1. *Let Ω be a $C^{1,1}$ domain or a convex domain, and let $f \in L^2(\Omega)$. Then there is a $u \in H^2(\Omega)$ satisfying the boundary value problem* (24b)*, with*

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}$$

for some constant C depending only on the domain Ω.

For a more general statement and proof, see the classical texts [29, 28]. For a non-convex polygonal domain, the solution will be in the linear span of $H^2(\Omega)$ and a finite number (equal to the number of reentrant corners) of basis functions that are H^2-unbounded in the neighbourhood of a reentrant corner.

The improved regularity allows for formulating the optimality system (24) in $L^2(\Omega) \times H^2(\Omega) \times L^2(\Omega)$, where $H^2(\Omega)$ is understood to have the boundary condition imposed, instead of "standard setting" of $L^2(\Omega) \times H^1(\Omega) \times H^1(\Omega)$.

In this new formulation, the optimality system assumes the form

$$\mathcal{A}_\alpha x = \begin{bmatrix} \alpha R_0 & & R_0 \\ & R_{0,\partial\Omega} & A^* \\ R_0 & A & \end{bmatrix} \begin{bmatrix} f \\ u \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ R_{0,\partial\Omega} d \\ 0 \end{bmatrix} \tag{26}$$

where the operator A is not symmetric, since integration by parts is not carried out as in the standard variational form of the Laplacian. We show that the operator $\mathcal{A}_\alpha$ is stable in norms induced by the symmetric positive isomorphism

$$\mathcal{R}_\alpha = \begin{bmatrix} \alpha R_0 & & \\ & R_{0,\partial\Omega} + \alpha R_2 & \\ & & R_0 \end{bmatrix} \tag{27}$$

where R_2 is an operator defining the standard inner product on $H^2(\Omega)$. A robust block-preconditioner for the optimality system (26) can be based on (27). For example, by using multigrid cycles to construct an approximate inverse to the (2,2) block of (27), and symmetric Gauss-Seidel iterations for approximate inversion of the (1,1) and (3,3) blocks. The analysis is supported with numerical results.

3.4 Conclusions

The method presented in the paper relies on a reformulation of the method with an assumption of more regularity on state variable. If the domain Ω satisfies the conditions of theorem 3.1, then solution does not change under this reformulation. On the other hand, if the conditions are not satisfied, then the new formulation excludes solutions that are $H^2(\Omega)$-defective in a neighbourhood of the reentrant corners. This is not necessarily a significant drawback. For instance, H^2-defective solutions may be considered discretisation artifacts if we are using a polygon to approximate a smooth domain Ω.

A more severe limitation appears when the PDE constraint (24b) is complicated with the introduction of spatially varying coefficients. In this case, we would have to impose smoothness requirements on the coefficients to ensure H^2-regularity of the state variable.

A more practical consequence is that finite-dimensional subspaces $H^2(\Omega)$ can be more complicated to construct. In particular, C^1 finite element methods with nodal polynomial basis functions becomes nearly intractable in three dimensions[73]. Alternative discretisation schemes such as iso-geometric analysis[17] and virtual element methods[13] may be more suitable, but it is not clear how to obtain effective multilevel preconditioners for the fourth order block in (27). Another possibility would be to use nonconforming methods as in [11], where a C^0 interior penalty method was used to discretise the model problem (19).

The technique presented in the paper can be adapted to similar problems with distributed control. In the abstract setting of (17), the fourth operator takes the form

$$R_\alpha = T^*T + \alpha A^*A.$$

An interesting example is the parabolic problem $A = \partial_t + (1 - \Delta)$, from which we obtain an anisotropic operator $A^*A = -\partial_t^2 + (1 - \Delta)^2$. Multilevel techniques may then be used to construct an approximate inverse to R_α as in [27]. This approach is interesting as the parabolic equation does not have to solved as part of the iterative scheme.

Finally, it is worth pointing out that the robust preconditioning techniques we have discussed in no way gets around the fundamental issue of the inverse problem being ill-posed in absence of regularisation. Such techniques shift the α-dependence of the condition number of the preconditioned system over to the stopping criterion of the iterative scheme. The merit of α-independent preconditioning is limited, since the cost of solving the problem to a parameter-independent tolerance stays roughly the same. See e.g. [49] for a discussion of minres performance with standard norms.

4 Parameter-robust preconditioning for a class of saddle point problems with trace constraints coupling problems of different dimensionality

In this section let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be an open and bounded Lipschitz domain, and let Γ be a be submanifold of Ω of topological dimension $n - k$, $k = 1, 2$. Consider the problem

$$-\Delta u + \sigma \delta_\Gamma p = f \quad \text{in } \Omega \tag{28a}$$

$$-\Delta v - p = g \quad \text{on } \Gamma \tag{28b}$$

$$\sigma u - v = 0 \quad \text{on } \Gamma. \tag{28c}$$

Here $\sigma > 0$ is a parameter and δ_Γ denotes Hausdorff measure concentrated on Γ, i.e. $\delta_\Gamma(U) = \mathcal{H}^{n-k}(U \cap \Gamma)$ for open $U \subset \mathbb{R}^n$. The measure is δ_Γ has properties similar to that of the atomic Dirac measure, and its presence in (28) implies that the problem must be understood in a variational sense. For simplicity, we assume homogeneous Dirichlet boundary conditions for u and v.

Problem (28) arises as a limit of the problem

$$-\nabla \cdot \left[(1 + \phi_\epsilon \sigma)\nabla u\right] = f + \phi_\epsilon g,$$

where $\phi_\epsilon(x) = \epsilon^{-1}\phi(\epsilon^{-1} \operatorname{dist}(x, \Gamma))$ and $\phi \in C^\infty(\mathbb{R}; \mathbb{R}_+)$ with $\operatorname{supp} \phi \subset [0, 1)$ and $\int_{\mathbb{R}} \phi \, dx = 1$. In other words, Γ can be considered a lower-dimension approximation of a thin, highly conductive structure. The problem is relevant to biomedical applications coupling fluid flow through tissue with flow along vascular networks, avoiding the difficulties of having to fully resolve the complex structure of the vasculature[19, 15, 18, 2].

The system of equations (28) can also be derived from the minimisation problem

$$\text{minimise:} \quad \int_\Omega \frac{1}{2}|\nabla u|^2 - fu \, dx + \int_\Gamma \frac{1}{2}|\nabla v|^2 - gv \, ds \tag{29a}$$

$$\text{subject to:} \quad \sigma \gamma_\Gamma u = v, \tag{29b}$$

where γ_Γ denotes a suitable trace operator, i.e. $\gamma_\Gamma u = u|_\Gamma$ for $u \in C(\overline{\Omega})$. This minimisation problem is a constrained minimisation problem similar to (3) treated in section 2, although the constraint is not a PDE. In particular, the method of Lagrange multipliers is applicable to (29), with the resulting optimality conditions (28), the variable p playing the role of the Lagrange multiplier for the trace constraints.

The problem (29) is relevant to domain decomposition methods with Lagrange multipliers [46] and problems coupling finite element and boundary element methods such as [25, 35]. A fourth-order problem similar to (29) can be used to model the bending of thin plates reinforced with ribs [70, chapter 9].

The mapping properties of the trace operator is central to this problem, and we briefly recall some of the relevant theory. The Besov space $B^s_{p,p}(\Omega)$, $s \in (0, 1)$, $1 \leq p < \infty$ is formally defined as an interpolation space [6, 1] and can be equipped with the norm

$$\|u\|_{B^s_{p,p}(\Omega)} = \left(\iint_{\Omega\times\Omega} \frac{|u(x)-u(y)|^p}{|x-y|^{n-sp}}\, dxdy \right)^{\frac{1}{p}} \tag{30}$$

Let $\gamma : C(\mathbb{R}^{n+1}) \to C(\mathbb{R}^n)$ be the trace operator defined $\gamma u(x_1, \dots, x_{n+1}) = u(x_1, \dots, x_n, 0)$. The following theorem characterising the traces of Sobolev and Bseov functions can be found textbooks, see e.g. [1, chapter 8] and [71, chapter 2].

Theorem 4.1. *Let $1 < p < \infty$.*

1. *The trace operator is bounded as a mapping of $W^{1,p}(\mathbb{R}^n)$ into the Besov space $B^{1-1/p}_{p,p}(\mathbb{R}^n)$ and has a bounded right inverse.*
2. *The trace operator is bounded as a mapping from $B^s_{p,p}(\mathbb{R}^{n+1})$ into the Besov space $B^{s-1/p}_{p,p}(\mathbb{R}^n)$ for $s > 1/p$, and it has a bounded right inverse for all $s \in (0, 1)$.*

For Γ with codimension 1, theorem 4.1 (along with standard extension theorems) establishes that $\gamma_\Gamma : H^1(\Omega) \to H^{1/2}(\Gamma)$ is bounded and surjective. In this case, the problem (28) is very closely related to an immersed boundary problem, where the Dirichlet boundary conditions are imposed using the method of Lagrange multipliers, see e.g. [4, 57].

For Γ with codimesion 2, the situation is more complicated. Theorem 4.1 does not allow for taking consecutive traces of H^1-functions into L^2, in fact this operation is known to be unbounded [44, chapter 3]. Boundedness can be be recovered by imposing additional smoothness or integrability on the function space. In [19] weighted Sobolev spaces $W^{1,p}_w$ were used to impose improved integrability locally around a one-dimensional Γ, allowing for well-defined traces from $W^{1,p}_w(\mathbb{R}^3)$ into $L^2(\Gamma)$.

4.1 Paper II

In this paper we develop robust preconditioners for the system (28). As remarked above, (28) defines the optimality conditions for the minimisation problem (29), hence naturally have a saddle point structure. Babuska-Brezzi theory (see e.g. [12, chapter 2]) provides framework for analysing such systems, and ideal preconditioners are derived from the theory of operator preconditioning[48].

Only the case of a 2D-1D coupling is considered, so that the trace is well-defined with standard spaces. The natural function space setting for (28) is then $(u, v) \in H^1_0(\Omega) \times H^1(\Gamma)$ for the primary variables, and for the Lagrange mulitplier p we identify a suitable fractional space.

The presence of homogeneous boundary conditions means that the image of the trace is characterized as $H^{1/2}_{00}(\Gamma)$, defined as the linear subspace of $H^{1/2}(\Gamma)$ consisting of functions that can, in a specific sense, be extended by zero outside of Γ. More precisely, if $\tilde{\Gamma}$ is a (Lipschitz) curve such that $\Gamma \subset \tilde{\Gamma}$, $H^{1/2}_{00}(\Gamma)$ is defined as the completion of $\mathcal{D}(\Gamma)$ in the norm

$$\|v\|_{H^{1/2}_{00}(\Gamma)} = \|\tilde{v}\|_{H^{1/2}(\tilde{\Gamma})} \quad \text{where} \quad \tilde{v}(x) = \begin{cases} v(x) & \text{if } x \in \Gamma \\ 0 & \text{otherwise .} \end{cases}$$

Note that space $H_{00}^{1/2}(\Gamma)$ is independent (up to norm equivalence) of the extension domain $\tilde{\Gamma}$ which is used only for theoretical considerations.

With the choice of spaces outlined above, the problem (28) takes the form of a saddle-point system

$$\mathcal{A}x \begin{bmatrix} A_\Omega & & \sigma\gamma^* \\ & A_\Gamma & j^* \\ \sigma\gamma & j & \end{bmatrix} \begin{bmatrix} u \\ v \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \\ 0 \end{bmatrix} = b \tag{31}$$

Where A_Ω and A_Γ denotes the variational Laplacian on Ω and Γ respectively, and j denotes the continuous embedding of $H_0^1(\Gamma)$ into $H_{00}^{1/2}(\Gamma)$. We define parameter-dependent spaces V and Q,

$$\begin{aligned} V &= H_0^1(\Omega) \times H_0^1(\Gamma) \\ Q &= H^{-1}(\Omega) \cap \sigma H_{00}^{1/2}(\Gamma)^*, \end{aligned}$$

where the norms are given by

$$\|(u,v)\|_V = \left(\|u\|_{H^1(\Omega)}^2 + \|v\|_{H^1(\Gamma)}^2\right)^{\frac{1}{2}} \tag{32a}$$

$$\|p\|_Q = \left(\|p\|_{H^{-1}(\Gamma)}^2 + \sigma^2\|p\|_{H^{1/2}(\Gamma)^*}^2\right)^{\frac{1}{2}}. \tag{32b}$$

The system (31) has standard saddle point form, and is readily analysed with Babuška-Brezzi theory[12, chapter 2]. In the paper we show that the coefficient matrix $\mathcal{A}$: in (31) defines an isomorphism mapping $V \times Q$ onto $V^* \times Q^*$, such that both $\|\mathcal{A}\|$ and $\|\mathcal{A}^{-1}\|$ are bounded independently of σ. This demonstrates the well-posedness of the problem, and additionally, in accordance with operator preconditioning techniques[48], an ideal preconditioner is provided by the Riesz operator associated with the norms (32).

The stability estimates for (31) do not in general apply when V and Q are replaced with discrete subspaces, making it necessary to identify stable subspaces $V_h \subset V$ and $Q_h \subset Q$. If V_h is a finite element space and Γ does not intersect element interiors, we can simply take take Q_h to be the exact trace space of V_h, and the numerical experiments in the paper are carried out in this setting. In the general case, one would have to more carefully select the subspace Q_h to ensure that the trace constraint is not too strong[4, 57].

For preconditioning the discretised systems, we need a computationally inexpensive operator that is spectrally equivalent with Riesz operator associated with the fractional norm (30). However, the non-local nature of the fractional norm makes it impractical to work with directly. Instead, we use a spectral construction of the fractional operator following [44, chapter 3]. This operator induces a norm equivalent to (30), a consequence of the fact that essentially all interpolation methods coincide on Hilbert spaces[16]. The smaller size and tri-diagonal structure of the discrete 1D Laplacian makes this approach viable in terms of computational expense. In the paper we present numerical results demonstrating effectiveness this technique.

4.2 Conclusions

In the paper only the 2D–1D problem is considered. While the analysis can also be carried out in the 3D–2D setting, a preconditioner based on matrix diagonalisation would scale poorly without a tri-diagonal structure to exploit. Therefore, the method presented in the paper is not suitable if the dimension of Q_h is large, and other methods for constructing an approximate fractional operator with better scaling properties should be used instead. For example an approximate Poincaré-Stekhlov operator can be constructed, using one multigrid V-cycle on A_Ω. Other approaches include approximating the fractional operator with contour integrals as in [34], or applying BPX preconditioning techniques[9, 7].

Another limitation is that the tridiagonal structure of A_Γ may break down even in the 2D–1D setting. This happens if the problem is discretised with polynomial elements of degree > 1, or if the 1D structure is branching. Possible remedies for this is making use of first order elements on a refined mesh for constructing the preconditioner, and using domain decomposition techniques for the branching.

The analysis presented in the paper does not easily generalise to the much more interesting 3D–1D setting. However, the smaller size of the 1D domain means the diagonalisation technique may useful. In [41] some results for this approach are presented without analysis.

5 Variational data assimilation for transient blood flow

Cerebral aneurysms are present in about 5% of the population[5]. The risk of rupture is fairly low, but when it happens, it fatal in more than half the cases, the half the survivors suffers morbidity. Detailed characteristics of the cerebral blood flow can potentially assist medical professionals in identifying the aneurysms having the great risk of rupture.

Imaging techniques, such as ultrasound and phase-contrast magnetic resonance imaging (PC-MRI), can provide non-invasive measurements of flow of the arterial bloodflow in the brain. However, the data obtained with these techniques may not have sufficient resolution to provide the detailed characteristics desired.

In recent years computational fluid mechanics (CFD) has received a lot of attention as a tool for investigating cerebral blood flow. The high temporal and spatial resolution afforded by CFD can reveal potentially important flow characteristic, for example turbulent or transitional flow, and allow for computation of quantities that cannot be directly measured in patient, such as pressure and wall shear stresses.

The accuracy of CFD simulation naturally depends on the computational model used and the inputs provided to it. Modelling the blood as a Newtonian fluid and disregarding the interaction between the fluid and the vessel walls been shown to be reasonable modelling simplifications. The uncertainty of the inputs, specifically the geometry and the boundary conditions, remains the main source of inaccuracy in simulations[59, 60, 21, 64]. In particular, this applies to the flow or pressure values prescribed on the artificial boundaries of the computational domain resulting from simulating only a part of the arterial system. Numerical simulations are highly sensitive to such values since they essentially determine the division of flow between the branches of arterial bifurcations.

One way of obtaining suitable boundary conditions is to combine the hemodynamical model with a simplified model for the cardiovascular system, providing averaged values pressure values or flow rates at the boundaries. A mathematical framework for this has been developed in [37, 72, 23, 24].

Another way to account for the unknown boundary conditions is to make use of measurements of the flow velocity field. Using the measurement values for the boundary conditions is likely to produce poor results, due to the poor resolution and presence of noise in the data.

A more robust approach is solving the inverse problem of recovering the boundary conditions that best reproduces the measured flow. This form of optimal control problem for the Navier-Stokes equations have been investigated in e.g. [26, 30, 43]. A theoretical framework in the hemodynamics context has been developed in [67, 31, 68], where the authors considered stationary flow with measurement data from crossections of the flow. A similar inverse problem have also been investigated in [39], and with reduced order methods in [47, 42]. Other related works includes compliance recovery[56] and shape optimisation[58] in a hemodynamical context.

5.1 Paper III

In this paper we apply variational data assimilation techniques to transient blood flow. We aim to minimise misfit between simulation and data, measured with the functional

$$J(f,u) = \frac{1}{2}\|Tu - d\|_H^2 + \frac{\alpha}{2}\|f\|_F^2, \tag{33}$$

where $d \in H$ is the observational data, belonging to some Hilbert space H, and u is a solution of the hemodynamic model given boundary conditions $f \in F$ as input parameters, for some Hilbert space F. The operator T, mapping velocity fields to data, can be thought of as a virtual measuring device, and ideally, mimics the MRI machine.

The model we use for the blood flow is the incompressible Navier-Stokes equations with Newtonian viscosity. In scaled quantities, the equations reads

$$u_t + (u \cdot \nabla)u - \mathrm{Re}^{-1}\,\Delta u - \nabla p = 0 \tag{34a}$$

$$\nabla \cdot u = 0, \tag{34b}$$

where Re is the Reynolds number, which here is typically around 350 for blood flow in cerebral arteries.

The PDE model must be closed with suitable boundary conditions. For this, we assume stationary vessel walls with noslip condition. The velocity field on the remaining parts of the boundary, the inlet and outlets of the computational domain, are unknowns that we want obtain from solving the inverse problem. However, prescribing a velocity to all inlets and outlets is known to adversely affect the stability of numerical simulation, and will result in an inconsistent problem if the net flow is nonzero.

A more practical approach is to prescribe velocity fields for the inlet and all but one outlets, and for the last outlet impose a traction-free condition. In summary, the boundary conditions reads

$$
\begin{aligned}
u &= 0 \quad \text{on } \Gamma_{\text{wall}} && \text{(35a)}\\
u &= f_i \quad \text{on } \Gamma_i,\ i = 0, \dots, m-1 && \text{(35b)}\\
(\mathrm{Re}^{-1}\,\nabla u + Ip)n &= 0 \quad \text{on } \Gamma_m, && \text{(35c)}
\end{aligned}
$$

where $\partial\Omega = \Gamma_{\text{wall}} \cup_{i=0}^{m} \Gamma_i$.

The resulting optimisation problem reads

$$
\begin{aligned}
\text{minimise:} &\quad J(f, u)\\
\text{subject to:} &\quad \text{(34)–(35)}
\end{aligned}
\tag{36}
$$

and can be solved with methods discussed in section 2, adapted to the nonlinear PDE constraint for this problem. Here, we solve the reduced problem using the LBFGS algorithm, a robust method for large unconstrained optimisation problems.

The LBFGS is a descent algorithm that can be written in general form (12), where $R_k : F^* \to F$ is an iteratively constructed approximation of the inverse Hessian. The constructed operators R_k build on an initial operator $R_0 = R_F$, the operator determining the inner product on F, which is such a way that R_F has an easily invertible tensor product structure. Combined by choosing step lengths ω_k that satisfy the strong Wolfe conditions, this ensures convergence of the method (see e.g. [50, chapter 6])

The PDE is linearised with NewtonÂt's method and discretised with stabilised P1-P1 elements using FEniCS[45]. The discretely consistent derivative is derived automatically using dolfin-adjoint[22].

The method is tested on a 2D and 3D test cases. For the 2D case, we generate synthetic data and examine the effect of noise and regularisation. For the 3D test case, we use geometry and measurement data obtained from 4D-PCMRI of an aneurysm in a dog. We compare the solution to (36) with a state of the art solver for the forward problem, where the flow boundary conditions are determined by vessel diameter and an idealised heart cycle.

5.2 Conclusions

We have demonstrated the feasibility of solving the inverse problem (36) for a transient flow in a realistic 3D geometry. However, there are several limitations to the method as we have implemented it. For one, to compare measurement data simulated velocity, we interpolated the data on the simulation mesh. Ideally, the simulated velocity would instead be subjected to an observation operator that mimics the MRI machine, but it not clear to us how this should be done in practice.

Another limitation is that for the 3D test case, the physicals models were not the same. To keep the computational cost of the two approaches roughly equal, the viscosity was increased for the inverse problem. This makes it difficult to draw a meaningful conclusion from the comparison.

Bibliography

[1] R. A. ADAMS AND J. J. FOURNIER, *Sobolev spaces*, vol. 140 of Pure and Applied Mathematics, Academic press, 2003.

[2] I. AMBARTSUMYAN, E. KHATTATOV, I. YOTOV, AND P. ZUNINO, *Simulation of flow in fractured poroelastic media: A comparison of different discretization approaches*, in International Conference on Finite Difference Methods, Springer, 2014, pp. 3–14.

[3] O. AXELSSON AND J. KARÁTSON, *Equivalent operator preconditioning for elliptic problems*, Numerical Algorithms, 50 (2009), pp. 297–380.

[4] I. BABUŠKA, *The finite element method with Lagrangian multipliers*, Numerische Mathematik, 20 (1973), pp. 179–192.

[5] J. B. BEDERSON, E. S. CONNOLLY, H. H. BATJER, R. G. DACEY, J. E. DION, M. N. DIRINGER, J. E. DULDNER, R. E. HARBAUGH, A. B. PATEL, AND R. H. ROSENWASSER, *Guidelines for the management of aneurysmal subarachnoid hemorrhage a statement for healthcare professionals from a special writing group of the stroke council, american heart association*, Stroke, 40 (2009), pp. 994–1025.

[6] J. BERGH AND J. LOFSTROM, *Interpolation spaces: An introduction*, vol. 223 of Grundlehren der matematischen Wissenschaften, Springer, 1976.

[7] J. BRAMBLE, J. PASCIAK, AND P. VASSILEVSKI, *Computational scales of Sobolev norms with application to preconditioning*, Mathematics of Computation of the American Mathematical Society, 69 (2000), pp. 463–480.

[8] J. H. BRAMBLE, *Multigrid methods*, vol. 294 of Pitman Research Notes in Mathematics, CRC Press, 1993.

[9] J. H. BRAMBLE, J. E. PASCIAK, AND J. XU, *Parallel multilevel preconditioners*, Mathematics of Computation, 55 (1990), pp. 1–22.

[10] S. BRENNER AND R. SCOTT, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer Science & Business Media, third ed., 2008.

[11] S. BRENNER, L.-Y. SUNG, AND Y. ZHANG, *A quadratic C^0 interior penalty method for an elliptic optimal control problem with state constraints*, in Recent Developments

in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations, Springer, 2014, pp. 97–132.

[12] F. BREZZI AND M. FORTIN, *Mixed and hybrid finite element methods*, vol. 15 of Springer series in Computational Mathematics, Springer Science & Business Media, 2012.

[13] F. BREZZI AND L. D. MARINI, *Virtual element methods for plate bending problems*, Computer Methods in Applied Mechanics and Engineering, 253 (2013), pp. 455–462.

[14] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM Journal on Control and Optimization, 24 (1986), pp. 1309–1318.

[15] L. CATTANEO AND P. ZUNINO, *Computational models for fluid exchange between microcirculation and tissue interstitium.*, Networks & Heterogeneous Media, 9 (2014).

[16] S. N. CHANDLER-WILDE, D. P. HEWETT, AND A. MOIOLA, *Interpolation of Hilbert and Sobolev spaces: Quantitative estimates and counterexamples*, Mathematika, 61 (2015), pp. 414–443.

[17] J. A. COTTRELL, T. J. R. HUGHES, AND Y. BAZILEVS, *Isogeometric analysis: Toward integration of CAD and FEA*, John Wiley & Sons, 2009.

[18] C. D'ANGELO, *Finite element approximation of elliptic problems with Dirac measure terms in weighted spaces: Applications to one-and three-dimensional coupled problems*, SIAM Journal on Numerical Analysis, 50 (2012), pp. 194–215.

[19] C. D'ANGELO AND A. QUARTERONI, *On the coupling of 1D and 3D diffusion-reaction equations: Application to tissue perfusion problems*, Mathematical Models and Methods in Applied Sciences, 18 (2008), pp. 1481–1504.

[20] O. L. ELVETUN AND B. F. NIELSEN, *PDE-constrained optimization with local control and boundary observations: Robust preconditioners*, SIAM Journal on Scientific Computing, 38 (2016), pp. A3461–A3491.

[21] Ø. EVJU, K. VALEN-SENDSTAD, AND K.-A. MARDAL, *A study of wall shear stress in 12 aneurysms with respect to different viscosity models and flow conditions*, Journal of biomechanics, 46 (2013), pp. 2802–2808.

[22] P. E. FARRELL, D. A. HAM, S. W. FUNKE, AND M. E. ROGNES, *Automated derivation of the adjoint of high-level transient finite element programs*, SIAM Journal on Scientific Computing, 35 (2013), pp. C369–C393.

[23] L. FORMAGGIA, A. VENEZIANI, AND C. VERGARA, *A new approach to numerical solution of defective boundary value problems in incompressible fluid dynamics*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2769–2794.

[24] ——, *Flow rate boundary problems for an incompressible fluid in deformable domains: formulations and solution methods*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 677–688.

[25] S. A. Funken and E. P. Stephan, *Hierarchical basis preconditioners for coupled FEM–BEM equations*, in Boundary Elements: Implementation and Analysis of Advanced Algorithms, Springer, 1996, pp. 92–101.

[26] A. V. Fursikov, M. D. Gunzburger, and L. Hou, *Boundary value problems and optimal boundary control for the Navier–Stokes system: The two-dimensional case*, SIAM Journal on Control and Optimization, 36 (1998), pp. 852–894.

[27] M. Griebel and P. Oswald, *Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems*, Advances in Computational Mathematics, 4 (1995), pp. 171–206.

[28] P. Grisvard, *Singularities in boundary value problems*, vol. 22, Springer, 1992.

[29] ——, *Elliptic problems in nonsmooth domains*, vol. 69, SIAM, 2011.

[30] T. Guerra, A. Sequeira, and J. Tiago, *Existence of optimal boundary control for the Navier–Stokes equations with mixed boundary conditions*, Portugaliae Mathematica, 72 (2015), pp. 267–283.

[31] T. Guerra, J. Tiago, and A. Sequeira, *Optimal control in blood flow simulations*, International Journal of Non-Linear Mechanics, 64 (2014), pp. 57–69.

[32] A. Günnel, R. Herzog, and E. Sachs, *A note on preconditioners and scalar products in Krylov subspace methods for self-adjoint problems in Hilbert space*, Electronic Transactions on Numerical Analysis, 41 (2014), pp. 13–20.

[33] J. Hadamard, *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton University Bulletin, 13 (1902), pp. 49–52.

[34] N. Hale, N. J. Higham, and L. N. Trefethen, *Computing aˆα,\log(a), and related matrix functions by contour integrals*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2505–2523.

[35] H. Harbrecht, F. Paiva, C. Pérez, and R. Schneider, *Multiscale preconditioning for the coupling of FEM–BEM*, Numerical linear algebra with applications, 10 (2003), pp. 197–222.

[36] R. Herzog and E. Sachs, *Superlinear convergence of Krylov subspace methods for self-adjoint problems in Hilbert space*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 1304–1324.

[37] J. Heywood, R. Rannacher, and S. Tureks, *Artificial boundary and flux and pressure conditions for the incompressible Navier–Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.

[38] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Computational Optimization and Applications, 30 (2005), pp. 45–61.

[39] L. J. John, *Optimal boundary Control in energy spaces: Preconditioning and applications*, Verlag der Technischen Universität, 2014.

[40] M. Kollmann and M. Kolmbauer, *A preconditioned minres solver for time-periodic parabolic optimal control problems*, Numerical Linear Algebra with Applications, 20 (2013), pp. 761–784.

[41] M. Kuchta, K.-A. Mardal, and M. Mortensen, *On preconditioning saddle point systems with trace constraints coupling 3d and 1d domains–applications to matching and nonmatching fem discretizations*, arXiv preprint arXiv:1612.03574, (2016).

[42] T. Lassila, A. Manzoni, A. Quarteroni, and G. Rozza, *A reduced computational and geometrical framework for inverse problems in hemodynamics*, International journal for numerical methods in biomedical engineering, 29 (2013), pp. 741–776.

[43] H. Lee, *Optimal control for quasi-Newtonian flows with defective boundary conditions*, Computer Methods in Applied Mechanics and Engineering, 200 (2011), pp. 2498–2506.

[44] J. L. Lions and E. Magenes, *Non-homogeneous boundary value problems and applications, vol. 1*, vol. 181 of Grundlehren der matematischen Wissenschaften, Springer, 1972.

[45] A. Logg, K.-A. Mardal, and G. Wells, *Automated solution of differential equations by the finite element method: The FEniCS book*, vol. 84 of Lecture Notes in Computational Science and Engineering, Springer Science & Business Media, 2012.

[46] F. Magoulès and F.-X. Roux, *Lagrangian formulation of domain decomposition methods: A unified theory*, Applied Mathematical Modelling, 30 (2006), pp. 593–615.

[47] A. Manzoni, T. Lassila, A. Quarteroni, and G. Rozza, *A reduced-order strategy for solving inverse bayesian shape identification problems in physiological flows*, in Modeling, Simulation and Optimization of Complex Processes-HPSC 2012, Springer, 2014, pp. 145–155.

[48] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numerical Linear Algebra with Applications, 18 (2011), pp. 1–40.

[49] B. F. Nielsen and K.-A. Mardal, *Analysis of the minimal residual method applied to ill-posed optimality systems*, SIAM Journal on Scientific Computing, 35 (2013), pp. A785–A814.

[50] J. NOCEDAL AND S. J. WRIGHT, *Sequential quadratic programming*, Springer Series in Operations Research and Financial Engineering, Springer Science+Business Media, second ed., 2006.

[51] P. OSWALD, *Multilevel finite element approximation: Theory and applications*, Teubner, 1994.

[52] J. W. PEARSON, *Preconditioned iterative methods for Navier–Stokes control problems*, Journal of Computational Physics, 292 (2015), pp. 194–207.

[53] 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 (2012), pp. 1126–1152.

[54] ——, *Preconditioners for state-constrained optimal control problems with Moreau–Yosida penalty function*, Numerical Linear Algebra with Applications, 21 (2014), pp. 81–97.

[55] J. W. PEARSON AND A. J. WATHEN, *Fast iterative solvers for convection-diffusion control problems*, Electronic Transactions on Numerical Analysis, 40 (2013), pp. 294–310.

[56] M. PEREGO, A. VENEZIANI, AND C. VERGARA, *A variational approach for estimating the compliance of the cardiovascular tissue: An inverse fluid-structure interaction problem*, SIAM Journal on Scientific Computing, 33 (2011), pp. 1181–1211.

[57] J. PITKÄRANTA, *Boundary subspaces for the finite element method with Lagrange multipliers*, Numerische Mathematik, 33 (1979), pp. 273–289.

[58] A. QUARTERONI AND G. ROZZA, *Optimal control and shape optimization of aorto-coronaric bypass anastomoses*, Mathematical Models and Methods in Applied Sciences, 13 (2003), pp. 1801–1823.

[59] S. RAMALHO, A. MOURA, A. GAMBARUTO, AND A. SEQUEIRA, *Sensitivity to outflow boundary conditions and level of geometry description for a cerebral aneurysm*, International Journal for Numerical Methods in Biomedical Engineering, 28 (2012), pp. 697–713.

[60] S. RAMALHO, A. B. MOURA, A. M. GAMBARUTO, AND A. SEQUEIRA, *Influence of blood rheology and outflow boundary conditions in numerical simulations of cerebral aneurysms*, in Mathematical Methods and Models in Biomedicine, Springer, 2013, pp. 149–175.

[61] 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.

[62] A. SCHIELA AND S. ULBRICH, *Operator preconditioning for a class of inequality constrained optimal control problems*, SIAM Journal on Optimization, 24 (2014), pp. 435–466.

[63] J. SCHÖBERL AND W. ZULEHNER, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 752–773.

[64] D. A. STEINMAN, *Assumptions in modelling of large artery hemodynamics*, in Modeling of physiological flows, Springer, 2012, pp. 1–18.

[65] M. STOLL AND A. WATHEN, *Preconditioning for partial differential equation constrained optimization with control constraints*, Numerical Linear Algebra with Applications, 19 (2012), pp. 53–71.

[66] ——, *All-at-once solution of time-dependent stokes control*, Journal of Computational Physics, 232 (2013), pp. 498–515.

[67] J. TIAGO, A. GAMBARUTO, AND A. SEQUEIRA, *Patient-specific blood flow simulations: setting dirichlet boundary conditions for minimal error with respect to measured data*, Mathematical Modelling of Natural Phenomena, 9 (2014), pp. 98–116.

[68] ——, *Patient-specific blood flow simulations: Setting dirichlet boundary conditions for minimal error with respect to measured data*, Mathematical Modelling of Natural Phenomena, 9 (2014), pp. 98–116.

[69] A. N. TIKHONOV, A. GONCHARSKY, V. STEPANOV, AND A. G. YAGOLA, *Numerical methods for the solution of ill-posed problems*, vol. 328, Springer Science & Business Media, 2013.

[70] S. P. TIMOSHENKO AND J. M. GERE, *Theory of elastic stability*, Engineering societies monographs, McGraw-Hill, 1961.

[71] H. TRIEBEL, *Interpolation theory, function spaces, differential operators*, Johann Ambrosius Barth, second ed., 1995.

[72] A. VENEZIANI AND C. VERGARA, *An approximate method for solving incompressible Navier–Stokes problems with flow rate conditions*, Computer methods in applied mechanics and engineering, 196 (2007), pp. 1685–1700.

[73] A. ŽENÍŠEK, *Polynomial approximation on tetrahedrons in the finite element method*, Journal of Approximation Theory, 7 (1973), pp. 334–351.

[74] W. ZULEHNER, *Nonstandard norms and robust estimates for saddle point problems*, SIAM Journal on Matrix Analysis and Applications, 32 (2011), pp. 536–560.

Papers

Paper I **K.-A. Mardal, B. F. Nielsen, M. Nordaas**
Robust preconditioners for PDE-constrained optimization with limited observations
BIT Numerical Mathematics (2016)

Paper II **M. Kuchta, M., Nordaas, J. C. G. Verschaeve, M. Mortensen, K.-A. Mardal**
Preconditioners for saddle point systems with trace constraints coupling 2D and 1D domains
SIAM Journal on Scientific Computing (2016)

Paper III **S. Funke, M. Nordaas, Ø. Evju, M.S. Alnæs, K.-A. Mardal**
Variational data assimilation for transient blood flow simulations
Subitted to Journal for Numerical Methods in Biomedical Engineering (2017)

Paper I

Robust preconditioners for PDE-constrained optimization with limited observations

BIT Numer Math
DOI 10.1007/s10543-016-0635-8

Robust preconditioners for PDE-constrained optimization with limited observations

Kent-André Mardal[1,2] · **Bjørn Fredrik Nielsen**[3] · **Magne Nordaas**[1]

Received: 6 October 2015 / Accepted: 3 October 2016

Abstract Regularization robust preconditioners for PDE-constrained optimization problems have been successfully developed. These methods, however, typically assume observation data and control throughout the entire domain of the state equation. For many inverse problems, this is an unrealistic assumption. In this paper we propose and analyze preconditioners for PDE-constrained optimization problems with limited observation data, e.g. observations are only available at the boundary of the solution domain. Our methods are robust with respect to both the regularization parameter and the mesh size. That is, the condition number of the preconditioned optimality system is uniformly bounded, independently of the size of these two parameters. The method does, however, require extra regularity. We first consider a prototypical elliptic control problem and thereafter more general PDE-constrained optimization problems. Our theoretical findings are illuminated by several numerical results.

Communicated by Rolf Stenberg.

✉ Magne Nordaas
magneano@simula.no

Kent-André Mardal
kent-and@simula.no

Bjørn Fredrik Nielsen
bjorn.f.nielsen@nmbu.no

[1] Center for Biomedical Computing, Simula Research Laboratory, Fornebu, Norway

[2] Department of Mathematics, University of Oslo, Oslo, Norway

[3] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway

Published online: 23 November 2016

Keywords PDE-constrained optimization · Preconditioning · Minimum residual method

Mathematics Subject Classification 65F08 · 65N21 · 65K10

1 Introduction

Consider the model problem:

$$\min_{f,u} \left\{ \frac{1}{2} \|u - d\|^2_{L^2(\partial\Omega)} + \frac{\alpha}{2} \|f\|^2_{L^2(\Omega)} \right\}, \tag{1}$$

on a Lipschitz domain $\Omega \subset \mathbb{R}^n$, subject to

$$-\Delta u + u + f = 0 \quad \text{in } \Omega, \tag{2}$$

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega. \tag{3}$$

This minimization task is similar to the standard example considered in PDE-constrained optimization. But instead of assuming that observation data is available everywhere in Ω, we consider the case where observations are only given at the boundary $\partial\Omega$ of Ω, that is $d \in L^2(\partial\Omega)$, see the first term in (1). For problems of the form (1)–(3), in which the objective functional is replaced by

$$\frac{1}{2} \|u - d\|^2_{L^2(\Omega)} + \frac{\alpha}{2} \|f\|^2_{L^2(\Omega)} \tag{4}$$

very efficient preconditioners have been developed for the associated KKT system. In fact, by employing proper α-dependent scalings of the involved Hilbert spaces [14], or by using a Schur complement approach [13], methods that are robust with respect to the size of the regularization parameter α have been developed. More specifically, the condition number of the preconditioned optimality system is small and bounded independently of $0 < \alpha \ll 1$ and the mesh size h. This ensures good performance for suitable Krylov subspace methods, e.g. the minimum residual method (MINRES), independently of both parameters. These techniques have been extended to handle time dependent problems [12] and PDE-constrained optimization with Stokes equations [17], but the rigorous analysis of α-independent bounds always requires that observations are available throughout all of Ω.

For cases with limited observations, for example with cost-functionals of the form (1), efficient preconditioners are also available for a rather large class of PDE-constrained optimization problems, see [10,11]. But these techniques do not yield convergence rates, for the preconditioned KKT-system, that are completely robust with respect to the size of the regularization parameter α. Instead, the number of preconditioned MINRES iterations grows logarithmically[1] with respect to the size of α^{-1}, as $\alpha \to 0$:

$$a + b \log_{10}\left(\alpha^{-1}\right) \tag{5}$$

for constants a, b independent of α. According to the numerical experiments presented in [11], the size of b may become significant. More specifically, $b \in [5, 50]$ for problems with simple elliptic state equations posed on rectangles. Thus, for small values of α, MINRES may require rather many iterations to converge—even though the growth in iteration numbers is only logarithmic.

In practice, observations are rarely available throughout the entire domain of the state equation. On the contrary, the purpose of solving an inverse problem is typically to use data recorded at the surface of an object to compute internal properties of that object: Impedance tomography, the inverse problem of electrocardiography (ECG), computerized tomography (CT), etc. This fact, combined with the discussion above, motivate the need for further improving numerical methods for solving KKT systems arising in connection with PDE-constrained optimization.

This paper is organized as follows. In the next section we derive the KKT system associated with the model problem (1)–(3). Our α robust preconditioner is presented in Sect. 3, along with a number of numerical experiments. Sections 4 and 5 contain our analysis, and the method is generalized in Sects. 6 and 7. In Sect. 8 we discuss the preconditioner when applied to a standard finite element approximation of the problem. Section 9 provides a discussion of our findings, including their limitations.

2 KKT system

Consider the PDE (2) with the boundary condition (3). A solution u to this elliptic PDE, with source term $f \in L^2(\Omega)$, is known to have improved regularity, i.e. $u \in H^s(\Omega)$, for some $s \in [1, 2]$, with s depending on the domain Ω. In the remainder of this paper we assume that the solution u is in $H^2(\Omega)$ for any source term $f \in L^2(\Omega)$. This assumption is known to hold if Ω is convex or if $\partial\Omega$ is C^2, see e.g. [5,7].

When solutions to (2) exhibit $H^2(\Omega)$-regularity, we can write the problem on the non-standard variational form: Find $u \in \bar{H}^2(\Omega)$ such that

$$(-\Delta u + u, w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall w \in L^2(\Omega), \tag{6}$$

where

$$\bar{H}^2(\Omega) = \left\{ \phi \in H^2(\Omega) \,\middle|\, \frac{\partial \phi}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \right\},$$

[1] In [10,11] it is proved that the number of needed preconditioned MINRES iterations cannot grow faster than

$$a + b\left[\log_{10}\left(\alpha^{-1}\right)\right]^2.$$

Furthermore, in [11] it is explained why iterations counts of the kind (5) often will occur in practice.

equipped with the inner product

$$\begin{aligned}(u,v)_{H^2(\Omega)} &= \int_\Omega \nabla^2 u : \nabla^2 v + \nabla u \cdot \nabla v + uv\,dx \\ &= \int_\Omega \Delta u \Delta v + \nabla u \cdot \nabla v + uv\,dx. \end{aligned} \tag{7}$$

Here $\nabla^2 u$ denotes the Hessian of u, and the second identity is due to the boundary condition $\frac{\partial u}{\partial \mathbf{n}} = 0$ imposed on the space $\bar{H}^2(\Omega)$.

We will see below that, in order to design a regularization robust preconditioner for (1)–(3), it is convenient to express the state equation in the form (6), instead of employing integration by parts/Green's formula to write it on the standard self-adjoint form.

2.1 Optimality system

We may express (1)–(3) in the form:

$$\min_{f\in L^2(\Omega),\, u\in \bar{H}^2(\Omega)} \left\{ \frac{1}{2}\|u-d\|^2_{L^2(\partial\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega)} \right\} \tag{8}$$

subject to

$$(-\Delta u + u, w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall w \in L^2(\Omega). \tag{9}$$

The associated Lagrangian reads

$$\mathcal{L}(f,u,w) = \frac{1}{2}\|u-d\|^2_{L^2(\partial\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega)} + (f - \Delta u + u, w)_{L^2(\Omega)},$$

with $f \in L^2(\Omega)$, $u \in \bar{H}^2(\Omega)$ and $w \in L^2(\Omega)$. From the first order optimality conditions

$$\frac{\partial \mathcal{L}}{\partial f} = 0, \quad \frac{\partial \mathcal{L}}{\partial u} = 0, \quad \frac{\partial \mathcal{L}}{\partial w} = 0,$$

we obtain the optimality system: determine $(f,u,w) \in L^2(\Omega) \times \bar{H}^2(\Omega) \times L^2(\Omega)$ such that

$$\alpha(f,\psi)_{L^2(\Omega)} + (\psi, w)_{L^2(\Omega)} = 0 \quad \forall \psi \in L^2(\Omega), \tag{10}$$

$$(u-d,\phi)_{L^2(\partial\Omega)} + (-\Delta\phi + \phi, w)_{L^2(\Omega)} = 0 \quad \forall \phi \in \bar{H}^2(\Omega), \tag{11}$$

$$(f,\xi)_{L^2(\Omega)} + (-\Delta u + u, \xi)_{L^2(\Omega)} = 0 \quad \forall \xi \in L^2(\Omega). \tag{12}$$

3 Numerical experiments

Prior to analyzing our model problem, we will consider some numerical experiments. Discretization of (10)–(12) yields an algebraic system of the form

$$\underbrace{\begin{bmatrix} \alpha M & 0 & M \\ 0 & M_\partial & A^T \\ M & A & 0 \end{bmatrix}}_{\mathcal{A}_\alpha} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{M}_\partial d \\ 0 \end{bmatrix}, \tag{13}$$

where M is a mass matrix, the discretization of the $L^2(\Omega)$ inner product. M_∂ is a mass matrix associated with the boundary $\partial\Omega$ of Ω. A is a matrix that arises upon discretization of the operator $(1-\Delta)$. Since we write the state equation on a non self-adjoint form, A will not be the usual sum of the stiffness and mass matrices. Instead, Eq. (6) is discretized with subspaces of $\bar{H}^2(\Omega)$ and $L^2(\Omega)$. Consequently, A will in general not be a square matrix.

In (13), we have implicitly used the same discretization for the control variable and the Lagrange multiplier. In (10)–(12), both variables belong to $L^2(\Omega)$, so it seems natural to preserve this correspondence in the discretization. In fact, we can see from (10) that $f = -\alpha^{-1}w$, so the control could be eliminated from the system prior to the discretization. This would result in a 2 × 2 block system in place of (13). While solving the smaller system is more practical in terms of computational costs, we find that the analysis is more clearly presented for the 3 × 3 system (13).

In the current numerical experiments, we employ the Bogner–Fox–Schmit (BFS) rectangle for discretizing the state variable $u \in \bar{H}^2(\Omega)$. That is, the finite element field consists of bicubic polynomials that are continuous, have continuous first order derivatives and mixed second order derivatives at each vertex of the mesh. BFS elements are C^1 on rectangles and therefore H^2-conforming. The control f and Lagrange multiplier w are discretized with discontinuous bicubic elements.

We propose to precondition (13) with the block-diagonal matrix

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha M & 0 & 0 \\ 0 & \alpha R + M_\partial & 0 \\ 0 & 0 & \frac{1}{\alpha} M \end{bmatrix}^{-1}, \tag{14}$$

where R results from a discretization of the bilinear form $b(\cdot,\cdot)$ on $\bar{H}^2(\Omega)$:

$$b(u, v) = (u, v)_{H^2(\Omega)} + \int_\Omega \nabla u \cdot \nabla v \, dx. \tag{15}$$

In the experiments presented below, we used this bilinear form to construct a multigrid approximation of $(\alpha R + M_\partial)^{-1}$.

Remark The bilinear form (15) is equivalent to the inner product on $\bar{H}^2(\Omega)$. The additional term stems from our choice of implementing a multigrid algorithm for the bilinear form associated with the operator $(\Delta - 1)^2 = \Delta^2 - 2\Delta + 1$. Indeed, the bilinear form $\alpha b(\cdot\,,\,\cdot) + (\cdot\,,\,\cdot)_{L^2(\partial\Omega)}$ can be seen to coincide with the variational form associated with the fourth order problem

$$\alpha(\Delta - 1)^2 u = f \quad \text{in } \Omega, \tag{16}$$

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega, \tag{17}$$

$$\alpha\frac{\partial \Delta u}{\partial \mathbf{n}} = u \quad \text{on } \partial\Omega. \tag{18}$$

To limit the technical complexity of the implementation, we considered the problem (1)–(3) on the unit square in two dimensions. The experiments were implemented in Python and SciPy. The meshes were uniform rectangular, with the coarsest level for the multigrid solver consisting of 8 × 8 rectangles. Figure 1 shows an example of a solution of the optimality system (13).

3.1 Eigenvalues

Let us first consider the exact preconditioner $\mathcal{B}_\alpha$ defined in (14). If $\mathcal{B}_\alpha$ is a good preconditioner for the discrete optimality system (13), then the spectral condition number of $\mathcal{B}_\alpha\mathcal{A}_\alpha$ should be small and bounded, independently of the size of both the regularization parameter α and the discretization parameter h.

The eigenvalues of this preconditioned system were computed by solving the generalized eigenvalue problem

$$\mathcal{A}_\alpha x = \lambda \mathcal{B}_\alpha^{-1} x.$$

We found that the absolute value of the eigenvalues λ were bounded, with

$$0.445 \leq |\lambda| \leq 1.809,$$

uniformly in $\alpha \in \{1, 10^{-1}, \ldots, 10^{-10}\}$ and $h \in \{2^{-2}, \ldots, 2^{-5}\}$. This yields a uniform condition number $k(\mathcal{B}_\alpha\mathcal{A}_\alpha) \approx 4.05$. The spectra of the preconditioned systems are pictured in Fig. 2 for some choices of α. The spectra are clearly divided into three bounded intervals, and the eigenvalues are more clustered for $\alpha \approx 1$ and for very small α.

3.2 Multilevel preconditioning

In practice, the action of $\mathcal{B}_\alpha$ is replaced with a less computationally expensive operation $\widehat{\mathcal{B}_\alpha}$. Note that $\mathcal{B}_\alpha$ has a block structure, and that computationally efficient approximations can be constructed for the individual blocks. The only challenging block of the preconditioner is the biharmonic operator $\alpha R + M_\partial$. Order optimal multilevel algorithms for forth order operators discretized with the Bogner–Fox–Schmit was developed in [16]. Specifically, it was shown that a multigrid V-cycle using a symmetric 4 × 4 block Gauss–Seidel smoother, where the blocks contain the matrix entries corresponding to all degrees of freedom associate with a vertex in the mesh, results

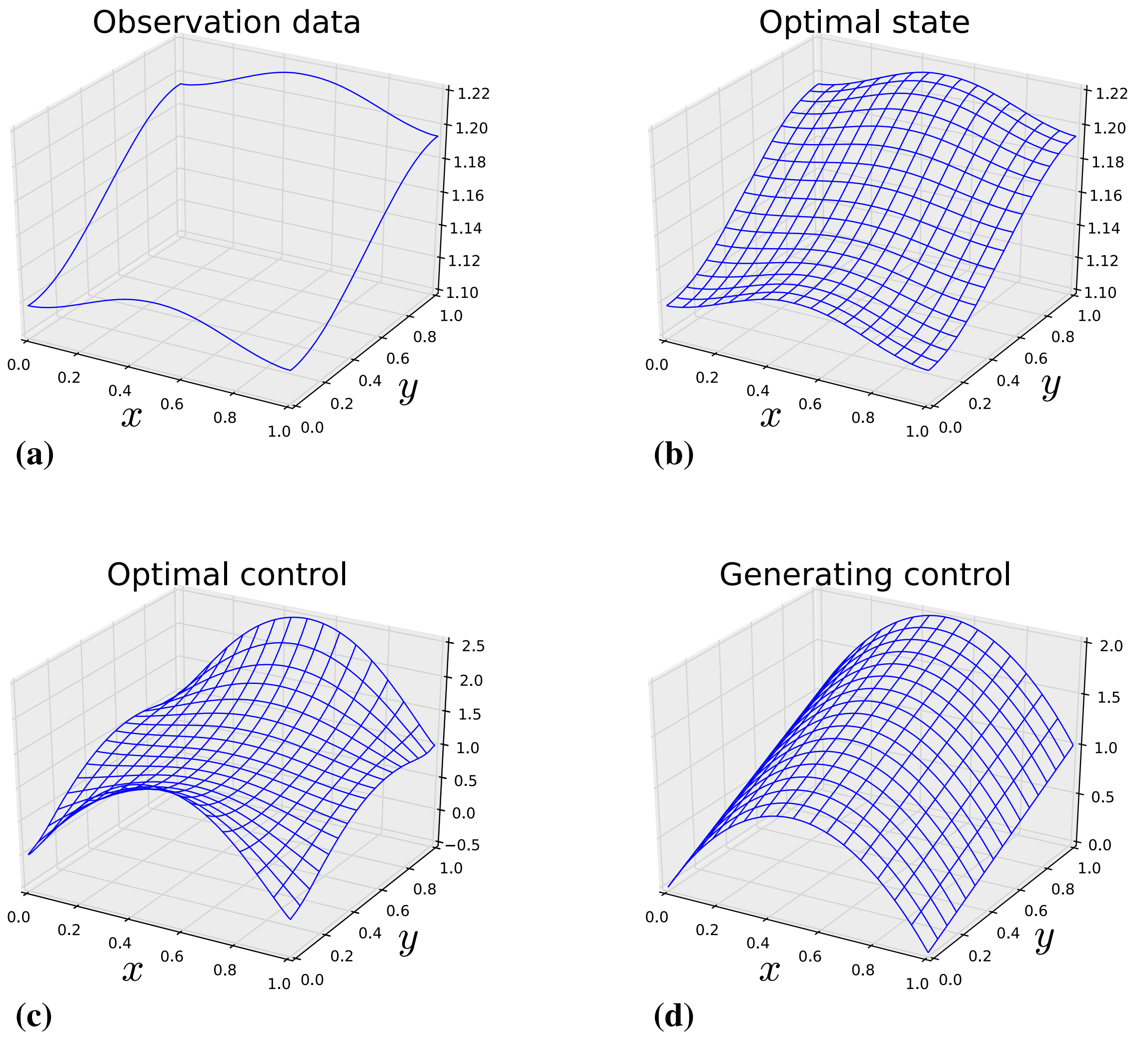


Fig. 1 An example of a solution of (13). The observation data d was generated with the forward model, using the "true" control $4x(1-x)+y$ shown in *panel* (**d**). Solutions to the unregularized problem are non-unique, and the generating control cannot be (exactly) recovered. The figures were generated with mesh parameter $h = 1/128$ and regularization parameter $\alpha = 10^{-6}$. **a** Observation data d. The forward model was solved for the control shown in **d**, but only the boundary values can be observed. **b** Computed optimal state u based on the observation data shown in **a**. **c** Computed optimal control f based on the observation data in **a**. **d** The "true" control function used to generate the observation data in **a**

in an order optimal approximation. The remaing blocks of the preconditioners are weighted mass matrices which are efficiently handled by two symmetric Gauss-Seidel iterations for the (1,1) and (3,3) blocks.

We estimated condition numbers of the individual blocks of $\mathcal{B}_\alpha^{-1}$ preconditioned with their respective approximations. The results are reported in Tables 1 and 2. A slight deterioration in the performance of the multigrid cycle can be seen for very small values of $\alpha > 0$.

3.3 Iteration numbers

To verify that also $\widehat{\mathcal{B}_\alpha}$ is an effective preconditioner for $\mathcal{A}_\alpha$, we applied the MINRES scheme to the system

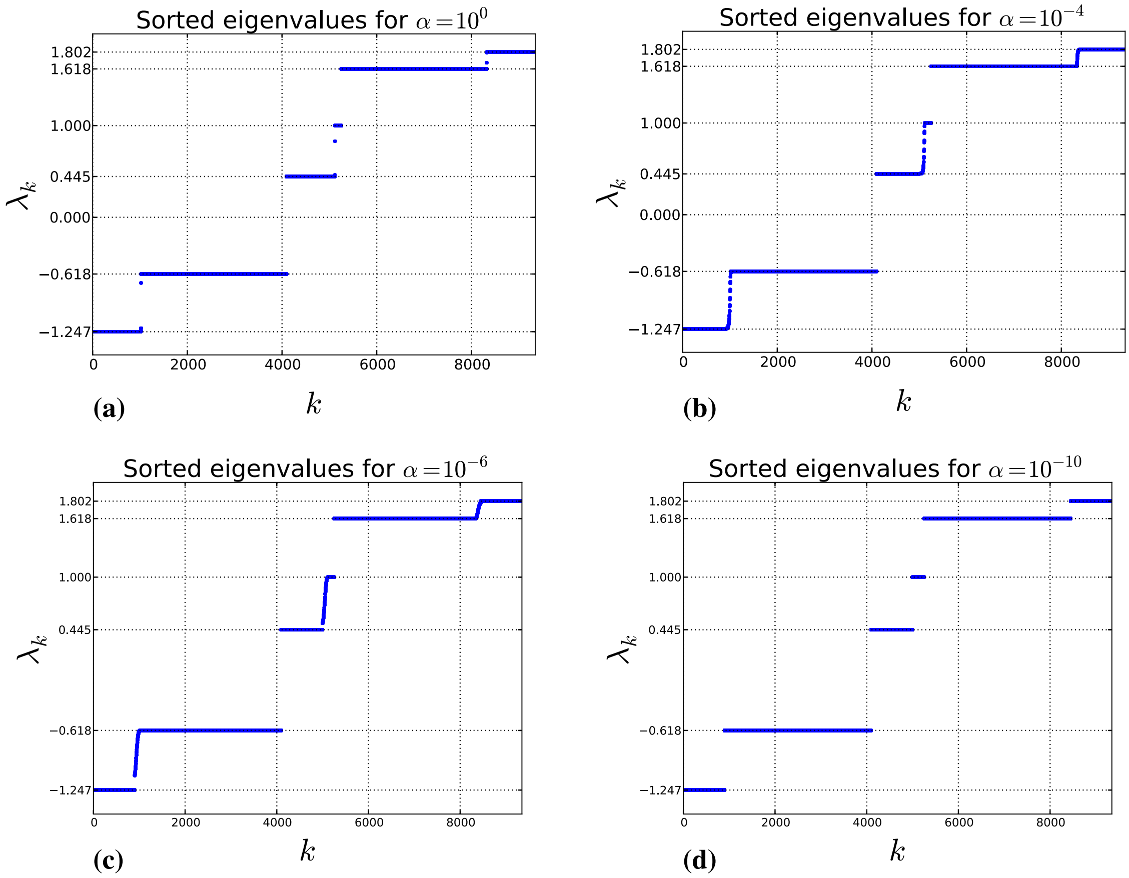


Fig. 2 Spectrum of $\mathcal{B}_\alpha \mathcal{A}_\alpha$ for different regularization parameters α. The discretization parameter was $h = 2^{-4}$ for all figures

Table 1 Condition numbers of M preconditioned with symmetric Gauss–Seidel iterations

Iterations	1	2	3
$(h = 2^{-8})$	1.931	1.303	1.126

$$\widehat{\mathcal{B}_\alpha} \mathcal{A}_\alpha x = \widehat{\mathcal{B}_\alpha} b.$$

For the results presented in Table 3, the MINRES iteration process was stopped as soon as

$$\frac{(r_k, \widehat{\mathcal{B}_\alpha} r_k)}{(r_0, \widehat{\mathcal{B}_\alpha} r_0)} = \frac{(\mathcal{A}_\alpha x_k - b, \widehat{\mathcal{B}_\alpha}\{\mathcal{A}_\alpha x_k - b\})}{(\mathcal{A}_\alpha x_0 - b, \widehat{\mathcal{B}_\alpha}\{\mathcal{A}_\alpha x_0 - b\})} \leq \varepsilon, \tag{19}$$

which is the standard termination criterion for the preconditioned MINRES scheme, provided that the preconditioner is SPD. A random initial guess x_0 was used, and the tolerance was set to $\varepsilon = 10^{-12}$.

4 Analysis of the KKT system

Recall that our optimality system reads:

$$\alpha(f, \psi)_{L^2(\Omega)} + (\psi, w)_{L^2(\Omega)} = 0 \quad \forall \psi \in L^2(\Omega),$$

Table 2 Estimated condition numbers of $\alpha R + M_\partial$ preconditioned with one V-cycle multigrid iteration

$\alpha \backslash h$	2^{-4}	2^{-6}	2^{-8}
1	1.130	1.136	1.140
10^{-4}	1.129	1.135	1.139
10^{-8}	1.237	1.150	1.149
10^{-12}	1.252	1.259	1.253

Table 3 Number of preconditioned MINRES iterations needed to solve the optimality system to a relative error tolerance $\varepsilon = 10^{-12}$

$\alpha \backslash h$	2^{-4}	2^{-5}	2^{-6}	2^{-7}
1	53(4.33)	53 (4.36)	53 (4.36)	53 (4.36)
10^{-1}	57 (4.31)	57 (4.34)	57 (4.35)	57 (4.35)
10^{-2}	75 (4.31)	72 (4.34)	70 (4.35)	68 (4.35)
10^{-3}	79 (4.31)	79 (4.34)	77 (4.35)	73 (4.35)
10^{-4}	81 (4.30)	81 (4.33)	79 (4.35)	77 (4.35)
10^{-5}	82 (4.33)	81 (4.33)	79 (4.35)	79 (4.35)
10^{-6}	81 (4.35)	79 (4.36)	79 (4.35)	81 (4.35)
10^{-7}	70 (4.35)	81 (4.37)	81 (4.36)	79 (4.35)
10^{-8}	62 (4.36)	70 (4.36)	79 (4.36)	81 (4.36)
10^{-9}	62 (4.36)	64 (4.37)	68 (4.37)	78 (4.36)
10^{-10}	62 (4.36)	63 (4.36)	64 (4.37)	67 (4.37)

Estimated condition numbers in parentheses, computed from conjugate gradient iterations on the normal equations for the preconditioned optimality system

$$
\begin{aligned}
(u-d,\phi)_{L^2(\partial\Omega)} + (-\Delta\phi + \phi, w)_{L^2(\Omega)} &= 0 \quad \forall \phi \in \bar{H}^2(\Omega),\\
(f,\xi)_{L^2(\Omega)} + (-\Delta u + u, \xi)_{L^2(\Omega)} &= 0 \quad \forall \xi \in L^2(\Omega),
\end{aligned}
$$

with unknowns $f \in L^2(\Omega)$, $u \in \bar{H}^2(\Omega)$ and $w \in L^2(\Omega)$. We may write this KKT system in the form:

Determine $(f, u, w) \in L^2(\Omega) \times \bar{H}^2(\Omega) \times L^2(\Omega)$ such that

$$
\underbrace{\begin{bmatrix} \alpha M & 0 & M' \\ 0 & M_\partial & A' \\ M & A & 0 \end{bmatrix}}_{\mathcal{A}_\alpha} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{M}_\partial d \\ 0 \end{bmatrix}, \tag{20}
$$

where

$$M : L^2(\Omega) \to L^2(\Omega)', \quad f \mapsto (f, \cdot)_{L^2(\Omega)}, \tag{21}$$

$$M_\partial : \bar{H}^2(\Omega) \to \bar{H}^2(\Omega)', \quad u \mapsto (u, \cdot)_{L^2(\partial\Omega)}, \tag{22}$$

$$\tilde{M}_\partial : L^2(\partial\Omega) \to \bar{H}^2(\Omega)', \quad d \mapsto (d, \cdot)_{L^2(\partial\Omega)}, \tag{23}$$

$$A : \bar{H}^2(\Omega) \to L^2(\Omega)', \quad u \mapsto (-\Delta u + u, \cdot)_{L^2(\Omega)}, \tag{24}$$

and the notation "′" is used to denote dual operators and dual spaces. In the rest of this paper, the symbols M, M_∂ and A will represent the mappings defined in (21), (22) and (24), respectively, and not (the associated) matrices, as was the case in Sect. 3. We believe that this mild ambiguity improves the readability of the present text.

By using standard techniques for saddle point problems, one can show that the system (20) satisfies the Brezzi conditions [1], provided that $\alpha > 0$. Therefore, for every $\alpha > 0$, this set of equations has a unique solution. Nevertheless, if the standard norms of $L^2(\Omega)$ and $H^2(\Omega)$ are employed in the analysis, then the constants in the Brezzi conditions will depend on α. More specifically, the constant in the coercivity condition will be of order $O(\alpha)$, and thus becomes very small for $0 < \alpha \ll 1$. This property is consistent with the ill posed nature of (1)–(3) for $\alpha = 0$, and makes it difficult to design α robust preconditioners for the algebraic system associated with (20).

Similar to the approach used in [9,10,14], we will now introduce weighted Hilbert spaces. The weights are constructed such that the constants appearing in the Brezzi conditions are independent of α. Thereafter, in Sect. 5, we will show how these scaled Hilbert spaces can be combined with simple maps to design α robust preconditioners for our model problem.

4.1 Weighted norms

Consider the α-weighted norms:

$$\|f\|^2_{L^2_\alpha(\Omega)} = \alpha\, \|f\|^2_{L^2(\Omega)}\,, \tag{25}$$

$$\|u\|^2_{H^2_\alpha(\Omega)} = \alpha\, \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)}\,, \tag{26}$$

$$\|w\|^2_{L^2_{\alpha^{-1}}(\Omega)} = \frac{1}{\alpha}\, \|w\|^2_{L^2(\Omega)}\,, \tag{27}$$

applied to the control f, the state u and the dual/Lagrange-multiplier w, respectively. Note that these norms become "meaningless" for $\alpha = 0$, but are well defined for positive α.

4.2 Brezzi conditions

We will now analyze the properties of

$$\mathcal{A}_\alpha : L^2_\alpha(\Omega) \times H^2_\alpha(\Omega) \times L^2_{\alpha^{-1}}(\Omega) \to L^2_\alpha(\Omega)' \times H^2_\alpha(\Omega)' \times L^2_{\alpha^{-1}}(\Omega)',$$

defined in (20). More specifically, we will show that the Brezzi conditions are satisfied with constants that do not depend on the size of the regularization parameter $\alpha > 0$. Note that we use the scaled Hilbert norms (25)–(27).

Lemma 1 *For all* $\alpha > 0$, *the following "inf-sup" condition holds:*

$$\inf_{w \in L^2_{\alpha^{-1}}(\Omega)} \sup_{(f,u) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \frac{(f, w)_{L^2(\Omega)} + (-\Delta u + u, w)_{L^2(\Omega)}}{\|(f, u)\|_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}} \geq 1.$$

Proof Note that $L^2_\alpha(\Omega)$ and $L^2_{\alpha^{-1}}(\Omega)$ contain the same functions, provided that $\alpha > 0$. Let $w \in L^2_{\alpha^{-1}}(\Omega)$ be arbitrary. By choosing $f = w$ and $u = 0$ we find that

$$\begin{aligned} &\sup_{(f,u) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \frac{(f, w)_{L^2(\Omega)} + (-\Delta u + u, w)_{L^2(\Omega)}}{\|(f, u)\|_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}} \\ &\geq \frac{(w, w)_{L^2(\Omega)}}{\|(w, 0)\|_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}} \\ &= \frac{\|w\|^2_{L^2(\Omega)}}{\sqrt{\alpha}\, \|w\|_{L^2(\Omega)} (\sqrt{\alpha})^{-1} \|w\|_{L^2(\Omega)}} \\ &= 1. \end{aligned}$$

Since $w \in L^2_{\alpha^{-1}}(\Omega)$ was arbitrary, this completes the proof. □

Expressed in terms of the operators that constitute $\mathcal{A}_\alpha$, Lemma 1 takes the form

$$\inf_{w \in L^2_{\alpha^{-1}}(\Omega)} \sup_{(f,u) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \frac{\langle Mf, w\rangle + \langle Au, w\rangle}{\|(f, u)\|_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}} \geq 1,$$

see (21) and (24).

Recall that we decided to write our state Eqs. (2)–(3) on the non-standard variational form (6). Throughout this paper we assume that problem (2)–(3) admits a unique solution $u \in \bar{H}^2(\Omega)$ for every $f \in L^2(\Omega)$, and that

$$\|u\|_{H^2(\Omega)} \leq c_1 \|f\|_{L^2(\Omega)}. \tag{28}$$

This assumption is valid if Ω is convex or if Ω has a C^2 boundary, see e.g. [5,7]. Inequality (28) is a key ingredient of the proof of our next lemma.

Lemma 2 *There exists a constant* c_2, *which is independent of* $\alpha > 0$, *such that*

$$\begin{aligned} \alpha \|f\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} &\geq c_2 \left(\alpha \|f\|^2_{L^2(\Omega)} + \alpha \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)}\right) \\ &= c_2 \|(f, u)\|^2_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \end{aligned}$$

for all $(f, u) \in L^2(\Omega) \times \bar{H}^2(\Omega)$ *such that*

$$(f, \phi)_{L^2(\Omega)} + (-\Delta u + u, \phi)_{L^2(\Omega)} = 0 \quad \forall \phi \in L^2(\Omega). \tag{29}$$

Proof If (f, u) satisfies (29), then

$$\|u\|_{H^2(\Omega)} \leq c_1 \|f\|_{L^2(\Omega)},$$

see the discussion of (28). Let $\theta = (1 + c_1^2)^{-1} \in (0, 1)$, and it follows that

$$\begin{aligned}\alpha \|f\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} &\geq \alpha\theta \|f\|^2_{L^2(\Omega)} + \alpha \frac{1-\theta}{c_1^2} \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} \\ &\geq \frac{1}{1+c_1^2} \left(\alpha \|f\|^2_{L^2(\Omega)} + \alpha \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} \right).\end{aligned}$$

□

This result may also be written in the form

$$\left\langle \begin{bmatrix} \alpha M & 0 \\ 0 & M_\partial \end{bmatrix} \begin{bmatrix} f \\ u \end{bmatrix}, \begin{bmatrix} f \\ u \end{bmatrix} \right\rangle \geq c_2 \|(f, u)\|^2_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)}$$

for all $(f, u) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)$ satisfying

$$Mf + Au = 0,$$

where M, M_∂ and A are the operators defined in (21), (22) and (24), respectively.

4.3 Boundedness

Having established that the Brezzi conditions hold, with constants that are independent of α, we next explore the boundedness of $\mathcal{A}_\alpha$.

Lemma 3 *For all* $(f, u), (\psi, \phi) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)$,

$$\left| \left\langle \begin{bmatrix} \alpha M & 0 \\ 0 & M_\partial \end{bmatrix} \begin{bmatrix} f \\ u \end{bmatrix}, \begin{bmatrix} \psi \\ \phi \end{bmatrix} \right\rangle \right| \leq \sqrt{2} \|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} \|(\psi, \phi)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)}.$$

Proof Recall the definitions (21) and (22) of M and M_∂, respectively. Since

$$\|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} = \sqrt{\alpha \|f\|^2_{L^2(\Omega)} + \alpha \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)}},$$

we find, by employing the Cauchy–Schwarz inequality, that

$$\begin{aligned}\left| \left\langle \begin{bmatrix} \alpha M & 0 \\ 0 & M_\partial \end{bmatrix} \begin{bmatrix} f \\ u \end{bmatrix}, \begin{bmatrix} \psi \\ \phi \end{bmatrix} \right\rangle \right| &= \left| \alpha(f, \psi)_{L^2(\Omega)} + (u, \phi)_{L^2(\partial\Omega)} \right| \\ &\leq \|f\|_{L^2_\alpha(\Omega)} \|\psi\|_{L^2_\alpha(\Omega)} + \|u\|_{L^2(\partial\Omega)} \|\phi\|_{L^2(\partial\Omega)} \\ &\leq \sqrt{2}\sqrt{\|f\|^2_{L^2_\alpha(\Omega)} \|\psi\|^2_{L^2_\alpha(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} \|\phi\|^2_{L^2(\partial\Omega)}} \\ &\leq \sqrt{2} \|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} \|(\psi, \phi)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)}.\end{aligned}$$

Lemma 4 *For all* $(f, u) \in L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)$, $w \in L^2_{\alpha^{-1}}(\Omega)$,

$$\left|\left\langle [M \;\; A] \begin{bmatrix} f \\ u \end{bmatrix}, w \right\rangle\right| \leq \sqrt{3}\, \|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}.$$

Proof Again, we note that

$$\begin{aligned} \|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} &= \sqrt{\alpha \|f\|^2_{L^2(\Omega)} + \alpha \|u\|^2_{H^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)}}, \\ \|w\|_{L^2_{\alpha^{-1}}(\Omega)} &= \frac{1}{\sqrt{\alpha}} \|w\|_{L^2(\Omega)}. \end{aligned}$$

From the definitions of M and A, see (21) and (24), and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} \left|\left\langle [M \;\; A] \begin{bmatrix} f \\ u \end{bmatrix}, w \right\rangle\right| &= |\langle Mf, w\rangle + \langle Au, w\rangle| \\ &= \left|(f, w)_{L^2(\Omega)} + (-\Delta u + u, w)_{L^2(\Omega)}\right| \\ &\leq \left(\|f\|_{L^2_\alpha(\Omega)} + \|\Delta u\|_{L^2_\alpha(\Omega)} + \|u\|_{L^2_\alpha(\Omega)}\right) \|w\|_{L^2_{\alpha^{-1}}(\Omega)} \\ &\leq \sqrt{3}\, \|(f, u)\|_{L^2_\alpha(\Omega)\times H^2_\alpha(\Omega)} \|w\|_{L^2_{\alpha^{-1}}(\Omega)}. \end{aligned}$$

For the last equality, recall from (7) that $\|\Delta u\|_{L^2(\Omega)} = \left\|\nabla^2 u\right\|_{L^2(\Omega)} \leq \|u\|_{H^2(\Omega)}$ for all $u \in \bar{H}^2(\Omega)$. □

4.4 Isomorphism

We have verified that the Brezzi conditions hold, and that $\mathcal{A}_\alpha$ is a bounded operator. Moreover, all constants appearing in the inequalities expressing these properties are independent of the regularization parameter $\alpha > 0$. Let

$$\mathcal{V} = L^2_\alpha(\Omega) \times H^2_\alpha(\Omega) \times L^2_{\alpha^{-1}}(\Omega), \tag{30}$$

$$\mathcal{V}' = L^2_\alpha(\Omega)' \times H^2_\alpha(\Omega)' \times L^2_{\alpha^{-1}}(\Omega)'. \tag{31}$$

Theorem 1 *The operator $\mathcal{A}_\alpha$, defined in* (20), *is bounded and continuously invertible for $\alpha > 0$ in the sense that for all nonzero $x \in \mathcal{V}$,*

$$c \leq \sup_{0\neq y\in\mathcal{V}} \frac{\langle \mathcal{A}_\alpha x, y\rangle}{\|y\|_{\mathcal{V}} \|x\|_{\mathcal{V}}} \leq C, \tag{32}$$

for some positive constants c and C that are independent of $\alpha > 0$. In particular,

$$\left\|\mathcal{A}_\alpha^{-1}\right\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} \leq c^{-1} \quad \text{and} \quad \|\mathcal{A}_\alpha\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \leq C.$$

Proof This result follows from Lemmas 1, 2, 3, and 4 and Brezzi theory for saddle point problems, see [1]. □

4.5 Estimates for the discretized problem

The stability properties (32) are not necessarily inherited by discretizations. However, the structure used to prove the so-called "inf-sup condition" in Lemma 1 is preserved in the discrete system provided that the same discretization is employed for the control and the Lagrange multiplier. Furthermore, the boundedness properties, Lemmas 3 and 4, certainly also hold for conforming discretizations.

It remains to adress the coercivity condition, Lemma 2, for the discretized problem. We consider finite dimensional subspaces $U_h \subset U = \bar{H}^2(\Omega)$ and $W_h \subset W = L^2(\Omega)$. For certain choices of U_h and W_h, the estimate of Lemma 2 carries over to the finite-dimensional setting.

Lemma 5 *Assume $U_h \subset U$ and $W_h \subset W$, such that $(1 - \Delta)U_h \subset W_h$. Then*

$$\alpha \left\| f_h \right\|^2_{L^2(\Omega)} + \left\| u_h \right\|^2_{L^2(\partial\Omega)} \geq c_2 \left\| (f_h, u_h) \right\|^2_{L^2_\alpha(\Omega) \times H^2_\alpha(\Omega)} \tag{33}$$

for all $(f_h, u_h) \in W_h \times U_h$ such that

$$(f_h, \phi_h)_{L^2(\Omega)} + (u_h - \Delta u_h, \phi_h)_{L^2(\Omega)} = 0 \quad \forall \phi_h \in W_h. \tag{34}$$

Proof Assume that $(1 - \Delta)U_h \subset W_h$, and that (34) holds for $(f_h, u_h) \in W_h \times U_h$. Then $f_h + (1 - \Delta)u_h \in W_h$, and (34) implies $f_h + (1 - \Delta)u_h = 0$. Therefore, (f_h, u_h) satisfies (29) and the estimate (33) follows from Lemma 2. □

If the discretization is chosen such that Lemma 5 is satisfied, then the estimates (32) carries over to discretized system. More precisely, we have

$$\left\| \mathcal{A}_{\alpha,h} \right\|_{\mathcal{L}(\mathcal{V}_h, \mathcal{V}_h')} \leq \left\| \mathcal{A}_\alpha \right\|_{\mathcal{L}(\mathcal{V}, \mathcal{V}')}, \quad \text{and} \quad \left\| \mathcal{A}^{-1}_{\alpha,h} \right\|_{\mathcal{L}(\mathcal{V}_h', \mathcal{V}_h)} \leq \left\| \mathcal{A}^{-1}_\alpha \right\|_{\mathcal{L}(\mathcal{V}', \mathcal{V})}, \tag{35}$$

where $\mathcal{V}_h = W_h \times U_h \times W_h \subset \mathcal{V}$, equipped with the inner prdocut of $\mathcal{V}$, and $\mathcal{A}_{\alpha,h}$ is discrete counterpart to $\mathcal{A}_\alpha$, defined by setting $\langle \mathcal{A}_{\alpha,h} x_h, y_h \rangle = \langle \mathcal{A}_\alpha x_h, y_h \rangle$ for all $x_h, y_h \in \mathcal{V}_h$.

If the state is discretized with C^1-conforming bicubic Bogner–Fox–Schmit rectangles, as in Sect. 3, then Lemma 5 is satisfied if the control and Lagrange multiplier is discretized with discontinuous bicubic elements on the same mesh. For triangular meshes, one could choose Argyris triangles for the state variable and piecewise quintic polynomials for the control and Lagrange multiplier variables.

We remark that Lemma 5 provides a sufficient, but not necessary criterion for stability of the discrete problem, and usually may imply far more degrees of freedom in the discrete space $W_h \subset W$ than is actually needed. The usefulness of Lemma 5 is that the estimates (35) can, in principle, always be obtained by choosing a sufficiently large space for the control and Lagrange multiplier.

5 Preconditioning

The linear problem (20) is of the form

$$\mathcal{A}x = b. \tag{36}$$

where x is sought in a Hilbert space $\mathcal{V}$, the right hand side b is in the dual space $\mathcal{V}'$, and $\mathcal{A}$ is a self-adjoint continuous mapping of $\mathcal{V}$ onto $\mathcal{V}'$. Iterative methods for linear problems are most often formulated for operators mapping $\mathcal{V}$ into itself, and can not be directly applied to the linear system (36), as described in [9]. If we want to apply such methods to (36), then we need to introduce a continuous operator mapping $\mathcal{V}'$ isomorphically back onto $\mathcal{V}$. More precisely, if we have a continuous operator

$$\mathcal{B} : \mathcal{V}' \to \mathcal{V},$$

then $\mathcal{M} = \mathcal{B}\mathcal{A} : \mathcal{V} \to \mathcal{V}$ is continuous and has the desired mapping properties, and if $\mathcal{B}$ is an isomorphism, the solutions to (36) coincides with the solutions to the problem

$$\mathcal{M}x = \mathcal{B}\mathcal{A}x = \mathcal{B}b. \tag{37}$$

In this paper we shall consider $\mathcal{B} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ a preconditioner if $\mathcal{B}$ is self-adjoint and positive definite. This implies that $\mathcal{B}^{-1}$ is self-adjoint and positive definite as well, and hence $\mathcal{B}^{-1}$ defines an inner product on $\mathcal{V}$ by setting

$$(x, y) = \left\langle \mathcal{B}^{-1}x, y \right\rangle, \qquad x, y \in \mathcal{V}. \tag{38}$$

This inner product has the crucial property of making $\mathcal{M}$ self-adjoint, in the sense that

$$(\mathcal{M}x, y) = \langle \mathcal{A}x, y \rangle = \langle \mathcal{A}y, x \rangle = (\mathcal{M}y, x). \tag{39}$$

Conversely, given any inner product on $(\cdot, \cdot)$ on $\mathcal{V}$, the Riesz–Fréchet theorem provides a self-adjoint positive definite isomorphism $\mathcal{B} : \mathcal{V}' \to \mathcal{V}$ such that (38) and (39) hold, and we say that $\mathcal{B}$ is the Riesz operator induced by $(\cdot, \cdot)$. This establishes a one-to-one correspondence between preconditioners and Riesz operators on $\mathcal{V}'$. Since the Riesz operator is an isometric isomorphism, the operator norm of $\mathcal{B}\mathcal{A}$ coincides with the operator norm of $\mathcal{A}$. We formulate this well-known fact here in a lemma for the sake of self-containedness. We refer to [6,9] for a more in-depth discussion of preconditioning and its relation to Riesz operators.

Lemma 6 *Let $\mathcal{V}$ be a Hilbert space, and let $\mathcal{A} : \mathcal{V} \to \mathcal{V}'$ be a self-adjoint isomorphism, and assume that $\mathcal{B}$ is the Riesz operator induced by the inner product on $\mathcal{V}$, or equivalently, that the inner product on $\mathcal{V}$ is defined by the self-adjoint positive definite isomorphism $\mathcal{B}^{-1} : \mathcal{V} \to \mathcal{V}'$. Then $\mathcal{B}\mathcal{A} : \mathcal{V} \to \mathcal{V}$ is an isomorphism, self-adjoint in the inner product on $\mathcal{V}$, with*

$$\|\mathcal{B}\mathcal{A}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} = \|\mathcal{A}\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \quad \textit{and} \quad \left\|(\mathcal{B}\mathcal{A})^{-1}\right\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} = \left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})}.$$

In particular, the condition number of $\mathcal{B}\mathcal{A}$ is given by

$$\kappa(\mathcal{B}\mathcal{A}) = \left\| \mathcal{A}^{-1} \right\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} \|\mathcal{A}\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} .$$

Proof Since $\mathcal{A}$ is self-adjoint, $\mathcal{M} = \mathcal{B}\mathcal{A}$ is self-adjoint with respect to the inner product on $\mathcal{V}$. From the Riesz–Fréchet theorem we have $\|\mathcal{A}x\|_{\mathcal{V}'} = \|\mathcal{B}\mathcal{A}x\| = \|\mathcal{M}x\|$, and we obtain following identity for the operator norm of $\mathcal{M}$.

$$\begin{aligned}\|\mathcal{M}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} &= \sup_{x\neq 0} \frac{\|\mathcal{M}x\|_{\mathcal{V}}}{\|x\|_{\mathcal{V}}} = \sup_{x\neq 0} \frac{\|\mathcal{A}x\|_{\mathcal{V}'}}{\|x\|_{\mathcal{V}}} \\ &= \sup_{x\neq 0} \sup_{y\neq 0} \frac{\langle \mathcal{A}x, y\rangle}{\|x\|_{\mathcal{V}} \|y\|_{\mathcal{V}}} = \|\mathcal{A}\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} .\end{aligned}$$

A similar identity is obtained for the norm of the inverse operator,

$$\begin{aligned}\left\|\mathcal{M}^{-1}\right\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} &= \sup_{x\neq 0} \frac{\left\|\mathcal{M}^{-1}x\right\|_{\mathcal{V}}}{\|x\|_{\mathcal{V}}} \\ &= \left(\inf_{x\neq 0} \frac{\|\mathcal{M}x\|_{\mathcal{V}}}{\|x\|_{\mathcal{V}}} \right)^{-1} \\ &= \left(\inf_{x\neq 0} \sup_{y\neq 0} \frac{\langle \mathcal{A}x, y\rangle}{\|x\|_{\mathcal{V}} \|y\|_{\mathcal{V}}} \right)^{-1} = \left\| \mathcal{A}^{-1} \right\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} .\end{aligned}$$

We say that a preconditioner $\mathcal{B}_\alpha$ for $\mathcal{A}_\alpha$ is robust with respect to the parameter α if $\kappa(\mathcal{B}_\alpha\mathcal{A}_\alpha)$ is bounded uniformly in α. The significance of Lemma 6 is that such a robust preconditioner can be found by identifying (parameter-dependent) norms in which $\mathcal{A}_\alpha$ and $\mathcal{A}_\alpha^{-1}$ are both uniformly bounded.

5.1 Parameter-robust minimum residual method

In Sect. 4 stability of $\mathcal{A}_\alpha$ was shown in the α-dependent norms defined in (25)–(27). The preconditioner provided by Lemma 6 is the Riesz operator induced by the weighted norms. This operator $\mathcal{B}_\alpha : \mathcal{V}' \to \mathcal{V}$ takes the form

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha M & 0 & 0 \\ 0 & \alpha R + M_\partial & 0 \\ 0 & 0 & \frac{1}{\alpha} M \end{bmatrix}^{-1} \tag{40}$$

where $R : \bar{H}^2(\Omega) \to \bar{H}^2(\Omega)'$ is the operator induced by the $H^2(\Omega)$ inner product, i.e. $\langle Ru, v\rangle = (u, v)_{H^2(\Omega)}$.

Since $\mathcal{A}_\alpha$ is self-adjoint, the preconditioned operator $\mathcal{B}_\alpha\mathcal{A}_\alpha : \mathcal{V} \to \mathcal{V}$ is self-adjoint in the inner product on $\mathcal{V}$. Consequently we can apply the minimum residual method (MINRES) to the problem

$$\mathcal{B}_\alpha \mathcal{A}_\alpha x = \mathcal{B}_\alpha b.$$

Theorem 2 *Let $\mathcal{A}_\alpha$ be the operator defined in (20) and $\mathcal{B}_\alpha$ the operator defined in (40). Then there exists an upper bound, independent of α, for the convergence rate of* MINRES *applied to the preconditioned system*

$$\mathcal{B}_\alpha \mathcal{A}_\alpha x = \mathcal{B}_\alpha b.$$

In particular there exists an upper bound, independent of α, for the number of iterations needed to reach the stopping criterion (19).

Proof A crude upper bound for the convergence rate (more precisely, the two-step convergence rate) of MINRES is given by

$$\|\mathcal{B}_\alpha \mathcal{A}_\alpha (x - x_{2m})\|_{\mathcal{V}} \leq \left(\frac{1-\kappa}{1+\kappa} \right)^m \|\mathcal{B}_\alpha \mathcal{A}_\alpha (x - x_0)\|_{\mathcal{V}}$$

where $\kappa = \kappa(\mathcal{B}_\alpha \mathcal{A}_\alpha)$ is the condition number of $\mathcal{B}_\alpha \mathcal{A}_\alpha$, see e.g. [9]. From Lemma 6 and (32) we determine that κ is bounded independently of α, with

$$\begin{aligned} \kappa &= \left\| (\mathcal{B}_\alpha \mathcal{A}_\alpha)^{-1} \right\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} \|\mathcal{B}_\alpha \mathcal{A}_\alpha\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} \\ &= \left\| \mathcal{A}_\alpha^{-1} \right\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} \|\mathcal{A}_\alpha\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \\ &\leq c^{-1} C. \end{aligned} \tag{41}$$

□

In practical applications, the operator $\mathcal{B}_\alpha$ will be replaced with a less computationally expensive approximation $\widehat{\mathcal{B}_\alpha}$. Ideally $\widehat{\mathcal{B}_\alpha}$ will be spectrally equivalent to $\mathcal{B}_\alpha$, in the sense that the condition number of $\widehat{\mathcal{B}_\alpha} \mathcal{B}_\alpha^{-1}$ is bounded, independently of α. Then the preconditioned system reads

$$\widehat{\mathcal{B}_\alpha} \mathcal{A}_\alpha x = \widehat{\mathcal{B}_\alpha} b,$$

and the upper bound for the convergence rate is determined by the conditioned number $\kappa(\widehat{\mathcal{B}_\alpha} \mathcal{A}_\alpha) \leq \kappa(\widehat{\mathcal{B}_\alpha} \mathcal{B}_\alpha^{-1}) \kappa(\mathcal{B}_\alpha \mathcal{A}_\alpha^{-1})$.

Remark In this paper we only consider the minimum residual method, and we therefore require that the preconditioner is self-adjoint and positive definite. More generally, if other Krylov subspace methods are to be applied to (20), then preconditioners lacking symmetry or definiteness may be considered.

We mention in particular that a preconditioned conjugate gradient method for problems similar to (20) was proposed in [14], based on a clever choice of inner product.

6 Generalization

Is our technique applicable to other problems than (1)–(3)? We will now briefly explore this issue, and show that the preconditioning scheme derived above yields α robust methods for a class of problems.

The scaling (25)–(27) was also investigated in [10], but for a family of abstract problems posed in terms of Hilbert spaces. More specifically, for general PDE-constrained optimization problems, subject to Tikhonov regularization, and with linear state equations. But in [10] no assumptions about the control, state or observation spaces were made, except that they were Hilbert spaces. Under these circumstances, it was proved that the coercivity and the boundedness, of the operator associated with the KKT system, hold with α-independent constants. Nevertheless, in this general setting, the inf-sup condition involved an α-dependent constant, which, eventually, yielded theoretical iteration bounds of order $O([\log\left(\alpha^{-1}\right)]^2)$ for MINRES.

In the present paper we were able to prove an α-robust inf-sup condition for the model problem (1)–(3). This is possible because both the control f and the dual/Lagrange-multiplier w belong to $L^2(\Omega)$. From a more general perspective, it turns out that this is the property that must be fulfilled in order for our approach to be successful: The control space and the dual space, associated with the state equation, must coincide. This will usually lead to additional regularity requirements for the state space.

Motivated by this discussion, let us consider an abstract problem of the form:

$$\min_{f\in W,\, u\in U} \left\{ \frac{1}{2} \|Tu - d\|_O^2 + \frac{1}{2}\alpha \, \|f\|_W^2 \right\} \tag{42}$$

subject to

$$\langle Au, w\rangle + (f, w)_W = 0, \quad \forall w \in W. \tag{43}$$

Here, W is the dual and control space, U is the state space, O is the observation space, W, U and O are Hilbert spaces.

Let us assume that

(A1) $A : U \to W'$ is a continuous linear operator with closed range. In particular, there is a constant c_1 such that for all $u \in U$,

$$\|u\|_{U/\operatorname{Ker} A} = \inf_{\tilde{u}\in \operatorname{Ker} A} \|u - \tilde{u}\|_U \le c_1 \, \|Au\|_{W'} .$$

(A2) $T : U \to O$ is linear and bounded, and invertible on the kernel of A. That is, there is a constant c_2 such that for all $u \in \operatorname{Ker} A$,

$$\|u\|_U \le c_2 \, \|Tu\|_O .$$

It then follows that the KKT system associated with (42)–(43) is well-posed for every $\alpha > 0$: Determine $(f, u, w) \in W \times U \times W$ such that

$$\underbrace{\begin{bmatrix} \alpha M & 0 & M' \\ 0 & K & A' \\ M & A & 0 \end{bmatrix}}_{=\mathcal{A}_\alpha} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{K} d \\ 0 \end{bmatrix}, \tag{44}$$

where

$$M : W \to W', \quad f \mapsto (f, \cdot)_W, \tag{45}$$

$$K : U \to U', \quad u \mapsto (Tu, T\,\cdot)_O, \tag{46}$$

$$\tilde{K} : O \to U', \quad d \mapsto (d, T\,\cdot)_O, \tag{47}$$

Note that, compared with (13), the boundary observation matrix M_∂ has been replaced with the general observation operator K in (44).

We introduce scaled norms as follows.

$$\begin{aligned} \|f\|_{W_\alpha}^2 &= \alpha\, \|f\|_W^2 \,, \\ \|u\|_{U_\alpha}^2 &= \alpha\, \|Au\|_{W'}^2 + \|Tu\|_O^2 \,, \\ \|w\|_{W_{\alpha^{-1}}}^2 &= \frac{1}{\alpha}\, \|w\|_W^2 \,. \end{aligned}$$

We first show that $\|\cdot\|_{U_\alpha}$ is indeed a norm on U when assumptions (A1) and (A2) hold. It suffices to show that $\|\cdot\|_{U_\alpha}$ is a norm equivalent to $\|\cdot\|_U$ when $\alpha = 1$. We have

$$\|Tu\|_O + \|Au\|_{W'} \le \left(\|T\|_{\mathcal{L}(U,O)} + \|A\|_{\mathcal{L}(U,W')} \right) \|u\|_U \,, \tag{48}$$

and letting π denote the orthogonal projection of U onto Ker A,

$$\begin{aligned} \|u\|_U &\le \|\pi u\|_U + \|u - \pi u\|_U \\ &\le c_2\, \|T\pi u\|_O + \|u - \pi u\|_U \\ &\le c_2\, \|Tu\|_O + \left(1 + c_2\, \|T\|_{\mathcal{L}(U,O)}\right) \|u - \pi u\|_U \\ &\le c_2\, \|Tu\|_O + c_1\left(1 + c_2\, \|T\|_{\mathcal{L}(U,O)}\right) \|Au\|_{W'} \,. \end{aligned} \tag{49}$$

Here the last inequality follows from $\|u - \pi u\|_U = \inf_{\tilde{u} \in \mathrm{Ker}\, A} \|u - \tilde{u}\|_U$ and assumption (A1).

We set $\mathcal{V} = W_\alpha \times U_\alpha \times W_{\alpha^{-1}}$. As in Sect. 4, $\mathcal{A}_\alpha : \mathcal{V} \to \mathcal{V}'$ can be shown to be an isomorphism, with parameter-independent estimates obtained in the weighted norms.

Theorem 3 *There exists positive constants c and C, independent of α, such that for all nonzero $x \in \mathcal{V}$,*

$$c \le \sup_{0 \ne y \in \mathcal{V}} \frac{\langle \mathcal{A}_\alpha x, y \rangle}{\|x\|_{\mathcal{V}}\, \|y\|_{\mathcal{V}}} \le C. \tag{50}$$

We omit the full proof, which is analogous to that of Theorem 1. The crucial part is the "inf-sup condition" of Lemma 1, which is easily shown to hold in the abstract setting:

$$\sup_{(f,u) \in W_\alpha \times U_\alpha} \frac{(f, w)_W + \langle Au, w \rangle}{\|(f, u)\|_{W_\alpha \times U_\alpha}\, \|w\|_{W_{\alpha^{-1}}}} \ge \frac{(w, w)_W}{\|(w, 0)\|_{W_\alpha \times U_\alpha}\, \|w\|_{W_{\alpha^{-1}}}} = 1.$$

The coercivity condition of Lemma 2 naturally holds in the prescribed norm on U_α, since for $(f, u) \in W \times U$ such that $Au = Mf$,

$$\alpha \|f\|_W^2 + \|Tu\|_O^2 = \frac{\alpha}{2}\|f\|_W^2 + \frac{\alpha}{2}\|Au\|_{W'}^2 + \|Tu\|_O^2 \geq \frac{1}{2}\left(\|f\|_{W_\alpha}^2 + \|u\|_{U_\alpha}^2\right).$$

Note that the weighted norm now depends on A, and as consequence, the estimates become A-independent. In fact, we obtain bounds for the constants c and C which are independent of α as well as the operators appearing in (42)–(43). This is postponed to the next section, where sharp estimates are obtained for (50).

With the estimates (50), Lemma 6 provides a preconditioner for the operator $\mathcal{A}_\alpha$, given as

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha M & 0 & 0 \\ 0 & \alpha A' M^{-1} A + K & 0 \\ 0 & 0 & \frac{1}{\alpha} M \end{bmatrix}^{-1}. \tag{51}$$

The condition number of $\mathcal{B}_\alpha \mathcal{A}_\alpha$ will be bounded independently of α. It is, however, not clear how to find a computationally efficient approximation of $\mathcal{B}_\alpha$ in the abstract setting of (42)–(43).

Example 1 The problem (1)–(3) fits in the abstract framework presented in this section when we assume that the state has $H^2(\Omega)$ regularity. We set $W = L^2(\Omega)$, $U = \bar{H}^2(\Omega)$, $A = 1 - \Delta$, and $T : \bar{H}^2(\Omega) \to L^2(\partial\Omega)$ is a trace operator, see (46). Since A is a continuous isomorphism, assumptions (A1) and (A2) are both valid. The inner product on U_α takes the form

$$\begin{aligned}(u, v)_{U_\alpha} &= \langle Ku, v\rangle + \alpha \left\langle AM^{-1}Au, v\right\rangle \\ &= \int_{\partial\Omega} uv\, ds + \alpha \int_\Omega (u - \Delta u)(v - \Delta v)\, dx \\ &= \int_{\partial\Omega} uv\, ds + \alpha \int_\Omega \nabla^2 u : \nabla^2 v + 2\nabla u \cdot \nabla v + uv\, dx,\end{aligned}$$

where $\nabla^2 u$ denotes the Hessian of u, and the last equality follows from the boundary condition $\partial u/\partial \mathbf{n} = 0$ imposed on $\bar{H}^2(\Omega)$. The resulting preconditioner is the one that was used in the numerical experiments, detailed in Sect. 3, and it is spectrally equivalent to the preconditioner defined in (40).

Example 2 Let U, W, and K be as in Example 1, but let us set $A = -\Delta$. Now A has non-trivial kernel, consisting of the a.e. constant functions, and for constant u we have

$$\|Tu\|_{L^2(\partial\Omega)} = \sqrt{\frac{|\partial\Omega|}{|\Omega|}}\, \|u\|_{\bar{H}^2(\Omega)}.$$

Since assumptions (A1) and (A2) are valid, the optimality system is still well-posed. In this case the inner product on U_α is given by

$$(u, v)_{U_\alpha} = \int_{\partial\Omega} uv\, ds + \alpha \int_\Omega D^2 u : D^2 v\, dx.$$

Example 3 Let us consider the "prototype" problem:

$$\min_{f,u} \left\{ \frac{1}{2} \|u - d\|^2_{L^2(\Omega)} + \frac{\alpha}{2} \|f\|^2_{L^2(\Omega)} \right\}$$

subject to

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

Note that we here consider the case in which observation data is assumed to be available throughout the entire domain Ω of the state equation.

If the usual variational form of the PDE is used, i.e.,

$$(u, w)_{H^1(\Omega)} + (f, w)_{L^2(\Omega)} = 0, \quad \forall w \in H^1(\Omega), \tag{52}$$

then the control space equals $L^2(\Omega)$, whereas the dual space is $H^1(\Omega)$. The preconditioning strategy presented in this section is therefore not applicable.

If instead we can assume $H^2(\Omega)$-regularity, we can use the variational form

$$(-\Delta u + u, w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0, \quad \forall w \in L^2(\Omega). \tag{53}$$

Now, the control and dual spaces both equal $L^2(\Omega)$. The methodology presented in this section can thus be applied, and a robust preconditioner is obtained. Compared with the preconditioner for the problem with boundary observations only, see Sect. 5, Eq. (40), the only change is the replacement of M_∂, in the (2, 2) block of $\mathcal{B}_\alpha$ with M.

We remark that in [13,14], parameter-robust preconditioners were proposed for the "prototype" problem, using the standard variational formulation (52) of the PDE. Those methods do not require improved regularity for the state space. Instead, they require that observations are available throughout the computational domain.

7 Eigenvalue analysis

In Sect. 6 it was shown that the condition number of $\mathcal{B}_\alpha \mathcal{A}_\alpha$, with $\mathcal{A}_\alpha$ defined in (44) and $\mathcal{B}_\alpha$ defined in (51), can be bounded independently of α, as well as independently of the operators appearing in (42)–(43). Moreover, the numerical experiments indicate that the eigenvalues are contained in three intervals, independently of the regularization parameter α, see Fig. 2. In this section we detail the structure of the spectrum of the preconditioned system considered in Sect. 6, and we obtain sharp estimates for the constants appearing in Theorem 3.

We consider self-adjoint linear operators $\mathcal{A}_\alpha$ and $\mathcal{B}_\alpha$,

$$\mathcal{A}_\alpha = \begin{bmatrix} \alpha M & 0 & M' \\ 0 & K & A' \\ M & A & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{B}_\alpha^{-1} = \begin{bmatrix} \alpha M & 0 & 0 \\ 0 & K + \alpha R & 0 \\ 0 & 0 & \alpha^{-1} M \end{bmatrix} \tag{54}$$

where R is defined by

$$R = A' M^{-1} A. \tag{55}$$

We assume that $A : U \to W'$ and $M : W \to W'$ are continuous operators, for some Hilbert spaces U and W. In addition we will make use of the following assumptions.

(B1) M is a self-adjoint and positive definite,
(B2) $K + R$ is positive definite,
(B3) K is self-adjoint and positive semi-definite.

Assumptions (B1)–(B3) ensure that $\mathcal{B}_\alpha$ is a self-adjoint and positive definite. In particular, assumptions (B1)–(B3) hold for $\mathcal{A}_\alpha$ as in (44), provided that the assumptions of Sect. 6 hold. For simplicity, we also assume that that $\mathcal{A}_\alpha$ and $\mathcal{B}_\alpha$ are finite-dimensional operators.

Theorem 4 *Let p, q, and r be the polynomials*

$$p(\lambda) = 1 - \lambda, \quad q(\lambda) = 1 + \lambda p(\lambda), \quad r(\lambda) = p - \lambda q(\lambda).$$

Let $q_1 < q_2$ and $r_1 < r_2 < r_3$ be the roots of q and r, respectively. The spectrum of $\mathcal{B}_\alpha \mathcal{A}_\alpha$ is contained within three intervals, determined by the roots of p and r, independently of α:

$$\operatorname{sp}(\mathcal{B}_\alpha \mathcal{A}_\alpha) \subset [r_1, q_1] \cup [r_2, 1] \cup [q_2, r_3]. \tag{56}$$

Consequently, the spectral condition number of $\mathcal{B}_\alpha \mathcal{A}_\alpha$ is bounded, uniformly in α,

$$k(\mathcal{B}_\alpha \mathcal{A}_\alpha) \leq \frac{r_3}{r_2} \approx 4.089. \tag{57}$$

If K has a nontrivial kernel, inequality (57) becomes an equality.

Proof Consider the equivalent generalized eigenvalue problem

$$\begin{bmatrix} \alpha M & 0 & M' \\ 0 & K & A' \\ M & A & 0 \end{bmatrix} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \lambda \begin{bmatrix} \alpha M & 0 & 0 \\ 0 & K + \alpha R & 0 \\ 0 & 0 & \alpha^{-1} M \end{bmatrix} \begin{bmatrix} f \\ u \\ w \end{bmatrix} \tag{58}$$

We show that (58) admits no nontrivial solutions unless λ is as in (56).

Since M is a self-adjoint isomorphism, by assumption (B1), we can rewrite (58) as the three identities

$$\alpha p f + w = 0, \tag{59}$$

$$p K u + A' w - \lambda \alpha R u = 0, \tag{60}$$

$$f + M^{-1} A u - \lambda \alpha^{-1} w = 0. \tag{61}$$

Assume that λ is not contained within the three closed intervals of (56). Then $p \neq 0$, and we can use (59) to eliminate f from (61).

$$\begin{aligned} 0 &= \alpha p (f + M^{-1} A u - \lambda \alpha^{-1} w) = \alpha p M^{-1} A u - (1 + \lambda p) w \\ &= \alpha p M^{-1} A u - q w. \end{aligned} \tag{62}$$

Since q is nonzero, we can use (62) to eliminate w from (60),

$$\begin{aligned} 0 &= q (p K u + A' w - \lambda \alpha R u) = q p K u + \alpha (p - \lambda q) R u \\ &= q p K u + r R u, \end{aligned} \tag{63}$$

where the identity (55) was used. By assumption, pq and r are both nonzero. Moreover, it can be easily seen that pq and r have the same sign outside of the bounded intervals of (56). From assumptions (B1)–(B3), we conclude that $qpK + rR$ is a self-adjoint definite operator. Then (63) only admits trivial solutions, hence λ can not be an eigenvalue of $\mathcal{B}_\alpha \mathcal{A}_\alpha$.

The estimate (57) follows from (56), noting that $|\operatorname{sp}(\mathcal{B}_\alpha \mathcal{A}_\alpha)| \subset [r_2, r_3]$. From (63) it can be seen that the roots of r are eigenvalues of $\mathcal{B}_\alpha \mathcal{A}_\alpha$ if Ker K is nontrivial. □

Remark If $A = (1 - \Delta) : \bar{H}^2(\Omega) \rightarrow L^2(\Omega)'$, then $R = A' M^{-1} A$ is characterized by a bilinear form $b(\cdot, \cdot)$ as in (15):

$$\begin{aligned} \langle A' M^{-1} A u, v \rangle &= \int_\Omega \Delta u \Delta v + 2 \nabla u \cdot \nabla v + u v \, dx \\ &= (u, v)_{H^2(\Omega)} + \int_\Omega \nabla u \cdot \nabla v \, dx = b(u, v) \end{aligned}$$

For discretizations $U_h \subset U$ and $W_h \subset W$ of A such that $A(U_h) \subset M(W_h)$, the discretization of b coincides with $A_h' M_h^{-1} A_h$. This follows from an argument similar to that in the proof of Lemma 5, and as a consequence, Theorem 4 can be applied to the preconditioned discrete systems considered in Sect. 3.

8 Discretization with H^1 conforming finite elements

The theory outlined in Sect. 4 provides a robust preconditioning technique for the optimality system (20) assuming additional regularity and making use of H^2 conforming elements. However, this additional regularity only appears relevant to the discretization of the (2, 2) block of the ideal preconditioner (51), since the coefficient matrix in

(44) only involves second order operators. It therefore seems reasonable that the use of sophisticated H^2 conforming elements could be avoided in favour of standard H^1 conforming elements, provided that we can implement an approximate inverse to the fourth operator appearing in the preconditioner.

To be precise, we can discretize the optimality system (10)–(12) with H^1-conforming piecewise linear Lagrange elements for all the unknown variables. Note that this requires that the integration by parts formula is applied to the state equation, resulting in the variation problem

$$\begin{aligned} \alpha(f_h, \psi_h)_{L^2(\Omega)} + (\psi_h, w_h)_{L^2(\Omega)} &= 0 \quad \forall \psi_h \in V_h, \\ (u_h - d, \phi_h)_{L^2(\partial\Omega)} + (\phi_h, w_h)_{H^1(\Omega)} &= 0 \quad \forall \phi_h \in V_h, \\ (f_h, \xi_h)_{L^2(\Omega)} + (u_h, \xi_h)_{H^1(\Omega)} &= 0 \quad \forall \xi_h \in V_h. \end{aligned}$$

for $(f_h, u_h, \xi_h) \in V_h \times V_h \times V_h$, where V_h is the space of continuous piecewise linear functions. Since all three unknowns belong to the same space, the eigenvalue analysis in Sect. 7 can be applied to the discretized coefficient matrix, which reads

$$\begin{bmatrix} \alpha M_h & 0 & M_h \\ 0 & K_h & A_h \\ M_h & A_h & 0 \end{bmatrix}, \tag{64}$$

where A_h and M_h are symmetric matrices. The analysis in Sect. 7 reveals that an ideal preconditioner is given by

$$\begin{bmatrix} \alpha M_h & 0 & 0 \\ 0 & K_h + \alpha A_h M_h^{-1} A_h & 0 \\ 0 & 0 & \alpha^{-1} M_h \end{bmatrix}^{-1}, \tag{65}$$

with condition numbers of the preconditioned system bounded independtly of α and the discretization parameter h.

The operator $K_h + A_h M_h^{-1} A_h$ in the (2,2) block of (65) coincides with Schur complement of a Ciarlet-Raviart mixed finite element formulation of the fourth order problem (16)–(18), and can be thought of as non-local fourth order operator. Multigrid techniques for a similar operator was studied in [8], where a multigrid W-cycle applied to a local operator approximating the Schur complement was shown to be an efficient preconditioner.

Table 4 presents iteration numbers and estimated condition numbers for a simplistic scheme where we replace the (2,2) block in (65) with $K_h + A_h \tilde{M}_h^{-1} A_h$, where $\tilde{M}_h$ is a lumped mass matrix. For the appxroximate inversion of (65), we applied an algebraic multigrid W-cycle for the (2,2) block and two symmetric Gauss–Seidel iterations to the remaining two diagonal blocks. The experiment was carried out on a unit square domain and an L-shaped domain, with both domains triangularized with structured meshes. For the L-shaped domain, the H^2-regularity discussed in the beginning of Sect. 2 is known not to hold.

Table 4 MINRES iteration counts with estimated condition numbers in paranthesis for the coefficient matrix (64), with a preconditioner based on (65)

$\alpha\backslash h$	2^{-6}	2^{-7}	2^{-8}	2^{-9}
Square domain				
10^{-10}	81 (6.80)	82 (6.80)	88 (6.87)	93 (6.90)
10^{-8}	90 (6.65)	93 (6.82)	91 (6.63)	89 (6.63)
10^{-6}	95 (7.10)	90 (6.63)	89 (6.66)	88 (6.71)
10^{-4}	89 (6.63)	89 (6.68)	88 (6.72)	86 (6.73)
10^{-2}	79 (6.63)	79 (6.70)	78 (6.73)	78 (6.73)
1	69 (6.68)	68 (6.74)	67 (6.75)	67 (6.75)
L-shaped domain				
10^{-10}	80 (6.80)	82 (6.76)	88 (6.81)	93 (6.87)
10^{-8}	90 (6.65)	93 (6.73)	91 (6.63)	89 (6.63)
10^{-6}	93 (6.92)	90 (6.63)	89 (6.64)	88 (6.71)
10^{-4}	89 (6.64)	89 (6.68)	92 (8.16)	92 (10.2)
10^{-2}	85 (8.25)	87 (10.2)	89 (13.1)	90 (17.4)
1	90 (16.4)	93 (22.5)	94 (32.0)	86 (46.9)

The (2,2) block of (65) was replaced by an AMG W-cycle (BoomerAMG from the Hypre[4] library), constructed from the operator $K_h + A_h\tilde{M}_h^{-1}A_h$, where $\tilde{M}_h$ is a lumped mass matrix. The AMG operator was constructed with 0.5 treshold parameter, with other parameters set to their default values. The remaining two diagonal blocks were each replaced two symmetric Gauss–Seidel iterations were applied. The optimality system was solved approximately with a random initial guess and relative convergence criterion (19) with $\varepsilon = 10^{-8}$

The iteration numbers reported in Table 4 appears bounded, although we observe an increase in the estimated condition number for the L-shaped domain as the mesh is refined. Although the condition number with an exact inverse (65) is bounded in accordance with the analysis in Sect. 7, this appears not to be the case when the exact inverse of the (2,2) block is replaced with an AMG cycle.

We remark that for unstructered meshes we observed iteration counts increasing with mesh refinement, indicating the need for a more sophisticated approach to the multilevel approximation of the (2,2) block, for example as in [8], for more complicated geometries.

9 Discussion

Previously, parameter robust preconditioners for PDE-constrained optimization problems have been successfully developed, provided that observation data is available throughout the entire domain of the state equation. For many important inverse problems, arising in industry and science, this is an unrealistic requirement. On the contrary, observation data will typically only be available in subregions, of the domain of the state variable, or at the boundary of this domain. We have therefore explored the possibility for also constructing robust preconditioners for PDE-constrained optimization problems with limited observation data.

For an elliptic control problem, with boundary observations only, we have developed a regularization robust preconditioner for the associated KKT system. Consequently, the number of MINRES iterations required to solve the problem is bounded independently of both regularization parameter α and the mesh size h. In order to achieve this, it was necessary to write the elliptic state equation on a non-standard, and non-self-adjoint, variational form. If this approach is employed, then the control and the Lagrange multiplier will belong to the same Hilbert space, which leads to extra regularity requirements for the state. This fact makes it possible to construct parameter weighted metrics such that the constants appearing in the Brezzi conditions, as well as the constants in the inequalities expressing the boundedness of the KKT system, are independent of α and h. Consequently, the spectrum of the preconditioned KKT system is uniformly bounded with respect to α and h, which is ideal for the MINRES scheme. These properties were illuminated through a series of numerical experiments, and the preconditioned MINRES scheme handled our model problem excellently.

The use of a non-self-adjoint form of the elliptic state equation leads to additional challenges for conforming discretization schemes and in multigrid implementations. For the numerical experiments, we employed a C^1 finite element discretization that is H^2-conforming, where the rectangular elements are tensor products of Hermite intervals. This discretization is limited to structured meshes. While there are other, more flexible C^1 finite element discretizations available in two dimensions (e.g. Argyris and Bell triangles), all of the methods suffer from high computational cost due the smoothness requirements imposed on the nodal basis functions. In three dimensions, the situation is even worse, and C^1 discretizations with tetrahedrons become nearly intractable, see e.g. [15].

Some of the difficulties with traditional C^1 finite element discretizations can be avoided with Galerkin methods making use of basis functions that naturally fulfill the smoothness requirements. Examples of such methods include discretization with spline basis functions, such as isogeometric analysis [3]. Another approach is the virtual element method [2]. However, the development of multilevel methods for the fourth order operator in the preconditioner (51) would remain a challenging problem.

We have also demonstrated that the technique is applicable also outside of H^2-conforming discretizations.

Our findings for the simple elliptic control problem were generalized to a broader class of KKT systems. It turns out that the methodology is applicable whenever the control and the Lagrange multiplier belong to the same space, and extra regularity properties are fulfilled by the state equation - these are the key issues. From a theoretical perspective, this is in many cases not a severe restriction, but it gives rise to new challenges for the discrete problems. This is even the case for the elliptic state equation considered in this text. Also, our approach will not yield α independent bounds if the control is only defined on a subdomain of the domain of the state equation. In such cases, the spaces for the control and the Lagrange multiplier will not coincide. How to design efficient parameter-robust preconditioners for such problems, is, as far as the authors know, still an open problem.

Acknowledgements The first author is grateful to Adrian Hope for conducting numerical experiments concerning Hermite elements.

References

1. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. RAIRO Numer. Anal. **8**, 129–151 (1974)
2. Brezzi, F., Marini, L.D.: Virtual element methods for plate bending problems. Comput. Methods Appl. Mech. Eng. **253**, 455–462 (2013)
3. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: Isogeometric analysis: toward integration of CAD and FEA. Wiley, USA (2009)
4. Falgout, R., Yang, U.: Hypre: a library of high performance preconditioners. Comput. Sci. ICCS **2002**, 632–641 (2002)
5. Grisvard, P.: Elliptic problems in nonsmooth domains. Pitman, Boston (1985)
6. Günnel, A., Herzog, R., Sachs, E.: A note on preconditioners and scalar products in Krylov subspace methods for self-adjoint problems in Hilbert space. Electron. Trans. Numer. Anal. **41**, 13–20 (2014)
7. Hackbusch, W.: Elliptic differential equations. Theory and numerical treatment. Springer-Verlag, Berlin (1992)
8. Hanisch, M.R.: Multigrid preconditioning for the biharmonic dirichlet problem. SIAM J. Numer. Anal. **30**(1), 184–214 (1993)
9. Mardal, K.A., Winther, R.: Preconditioning discretizations of systems of partial differential equations. Numer. Linear Algebra Appl. **18**(1), 1–40 (2011)
10. Nielsen, B.F., Mardal, K.A.: Efficient preconditioners for optimality systems arising in connection with inverse problems. SIAM J. Control Optim. **48**(8) (2010)
11. Nielsen, B.F., Mardal, K.A.: Analysis of the minimal residual method applied to ill-posed optimality systems. SIAM J. Sci. Comput. **35**(2), A785–A814 (2013)
12. Pearson, J.W., Stoll, M., Wathen, A.J.: Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems. SIAM J. Matrix Anal. Appl. **33**, 1126–1152 (2012)
13. Pearson, J.W., Wathen, A.J.: A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. Numer. Linear Algebra Appl. **19**, 816–829 (2012)
14. Schöberl, J., Zulehner, W.: Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. SIAM J. Matrix Anal. Appl. **29**(3), 752–773 (2007)
15. Ženíšek, Alexander: Polynomial approximation on tetrahedrons in the finite element method. J. Approx. Theory **7**(4), 334–351 (1973)
16. Zhang, X.: Multilevel schwarz methods for the biharmonic dirichlet problem. SIAM J. Sci. Comput. **15**(3), 621–644 (1994)
17. Zulehner, W.: Nonstandard norms and robust estimates for saddle point problems. SIAM J. Matrix Anal. Appl. **32**, 536–560 (2011)

Paper I

Preconditioners for saddle point systems with trace constraints coupling 2D and 1D domains

SIAM J. Sci. Comput.
Vol. 38, No. 6, pp. B962–B987

PRECONDITIONERS FOR SADDLE POINT SYSTEMS WITH TRACE CONSTRAINTS COUPLING 2D AND 1D DOMAINS[*]

MIROSLAV KUCHTA[†], MAGNE NORDAAS[‡], JORIS C. G. VERSCHAEVE[†], MIKAEL MORTENSEN[†‡], AND KENT-ANDRE MARDAL[†‡]

Abstract. We study preconditioners for a model problem describing the coupling of two elliptic subproblems posed over domains with different topological dimension by a parameter dependent constraint. A pair of parameter robust and efficient preconditioners is proposed and analyzed. Robustness and efficiency of the preconditioners is demonstrated by numerical experiments.

Key words. preconditioning, saddle-point problem, Lagrange multipliers

AMS subject classification. 65F08

DOI. 10.1137/15M1052822

1. Introduction.

This paper is concerned with preconditioning of multiphysics problems where two subproblems of different dimensionality are coupled. We assume that Γ is a submanifold contained within $\Omega \in \mathbb{R}^n$ and consider the following problem:

$$-\Delta u + \epsilon\delta_\Gamma p = f \quad \text{in } \Omega, \tag{1a}$$

$$-\Delta v - p = g \quad \text{on } \Gamma, \tag{1b}$$

$$\epsilon u - v = 0 \quad \text{on } \Gamma, \tag{1c}$$

where δ_Γ is a function with properties similar to the Dirac delta function, as will be discussed later. To allow for a unique solution (u, v, p), the system must be equipped with suitable boundary conditions, and we shall here, for simplicity, consider homogeneous Dirichlet boundary conditions for u and v on $\partial\Omega$ and $\partial\Gamma$, respectively. We note that the unknowns u, v are here the primary variables, while the unknown p should be interpreted as a Lagrange multiplier associated with the constraint (1c).

The two elliptic equations that are stated on two different domains, Ω and Γ, are coupled, and therefore the restriction of u to Γ and the extension of p to Ω are crucial. When the codimension of Γ is one, the restriction operator is a trace operator, and the extension operator is similar to the Dirac delta function. We note that $\epsilon \in (0, 1)$ and that the typical scenario will be that $\epsilon \ll 1$. We will therefore focus on methods that are robust in ϵ.

The problem (1a)–(1c) is relevant to biomedical applications [18, 15, 2, 17] where it models the coupling of the porous media flow inside tissue to the vascular bed through Starling's law. Further, problems involving coupling of the finite element method

[*] Submitted to the journal's Computational Methods in Science and Engineering section December 14, 2015; accepted for publication (in revised form) August 24, 2016; published electronically November 1, 2016.

http://www.siam.org/journals/sisc/38-6/M105282.html

Funding: The work of Magne Nordaas, Mikael Mortensen, and Kent-Andre Mardal has been supported by Center of Excellence grant 179578 from the Research Council of Norway to the Center for Biomedical Computing at Simula Research Laboratory. The work of Kent-Andre Mardal has also been supported by the Research Council of Norway through grant 209951.

[†] Department of Mathematics, Division of Mechanics, University of Oslo, Oslo, Norway (mirok@math.uio.no, joris@math.uio.no, mikaem@math.uio.no).

[‡] Center for Biomedical Computing, Simula Research Laboratory, 1325 Lysaker, Norway (magneano@simula.no, kent-and@simula.no).

and the boundary element method, e.g., [24, 26], are of the form (1). The system is also relevant for domain decomposition methods based on Lagrange multipliers [32]. Finally, in solid mechanics, the problem of plates reinforced with ribs (cf., for example, [44, Ch. 9.11]) can be recast into a related fourth order problem. We also note that the techniques developed here to address the constraint (1c) are applicable in preconditioning fluid-structure interaction problems involving interactions with thin structures, e.g., filaments [22].

One way of deriving equations (1) is to consider the following minimization problem:

$$\left.\begin{aligned}\int_\Omega (\nabla u)^2 - 2uf\,\mathrm{d}x\\ \int_\Gamma (\nabla v)^2 - 2vg\,\mathrm{d}s\end{aligned}\right\} \to \min \tag{2}$$

subject to the constraint

$$\epsilon u - v = 0 \qquad \text{on } \Gamma. \tag{3}$$

Using the method of Lagrange multipliers, the constrained minimization problem will be recast as a saddle-point problem. The saddle-point problem is then analyzed in terms of the Brezzi conditions [13], and efficient solution algorithms are obtained using operator preconditioning [35]. A main challenge is the fact that the constraint (3) necessitates the use of trace operators, which leads to operators in fractional Sobolev spaces on Γ.

An outline of the paper is as follows: Section 2 presents the necessary notation and mathematical framework needed for the analysis. Then the mathemathical analysis as well as the numerical experiments of two different preconditioners are presented in sections 3 and 4, respectively. Section 5 discusses the computational efficiency of both methods.

2. Preliminaries.

Let X be a Hilbert space of functions defined on a domain D, and let $\|\cdot\|_X$ denote its norm. The L^2 inner product on a domain D is denoted $(\cdot,\cdot)_D$ or $\int_D \cdot$, while $\langle\cdot,\cdot\rangle_D$ denotes the corresponding duality pairing between a Hilbert space X and its dual space X^*. We will use $H^m = H^m(D)$ to denote the Sobolev space of functions on D with m derivatives in $L^2 = L^2(D)$. The corresponding norm is denoted $\|\cdot\|_{m,D}$. In general, we will use H^m_0 to denote the closure in H^m of the space of smooth functions with compact support in D, and the seminorm is denoted as $|\cdot|_{m,D}$.

The space of bounded linear operators mapping elements of X to Y is denoted $\mathcal{L}(X,Y)$, and if $Y = X$, we simply write $\mathcal{L}(X)$ instead of $\mathcal{L}(X,X)$. If X and Y are Hilbert spaces, both continuously contained in some larger Hilbert space, then the intersection $X \cap Y$ and the sum $X + Y$ are both Hilbert spaces with norms given by

$$\|x\|^2_{X\cap Y} = \|x\|^2_X + \|x\|^2_Y \quad \text{and} \quad \|z\|^2_{X+Y} = \inf_{\substack{x\in X, y\in Y\\ z=x+y}} (\|x\|^2_X + \|y\|^2_Y).$$

In the following $\Omega \subset \mathbb{R}^n$ is an open connected domain with Lipschitz boundary $\partial\Omega$. The trace operator T is defined by $Tu = u|_\Gamma$ for $u \in C(\overline{\Omega})$ and Γ a Lipschitz submanifold of codimension one in Ω. The trace operator extends to bounded and surjective linear operator $T : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma)$; see, e.g., [1, Ch. 7]. The fractional Sobolev space $H^{\frac{1}{2}}(\Gamma)$ can be equipped with the norm

$$\|u\|^2_{H^{\frac{1}{2}}(\Gamma)} = \|u\|^2_{L^2(\Gamma)} + \int_{\Gamma\times\Gamma} \frac{|u(x)-u(y)|^2}{|x-y|^{n+1}}\,\mathrm{d}x\mathrm{d}y. \tag{4}$$

However, the trace is not surjective as an operator from $H^1_0(\Omega)$ into $H^{\frac{1}{2}}(\Gamma)$; in particular, the constant function $1 \in H^{\frac{1}{2}}(\Gamma)$ is not in the image of the trace operator. Note that $H^{\frac{1}{2}}_0(\Gamma)$ does not characterize the trace space, since $H^{\frac{1}{2}}_0(\Gamma) = H^{\frac{1}{2}}(\Gamma)$; see [30, Ch. 2, Thm. 11.1]. Instead, the trace space can be identified as $H^{\frac{1}{2}}_{00}(\Gamma)$, defined as the subspace of $H^{\frac{1}{2}}(\Gamma)$ for which extension by zero into $H^{\frac{1}{2}}(\tilde{\Gamma})$ is continuous, for some suitable extension domain $\tilde{\Gamma}$ extending Γ (e.g., $\tilde{\Gamma} = \Gamma \cup \partial\Omega$). To be precise, the space $H^{\frac{1}{2}}_{00}(\Gamma)$ can be characterized with the norm

$$\|u\|_{H^{\frac{1}{2}}_{00}(\Gamma)} = \|\tilde{u}\|_{H^{\frac{1}{2}}(\tilde{\Gamma})}, \quad \tilde{u}(x) = \begin{cases} u(x), & x \in \Gamma, \\ 0, & x \notin \Gamma. \end{cases} \tag{5}$$

The space $H^{\frac{1}{2}}_{00}(\Gamma)$ does not depend on the extension domain $\tilde{\Gamma}$, since the norms induced by different choices of $\tilde{\Gamma}$ will be equivalent.

The above norms (4)–(5) for the fractional spaces are impractical from an implementation point of view, and we will therefore consider the alternative construction following [30, Ch. 2.1] and [16]. For $u, v \in H^1_0(\Gamma)$, set $L_u(v) = (u, v)_\Gamma$. Then L_u is a bounded linear functional on $H^1_0(\Gamma)$, and in accordance with the Riesz–Fréchet theorem there is an operator $S \in \mathcal{L}\big(H^1_0(\Gamma)\big)$ such that

$$(Su, w)_{H^1_0(\Omega)} = L_u(w) = (u, w)_\Gamma, \qquad u, w \in H^1_0(\Gamma). \tag{6}$$

The operator S is self-adjoint, positive definite, injective, and compact. Therefore, the spectrum of S consists of a nonincreasing sequence of positive eigenvalues $\{\lambda_k\}_{k=1}^\infty$ such that $0 < \lambda_{k+1} \le \lambda_k$ and $\lambda_k \to 0$; see, e.g., [48, Ch. X.5, Thm. 2]. The eigenvectors $\{\phi_k\}_{k=1}^\infty$ of S satisfy the generalized eigenvalue problem

$$A\phi_k = \lambda_k^{-1} M \phi_k,$$

where operators A, M are such that $\langle Au, v\rangle_\Gamma = (\nabla u, \nabla v)_\Gamma$ and $\langle Mu, v\rangle_\Gamma = (u, v)_\Gamma$. The set of eigenvectors $\{\phi_k\}_{k=1}^\infty$ forms a basis of $H^1_0(\Gamma)$ orthogonal with respect to the inner product of $H^1_0(\Gamma)$ and orthonormal with respect to the inner product on $L^2(\Gamma)$. Then for $u = \sum_k c_k \phi_k \in \text{span}\ \{\phi_k\}_{k=1}^\infty$ and $s \in [-1, 1]$, we set

$$\|u\|_{H_s} = \sqrt{\sum_k c_k^2 \lambda_k^{-s}} \tag{7}$$

and define H_s to be the closure of span $\{\phi_k\}_{k=1}^\infty$ in the above norm. Then $H_0 = L^2(\Gamma)$ and $H_1 = H^1_0(\Gamma)$ with equality of norms. Moreover, we have $H_{\frac{1}{2}} = H^{\frac{1}{2}}_{00}(\Gamma)$ with equivalence of norms. This essentially follows from the fact that $H_{\frac{1}{2}}$ and $H^{\frac{1}{2}}_{00}(\Gamma)$ are closely related interpolation spaces; see [16, Thm. 3.4]. Note that we also have $H_{-1} = (H^1_0(\Gamma))^* = H^{-1}(\Gamma)$ and $H_{-\frac{1}{2}} = (H^{\frac{1}{2}}_{00}(\Gamma))^* = H^{-\frac{1}{2}}(\Gamma)$.

As the preceding paragraph suggests, we shall use the normal font to denote linear operators, e.g., A. To signify that the particular operator acts on a vector space with multiple components, we employ the calligraphic font, e.g., $\mathcal{A}$. Vectors and matrices are denoted by the sans serif font, e.g., A and x. In the case when the matrix has a block structure, it is typeset with the blackboard bold font, e.g., $\mathbb{A}$. Matrices and vectors are related to the discrete problems as follows (see also [35, Ch. 6]). Let $V_h \subset H^1_0(D)$, and let the discrete operator $A_h : V_h \to V_h^*$ be defined in terms of the Galerkin method:

$$\langle A_h u_h, v_h\rangle_D = \langle Au, v_h\rangle_D \quad \text{for } u_h, v_h \in V_h \text{ and } u \in H^1_0(D).$$

Let $\psi_j, j \in [1, m]$ be the basis functions of V_h. The matrix equation,

$$\mathsf{A}\mathsf{u} = \mathsf{f}, \quad \mathsf{u} \in \mathbb{R}^m \text{ and } \mathsf{f} \in \mathbb{R}^m,$$

is obtained as follows: Let $\pi_h : V_h \to \mathbb{R}^m$ and $\mu_h : V_h^* \to \mathbb{R}^m$ be given by

$$v_h = \sum_j (\pi_h v_h)_j \, \psi_j, \quad v_h \in V_h, \qquad \text{and} \qquad (\mu_h f_h)_j = \langle f_h, \psi_j \rangle_D, \quad f_h \in V_h^*.$$

Then

$$\mathsf{A} = \mu_h A_h \pi_h^{-1}, \quad \mathsf{v} = \pi_h v_h, \quad \mathsf{f} = \mu_h f_h.$$

A discrete equivalent to the H_s inner product (7) is constructed in the following manner, similarly to the continuous case. There exist a complete set of eigenvectors $\mathsf{u}_i \in \mathbb{R}^m$ with the property $\mathsf{u}_j{}^\top \mathsf{M} \mathsf{u}_i = \delta_{ij}$ and m positive definite (not necessarily distinct) eigenvalues λ_i of the generalized eigenvalue problem $\mathsf{A}\mathsf{u}_i = \lambda_i \mathsf{M}\mathsf{u}_i$. Equivalently, the matrix A can be decomposed as $\mathsf{A} = (\mathsf{MU})\,\mathsf{\Lambda}(\mathsf{MU})^\top$ with $\mathsf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_m)$ and $\operatorname{col}_i \mathsf{U} = \mathsf{u}_i$ so that $\mathsf{U}^\top \mathsf{M} \mathsf{U} = \mathsf{I}$ and $\mathsf{U}^\top \mathsf{A} \mathsf{U} = \Lambda$. We remark that A is the stiffness matrix, while M is the mass matrix.

Let now $\mathsf{H} : \mathbb{R} \to \mathsf{P}_{\text{sym}}$, where P_{sym} denotes the space of symmetric positive definite matrices, be defined as

$$\mathsf{H}(s) = (\mathsf{MU})\,\Lambda^s (\mathsf{MU})^\top. \tag{8}$$

Note that, due to M orthonormality of the eigenvectors, the inverse of $\mathsf{H}(s)$ is given as $\mathsf{H}(s)^{-1} = \mathsf{U}\mathsf{\Lambda}^{-s}\mathsf{U}^\top$. To motivate the definition of the mapping, we shall in the following example consider several values $\mathsf{H}(s)$ and show the relation of the matrices to different Sobolev (semi)norms of functions in V_h.

Example 1 (L_2, H_0^1, and H^{-1} norms in terms of matrices). Let $V_h \subset H_0^1(\Gamma)$, $\dim V_h = m$, $v_h \in V_h$, and $\mathsf{v} \in \mathbb{R}^m$ be the representation of v_h in the basis of V_h, i.e., $\mathsf{v} = \pi_h v_h$. The L^2 norm of v_h is given through the mass matrix M as $\|v_h\|_{0,\Gamma}^2 = \mathsf{v}^\top \mathsf{M} \mathsf{v}$ and $\mathsf{M} = \mathsf{H}(0)$. Similarly for the H_0^1 (semi)norm, it holds that $|v_h|_{1,\Gamma}^2 = \mathsf{v}^\top \mathsf{A} \mathsf{v}$, where A is the stiffness matrix, and $\mathsf{A} = \mathsf{H}(1)$. Finally, for a less trivial example, let $f_h \in V_h$, and consider f_h as a bounded linear functional, $\langle f_h, v_h \rangle_\Gamma = (f_h, v_h)_\Gamma$ for $v_h \in V_h$. Then $\|f_h\|_{-1,\Gamma}^2 = \mathsf{f}^\top \mathsf{H}(-1)\,\mathsf{f}$. By the Riesz representation theorem there exists a unique $u_h \in V_h$ such that $(\nabla u_h, \nabla v_h)_\Gamma = \langle f_h, v_h \rangle_\Gamma$ for all $v_h \in V_h$ and $\|f_h\|_{-1,\Gamma} = |u_h|_{1,\Gamma}$. The latter equality yields $\|f_h\|_{-1,\Gamma}^2 = \mathsf{u}^\top \mathsf{A} \mathsf{u}$, but since $u_h \in V_h$ is given by the Riesz map, the coordinate vector comes as a unique solution of the system $\mathsf{A}\mathsf{u} = \mathsf{M}\mathsf{f}$, i.e., $\mathsf{u} = \mathsf{A}^{-1}\mathsf{M}\mathsf{f}$ (see, e.g., [33, Ch. 3]). Thus $\|f_h\|_{-1,\Gamma}^2 = \mathsf{f}^\top \mathsf{M}\mathsf{A}^{-1}\mathsf{M}\mathsf{f}$. The matrix product in the expression is then $\mathsf{H}(-1)$.

In general, let c be the representation of vector $\mathsf{u} \in \mathbb{R}^m$ in the basis of eigenvectors u_i, $\mathsf{u} = \mathsf{U}\mathsf{c}$. Then

$$\mathsf{u}^\top \mathsf{H}(s)\,\mathsf{u} = \mathsf{c}^\top \mathsf{\Lambda}^s \mathsf{c} = \sum_j c_j^2 \lambda_j^s,$$

and so $\mathsf{u}^\top \mathsf{H}(s)\,\mathsf{u} = \|u_h\|_{H_s}^2$ for $u_h \in V_h$ such that $u_h = \pi_h^{-1}\mathsf{u}$. Similarly to the continuous case, the norm can be obtained in terms of powers of an operator

$$\mathsf{u}^\top \mathsf{H}(s)\,\mathsf{u} = \left[\mathsf{U}\mathsf{\Lambda}^{\frac{s}{2}}(\mathsf{MU})^\top \mathsf{u}\right]^\top \mathsf{M} \left[\mathsf{U}\mathsf{\Lambda}^{\frac{s}{2}}(\mathsf{MU})^\top \mathsf{u}\right] = \left[\mathsf{S}^{-\frac{s}{2}}\mathsf{u}\right]^\top \mathsf{M} \left[\mathsf{S}^{-\frac{s}{2}}\mathsf{u}\right],$$

where $\mathsf{S} = \mathsf{A}^{-1}\mathsf{M}$ is the matrix representation of the Riesz map $H^{-1}(\Gamma) \to H_0^1(\Gamma)$ in the basis of V_h.

Remark 2. The norms constructed above for the discrete space are equivalent to, but not identical to, the H_s-norm from the continuous case.

Before considering proper preconditioning of the weak formulation of problem (1), we illustrate the use of operator preconditioning with an example of a boundary value problem where operators in fractional spaces are utilized to weakly enforce the Dirichlet boundary conditions by Lagrange multipliers [6].

Example 3 (Dirichlet boundary conditions using the Lagrange multiplier). The problem considered in [6] reads as follows: Find u such that

$$\begin{aligned} -\Delta u + u &= f && \text{in } \Omega, \\ u &= g && \text{on } \Gamma \subset \partial\Omega, \\ \partial_n u &= 0 && \text{on } \partial\Omega \setminus \Gamma. \end{aligned} \tag{9}$$

Introducing a Lagrange multiplier p for the boundary value constraint and a trace operator $T : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma)$ leads to a variational problem for $(u, p) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\Gamma)$ satisfying

$$\begin{aligned} &(\nabla u, \nabla v)_\Omega + (u, v)_\Omega + \langle p, Tv\rangle_\Gamma = (f, v)_\Omega, && v \in H^1(\Omega), \\ &\langle q, Tu\rangle_\Gamma = \langle q, g\rangle_\Gamma, && q \in H^{-\frac{1}{2}}(\Gamma). \end{aligned} \tag{10}$$

In terms of the framework of operator preconditioning, the variational problem (10) defines an equation

$$\mathcal{A}x = b, \quad \text{where} \quad \mathcal{A} = \begin{bmatrix} -\Delta_\Omega + I & T' \\ T & 0 \end{bmatrix}. \tag{11}$$

In [6] the problem is proved to be well-posed, and therefore $\mathcal{A} : V \to V^*$ is a symmetric isomorphism, where $V = H^1(\Omega) \times H^{-\frac{1}{2}}(\Gamma)$ and $x \in V$, $b \in V^*$. A preconditioner is then $\mathcal{B} \in \mathcal{L}(V^*, V)$, constructed such that $\mathcal{B}$ is a positive, self-adjoint isomorphism. Then $\mathcal{BA} \in \mathcal{L}(V)$ is an isomorphism.

To discretize (11) we shall here employ finite element spaces V_h consisting of linear continuous finite elements where Γ_h is formed by the facets of Ω_h; cf. Figure 1. The stability of discretizations of (10) (for the more general case where the discretization of Ω and Γ are independent) is studied, e.g., in [40] and [42, Ch. 11.3].

The linear system resulting from discretization leads to the following system of equations:

$$\mathbb{BA}\mathsf{x} = \mathbb{B}\mathsf{b}, \tag{12}$$

where

$$\mathbb{B} = \begin{bmatrix} \mathsf{A}^{-1} & \\ & \mathsf{H}(-\frac{1}{2})^{-1} \end{bmatrix} \quad \text{and} \quad \mathbb{A} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^\top \\ \mathsf{B} & \end{bmatrix}.$$

The last block of the matrix preconditioner $\mathbb{B}$ is the inverse of the matrix constructed by (8) (using discretization of an operator inducing the $H^1(\Gamma)$ norm on the second subspace of V_h), and matrix $\mathbb{BA}$ has the same eigenvalues as operator $\mathcal{B}_h\mathcal{A}_h$.

Tables 1 and 2 consider the problem (10) with Ω the unit square and Γ its left edge. In Table 1 we show the spectral condition number of the matrix $\mathbb{BA}$ as a function of the discretization parameter h. It is evident that the condition number is bounded by a constant.

TABLE 1
The smallest and the largest eigenvalues and the spectral condition number of matrix $\mathbb{BA}$ *from system* (12).

h	$\lambda_{\min}$	$\lambda_{\max}$	κ
1.77×10^{-1}	0.311	1.750	5.622
8.84×10^{-2}	0.311	1.750	5.622
4.42×10^{-2}	0.311	1.750	5.622
2.21×10^{-2}	0.311	1.750	5.622
1.11×10^{-2}	0.311	1.750	5.622

TABLE 2
The number of iterations required for convergence of the minimal residual method for system (12) *with* $\mathbb{B}$ *replaced by the approximation* (13).

Size	n_{iters}	$\Vert u - u_h \Vert_{1,\Omega}$
4290	38	$6.76 \times 10^{-2}(1.00)$
16770	40	$3.38 \times 10^{-2}(1.00)$
66306	38	$1.69 \times 10^{-2}(1.00)$
263682	38	$8.45 \times 10^{-3}(1.00)$
1051650	39	$4.23 \times 10^{-3}(1.00)$

Table 2 then reports the number of iterations required for convergence of the minimal residual method [38] with the system (12) of different sizes. The iterations are started from a random initial vector, and for convergence it is required that r_k, the kth residuum, satisfy $\mathsf{r}_k{}^{\top}\overline{\mathbb{B}}\mathsf{r}_k < 10^{-10}$. The operator $\overline{\mathbb{B}}$ is the spectrally equivalent approximation of $\mathbb{B}$ given as[1]

$$\overline{\mathbb{B}} = \operatorname{diag}\left(\mathrm{AMG}\left(\mathsf{A}\right), \mathrm{LU}\left(\mathsf{H}\left(-\tfrac{1}{2}\right)\right)\right). \tag{13}$$

The iteration count appears to be bounded independently of the size of the linear system.

Together the presented results indicate that the constructed preconditioner whose discrete approximation utilizes matrices (8) is a good preconditioner for system (9).

Finally, with $\Omega \in \mathbb{R}^2$, $\Gamma \subset \Omega$ of codimension one, we consider problem (1). The weak formulation of (1a)–(1c), using the method of Lagrange multipliers, defines a variational problem for the triplet $(u, v, p) \in U \times V \times Q$,

$$\begin{aligned} (\nabla u, \nabla \phi)_{\Omega} + \langle p, \epsilon T_{\Gamma}\phi\rangle_{\Gamma} &= (f, \phi)_{\Omega}, &\quad \phi \in U,\\ (\nabla v, \nabla \psi)_{\Gamma} - \langle p, \psi\rangle_{\Gamma} &= (g, \psi)_{\Gamma}, &\quad \psi \in V,\\ \langle \chi, \epsilon T_{\Gamma} u - v\rangle_{\Gamma} &= 0, &\quad \chi \in Q, \end{aligned} \tag{14}$$

where U, V, Q are Hilbert spaces to be specified later. The well-posedness of (14) is guaranteed provided that the celebrated Brezzi conditions (see Appendix A) are fulfilled. We remark that

$$\langle p, T_{\Gamma}\phi\rangle_{\Gamma} = \langle \delta_{\Gamma} p, \phi\rangle_{\Omega}.$$

Hence δ_{Γ} is in our context the dual operator to the trace operator T_{Γ}. Since $T_{\Gamma}: H_0^1(\Omega) \to H_{00}^{\frac{1}{2}}(\Gamma)$, then $\delta_{\Gamma} : H^{-\frac{1}{2}}(\Gamma) \to H^{-1}(\Omega)$.

For our discussion of preconditioners it is suitable to recast (14) as an operator equation for the self-adjoint operator $\mathcal{A}$,

$$\mathcal{A}\begin{bmatrix} u \\ v \\ p \end{bmatrix} = \begin{bmatrix} A_U & & B_U^* \\ & A_V & B_V^* \\ B_U & B_V & \end{bmatrix}\begin{bmatrix} u \\ v \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \\ \\ \end{bmatrix}, \tag{15}$$

[1] Here and in the subsequent numerical experiments AMG is the algebraic multigrid BOOMERAMG from the Hypre library [23], and LU is the direct solver from the UMFPACK library [19]. The libraries were accessed through the interface provided by PETSc [7] version 3.5.3. To assemble the relevant matrices FEniCS library [31] version 1.6.0 and its extension for block-structured systems cbc.block [34] were used. The AMG preconditioner was used with the default options, except for coarsening, which was set to Ruge–Stueben algorithm.

with the operators A_i, B_i, $i \in \{U, V\}$, given by

$$\begin{aligned} \langle A_U u, \phi\rangle_\Omega &= (\nabla u, \nabla\phi)_\Omega\,, \quad \langle A_V v, \psi\rangle_\Gamma = (\nabla v, \nabla\psi)_\Gamma\,, \\ \langle B_U u, \chi\rangle_\Gamma &= \langle \chi, \epsilon T_\Gamma u\rangle_\Gamma, \quad\ \ \langle B_V v, \chi\rangle_\Gamma = -\langle \chi, v\rangle_\Gamma. \end{aligned}$$

Further, for discussion of mapping properties of $\mathcal{A}$ it will be advantageous to consider the operator as a map defined over space $W \times Q$, $W = U \times V$ as

$$\mathcal{A} = \begin{bmatrix} A & B^* \\ B & \end{bmatrix} \quad \text{with} \quad A = \begin{bmatrix} A_U & \\ & A_V \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_U & B_V \end{bmatrix}. \tag{16}$$

Considering two different choices of spaces U, V, and Q, we will propose two formulations that lead to different preconditioners:

$$\mathcal{B}_Q^{-1} = \begin{bmatrix} A_U & & \\ & A_V & \\ & & B_U A_U^{-1} B_U^* + B_V A_V^{-1} B_V^* \end{bmatrix} \tag{17}$$

and

$$\mathcal{B}_W^{-1} = \begin{bmatrix} A_U + B_U^* R B_U & & \\ & A_V & \\ & & B_V A_V^{-1} B_V^* \end{bmatrix}. \tag{18}$$

Here R is the Riesz map from Q^* to Q. Preconditioners of the form (17)–(18) will be referred to as the Q-cap and the W-cap preconditioners. This naming convention reflects the role intersection spaces play in the respected formulations. We remark that the definitions should be understood as templates identifying the correct structure of the preconditioner.

3. Q-cap preconditioner.

Consider operator $\mathcal{A}$ from problem (15) as a mapping $W \times Q \to W^* \times Q^*$,

$$\begin{aligned} W &= H_0^1(\Omega) \times H_0^1(\Gamma)\,, \\ Q &= \epsilon H^{-\frac{1}{2}}(\Gamma) \cap H^{-1}(\Gamma)\,. \end{aligned} \tag{19}$$

The spaces are equipped with norms

$$\|w\|_W^2 = |u|_{1,\Omega}^2 + |v|_{1,\Gamma}^2 \quad \text{and} \quad \|p\|_Q^2 = \epsilon^2 \|p\|_{-\frac{1}{2},\Gamma}^2 + \|p\|_{-1,\Gamma}^2. \tag{20}$$

Since $H^{-\frac{1}{2}}(\Gamma)$ is continuously embedded in $H^{-1}(\Gamma)$, the space Q is the same topological vector space as $H^{-\frac{1}{2}}(\Gamma)$, but equipped with an equivalent, ϵ-dependent inner product. See also [9, Ch. 2]. The next theorem shows that this definition leads to a well-posed problem.

We will need a right inverse of the trace operator and employ the following harmonic extension. Let $q \in H_{00}^{\frac{1}{2}}(\Gamma)$, and let u be the solution of the problem

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega \setminus \Gamma, \\ u &= 0 \quad \text{on } \partial\Omega, \\ u &= q \quad \text{on } \Gamma. \end{aligned} \tag{21}$$

Since the trace is surjective onto $H_{00}^{\frac{1}{2}}(\Gamma)$, (21) has a solution $u \in H_0^1(\Omega)$ and $|u|_{1,\Omega} \leq C|q|_{\frac{1}{2},\Gamma}$ for some constant C. We denote the harmonic extension operator by E, i.e., $u = Eq$ with $\|E\| \leq C$.

THEOREM 4. *Let W and Q be the spaces* (19). *The operator $\mathcal{A} : W \times Q \to W^* \times Q^*$, defined in* (15), *is an isomorphism, and the condition number of $\mathcal{A}$ is bounded independently of $\epsilon > 0$.*

Proof. The statement follows from the Brezzi theorem, Theorem 13, once its assumptions are verified. Since A induces the inner product on W, A is continuous and coercive, and the conditions (51a) and (51b) hold. Next, we see that B is bounded:

$$\begin{aligned}\langle Bw, q\rangle_\Gamma &= \langle q, \epsilon T_\Gamma u - v\rangle_\Gamma\\ &\le \|q\|_{-\frac{1}{2},\Gamma}\|\epsilon T_\Gamma u\|_{\frac{1}{2},\Gamma} + \|q\|_{-1,\Gamma}|v|_{1,\Gamma}\\ &\le \big(1 + \|T_\Gamma\|\big)\sqrt{\epsilon^2\|q\|^2_{-\frac{1}{2},\Gamma} + \|q\|^2_{-1,\Gamma}}\sqrt{|u|^2_{1,\Omega} + |v|^2_{1,\Gamma}}\\ &= \big(1 + \|T_\Gamma\|\big)\|q\|_Q\|w\|_W.\end{aligned}$$

It remains to show the inf-sup condition (51d). Since the trace is bounded and surjective, for all $\xi \in H^{\frac{1}{2}}_{00}(\Gamma)$ we let u be defined in terms of the harmonic extension (21) such that $u = \epsilon^{-1}E\xi$ and $|u|_{1,\Omega} \le \epsilon^{-1}\|E\|\|\xi\|_{\frac{1}{2},\Gamma}$. Hence,

$$\begin{aligned}\sup_{w\in W}\frac{\langle Bw, q\rangle_\Gamma}{\|w\|_W} &= \sup_{w\in W}\frac{\langle q, \epsilon T_\Gamma u - v\rangle_\Gamma}{\sqrt{|u|^2_{1,\Omega} + |v|^2_{1,\Gamma}}}\\ &\ge \big(1 + \|E\|\big)^{-1}\sup_{(\xi,v)\in H^{\frac{1}{2}}_{00}(\Gamma)\times H^1_0(\Gamma)}\frac{\langle q, \xi + v\rangle_\Gamma}{\sqrt{\epsilon^{-2}\|\xi\|^2_{\frac{1}{2},\Gamma} + \|v\|^2_{1,\Gamma}}}.\end{aligned}$$

Note that we have the identity

$$Q^* = \big(\epsilon H^{-\frac{1}{2}}(\Gamma) \cap H^{-1}(\Gamma)\big)^* = \epsilon^{-1}H^{\frac{1}{2}}_{00}(\Gamma) + H^1_0(\Gamma),$$

equipped with the norm

$$\|q^*\|_{Q^*} = \inf_{q^*=q_1^*+q_2^*}\epsilon^{-2}\|q_1^*\|^2_{\frac{1}{2},\Gamma} + |q_2^*|^2_{1,\Gamma}.$$

See also [9]. It follows that

$$\begin{aligned}\sup_{(\xi,v)\in H^{\frac{1}{2}}(\Gamma)\times H^1_0(\Gamma)}\frac{\langle q, \xi + v\rangle_\Gamma}{\sqrt{\epsilon^{-2}\|\xi\|^2_{\frac{1}{2},\Gamma} + |v|^2_{1,\Gamma}}} &= \sup_{\zeta\in Q^*}\sup_{\substack{\xi+v=\zeta\\ v\in H^1_0(\Gamma)}}\frac{\langle q, \xi + v\rangle_\Gamma}{\sqrt{\epsilon^{-2}\|\xi\|^2_{\frac{1}{2},\Gamma} + |v|^2_{1,\Gamma}}}\\ &= \sup_{\zeta\in Q^*}\frac{\langle q, \zeta\rangle_\Gamma}{\inf\limits_{\substack{\xi+v=\zeta\\ v\in H^1_0(\Gamma)}}\sqrt{\epsilon^{-2}\|\xi\|^2_{\frac{1}{2},\Gamma} + |v|^2_{1,\Gamma}}}\\ &= \|q\|_{Q^{**}} = \|q\|_Q.\end{aligned}$$

Consequently, condition (51d) holds with a constant independent of ϵ. □

Following Theorem 4 and [35], a preconditioner for the symmetric isomorphic operator $\mathcal{A}$ is the Riesz mapping $W^* \times Q^*$ to $W \times Q$:

$$\mathcal{B}_Q = \begin{bmatrix}-\Delta_\Omega & & \\ & -\Delta_\Gamma & \\ & & \epsilon^2\Delta_\Gamma^{-\frac{1}{2}} + \Delta_\Gamma^{-1}\end{bmatrix}^{-1}. \tag{22}$$

Here Δ_Γ^s is defined by $\langle \Delta_\Gamma^s v, w\rangle_\Gamma = (v,w)_{H_s}$, with the H_s-inner product defined by (7). Hence the norm induced on $W \times Q$ by the operator $\mathcal{B}_Q^{-1}$ is not (20) but an equivalent norm

$$\langle {\mathcal{B}_Q}^{-1} x, x\rangle = |u|_{1,\Omega}^2 + |v|_{1,\Gamma}^2 + \epsilon^2 \|p\|_{H_{-\frac{1}{2}}(\Gamma)}^2 + \|p\|_{H_{-1}(\Gamma)}^2$$

for any $x = (u, v, p) \in W \times Q$. Note that $\mathcal{B}_Q$ fits the template defined in (17).

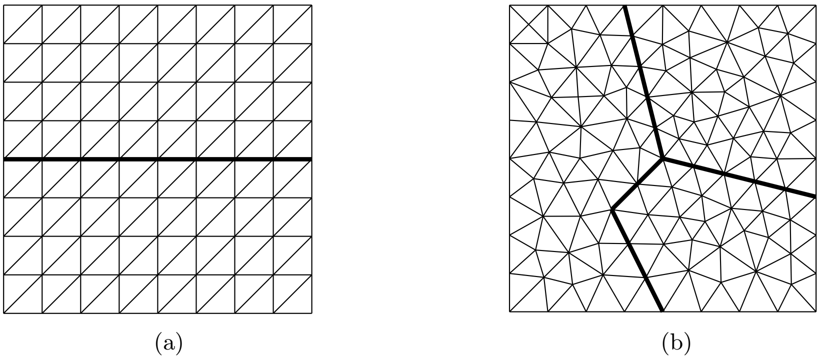


Fig. 1. *Geometrical configurations and their sample triangulations considered in the numerical experiments.*

3.1. Discrete Q-cap preconditioner. Following Theorem 4, the Q-cap preconditioner (22) is a good preconditioner for operator equation $\mathcal{A}x = b$ with the condition number independent of the material parameter ϵ. To translate the preconditioned operator equation $\mathcal{B}_Q\mathcal{A}x = \mathcal{B}_Q b$ into a stable linear system it is necessary to employ suitable discretization. In particular, the Brezzi conditions must hold on each approximation space $W_h \times Q_h$ with constants independent of the discretization parameter h. Such a suitable discretization will be referred to as stable.

Let us consider a stable discretization of operator $\mathcal{A}$ from Theorem 4 by finite dimensional spaces U_h, V_h, and Q_h defined as

$$U_h = \text{span }\{\phi_i\}_{i=1}^{n_U}, \quad V_h = \text{span }\{\psi_i\}_{i=1}^{n_V}, \quad Q_h = \text{span }\{\chi_i\}_{i=1}^{n_Q}.$$

Then the Galerkin method for problem (15) reads as follows: Find $(u_h, v_h, p_h) \in U_h \times V_h \times Q_h$ such that

$$\begin{aligned} (\nabla u_h, \nabla \phi)_\Omega + \langle p_h, \epsilon T_\Gamma \phi\rangle_\Gamma &= (f,\phi)_\Omega\,, && \phi \in U_h,\\ (\nabla v_h, \nabla \psi)_\Gamma - \langle p_h, \psi\rangle_\Gamma &= (g,\psi)_\Gamma\,, && \psi \in V_h,\\ \langle \chi, \epsilon T_\Gamma u_h - v_h\rangle_\Gamma &= 0, && \chi \in Q_h. \end{aligned}$$

Further, we shall define matrices A_U, A_V and B_U, B_V in the following way:

$$\begin{aligned} &\mathsf{A}_U \in \mathbb{R}^{n_U \times n_U}, \quad (\mathsf{A}_U)_{i,j} = (\nabla \phi_j, \nabla \phi_i)_\Omega\,,\\ &\mathsf{A}_V \in \mathbb{R}^{n_V \times n_V}, \quad (\mathsf{A}_V)_{i,j} = (\nabla \psi_j, \nabla \psi_i)_\Gamma\,,\\ &\mathsf{B}_U \in \mathbb{R}^{n_Q \times n_U}, \quad (\mathsf{B}_U)_{i,j} = \langle \epsilon T_\Gamma \phi_j, \chi_i\rangle_\Gamma,\\ &\mathsf{B}_V \in \mathbb{R}^{n_Q \times n_V}, \quad (\mathsf{B}_V)_{i,j} = -\langle \psi_j, \chi_i\rangle_\Gamma. \end{aligned} \tag{23}$$

We note that B_V can be viewed as a representation of the negative identity mapping between spaces V_h and Q_h. Similarly, matrix B_U can be viewed as a composite, $\mathsf{B}_U = \mathsf{M}_{\overline{U}Q}\mathsf{T}$. Here $\mathsf{M}_{\overline{U}Q}$ is the representation of an identity map from space $\overline{U}_h$ to space Q_h. The space $\overline{U}_h$ is the image of U_h under the trace mapping T_Γ. We shall respectively denote the dimension of the space and its basis functions $n_{\overline{U}}$ and $\bar{\phi}_i$, $i \in [1, n_{\overline{U}}]$. Matrix $\mathsf{T} \in \mathbb{R}^{n_{\overline{U}} \times n_U}$ is then a representation of the trace mapping $T_\Gamma : U_h \to \overline{U}_h$.

We note that the rank of T is n_Q, and mirroring the continuous operator T_Γ, the matrix has a unique right inverse T^+. We refer the reader to [36] for the continuous case. The matrix T^+ can be computed as a pseudoinverse via the reduced singular value decomposition $\mathsf{TU} = \mathsf{Q}\Sigma$; see, e.g., [45, Ch. 11]. Then $\mathsf{T}^+ = \mathsf{U}\Sigma^{-1}\mathsf{Q}$. Here, the columns of U can be viewed as coordinates of functions $\bar{\phi}_i$ zero-extended to Ω such that they form the l^2 orthonormal basis of the subspace of $\mathbb{R}^{n_U}$ where the problem $\mathsf{Tu} = \bar{\mathsf{u}}$ is solvable. Further, the kernel of T is spanned by n_U-vectors representing those functions in U_h whose trace on Γ is zero.

For the space U_h constructed by the finite element method with the triangulation of Ω such that Γ is aligned with the element boundaries (cf. Figure 1), it is a consequence of the nodality of the basis that $\mathsf{T}^+ = \mathsf{T}^\top$.

With definitions (23) we use $\mathbb{A}$ to represent the operator $\mathcal{A}$ from (15) in the basis of $W_h \times Q_h$:

$$\mathbb{A} = \begin{bmatrix} \mathsf{A}_U & & {\mathsf{B}_U}^\top \\ & \mathsf{A}_V & {\mathsf{B}_V}^\top \\ \mathsf{B}_U & \mathsf{B}_V & \end{bmatrix}. \tag{24}$$

Finally, a discrete Q-cap preconditioner is defined as a matrix representation of (22) with respect to the basis of $W_h \times Q_h$:

$$\mathbb{B}_Q = \begin{bmatrix} \mathsf{A}_U & & \\ & \mathsf{A}_V & \\ & & \epsilon^2\mathsf{H}(-\tfrac{1}{2}) + \mathsf{H}(-1) \end{bmatrix}^{-1}. \tag{25}$$

The matrices A, M which are used to compute the values $\mathsf{H}(\cdot)$ through the definition (8) have the properties $|p|^2_{1,\Gamma} = \mathsf{p}^\top\mathsf{A}\mathsf{p}$ and $\|p\|^2_{0,\Gamma} = \mathsf{p}^\top\mathsf{M}\mathsf{p}$ for every $p \in Q_h$ and $\mathsf{p} \in \mathbb{R}^{n_Q}$ its coordinate vector. Note that due to properties of matrices $\mathsf{H}(\cdot)$, the matrix N_Q,

$$\mathsf{N}_Q = \left[\epsilon^2\mathsf{H}(-\tfrac{1}{2}) + \mathsf{H}(-1)\right]^{-1} = \mathsf{U}\left[\epsilon^2\Lambda^{-\frac{1}{2}} + \Lambda^{-1}\right]^{-1}\mathsf{U}^\top, \tag{26}$$

is the inverse of the final block of $\mathbb{B}_Q$.

By Theorem 4 and the assumption on spaces $W_h \times Q_h$ being stable, the matrix $\mathbb{B}_Q\mathbb{A}$ has a spectrum bounded independently of the parameter ϵ and the size of the system or equivalently discretization parameter h. In turn, $\mathbb{B}_Q$ is a good preconditioner for matrix $\mathbb{A}$. To demonstrate this property we shall now construct a stable discretization of the space $W \times Q$ using the finite element method.

3.2. Stable subspaces for Q-cap preconditioner.

For $h > 0$ fixed, let Ω_h be the polygonal approximation of Ω. For the set $\bar{\Omega}_h$, we construct a shape-regular triangulation consisting of closed triangles K_i such that $\Gamma \cap K_i$ is an edge e_i of the triangle. Let Γ_h be a union of such edges. The discrete spaces $W_h \subset W$ and $Q_h \subset Q$ shall be defined in the following way. Let

$$\begin{aligned} U_h &= \{v \in C\left(\bar{\Omega}_h\right) : v|_K = \mathbb{P}_1(K)\}, \\ V_h &= \{v \in C\left(\bar{\Gamma}_h\right) : v|_e = \mathbb{P}_1(e)\}, \end{aligned} \tag{27}$$

where $\mathbb{P}_1(D)$ are linear polynomials on the simplex D. Then we set

$$\begin{aligned} W_h &= \left(U_h \cap H_0^1(\Omega)\right) \times \left(V_h \cap H_0^1(\Gamma)\right), \\ Q_h &= V_h \cap H_0^1(\Gamma). \end{aligned} \tag{28}$$

Let A_h, B_h be the finite dimensional operators defined on the approximation spaces (28) in terms of the Galerkin method for operators A, B in (16). Since the constructed spaces are conforming, the operators A_h, B_h are continuous with respect to the norms (20). Further, A_h is W-elliptic on W_h since the operator defines an inner product on the discrete space. Thus, to show that the spaces $W_h \times Q_h$ are stable, it remains to show that the discrete inf-sup condition holds.

LEMMA 5. *Let $W_h \subset W$, $Q_h \subset Q$ be the spaces* (28). *Further, let $\|\cdot\|_W$, $\|\cdot\|_Q$ be the norms* (20). *Finally, let B_h be such that $\langle B_h w_h, q_h\rangle_\Gamma = \langle Bw, q_h\rangle_\Gamma$, $w \in W$. There exists a constant $\beta > 0$ such that*

$$\inf_{q_h \in Q_h} \sup_{w_h \in W_h} \frac{\langle B_h w_h, q_h\rangle_\Gamma}{\|w_h\|_W \|q_h\|_Q} \geq \beta. \tag{29}$$

Proof. Recall that $Q = \epsilon H^{-\frac{1}{2}}(\Gamma) \cap H^{-1}(\Gamma)$. We follow the steps of the continuous inf-sup condition in reverse order. By definition,

$$\begin{aligned} \|q_h\|_Q &= \sup_{p \in \epsilon H_{00}^{\frac{1}{2}}(\Gamma) + H_0^1(\Gamma)} \frac{\langle q_h, p\rangle_\Gamma}{\inf\limits_{p = p_1 + p_2} \sqrt{\epsilon^{-2}\|p_1\|_{\frac{1}{2},\Gamma}^2 + |p_2|_{1,\Gamma}^2}} \\ &= \sup_{p} \sup_{p = p_1 + p_2} \frac{\langle q_h, p_1\rangle_\Gamma + \langle q_h, p_2\rangle_\Gamma}{\sqrt{\epsilon^{-2}\|p_1\|_{\frac{1}{2},\Gamma}^2 + |p_2|_{1,\Gamma}^2}}. \end{aligned} \tag{30}$$

For each $p_1 \in H_{00}^{\frac{1}{2}}(\Gamma)$, let $u_h \in U_h$ be the weak solution of the boundary value problem

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega, \\ \epsilon u &= p_1 && \text{on } \Gamma, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Then $\epsilon T_\Gamma u_h = p_1$ in $H_{00}^{\frac{1}{2}}(\Gamma)$ and $\epsilon |u_h|_{1,\Omega} \leq C\|p_1\|_{\frac{1}{2},\Gamma}$ for some constant C depending only on Ω and Γ. For each $p_2 \in H_0^1(\Gamma)$, let $v_h \in V_h$ be the L^2 projection of p_2 onto the space V_h:

$$\langle v_h - p_2, z\rangle_\Gamma = 0, \quad z \in V_h. \tag{31}$$

By construction we then have $\langle q_h, p_2 - v_h\rangle_\Gamma = 0$ for all $q_h \in Q_h$ and $\|v_h\|_{0,\Gamma} \leq \|p_2\|_{0,\Gamma}$. Moreover, for shape-regular triangulation, the projection $\Pi : H_0^1(\Gamma) \to V_h$, $v_h = \Pi p_2$ is bounded in the H_0^1 norm:

$$|v_h|_{1,\Gamma} \leq |p_2|_{1,\Gamma}. \tag{32}$$

We refer the reader to [10, Ch. 7] for this result. For constructed u_h, v_h it follows from (30) that

$$\begin{aligned} \|q_h\|_Q &\lesssim \sup_{w_h \in U_h + V_h} \sup_{w_h = u_h + v_h} \frac{\langle q_h, \epsilon T_\Gamma u_h + v_h\rangle_\Gamma}{\sqrt{|u_h|_{1,\Omega}^2 + |v_h|_{1,\Gamma}^2}} \\ &= \sup_{(u_h, v_h) \in U_h \times V_h} \frac{\langle q_h, \epsilon T_\Gamma u_h + v_h\rangle_\Gamma}{\|(u_h, v_h)\|_W} = \sup_{w_h \in W_h} \frac{\langle B_h w_h, q_h\rangle_\Gamma}{\|w_h\|_W}. \end{aligned}$$

□

The constructed stable discretizations (28) are a special case of conforming spaces built from $U_{h;k} \subset H^1(\Omega)$ and $V_{h;l} \subset H^1(\Gamma)$ defined as

$$
\begin{aligned}
U_{h;k} &= \{v \in C\left(\bar{\Omega}_h\right) \,:\, v|_K = \mathbb{P}_k(K)\},\\
V_{h;l} &= \{v \in C\left(\bar{\Gamma}_h\right) \,:\, v|_e = \mathbb{P}_l(e)\}.
\end{aligned}
\tag{33}
$$

The following corollary gives a necessary compatibility condition on polynomial degrees in order to build inf-sup stable spaces from components (33).

Corollary 6. *Let* $W_{h;k,l} = \left(U_{h;k} \cap H_0^1(\Omega)\right) \times \left(V_{h;l} \cap H_0^1(\Gamma)\right)$ *and* $Q_{h;m} = V_{h;m} \cap H_0^1(\Gamma)$*. The necessary condition for* (29) *to hold with space* $W_{h;k,l} \times Q_{h;m}$ *is that* $m \leq \max(k,l)$*.*

Proof. Note that $T_\Gamma u_h - v_h$ is a piecewise polynomial of degree $\max(k,l)$. Suppose $m > \max(k,l)$. Then for each $(u_h, v_h) \in W_{h;k,l}$ we can find an orthogonal polynomial $0 \neq q_h \in Q_{h;m}$ such that

$$\langle q_h, T_\Gamma u_h - v_h \rangle_\Gamma = 0.$$

In turn, $\beta = 0$ in (29), and the discrete inf-sup condition cannot hold. □

3.3. Numerical experiments.

Let now $\mathbb{A}$, $\mathbb{B}_Q$ be the matrices (24), (25) assembled over the constructed stable spaces (28). We demonstrate the robustness of the Q-cap preconditioner (22) through a pair of numerical experiments. First, the *exact* preconditioner represented by the matrix $\mathbb{B}_Q$ is considered, and we are interested in the condition number of $\mathbb{B}_Q\mathbb{A}$ for different values of the parameter ϵ. The spectral condition number is computed from the smallest and largest (in magnitude) eigenvalues of the generalized eigenvalue problem $\mathbb{A}\mathsf{x} = \lambda\mathbb{B}_Q^{-1}\mathsf{x}$, which is here solved by SLEPc [27].[2] The obtained results are reported in Table 3. In general, the condition numbers are well behaved, indicating that $\mathbb{B}_Q$ defines a parameter robust preconditioner. We note that for $\epsilon \ll 1$ the spectral condition number is close to $\left(1+\sqrt{5}\right)/\left(\sqrt{5}-1\right) \approx 2.618$. In section 3.4 this observation is explained by the relation of the proposed preconditioner $\mathbb{B}_Q$ and the matrix preconditioner of Murphy, Golub, and Wathen [37].

Table 3
Spectral condition numbers of matrices $\mathbb{B}_Q\mathbb{A}$ *for the system assembled on geometry* (a) *in Figure* 1.

Size	n_Q	$\log_{10}\epsilon$						
		-3	-2	-1	0	1	2	3
99	9	2.655	2.969	4.786	6.979	7.328	7.357	7.360
323	17	2.698	3.323	5.966	7.597	7.697	7.715	7.717
1155	33	2.778	3.905	7.031	7.882	7.818	7.816	7.816
4355	65	2.932	4.769	7.830	8.016	7.855	7.843	7.843
16899	129	3.217	5.857	8.343	8.081	7.868	7.854	7.852
66563	257	3.710	6.964	8.637	8.113	7.872	7.856	7.855

In the second experiment, we monitor the number of iterations required for convergence of the MinRes method [38] (the implementation is provided by cbc.block [34]) applied to the preconditioned equation $\overline{\mathbb{B}}_Q\mathbb{A}\mathsf{x} = \overline{\mathbb{B}}_Q\mathsf{b}$. The operator $\overline{\mathbb{B}}_Q$ is an

[2] We use the generalized Davidson method with the Cholesky preconditioner and convergence tolerance 10^{-8}.

efficient and spectrally equivalent approximation of $\mathbb{B}_Q$,

$$\bar{\mathbb{B}}_Q = \begin{bmatrix} \mathrm{AMG}\left(\mathsf{A}_U\right) & & \\ & \mathrm{LU}(\mathsf{A}_V) & \\ & & \mathsf{N}_Q \end{bmatrix}, \tag{34}$$

with N_Q defined in (26). The iterations are started from a random initial vector, and as a stopping criterion a condition on the magnitude of the kth preconditioned residual r_k, $\mathsf{r}_k{}^{\top}\bar{\mathbb{B}}_Q\mathsf{r}_k < 10^{-12}$ is used. The observed number of iterations is shown in Table 4. Robustness with respect to size of the system and the material parameter is evident as the iteration count is bounded for all the considered discretizations and values of ϵ.

Table 4

Iteration count for convergence of $\bar{\mathbb{B}}_Q\mathbb{A}\mathsf{x} = \bar{\mathbb{B}}_Q\mathsf{b}$ *solved with the minimal residual method. The problem is assembled on geometry* (a) *from Figure* 1.

Size	n_Q	$\log_{10}\epsilon$						
		-3	-2	-1	0	1	2	3
66563	257	20	34	37	32	28	24	21
264195	513	22	34	34	30	26	24	20
1052675	1025	24	33	32	28	26	22	18
4202499	2049	26	32	30	26	24	20	17
8398403	2897	26	30	30	26	22	19	15
11075583	3327	26	30	30	26	22	19	15

Comparing Tables 3 and 4, we observe that the ϵ-behavior of the condition number and the iteration counts are different. In particular, fewer iterations are required for $\epsilon = 10^3$ than for $\epsilon = 10^{-3}$, while the condition number in the former case is larger. Moreover, the condition numbers for $\epsilon > 1$ are almost identical, whereas the iteration counts decrease as the parameter grows. We note that these observations should be viewed in light of the fact that the convergence of the minimal residual method in general does not depend solely on the condition number (e.g., [29]), and a more detailed knowledge of the eigenvalues is required to understand the behavior.

Having proved and numerically verified the properties of the Q-cap preconditioner, we shall in the next section link $\mathbb{B}_Q$ to a block diagonal matrix preconditioner suggested by Murphy, Golub, and Wathen [37]. Both matrices are assumed to be assembled on the spaces (28), and the main objective of the section is to prove spectral equivalence of the two preconditioners.

3.4. Relation to Schur complement preconditioner. Consider a linear system $\mathbb{A}\mathsf{x} = \mathsf{b}$ with an indefinite matrix (24) which shall be preconditioned by a block diagonal matrix

$$\mathbb{B} = \operatorname{diag}\left(\mathsf{A}_U, \mathsf{A}_V, \mathsf{S}\right)^{-1}, \quad \mathsf{S} = \mathsf{B}_U\mathsf{A}_U{}^{-1}\mathsf{B}_U{}^{\top} + \mathsf{B}_V\mathsf{A}_V{}^{-1}\mathsf{B}_V{}^{\top}, \tag{35}$$

where S is the negative Schur complement of $\mathbb{A}$. Following [37], the spectrum of $\mathbb{B}\mathbb{A}$ consists of three distinct eigenvalues. In fact, $\rho\left(\mathbb{B}\mathbb{A}\right) = \{1, \frac{1}{2} \pm \frac{1}{2}\sqrt{5}\}$. A suitable Krylov method is thus expected to converge in no more than three iterations. However, in its presented form, $\mathbb{B}$ does not define an efficient preconditioner. In particular, the cost of setting up the Schur complement comes close to inverting the system matrix $\mathbb{A}$. Therefore, a cheaply computable approximation of S is needed to make the preconditioner practical (see, e.g., [8, Ch. 10.1] for an overview of generic methods for constructing the approximation). We proceed to show that if spaces (28) are used for discretization, the Schur complement is more efficiently approximated with the inverse of the matrix N_Q defined in (26).

Let W_h, Q_h be the spaces (28). Then the mass matrix $\mathsf{M}_{\overline{U}Q} = \mathsf{M}_{VQ}$ (cf. the discussion prior to (23)), and the matrix will be referred to as M. Moreover, let us set $\mathsf{A}_V = \mathsf{A}$. With these definitions the Schur complement of $\mathbb{A}$ reads

$$\mathsf{S} = \epsilon^2 \mathsf{M}\mathsf{T}{\mathsf{A}_U}^{-1}\mathsf{T}^\top\mathsf{M} + \mathsf{M}\mathsf{A}^{-1}\mathsf{M}. \tag{36}$$

Further, note that such matrices A, M are suitable for constructing the approximation of the H_s norm on the space Q_h by the mapping (8). In particular, A is such that $|p|^2_{1,\Gamma} = \mathsf{p}^\top\mathsf{A}\mathsf{p}$ with $p \in Q_h$ and $\mathsf{p} \in \mathbb{R}^{n_Q}$ its coordinate vector. In turn, the inverse of the matrix N_Q reads

$${\mathsf{N}_Q}^{-1} = (\mathsf{M}\mathsf{U})\left(\epsilon^2\Lambda^{-\frac{1}{2}} + \Lambda^{-1}\right)(\mathsf{M}\mathsf{U})^\top = \epsilon^2\mathsf{H}\left(-\tfrac{1}{2}\right) + \mathsf{H}(-1)\,. \tag{37}$$

Recalling that $\mathsf{H}(-1) = \mathsf{M}\mathsf{A}^{-1}\mathsf{M}$ and contrasting (36) with (37), we see that the matrices differ only in the first terms. We shall first show that if the terms are spectrally equivalent, then so are S and ${\mathsf{N}_Q}^{-1}$.

Theorem 7. *Let S, ${\mathsf{N}_Q}^{-1}$ be the matrices defined, respectively, in* (36) *and* (37)*, and let n_Q be their size. Assume that there exist positive constants c_1, c_2 dependent only on Ω and Γ such that for every $n_Q > 0$ and any $\mathsf{p} \in \mathbb{R}^{n_Q}$*

$$c_1\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p} \le \mathsf{p}^\top\mathsf{M}\mathsf{T}{\mathsf{A}_U}^{-1}\mathsf{T}^\top\mathsf{M}\mathsf{p} \le c_2\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p}.$$

Then, for each $n_Q > 0$, matrix S is spectrally equivalent with ${\mathsf{N}_Q}^{-1}$.

Proof. By direct calculation we have

$$\begin{aligned}\mathsf{p}^\top\mathsf{S}\mathsf{p} &= \epsilon^2\mathsf{p}^\top\mathsf{M}\mathsf{T}{\mathsf{A}_U}^{-1}\mathsf{T}^\top\mathsf{M}\mathsf{p} + \mathsf{p}^\top\mathsf{H}(-1)\,\mathsf{p}\\ &\le c_2\epsilon^2\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p} + \mathsf{p}^\top\mathsf{H}(-1)\,\mathsf{p}\\ &\le C_2\mathsf{p}^\top{\mathsf{N}_Q}^{-1}\mathsf{p}\end{aligned}$$

for $C_2 = \sqrt{1 + c_2^2}$. The existence of the lower bound follows from the estimate

$$\mathsf{p}^\top\mathsf{S}\mathsf{p} \ge c_1\epsilon^2\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p} + \mathsf{p}^\top\mathsf{H}(-1)\,\mathsf{p} \ge C_1\mathsf{p}^\top{\mathsf{N}_Q}^{-1}\mathsf{p}$$

with $C_1 = \min(1, c_1)$. □

The spectral equivalence of preconditioners $\mathbb{B}_Q$ and $\mathbb{B}$ now follows immediately from Theorem 7. Note that for $\epsilon \ll 1$ the term $\mathsf{H}(-1)$ dominates both S and ${\mathsf{N}_Q}^{-1}$. In turn, the spectrum of $\mathbb{B}\mathbb{A}$ is expected to approximate well the eigenvalues of $\mathbb{B}_Q\mathbb{A}$. This is then a qualitative explanation of why the spectral condition numbers of $\mathbb{B}_Q\mathbb{A}$ observed for $\epsilon = 10^{-3}$ in Table 3 are close to $\left(1+\sqrt{5}\right)/\left(\sqrt{5}-1\right)$. It remains to prove that the assumption of Theorem 7 holds.

Lemma 8. *There exist constants $c_1, c_2 > 0$ depending only on Ω, Γ such that for all $n_Q > 0$ and $p \in \mathbb{R}^{n_Q}$*

$$c_1\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p} \le \mathsf{p}^\top\mathsf{M}\mathsf{T}{\mathsf{A}_U}^{-1}\mathsf{T}^\top\mathsf{M}\mathsf{p} \le c_2\mathsf{p}^\top\mathsf{H}\left(-\tfrac{1}{2}\right)\mathsf{p}.$$

Proof. For the sake of readability let $n = n_Q$ and $m = n_U$. Since M is symmetric and invertible, $\mathsf{H}\left(-\frac{1}{2}\right) = \mathsf{M}\mathsf{U}\Lambda^{-\frac{1}{2}}\mathsf{U}^\top\mathsf{M}$ and $\mathsf{U}\Lambda^{-\frac{1}{2}}\mathsf{U}^\top = \mathsf{H}\left(\frac{1}{2}\right)^{-1}$, the statement is equivalent to

$$c_1\mathsf{y}^\top\mathsf{H}\left(\tfrac{1}{2}\right)^{-1}\mathsf{y} \leq \mathsf{y}^\top\mathsf{T}\mathsf{A}_U{}^{-1}\mathsf{T}^\top\mathsf{y} \leq c_2\mathsf{y}^\top\mathsf{H}\left(\tfrac{1}{2}\right)^{-1}\mathsf{y} \quad \text{for all } y \in \mathbb{R}^m. \tag{38}$$

The proof is based on properties of the continuous trace operator T_Γ. Recall the trace inequality: There exists a positive constant $K_2 = K_2(\Omega, \Gamma)$ such that $\|T_\Gamma u\|_{\frac{1}{2},\Gamma} \leq K_2|u|_{1,\Omega}$ for all $u \in H_0^1(\Omega)$. From here it follows that the sequence $\{\lambda_m^{\max}\}$, where for each m value $\lambda_m^{\max}$ is the largest eigenvalue of the eigenvalue problem

$$\mathsf{T}^\top\mathsf{H}\left(\tfrac{1}{2}\right)\mathsf{T}\mathsf{u} = \lambda\mathsf{A}_U\mathsf{u}, \tag{39}$$

is bounded from above by K_2. Note that the eigenvalue problem can be solved with a nontrivial eigenvalue only for $\mathsf{u} \in \mathbb{R}^n$ for which there exists some $\mathsf{q} \in \mathbb{R}^m$ such that $\mathsf{u} = \mathsf{T}^\top\mathsf{q}$. Consequently, the eigenvalue problem becomes $\mathsf{T}^\top\mathsf{H}\left(\frac{1}{2}\right)\mathsf{q} = \lambda\mathsf{A}_U\mathsf{T}^\top\mathsf{q}$. Next, applying the inverse of A_U and the trace matrix yields $\mathsf{T}\mathsf{A}_U{}^{-1}\mathsf{T}^\top\mathsf{H}\left(\frac{1}{2}\right)\mathsf{q} = \lambda\mathsf{q}$. Finally, setting $\mathsf{q} = \mathsf{H}\left(\frac{1}{2}\right)^{-1}\mathsf{p}$ yields

$$\mathsf{T}\mathsf{A}_U{}^{-1}\mathsf{T}^\top\mathsf{p} = \lambda\mathsf{H}\left(\tfrac{1}{2}\right)^{-1}\mathsf{p}. \tag{40}$$

Thus the largest eigenvalues of (39) and (40) coincide, and, in turn, $C_2 = K_2$. Further, (40) has only positive eigenvalues, and the smallest nonzero eigenvalue of (39) is the smallest eigenvalue $\lambda_m^{\min}$ of (40). Therefore, for all $\mathsf{y} \in \mathbb{R}^m$ it holds that $\lambda_m^{\min}\mathsf{y}^\top\mathsf{H}\left(\frac{1}{2}\right)^{-1}\mathsf{y} \leq \mathsf{y}^\top\mathsf{T}\mathsf{A}_U{}^{-1}\mathsf{T}^\top\mathsf{y}$. But the sequence $\{\lambda_m^{\min}\}$ is bounded from below since the right-inverse of the trace operator is bounded [36]. □

The proof of Lemma 8 suggests that the constants c_1, c_2 for spectral equivalence are computable as the limit of convergent sequences $\{\lambda_m^{\min}\}$, $\{\lambda_m^{\max}\}$ consisting of the smallest and largest eigenvalues of the generalized eigenvalue problem (40). Convergence of such sequences for the two geometries in Figure 1 is shown in Figure 2. For the simple geometry (a), the sequences converge rather fast, and the equivalence constants c_1, c_2 are clearly visible in the figure. Convergence on the more complex geometry (b) is slower.

So far we have by Theorem 4 and Lemma 5 that the condition numbers of matrices $\mathbb{B}_Q\mathbb{A}$ assembled over spaces (28) are bounded by constants independent of $\{h, \epsilon\}$. A more detailed characterization of the spectrum of the system preconditioned by the Q-cap preconditioner is given next. In particular, we relate the spectrum to computable bounds C_1, C_2 and characterize the distribution of eigenvalues. Further, the effect of varying ϵ (cf. Tables 3–4) is illustrated by numerical experiment.

3.5. Spectrum of the *Q*-cap preconditioned system.

In the following, the left-right preconditioning of $\mathbb{A}$ based on $\mathbb{B}_Q$ is considered, and we are interested in the spectrum of

$$\mathbb{B}_Q^{\frac{1}{2}}\mathbb{A}\mathbb{B}_Q^{\frac{1}{2}} = \begin{bmatrix} \mathsf{I}_U & & \mathsf{A}_U^{-\frac{1}{2}}\mathsf{B}_U{}^\top\mathsf{N}_Q{}^{\frac{1}{2}} \\ & \mathsf{I}_V & \mathsf{A}_V^{-\frac{1}{2}}\mathsf{B}_V{}^\top\mathsf{N}_Q{}^{\frac{1}{2}} \\ \mathsf{N}_Q^{\frac{1}{2}}\mathsf{B}_U\mathsf{A}_U^{-\frac{1}{2}} & \mathsf{N}_Q^{\frac{1}{2}}\mathsf{B}_V\mathsf{A}_V^{-\frac{1}{2}} & \end{bmatrix}. \tag{41}$$

The spectra of the left preconditioner system $\mathbb{B}_Q\mathbb{A}$ and the left-right preconditioned system $\mathbb{B}_Q^{\frac{1}{2}}\mathbb{A}\mathbb{B}_Q^{\frac{1}{2}}$ are identical. Using results of [41] the spectrum ρ of (41) is such that

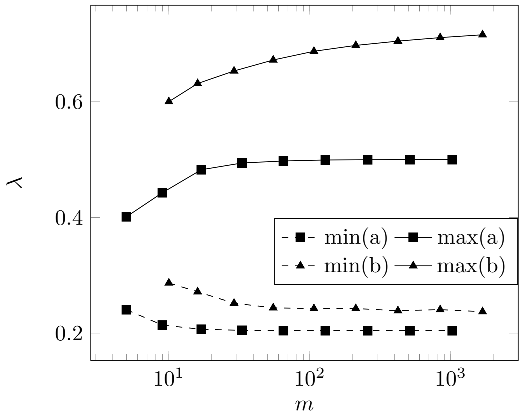


FIG. 2. *Convergence of sequences* $\{\lambda_m^{max}\}$ $\{\lambda_m^{min}\}$ *from Lemma* 8 *for geometries in Figure* 1*. For all sequences but max*(*b*) *the constant bound is reached within the considered range of discretization parameter* $m = n_Q$.

$\rho = I^- \cup I^+$ with

$$I^- = \left[\frac{1-\sqrt{1+4\sigma_{\max}^2}}{2}, \frac{1-\sqrt{1+4\sigma_{\min}^2}}{2}\right], \qquad I^+ = \left[1, \frac{1+\sqrt{1+4\sigma_{\max}^2}}{2}\right], \tag{42}$$

and $\sigma_{\min}, \sigma_{\max}$ the smallest and largest singular values of the block matrix formed by the first two row blocks in the last column of $\mathbb{B}_Q^{\frac{1}{2}}\mathbb{A}\mathbb{B}_Q^{\frac{1}{2}}$. We shall denote the matrix as $\mathbb{D}$:

$$\mathbb{D} = \begin{bmatrix} \mathsf{A}_U^{-\frac{1}{2}}\mathsf{B}_U{}^\top\mathsf{N}_Q{}^{\frac{1}{2}} \\ \mathsf{A}_V^{-\frac{1}{2}}\mathsf{B}_V{}^\top\mathsf{N}_Q{}^{\frac{1}{2}} \end{bmatrix}.$$

PROPOSITION 9. *The condition number* $\kappa\left(\mathbb{B}_Q\mathbb{A}\right)$ *is bounded such that*

$$\kappa\left(\mathbb{B}_Q\mathbb{A}\right) \leq \frac{1+\sqrt{1+4C_2}}{1-\sqrt{1+4C_1}},$$

where C_1, C_2 *are the spectral equivalence bounds from Theorem* 7*.*

Proof. Note that the singular values of matrix $\mathbb{D}$ and the eigenvalues of matrix $\mathsf{N}_Q{}^{\frac{1}{2}}\mathsf{S}\mathsf{N}_Q{}^{\frac{1}{2}}$ are identical. Further, using Theorem 7 with $\mathsf{p} = \mathsf{N}_Q{}^{\frac{1}{2}}\mathsf{q}$, $\mathsf{q} \in \mathbb{R}^{n_Q}$ yields

$$C_1\mathsf{q}^\top\mathsf{q} \leq \mathsf{q}^\top\mathsf{N}_Q{}^{\frac{1}{2}}\mathsf{S}\mathsf{N}_Q{}^{\frac{1}{2}}\mathsf{q} \leq C_2\mathsf{q}^\top\mathsf{q} \quad \text{for all } q \in \mathbb{R}^{n_Q}.$$

In turn, the spectrum of matrices $\mathsf{N}_Q{}^{\frac{1}{2}}\mathsf{S}\mathsf{N}_Q^{\frac{1}{2}}$ is contained in the interval $[C_1, C_2]$. The statement now follows from (42). ◻

From numerical experiments we observe that the bound due to Proposition 9 slightly overestimates the condition number of the system. For example, using numerical trace bounds (cf. Figure 2) of geometry (a) in Figure 1, $c_1 = 0.204, c_2 = 0.499$, and Theorem 7, the formula yields 9.607 as the upper bound on the condition number. On the other hand, condition numbers reported in Table 3 do not exceed 8.637. Similarly, using estimated bounds for geometry (b), $c_1 = 0.237, c_2 = 0.716$, the formula gives the upper bound 8.676. The largest condition number in our experiments (not reported here) was 7.404.

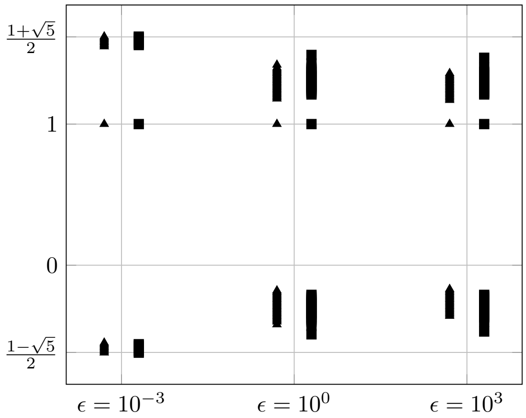


FIG. 3. *Eigenvalues of matrices $\mathbb{B}_Q\mathbb{A}$ assembled on geometries from Figure* 1 *for three different values of ϵ. The value of ϵ is indicated by gray vertical lines. On the left side of the lines is the spectrum for configuration* (a). *The spectrum for geometry* (b) *is then plotted on the right side. For $\epsilon \ll 1$ the eigenvalues cluster near $\lambda = 1$ and $\lambda = \frac{1}{2} \pm \frac{1}{2}\sqrt{5}$ (indicated by gray horizontal lines), which form the spectrum of $\mathbb{B}\mathbb{A}$.*

It is clear that (42) could be used to analyze the effect of the parameter ϵ on the spectrum provided that the singular values $\sigma_{\min}$, $\sigma_{\max}$ were given as functions of ϵ. We do not attempt to give this characterization here. Instead the effect of ϵ is illustrated by a numerical experiment. Figure 3 considers the spectrum of $\mathbb{B}_Q\mathbb{A}$ assembled on geometries from Figure 1 and three different values of the parameter. The systems from the two geometrical configurations are similar in size: 4355 for (a) and 4493 for (b). Note that for $\epsilon \ll 1$ the eigenvalues for both configurations cluster near $\lambda = 1$ and $\lambda = \frac{1}{2} \pm \frac{1}{2}\sqrt{5}$, that is, near the eigenvalues of $\mathbb{B}\mathbb{A}$. This observation is expected in light of the discussion following Theorem 7. With ϵ increasing, the difference between $\mathbb{B}_Q$ and $\mathbb{B}$ caused by $\mathsf{H}\big(-\frac{1}{2}\big)$ becomes visible as the eigenvalues are no longer clustered. Observe that in these cases the lengths of intervals I^-, I^+ are greater for geometry (b). This observation can be qualitatively understood via Proposition 9, Theorem 7, and Figure 2, where the trace map constants c_1, c_2 of configuration (a) are more widely spread than those of (b).

4. *W*-cap preconditioner.

To circumvent the need for mappings involving fractional Sobolev spaces, we shall next study a different preconditioner for (14). As will be seen, the new W-cap preconditioner (18) is still robust with respect to the material and discretization parameters.

Consider operator $\mathcal{A}$ from problem (15) as a mapping $W \times Q \to W^* \times Q^*$, with spaces W, Q defined as

$$\begin{aligned} W &= \left(H^1_0(\Omega) \cap \epsilon H^1_0(\Gamma)\right) \times H^1_0(\Gamma), \\ Q &= H^{-1}(\Gamma). \end{aligned} \tag{43}$$

The spaces are equipped with norms

$$\|w\|^2_W = |u|^2_{1,\Omega} + \epsilon^2 |T_\Gamma u|^2_{1,\Gamma} + |v|^2_{1,\Gamma} \quad \text{and} \quad \|p\|^2_Q = \|p\|^2_{-1,\Gamma}. \tag{44}$$

Note that the trace of functions from space U is here controlled in the norm $|\cdot|_{1,\Gamma}$ and not the fractional norm $\|\cdot\|_{\frac{1}{2},\Gamma}$, as was the case in section 3. Also note that the space

W now is dependent on ϵ while Q is not. The following result establishes the well-posedness of (14) with the above spaces.

THEOREM 10. *Let W and Q be the spaces* (43). *The operator $\mathcal{A} : W \times Q \to W^* \times Q^*$, defined in* (15)*, is an isomorphism, and the condition number of $\mathcal{A}$ is bounded independently of $\epsilon > 0$.*

Proof. The proof proceeds by verifying the Brezzi conditions in Theorem 13. With $w = (u, v)$, $\omega = (\phi, \psi)$, application of the Cauchy–Schwarz inequality yields

$$\begin{aligned}\langle Aw, \omega\rangle_\Omega &= (\nabla u, \nabla\phi)_\Omega + (\nabla v, \nabla\psi)_\Gamma \\ &\le |u|_{1,\Omega}|\phi|_{1,\Omega} + |v|_{1,\Gamma}|\psi|_{1,\Gamma} \\ &\le |u|_{1,\Omega}|\phi|_{1,\Omega} + \epsilon^2|T_\Gamma u|_{1,\Gamma}|\phi|_{1,\Gamma} + |v|_{1,\Gamma}|\psi|_{1,\Gamma} \\ &\le \|w\|_W\|\omega\|_W.\end{aligned}$$

Therefore, A is bounded with $\|A\| = 1$, and (51a) holds. The coercivity of A on $\ker B$ for (51b) is obtained from

$$\begin{aligned}\inf_{w\in\ker B}\frac{\langle Aw, w\rangle_\Omega}{\|w\|_W^2} &= \inf_{w\in\ker B}\frac{|u|_{1,\Omega}^2 + |v|_{1,\Gamma}^2}{|u|_{1,\Omega}^2 + \epsilon^2|T_\Gamma u|_{1,\Gamma}^2 + |v|_{1,\Gamma}^2} \\ &= \inf_{w\in\ker B}\frac{|u|_{1,\Omega}^2 + |v|_{1,\Gamma}^2}{|u|_{1,\Omega}^2 + 2|v|_{1,\Gamma}^2} \ge \frac{1}{2},\end{aligned}$$

where we used that $\epsilon T_\Gamma u = v$ a.e. on the kernel. Consequently, $\alpha = \frac{1}{2}$. Boundedness of B in (51c) with a constant $\|B\| = \sqrt{2}$ follows from the Cauchy–Schwarz inequality:

$$\begin{aligned}\langle Bw, q\rangle_\Gamma &\le \|q\|_{-1,\Gamma}\epsilon|T_\Gamma u|_{1,\Gamma} + \|q\|_{-1,\Gamma}|v|_{1,\Gamma} \\ &\le \sqrt{2}\|q\|_Q\sqrt{\epsilon^2|T_\Gamma u|_{1,\Gamma}^2 + |v|_{1,\Gamma}^2} \\ &\le \sqrt{2}\|q\|_Q\sqrt{|u|_{1,\Omega}^2 + \epsilon^2|T_\Gamma u|_{1,\Gamma}^2 + |v|_{1,\Gamma}^2} \\ &\le \sqrt{2}\|q\|_Q\|w\|_W.\end{aligned}$$

To show that the inf-sup condition holds, compute

$$\begin{aligned}\sup_{w\in W}\frac{\langle Bw, q\rangle_\Gamma}{\|w\|_W} &= \sup_{w\in W}\frac{\langle q, \epsilon T_\Gamma u - v\rangle_\Gamma}{\sqrt{|u|_{1,\Omega}^2 + \epsilon^2|T_\Gamma u|_{1,\Gamma}^2 + |v|_{1,\Gamma}^2}} \\ &\overset{u=0}{\ge} \sup_{v\in V}\frac{\langle q, v\rangle_\Gamma}{|v|_{1,\Gamma}} = \|q\|_Q.\end{aligned}$$

Thus $\beta = 1$ in condition (51d). □

Following Theorem 10, the operator $\mathcal{A}$ is a symmetric isomorphism between spaces $W \times Q$ and $W^* \times Q^*$. As a preconditioner we shall consider a symmetric positive-definite isomorphism $W^* \times Q^* \to W \times Q$:

$$\mathcal{B}_W = \begin{bmatrix}\left(-\Delta_\Omega + T_\Gamma^*\left(-\epsilon^2\Delta_\Gamma\right)T_\Gamma\right)^{-1} & & \\ & \left(-\Delta_\Gamma\right)^{-1} & \\ & & -\Delta_\Gamma\end{bmatrix}. \tag{45}$$

4.1. Discrete preconditioner. Similar to section 3.1, we shall construct discretizations $W_h \times Q_h$ of space $W \times Q$ (43) such that the finite dimensional operator $\mathcal{A}_h$ defined by considering $\mathcal{A}$ from (15) on the constructed spaces satisfies the Brezzi conditions in Theorem 13.

Let $W_h \subset W$ and $Q_h \subset Q$ be the spaces (28) of continuous piecewise linear polynomials. Then A_h, B_h are continuous with respect to norms (44), and it remains to verify conditions (51a) and (51d). First, coercivity of A_h is considered.

LEMMA 11. *Let W_h, Q_h be the spaces (28), and let A_h, B_h be such that $\langle Aw, \omega_h\rangle_\Omega = \langle A_h w_h, \omega_h\rangle_\Omega$, $\langle Bw, q_h\rangle_\Gamma = \langle B_h w_h, q_h\rangle_\Gamma$ for $\omega_h, w_h \in W_h$, $w \in W$, and $q_h \in Q_h$. Then there exists a constant $\alpha > 0$ such that, for all $z_h \in \ker B_h$,*

$$\langle A_h z_h, z_h\rangle \geq \alpha \|z_h\|_W,$$

where $\|\cdot\|_W$ is defined in (44).

Proof. The claim follows from coercivity of A over $\ker B$ (cf. Theorem 10) and the property $\ker B_h \subset \ker B$. To see that the inclusion holds, let $z_h \in \ker B_h$. Since z_h is continuous on Γ, we have from definition $\langle z_h, q_h\rangle_\Gamma = 0$ for all $q_h \in Q_h$ that $z_h|_\Gamma = 0$. But then $\langle z_h, q\rangle = 0$ for all $q \in Q$, and therefore $z_h \in \ker B$. □

Finally, to show that the discretization $W_h \times Q_h$ is stable, we show that the inf-sup condition for B_h holds.

LEMMA 12. *Let spaces W_h, Q_h and operator B_h from Lemma 11 be given. Then there exists $\beta > 0$ such that*

$$\inf_{q_h \in Q_h} \sup_{w_h \in W_h} \frac{\langle B_h w_h, q_h\rangle_\Gamma}{\|w_h\|_W \|q_h\|_Q} \geq \beta, \tag{46}$$

where $\|\cdot\|_Q$ is defined in (44).

Proof. We first proceed as in the proof of Theorem 10 and compute

$$\sup_{w_h \in W_h} \frac{\langle q_h, \epsilon T_\Gamma u_h - v_h\rangle_\Gamma}{\|w_h\|_W} \overset{u_h=0}{\geq} \sup_{v_h \in V_h} \frac{\langle v_h, q_h\rangle_\Gamma}{|v_h|_{1,\Gamma}}. \tag{47}$$

Next, for each $p \in H_0^1(\Gamma)$, let $v_h = \Pi p$ be the element of V_h defined in the proof of Lemma 5. In particular, it holds that

$$\langle p - v_h, q_h\rangle_\Gamma = 0, \quad q_h \in Q_h,$$

and $|v_h|_{1,\Gamma} \leq C|p|_{1,\Gamma}$ for some constant C depending only on Ω and Γ. Then

$$\|q_h\|_{-1,\Gamma} = \sup_{p \in H_0^1(\Gamma)} \frac{\langle q_h, p\rangle_\Gamma}{|p|_{1,\Gamma}} \leq C \sup_{v_h \in V_h} \frac{\langle q_h, v_h\rangle_\Gamma}{|v_h|_{1,\Gamma}}.$$

The estimate together with (47) proves the claim of the lemma. □

Let now $\mathsf{A}_U, \mathsf{A}_V$ and $\mathsf{B}_U, \mathsf{B}_V$ be the matrices defined in (23) as representations of the corresponding finite dimensional operators in the basis of the stable spaces W_h and Q_h. We shall represent the preconditioner $\mathcal{B}_W$ by a matrix

$$\mathbb{B}_W = \begin{bmatrix} \left(\mathsf{A}_U + \epsilon^2 \mathsf{T}^\top \mathsf{A} \mathsf{T}\right)^{-1} & & \\ & \left(\mathsf{A}_V\right)^{-1} & \\ & & \mathsf{H}(-1)^{-1} \end{bmatrix}, \tag{48}$$

where $\mathsf{H}(-1)^{-1} = \mathsf{M}^{-1}\mathsf{A}\mathsf{M}^{-1}$ (cf. (8)) and M, A are the matrices inducing L^2 and H^1_0 inner products on Q_h. Let us point out that there is an obvious correspondence between the matrix preconditioner $\mathbb{B}_W$ and the operator $\mathcal{B}_W$ defined in (18). On the other hand, it is not entirely straightforward that the matrix $\mathbb{B}_W$ represents the W-cap preconditioner defined in (45). In particular, since the isomorphism from $Q^* = H^1_0(\Gamma)$ to $Q = H^{-1}(\Gamma)$ is realized by the Laplacian, a case could be made for using the stiffness matrix A as a suitable representation of the operator.

Let us first argue for A not being a suitable representation for preconditioning. Note that the role of matrix $\mathbb{A} \in \mathbb{R}^{m\times n}$ in a linear system $\mathbb{A}\mathsf{x} = \mathsf{b}$ is to transform vectors from the solution space $\mathbb{R}^n$ to the residual space $\mathbb{R}^m$. In the case when the matrix is invertible, the spaces coincide. However, to emphasize the conceptual difference between the spaces, let us write $\mathbb{A} : \mathbb{R}^n \to \mathbb{R}^{n*}$. Then a preconditioner matrix is a mapping $\mathbb{B} : \mathbb{R}^{n*} \to \mathbb{R}^n$. The stiffness matrix A, however, is such that $\mathsf{A} : \mathbb{R}^{n_Q} \to \mathbb{R}^{n_Q*}$.

It remains to show that $\mathsf{M}^{-1}\mathsf{A}\mathsf{M}^{-1}$ is the correct representation of $A = -\Delta_\Gamma$. Recall that $Q_h \subset Q^*$ and A is the matrix representation of operator $A_h : Q_h \to Q^*_h$. Further, mappings $\pi_h : Q_h \to \mathbb{R}^{n_Q}$, $\mu_h : Q^*_h \to \mathbb{R}^{n_Q*}$,

$$p_h = \sum_j (\pi_h p_h)_j \chi_j, \quad p_h \in Q_h, \qquad \text{and} \qquad (\mu_h f_h)_j = \langle f_j, \chi_j\rangle, \quad f_h \in Q^*_h,$$

define isomorphisms between[3] spaces Q_h, $\mathbb{R}^{n_Q}$ and Q^*_h, $\mathbb{R}^{n_Q*}$, respectively. We can uniquely associate each $p_h \in Q_h$ with a functional in Q^*_h via the Riesz map $I_h : Q_h \to Q^*_h$ defined as $\langle I_h p_h, q_h\rangle_\Gamma = (p_h, q_h)_\Gamma$. Since

$$(\mu_h I_h p_h)_j = (I_h p_h, \chi_j)_\Gamma = \sum_i (\pi_h p_h)_i (\chi_i, \chi_j)_\Gamma,$$

the operator I_h is represented as the mass matrix M. The matrix then provides a natural isomorphism from $\mathbb{R}^{n_Q}$ to $\mathbb{R}^{n_Q*}$. In turn, $\mathsf{M}^{-1}\mathsf{A}\mathsf{M}^{-1} : \mathbb{R}^{n_Q*} \to \mathbb{R}^{n_Q}$ has the desired mapping properties. In conclusion, the inverse of the mass matrix was used in (48) as a natural adapter to obtain a matrix operating between spaces suitable for preconditioning.

Finally, we make a few observations about the matrix preconditioner $\mathbb{B}_W$. Recall that the Q-cap preconditioner $\mathbb{B}_Q$ could be related to the Schur complement based preconditioner (35) obtained by factorizing $\mathbb{A}$ in (24). The relation of $\mathbb{A}$ to the W-cap preconditioner matrix (48) is revealed in the following calculation:

$$\mathbb{U}\mathbb{L}\mathbb{A} = \begin{bmatrix} \mathsf{A}_V + \epsilon^2\mathsf{T}^\top\mathsf{A}\mathsf{T} & & \\ & \tau^2\mathsf{A} & -\mathsf{M} \\ -\epsilon\mathsf{M}\mathsf{T} & & \mathsf{M}\mathsf{A}^{-1}\mathsf{M} \end{bmatrix}, \tag{49}$$

where

$$\mathbb{U} = \begin{bmatrix} \mathsf{I} & & -\mathsf{T}^\top\epsilon\mathsf{A}\mathsf{M}^{-1} \\ & \mathsf{I} & \\ & & \mathsf{I} \end{bmatrix} \quad \text{and} \quad \mathbb{L} = \begin{bmatrix} \mathsf{I} & & \\ & \mathsf{I} & \\ & -\mathsf{M}\mathsf{A}^{-1} & -\mathsf{I} \end{bmatrix}.$$

[3] Note that in section 1 the mapping μ_h was considered as $\mu_h : Q^*_h \to \mathbb{R}^{n_Q}$. The definition used here reflects the conceptual distinction between spaces $\mathbb{R}^{n_Q}$ and $\mathbb{R}^{n_Q*}$. That is, μ_h is viewed as a map from the space of right-hand sides of the operator equation $A_h p_h = L_h$ to the space of right-hand sides of the corresponding matrix equation $\mathsf{A}\mathsf{p} = \mathsf{b}$.

Here the matrix $\mathbb{L}$ introduces a Schur complement of a submatrix of $\mathbb{A}$ corresponding to spaces V_h, Q_h.The matrix $\mathbb{U}$ then eliminates the constraint on the space U_h. Preconditioner $\mathbb{B}_W$ could now be interpreted as coming from the diagonal of the resulting matrix in (49). Further, note that the action of the Q_h-block can be computed cheaply by Jacobi iterations with a diagonally preconditioned mass matrix (cf. [47]).

Table 5

Spectral condition numbers of matrices $\mathbb{B}_W\mathbb{A}$ *for the system assembled on geometry* (a) *in Figure* 1.

Size	$\log_{10}\epsilon$						
	-3	-2	-1	0	1	2	3
99	2.619	2.627	2.546	3.615	3.998	4.044	4.048
323	2.623	2.653	2.780	3.813	4.023	4.046	4.049
1155	2.631	2.692	3.194	3.925	4.036	4.048	4.049
4355	2.644	2.740	3.533	3.986	4.042	4.048	4.049
16899	2.668	2.788	3.761	4.017	4.046	4.049	4.049
66563	2.703	3.066	3.896	4.033	4.047	4.049	4.049

4.2. Numerical experiments.

Parameter robust properties of the W-cap preconditioner are demonstrated by the two numerical experiments used to validate the Q-cap preconditioner in section 3.3. Both experiments use discretization of domain (a) from Figure 1. First, using the *exact* preconditioner, we consider the spectral condition numbers of matrices $\mathbb{B}_W\mathbb{A}$. Next, using an approximation of $\mathbb{B}_W$, the linear system $\bar{\mathbb{B}}_W\mathbb{A}\mathsf{x} = \bar{\mathbb{B}}_W\mathsf{f}$ is solved with the minimal residual method. The operator $\bar{\mathbb{B}}_W$ is defined as

$$\bar{\mathbb{B}}_W = \begin{bmatrix} \mathrm{AMG}\big(\mathsf{A}_U + \epsilon^2\mathsf{T}^\top\mathsf{A}\mathsf{T}\big) & & \\ & \mathrm{LU}(\mathsf{A}) & \\ & & \mathrm{LU}\left(\mathsf{M}\right)\,\mathsf{A}\,\mathrm{LU}\left(\mathsf{M}\right) \end{bmatrix}. \tag{50}$$

The spectral condition numbers of matrices $\mathbb{B}_W\mathbb{A}$ for different values of material parameter ϵ are listed in Table 5. For all the considered discretizations, the condition numbers are bounded with respect to ϵ. We note that the mesh convergence of the condition numbers appears to be faster and the obtained values are in general smaller than in case of the Q-cap preconditioner (cf. Table 3).

Table 6 reports the number of iterations required for convergence of the minimal residual method for the linear system $\bar{\mathbb{B}}_W\mathbb{A}\mathsf{x} = \bar{\mathbb{B}}_W\mathsf{f}$. Like for the Q-cap preconditioner, the method is started from a random initial vector, and the condition ${\mathsf{r}_k}^\top\bar{\mathbb{B}}_W\mathsf{r}_k < 10^{-12}$ is used as a stopping criterion. We find that the iteration counts with the W-cap preconditioner are again bounded for all the values of the parameter ϵ. Consistent with the observations about the spectral condition number, the iteration count is in general smaller than for the system preconditioned with the Q-cap preconditioner.

We note that the observations from section 3.3 about the difference in ϵ-dependence of condition numbers and iteration counts of the Q-cap preconditioner apply to the W-cap preconditioner as well.

Before addressing the question of computational costs of the proposed preconditioners, let us remark that the Q-cap preconditioner and the W-cap preconditioner are not spectrally equivalent. Further, both preconditioners yield numerical solutions with linearly (optimally) converging error; see Appendix B.

5. Computational costs.

We conclude by assessing computational efficiency of the proposed preconditioners. In particular, the setup cost and its relation to the

TABLE 6

Iteration count for system $\overline{\mathbb{B}}_W \mathbb{A}\mathsf{x} = \overline{\mathbb{B}}_W \mathsf{f}$ solved with the minimal residual method. The problem is assembled on geometry (a) *from Figure* 1. *A comparison to the number of iterations with the Q-cap preconditioned system is shown in the brackets (cf. also Table* 4*).*

Size	$\log_{10}\epsilon$						
	-3	-2	-1	0	1	2	3
66563	17(-3)	33(-1)	40(3)	30(-2)	20(-8)	14(-10)	12(-9)
264195	19(-3)	35(1)	39(5)	28(-2)	19(-7)	14(-10)	11(-9)
1052675	22(-2)	34(1)	37(5)	27(-1)	19(-7)	14(-8)	11(-7)
4202499	24(-2)	34(2)	34(4)	25(-1)	17(-7)	12(-8)	9(-8)
8398403	25(-1)	32(2)	32(2)	24(-2)	16(-6)	11(-8)	8(-7)
11075583	25(-1)	32(2)	32(2)	25(-1)	16(-6)	13(-6)	11(-4)

aggregate solution time of the Krylov method is of interest. For simplicity we let $\epsilon = 1$.

In case of the Q-cap preconditioner discretized as (34) the setup cost is determined by the construction of algebraic multigrid (AMG) and the solution of the generalized eigenvalue problem $\mathsf{A}\mathsf{x} = \lambda\mathsf{M}\mathsf{x}$ (GEVP). The problem is here solved by calling the OpenBLAS [46] implementation of the LAPACK [3] routine DSYGVD. The setup cost of the W-cap preconditioner is dominated by the construction of multigrid for operator $\mathsf{A}_U + \mathsf{T}^\top\mathsf{A}\mathsf{T}$. We found that the operator can be assembled with negligible costs and therefore do not report the timings of this operation.

The setup costs of the preconditioners obtained on a Linux machine with 16GB RAM and a single Intel Core i5-2500 CPU clocking at 3.3 GHz are reported in Table 7. We remark that the timings on the finest discretization deviate from the trend set by the predecessors. This is due to SWAP memory being required to complete the operations and the case should therefore be omitted from the discussion. On the remaining discretizations the following observations can be made: (i) the solution time always dominates the construction time by a factor 5.5 for W-cap and 3.5 for Q-cap; (ii) W-cap preconditioner is close to two times cheaper to construct than the Q-cap preconditioner in the form (34); (iii) the eigenvalue problem always takes fewer seconds to solve than the construction of multigrid.

For our problems of about 11 million nodes in the $2d$ domain, the strategy of solving the generalized eigenvalue problem using a standard LAPACK routine provided an adequate solution. However, the DSYGVD routine appears to be nearly cubic in complexity ($\mathcal{O}(n_Q^{2.70})$ or $\mathcal{O}(n_U^{1.35})$; cf. Table 7), which may represent a bottleneck for larger problems. However, the transformation $\mathsf{M}_l^{-\frac{1}{2}}\mathsf{A}\mathsf{M}_l^{-\frac{1}{2}}$ with M_l the lumped mass matrix presents a simple trick providing significant speed-up. In fact, the resulting eigenvalue problem is symmetric and tridiagonal and can be solved with fast algorithms of nearly quadratic complexity [20, 21]. We note that due to the spectral equivalence of M and M_l (e.g., [47]), the trick leads to a preconditioner spectrally equivalent to (25). In particular, the iteration count with lumping is expected to remain bounded. In our experiments (not reported here) the lumped preconditioner leads to convergence in 3–10 fewer iterations than (34). However, the savings should be interpreted in light of the fact that convergence in the two cases is measured with respect to different norms. Note also that the tridiagonal property holds under the assumption of Γ having no bifurcations and that the elements are linear. To illustrate the potential gains with mass lumping, using the transformation and applying the dedicated LAPACK routine DSTEGR, we were able to compute eigenpairs for systems of order 16,000 in about 50 seconds. This presents more than a factor 10 speed-up relative to the original generalized eigenvalue problem. The value should also be viewed in light of the fact that the relevant space U_h has in this case about a quarter billion degrees of freedom. We remark that [28] presents a method for computing all the eigenpairs of the generalized symmetric tridiagonal eigenvalue problem with an estimated quadratic complexity.

Let us briefly mention a few alternative methods for realizing the mapping between fractional Sobolev spaces needed by the Q-cap preconditioner. The methods have a common feature of computing the action of operators rather than constructing the operators themselves. Taking advantage of the fact that $\mathsf{H}(s) = \mathsf{M}\mathsf{S}^{-s}$, $\mathsf{S} = \mathsf{A}^{-1}\mathsf{M}$, the action of the powers of the matrix S is efficiently computable by contour integrals [25], by the symmetric Lanczos process [4, 5], or, in cases when the matrices A, M are structured, by fast Fourier transform [39]. Alternatively, the mapping can be realized by the BPX preconditioner [12, 11] or integral operator based preconditioners (e.g., [43]). The above-mentioned techniques are all less than $\mathcal{O}(n_Q^2)$ in complexity.

In summary, for linear elements and geometrical configurations where Γ is free of bifurcations, the eigenvalue problem required for (8) lends itself to solution methods with complexity nearing that of the multigrid construction. In such cases the Q-cap preconditioner (34) is feasible whenever the methods deliver acceptable performance ($n_Q \sim 10^4$). For larger spaces Q_h, a practical realization of the Q-cap preconditioner could be achieved by one of the listed alternatives.

Table 7

Timings of elements of construction of the Q, W-cap for $\epsilon = 1$ and discretizations from Tables 4 and 6. Estimated complexity of computing quantity v at the ith row, $r_i = \log v_i - \log v_{i-1}/\log m_i - \log m_{i-1}$, is shown in the brackets. Fitted complexity of computing v, $\mathcal{O}(n_Q^r)$ is obtained by least-squares. All fits but GEVP ignore the SWAP-affected final discretization.

n_U	n_Q	Q-cap			W-cap	
		AMG[s]	GEVP[s]	MinRes[s]	AMG[s]	MinRes[s]
66049	257	0.075(1.98)	0.014(1.81)	0.579(1.69)	0.078(1.94)	0.514(1.73)
263169	513	0.299(2.01)	0.066(2.27)	2.286(1.99)	0.309(1.99)	2.019(1.98)
1050625	1025	1.201(2.01)	0.477(2.87)	8.032(1.82)	1.228(1.99)	7.909(1.97)
4198401	2049	4.983(2.05)	3.311(2.80)	30.81(1.94)	4.930(2.01)	30.31(1.94)
8392609	2897	9.686(1.92)	8.384(2.68)	62.67(2.05)	10.64(2.22)	59.13(1.93)
11068929	3327	15.94(3.60)	12.25(2.74)	84.43(2.15)	15.65(2.79)	82.13(2.37)
Fitted complexity		(2.02)	(2.70)	(1.92)	(2.02)	(1.96)

6. Conclusions.

We have studied preconditioning of model multiphysics problem (1) with Γ being the subdomain of Ω having codimension one. Using operator preconditioning [35], two robust preconditioners were proposed and analyzed. Theoretical findings obtained in the present treatise about robustness of preconditioners with respect to material and discretization parameter were demonstrated by numerical experiments using a stable finite element approximation for the related saddle-point problem developed herein. Computational efficiency of the preconditioners was assessed revealing that the W-cap preconditioner is more practical. The Q-cap preconditioner with discretization based on eigenvalue factorization is efficient for smaller problems, and its application to large scale computing possibly requires different means of realizing the mapping between the fractional Sobolev spaces.

Possible future work based on the presented ideas includes extending the preconditioners to problems coupling three-dimensional and one-dimensional domains, problems with multiple disjoint subdomains, and problems describing different physics on the coupled domains. In addition, a finite element discretization of the problem which avoids the constraint of Γ_h being aligned with facets of Ω_h is of general interest.

Appendix A. Brezzi theory.

THEOREM 13 (Brezzi). *The operator $\mathcal{A} : V \times Q \to V^* \times Q^*$ in* (16) *is an isomorphism if the following conditions are satisfied:*

(a) *A is bounded,*

$$\sup_{u\in V}\sup_{v\in V}\frac{\langle Au, v\rangle}{\|u\|_V\|v\|_V} = c_A \equiv \|A\| < \infty; \tag{51a}$$

(b) *A is invertible on* $\ker B$, *with*

$$\inf_{u\in\ker B}\frac{\langle Au, u\rangle}{\|u\|_V^2} \geq \alpha > 0; \tag{51b}$$

(c) *B is bounded,*

$$\sup_{q\in Q}\sup_{v\in V}\frac{\langle Bv, q\rangle}{\|v\|_V\|q\|_Q} = c_B \equiv \|B\| < \infty; \tag{51c}$$

(d) *B is surjective (this is also the inf-sup or LBB condition), with*

$$\inf_{q\in Q}\sup_{v\in V}\frac{\langle Bv, q\rangle}{\|v\|_V\|q\|_Q} \geq \beta > 0. \tag{51d}$$

The operator norms $\|\mathcal{A}\|$ and $\|\mathcal{A}^{-1}\|$ are bounded in terms of the constants appearing in (a)–(d).

Proof. See, for example, [14]. □

Appendix B. Estimated order of convergence. Refinements of a uniform discretization of geometry (a) in Figure 1 are used to establish order of convergence of numerical solutions of a manufactured problem obtained using Q-cap and W-cap preconditioners. The error of discrete solutions u_h and v_h is interpolated by discontinuous piecewise cubic polynomials and measured in the H_0^1 norm. The observed convergence rate is linear (optimal).

Size	Q-cap		W-cap	
	$\lvert u-u_h\rvert_{1,\Omega}$	$\lvert v-v_h\rvert_{1,\Gamma}$	$\lvert u-u_h\rvert_{1,\Omega}$	$\lvert v-v_h\rvert_{1,\Gamma}$
16899	$3.76\times10^{-2}(1.00)$	$1.32\times10^{-2}(1.00)$	$3.76\times10^{-2}(1.00)$	$1.32\times10^{-2}(1.00)$
66563	$1.88\times10^{-2}(1.00)$	$6.58\times10^{-3}(1.00)$	$1.88\times10^{-2}(1.00)$	$6.58\times10^{-3}(1.00)$
264195	$9.39\times10^{-3}(1.00)$	$3.29\times10^{-3}(1.00)$	$9.39\times10^{-3}(1.00)$	$3.29\times10^{-3}(1.00)$
1052675	$4.70\times10^{-3}(1.00)$	$1.64\times10^{-3}(1.00)$	$4.70\times10^{-3}(1.00)$	$1.64\times10^{-3}(1.00)$
4202499	$2.35\times10^{-3}(1.00)$	$8.22\times10^{-4}(1.00)$	$2.35\times10^{-3}(1.00)$	$8.22\times10^{-4}(1.00)$

Acknowledgment. We would like to thank the anonymous referees for their valuable and constructive comments, which improved the presentation of this paper.

REFERENCES

[1] R. A. ADAMS AND J. F. FOURNIER, *Sobolev Spaces*, 2nd ed., Pure Appl. Math. 140, Elsevier, Academic Press, Amsterdam, 2003.

[2] I. AMBARTSUMYAN, E. KHATTATOV, I. YOTOV, AND P. ZUNINO, *Simulation of flow in fractured poroelastic media: A comparison of different discretization approaches*, in Finite Difference Methods, Theory and Applications, I. Dimov, I. Faragó, and L. Vulkov, eds., Lecture Notes in Comput. Sci. 9045, Springer, Berlin, 2015, pp. 3–14.

[3] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[4] M. Arioli, D. Kourounis, and D. Loghin, *Discrete fractional Sobolev norms for domain decomposition preconditioning*, IMA J. Numer. Anal., 33 (2012), pp. 318–342.

[5] M. Arioli and D. Loghin, *Discrete interpolation norms with applications*, SIAM J. Numer. Anal., 47 (2009), pp. 2924–2951, https://doi.org/10.1137/080729360.

[6] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.

[7] S. Balay, J. Brown, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang, *PETSc Users' Manual*, Tech. Report ANL-95/11, Revision 3.4, Argonne National Laboratory, Lemont, IL, 2013.

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

[9] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer, Berlin, 1976.

[10] D. Braess, *Finite Elements*, 3rd ed., Cambridge University Press, Cambridge, UK, 2007.

[11] J. Bramble, J. Pasciak, and P. Vassilevski, *Computational scales of Sobolev norms with application to preconditioning*, Math. Comp., 69 (2000), pp. 463–480.

[12] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[13] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Ser. Rouge, 8 (1974), pp. 129–151.

[14] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer, New York, 1991.

[15] L. Cattaneo and P. Zunino, *Computational models for fluid exchange between microcirculation and tissue interstitium*, Netw. Heterog. Media, 9 (2014), pp. 135–159.

[16] S. N. Chandler-Wilde, D. P. Hewett, and A. Moiola, *Interpolation of Hilbert and Sobolev spaces: Quantitative estimates and counterexamples*, Mathematika, 61 (2015), pp. 414–443.

[17] C. D'Angelo, *Finite element approximation of elliptic problems with Dirac measure terms in weighted spaces: Applications to one- and three-dimensional coupled problems*, SIAM J. Numer. Anal., 50 (2012), pp. 194–215, https://doi.org/10.1137/100813853.

[18] C. D'Angelo and A. Quarteroni, *On the coupling of* $1D$ *and* $3D$ *diffusion-reaction equations: Application to tissue perfusion problems*, Math. Models Methods Appl. Sci., 18 (2008), pp. 1481–1504.

[19] T. A. Davis, *Algorithm* 832*: Umfpack V*4.3*—an unsymmetric-pattern multifrontal method*, ACM Trans. Math. Software, 30 (2004), pp. 196–199.

[20] J. W. Demmel, O. A. Marques, B. N. Parlett, and C. Vömel, *Performance and accuracy of LAPACK's symmetric tridiagonal eigensolvers*, SIAM J. Sci. Comput., 30 (2008), pp. 1508–1526, https://doi.org/10.1137/070688778.

[21] S. I. Dhillon and B. N. Parlett, *Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices*, Linear Algebra Appl., 387 (2004), pp. 1–28.

[22] J. Etienne, J. Lohéac, and P. Saramito, *A Lagrange-Multiplier Approach for the Numerical Simulation of an Inextensible Membrane or Thread Immersed in a Fluid*, preprint, https://hal.inria.fr/inria-00449805, 2010.

[23] R. D. Falgout and U. Meier Yang, *Hypre: A library of high performance preconditioners*, in Computational Science, ICCS 2002, P. M. A. Sloot, A. G. Hoekstra, C. J. K. Tan, and J. J. Dongarra, eds., Lecture Notes in Comput. Sci. 2331, Springer, Berlin, Heidelberg, 2002, pp. 632–641.

[24] S. A. Funken and E. P. Stephan, *Hierarchical basis preconditioners for coupled FEM-BEM equations*, in Boundary Elements: Implementation and Analysis of Advanced Algorithms, W. Hackbusch and G. Wittum, eds., Notes Numer. Fluid Mech. 50, Vieweg+Teubner Verlag, Berlin, 1996, pp. 92–101.

[25] N. Hale, N. J. Higham, and L. N. Trefethen, *Computing* A^{α}*,* $\log(A)$*, and related matrix functions by contour integrals*, SIAM J. Numer. Anal., 46 (2008), pp. 2505–2523, https://doi.org/10.1137/070700607.

[26] H. Harbrecht, F. Paiva, C. Pérez, and R. Schneider, *Multiscale preconditioning for the coupling of FEM-BEM*, Numer. Linear Algebra Appl., 10 (2003), pp. 197–222.

[27] V. Hernandez, J. S. Roman, and V. Vidal, *SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems*, ACM Trans. Math. Software, 31 (2005), pp. 351–362.

[28] K. Li, T.-Y. Li, and Z. Zeng, *An algorithm for the generalized symmetric tridiagonal eigenvalue problem*, Numer. Algorithms, 8 (1994), pp. 269–291.

[29] J. Liesen and . Tichý, *Convergence analysis of Krylov subspace methods*, GAMM Mitt. Ges. Angew. Math. Mech., 27 (2004), pp. 153–173.

[30] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications, Vol.* 1, Grundlehren Math. Wiss. 181, Springer, Berlin, 1972.

[31] A. Logg, K.-A. Mardal, and G. N. Wells, eds., *Automated Solution of Differential Equations by the Finite Element Method*, Lect. Notes Comput. Sci. Eng. 84, Springer, Berlin, 2012.

[32] F. Magouls and F. X. Roux, *Lagrangian formulation of domain decomposition methods: A unified theory*, Appl. Math. Model., 30 (2006), pp. 593–615.

[33] J. Málek and Z. Strakoš, *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*, SIAM Spotlights 1, SIAM, Philadelphia, 2015.

[34] K.-A. Mardal and J. B. Haga, *Block preconditioning of systems of PDEs*, in Automated Solution of Differential Equations by the Finite Element Method, A. Logg, K.-A. Mardal, and G. N. Wells, eds., Lect. Notes Comput. Sci. Eng. 84, Springer, Berlin, 2012, pp. 643–655.

[35] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.

[36] J. Marschall, *The trace of Sobolev-Slobodeckij spaces on Lipschitz domains*, Manuscripta Math., 58 (1987), pp. 47–65.

[37] 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, https://doi.org/10.1137/S1064827599355153.

[38] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629, https://doi.org/10.1137/0712047.

[39] P. Peisker, *On the numerical solution of the first biharmonic equation*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 655–676.

[40] J. Pitkäranta, *Boundary subspaces for the finite element method with Lagrange multipliers*, Numer. Math., 33 (1979), pp. 273–289.

[41] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904, https://doi.org/10.1137/0613054.

[42] O. Steinbach, *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*, Texts Appl. Math., Springer New York, 2008.

[43] O. Steinbach and W. L. Wendland, *The construction of some efficient preconditioners in the boundary element method*, Adv. Comput. Math., 9 (1998), pp. 191–216.

[44] S. Timoshenko, *Theory of Elastic Stability*, 2nd ed., Engineering Societies Monographs, McGraw–Hill, New York, 1961.

[45] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[46] Q. Wang, X. Zhang, Y. Zhang, and Q. Yi, *AUGEM: Automatically generate high performance dense linear algebra kernels on* x86 *CPUs*, in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, SC '13, ACM, New York, 2013, 25, https://doi.org/10.1145/2503210.2503219.

[47] A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.

[48] K. Yosida, *Functional Analysis*, 6th ed., Springer, New York, 1980.

Paper I

Variational data assimilation for transient blood flow simulations

VARIATIONAL DATA ASSIMILATION FOR TRANSIENT BLOOD FLOW SIMULATIONS

S. W. FUNKE*, M. NORDAAS†, Ø. EVJU‡, M. S. ALNÆS§, AND K.-A. MARDAL¶

Abstract. Several cardiovascular diseases are caused from localised abnormal blood flow such as in the case of stenosis or aneurysms. Prevailing theories propose that the development is caused by abnormal wall-shear stress in focused areas. Computational fluid mechanics have arisen as a promising tool for a more precise and quantitative analysis, in particular because the anatomy is often readily available even by standard imaging techniques such as magnetic resolution and computed tomography angiography. However, computational fluid mechanics rely on accurate initial and boundary conditions which is difficult to obtain. In this paper we address the problem of recovering high resolution information from noisy, low-resolution measurements of blood flow using variational data assimilation based on a transient Navier-Stokes model. Numerical experiments are performed in both 2D and 3D and with pulsatile flow relevant for physiological flow in cerebral aneurysms. The results demonstrate that, with suitable regularisation, the model accurately reconstructs flow, even in the presence of significant noise.

Key words. blood flow, variational data assimilation, finite element method, adjoint equations, Navier-Stokes, BFGS

AMS subject classifications. 35Q92, 35Q93, 65K10, 76D55, 35Q30

1. Introduction.

Detailed insight of blood flow has the potential to assist clinical decisions, for example when evaluating the risk of rupture of an aneurysm [5, 47, 43]. However, non-invasive measurement techniques, such as ultrasound or phase-contrast magnetic resonance imaging (PC-MRI), are too still coarse in space and time to unveil potentially important flow details. For example, PC-MRI provides 4D (3D space and time) images of velocity, but is subject to noise and coarse resolution. A promising remedy are computational blood flow models which yield blood flow data at high temporal and spatial resolutions. In addition, they allow for computation of non-observable variables, such as the blood pressure and wall-shear stresses, both of which are considered important factors in vascular diseases [8, 28, 40, 43]. One major challenge is that the validity and accuracy of the results depend on the underlying model assumptions, the model parameters [26, 16], the boundary conditions [33, 38], and the segmentation of the vascular geometry [4, 15, 17]. These parameters are typically partially or fully unknown and specific to each patient [38, 39]. As a result, there is often a notable discrepancy between CFD simulations and PC-MRI measurements [23, 24, 13].

The discrepancy can be reduced by assimilating the physical measurements into the blood flow model, such that the high-resolution simulation the available measurements. This idea of applying data assimilating techniques to blood flow models has received significant attention over the recent years, see [2] for an overview. In particular the variational approach, which identifies the unknown model inputs that minimise the difference between observations and model results, has been studied in the general setting for the optimal control of the Navier-Stokes equations [14, 30, 18] and in the specific case of blood-flow simulations [6, 7, 25, 44]. The mathematical theory

[*] Center for Biomedical Computing, Simula Research Laboratory, Norway (simon@simula.no)

[†] Center for Biomedical Computing, Simula Research Laboratory, Norway

[‡] Center for Biomedical Computing, Simula Research Laboratory, Norway

[§] Center for Biomedical Computing, Simula Research Laboratory, Norway

[¶] Department of Mathematics, University of Oslo, Norway

behind the variational approach is well developed, and in particular well-posedness of the (regularized) inverse minimization problem for both flow and fluid–structure interaction problem has been addressed in, e.g., [18, 19, 36], but the computational complexity presents a problem. Hence, numerical studies are mostly performed on simplified setups, in the sense that they assume steady-state flow and/or 2D geometries. Alternative, more advanced strategies use reduced basis methods and/or Bayesian parameter estimating, c.f. e.g., [9, 29, 32, 37]. In addition, more general strategies of setting boundary conditions have been developed in [22, 12, 46], in particular since the main problem is the determination of the flow division rather than a complete set of boundary conditions as long as flow extensions are appropriate.

In this paper, we investigate the feasibility of variational based data assimilation for a real-world hemodynamics case, in the sense that a 4D nonlinear flow model in a complex geometry is solved with coarse and noisy 4D data. We formulate the data-assimilation problem as a mathematical optimisation problem constrained by the Navier-Stokes equations (section 2). Special considerations will be put on the regularisation, and the inclusion of data that are coarse with respect to the time resolution. Latter is important because the number of samples per cardiac cycle is typically in the order 20-40 while the number of time steps in a CFD simulation typically is 100-10,000. We then formulate present numerical details based on the reduced problem (section 3). Here, the focus is on efficiency by achieving optimisation convergence which is independent of the Navier-Stokes discretisation. We apply the developed scheme to case studies in two and three dimensions (section 4). An idealistic 2D example is used to verify the data assimilation, and test its robustness against noisy data, the regularisation amplitude, the sparsity in the observations and the choice of the controlled boundaries. Finally, a 3D case is considered, based on 4D data obtained from a PC-MRI scan of an arterial bifurcation with aneurysm in a dog. This example is used to test the quality of the data assimilation by comparing the results to a "traditional" blood flow simulation, whose boundary conditions are interpolated directly from the observations.

2. Mathematical formulation.

2.1. Blood flow model. Most computational modelling in cerebral aneurysm studies assume Newtonian flow with rigid walls, which appear to be adequate [42, 10]. Therefore, we model the blood flow through a vessel with the incompressible Navier-Stokes equations

$$\begin{aligned} u_t + (u \cdot \nabla)u - \nu \Delta u + \nabla p &= f \qquad \text{in } \Omega \times (0,T], \\ \nabla \cdot u &= 0 \qquad \text{in } \Omega \times (0,T]. \end{aligned} \tag{1}$$

Here, $\Omega \times (0,T]$ is the space-time domain, u and p are the blood velocity and (scaled) pressure fields, ν is the (kinematic) viscosity and f describes external body forces. A more complete blood flow model could incorporate non-Newtonian effects and the fluid structure interactions between the blood and the vessel wall [45]. However, for the purpose of this paper it is sufficient to consider (1) and to note that the proposed techniques also apply to more complex models.

We only model a small subset of the artery system and the boundaries of the computational domain consists of a physical boundary, the vessel walls, as well non-physical boundaries at inlets and outlets. Again for simplicity, we consider the common scenario of one inlet and two outlets as sketched in figure 1. To close the system,

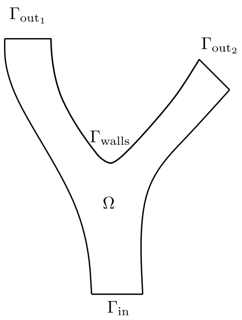


FIG. 1. *The model scenario considered in this paper: a small subset of the artery system with one inlet and two outlet boundaries in 2D and 3D.*

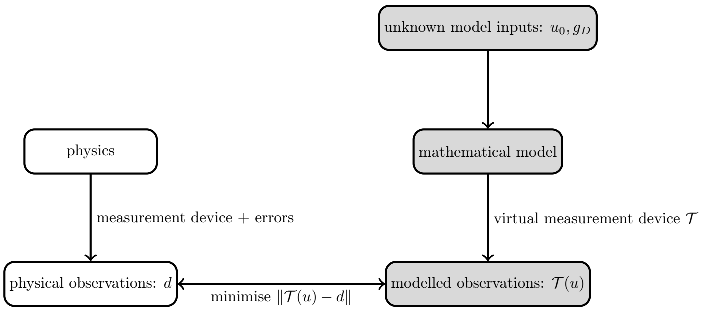

unknown model inputs: u_0, g_D
physics
mathematical model
measurement device + errors
virtual measurement device $\mathcal{T}$
physical observations: d
modelled observations: $\mathcal{T}(u)$
minimise $\|\mathcal{T}(u) - d\|$

FIG. 2. *Variational data assimilation replicates the observation steps in a mathematical model and minimises the discrepancy between measured and modelled observations by varying the model inputs.*

we specify suitable initial and boundary conditions

$$u = u_0, \qquad \text{on } \Omega \times \{t = 0\}, \tag{2a}$$

$$u = g_D, \qquad \text{on } \Gamma_D \times (0, T], \tag{2b}$$

$$pn - \mu \partial_n u = 0, \qquad \text{on } \Gamma_{\text{out}_2} \times (0, T], \tag{2c}$$

$$u = 0, \qquad \text{on } \Gamma_{\text{walls}} \times (0, T]. \tag{2d}$$

with normal vector n and a Dirichlet boundary $\Gamma_D := \Gamma_{\text{in}} \cup \Gamma_{\text{out}_1}$. A traction free boundary is assumed on the outlet Γ_{out_2}, which implies that the vessel is straight in the surroundings of this outlet. If this assumption is not valid, other choices for (2c) are possible, but mass conservation must be ensured, i.e. that $\int_\Gamma u \cdot n \, \mathrm{d}x = 0$. Finally, in some cases it might be beneficial to replace the initial condition by a time-periodic condition, for example if the temporal range of the observations spans one heart cycle.

2.2. Variational data assimilation. Variational data assimilation is a technique to recover unknown data from given observation. The idea is to build a model that replicates the steps of the physical measurement acquisition, and to tweak the free model parameter so that the discrepancy between observed and modelled measurements is minimised (figure 2). Specifically for blood flow, the aim is to recover flow velocity and pressure fields from velocity observations that is has limited spatial and temporal resolution, and is noisy.

The model parameters for equations (1) are the initial condition u_0 and the Dirichlet boundary condition g_D. In the data assimilation setting these inputs are unknown, and instead observational data d is available. We assume that the observations consist of N velocity fields, $d = (d_1, \ldots, d_N) \in L^2(\Omega_{\text{obs}})^N$, on a subdomain $\Omega_{\text{obs}} \subseteq \Omega$. For example, each observation might be an instantaneous snapshot of the blood flow velocity, or the average over a time interval. Since the observations might be available only in parts of the computational domain, we introduced the observation subdomain $\Omega_{\text{obs}} \subseteq \Omega$.

The objective is to recover the initial and boundary conditions by minimising the misfit between simulation and measurement data. Hence we define the goal quantity

$$J(u) = \|\mathcal{T}u - d\|^2 = \sum_{n=1}^{N} \int_{\Omega_{\text{obs}}} |\mathcal{T}_n u - d_n|^2 \, \mathrm{d}x. \tag{3}$$

The observation operators $\mathcal{T}_n$ model the physical measurement procedure, that is they map the simulated velocity u to a simulated measurement outcome. We consider two cases: first, the measurement device takes instantaneously measurements at N timelevels $t_1, .., t_N$. In this case, the observation operators are pointwise evaluations in time, that is $\mathcal{T}_n u = u(t_n)$. Second, the measurement device takes averaged measurements over a time period. In this case, the observation operators are time averaged evaluations of the velocity state, that is $\mathcal{T}_n u = \int_{t_{n-1}}^{t_n} u(t) \, \mathrm{d}t / (t_n - t_{n-1})$.

The data assimilation problem can now be stated as a minimisation problem constrained by the Navier-Stokes equations:

$$\min_{\substack{(u,p)\in Y \\ (u_0, g_D)\in M}} \quad J(u) + \mathcal{R}(u_0, g_D) \quad \text{subject to (1)-(2)}, \tag{4}$$

where M and Y are suitable function spaces, to be determined later in section 3.3 below. The regularisation term $\mathcal{R}$ is required to ensure well-posedeness of the problem, and depends on the amount of observations [6]. In particular, [7] showed that a linearised variation of (2) is well-defined if sufficient observations are available. For the following numerical examples, we apply a Tikhonov regularisation:

$$\mathcal{R}(u_0, g_D) = \frac{\alpha}{2} \|g_D\|^2_{\Gamma_D \times (0,T]} + \frac{\gamma}{2} \|u_0\|^2_{\Omega}, \tag{5}$$

where the coefficients α and γ determine how strongly the problem is regularised in the given norms.

2.2.1. Choice of norms. The choice of norms in the regularisation term (5) specifies the expected regularity of the reconstructed blood flow. For instance, [25] has shown that a unsuitable choice can have a negative impact on the quality of on the reconstructed data.

The norm used for the initial data is

$$\|u_0\|_\Omega = \|u_0\|_{H^1(\Omega)} = \left(\int_\Omega |u_0|^2 + |\nabla u_0|^2 \, \mathrm{d}x \right)^{\frac{1}{2}},$$

and for the boundary

$$\|g_D\|_{\Gamma_D \times (0,T]} = \left(\int_0^T \int_\Omega |g_D|^2 + |\nabla g_D|^2 + |\dot{g}_D|^2 + |\nabla \dot{g}_D|^2 \, \mathrm{d}x \, \mathrm{d}t \right)^{\frac{1}{2}}.$$

We remark that the norms used require more smoothness on the boundary and initial data than is usually required for the variational formulation of the Navier-Stokes equations, in particular for the time derivative $\dot{g}_D$.

3. Numerical solution.

3.1. Formulation of the reduced problem. Two common strategies exist for solving the inverse problem (4) [20, 21]. One strategy is to derive and solve the first-order optimality system. This leads to a large, non-linear system that couples all spatial and temporal degrees of freedoms of the discretised Navier-Stokes and adjoint Navier-Stokes equations. Solving this system is numerically challenging and requires the development of specialised solvers.

The second strategy, taken here, is based on the reduced optimisation problem of (4). The reduced problem is formed by considering the velocity solution as an implicit function of the initial and boundary controls by solving the Navier-Stokes equations (1) and (2). We denote this velocity operator as $u(u_0, g_D)$.

To simplify notation, let $m = (u_0, g_D) \in M$ denote the controlled variable. The functional (3) now has the reduced form

$$\hat{J}(m) := J(u(m)) + \mathcal{R}(m). \tag{6}$$

The *reduced optimisation problem* reads

$$\min_{m \in M} \hat{J}(m). \tag{7}$$

Note that in contrast to (4), the reduced problem is an unconstrained optimisation problem and can hence be solved with established unconstrained optimisation methods. The evaluating of the reduced functional requires the solution of a Navier-Stokes system for which standard solver techniques can be directly applied.

3.2. Optimisation . The reduced minimisation problem (7) is solved with the Broyden-Fletcher-Goldbarb-Shanno (BFGS) algorithm. In this section we present brief overview of the method and its implementation.

The minimisation problem (7) is iteratively solved by generating a sequence of points $m_0, m_1, \ldots,$ approximating a miniser of $\hat{J}$. In each iteration, evaluations of the derivative $D\hat{J}(m_k) \in M^*$ are used to determine a direction $d_k \in M$ in which the functional is decreasing. This general descent algorithm in Hilbert spaces is formulated in algorithm 1.

Algorithm 1 A general descent algorithm in Hilbert spaces, applied to the reduced minimisation problem (7).

```
Choose an initial point $x_0$
for $x = 0, 1, \ldots$ do
  Choose a search direction $d_k = -H_k D\hat{J}(x_k)$
  Choose a step length $\alpha_k > 0$ such that $\hat{J}(x_k + \alpha_k d_k) < \hat{J}(x_k)$
  Set $x_{k+1} = x_k + \alpha_k d_k$
  if converged then
    return
  end if
end for
```

The search direction $d_k = H_k D\hat{J}(m_k)$ is a descent direction if the operator $H_k : M^* \to M$ is positive definite and self-adjoint. The choice of operators H_k mapping derivatives to search directions essentially characterises the method. For example, taking $H_k = H$ as the Riesz operator for M (i.e. choosing d_k to be the gradient of

$\hat{J}$ at m_k) results in a steepest algorithm. Setting $H_k = D^2\hat{J}(m_k)^{-1}$, assuming $\hat{J}$ is convex, results in a Newton algorithm.

In the present paper, the algorithm used is the Broyden–Fletcher–Goldbarb–Shanno (BFGS) algorithm, which is a descent method of quasi-Newton type. Quasi-Newton methods have good convergence properties and do not require evaluations of the Hessian. Instead, such methods maintains an iteratively constructed approximation to the inverse of the Hessian. The update formula specific to the BFGS algorithm is

$$H_{k+1} = \left(1 - \frac{s_{k+1} \otimes y_{k+1}}{\rho_{k+1}}\right) H_k \left(1 - \frac{y_{k+1} \otimes s_{k+1}}{\rho_{k+1}}\right) + \frac{s_{k+1} \otimes s_{k+1}}{\rho_{k+1}} \tag{8}$$

see e.g. [35, Chapter 6]. Here, $\otimes : X \times Y \to \mathcal{B}(Y^*, X)$ denotes the outer product defined by $(x \otimes y)(z) = x\langle z, y\rangle_{Y^*,Y}$, for $x \in X$ and $y \in Y$, where $\langle \cdot, \cdot \rangle_{Y^*,Y}$ denotes duality coupling, and

$$\begin{aligned} s_{k+1} &= m_{k+1} - m_k, \\ y_{k+1} &= D\hat{J}(m_{k+1}) - D\hat{J}(m_k), \\ \rho_{k+1} &= \langle y_{k+1}, s_k \rangle_{M^*,M}. \end{aligned}$$

Note that the initial $H_0 : M^* \to M$ has to prescribed. A natural choice is to take H_0 to be Riesz operator for the space M. That is, H_0 is the unique operators such that

$$(m_0, m_1)_M = \langle H_0^{-1} m_0, m_1 \rangle_{M^*,M} \tag{9}$$

for all $m_0, m_1 \in M$. This definition of H_0 allows for mesh-independent convergence [41], and is readily seen to coincide with the second order partial derivative of $\hat{J}$ with respect to M. If $D^2\hat{J}(m) - H_0^{-1}$ is compact, the method converges superlinearly, see e.g. [27].

For practical implementations, it is common to truncate the update formula (8) and store only the last $3 - 10$ pairs of vectors y_k and s_k. The step lengths α_k in algorithm 1 should chosen to satisfy the Wolfe conditions, which ensures the convergence of the method [35, chapter 6].

3.3. Discretisation. The optimisation method in section 3.2 requires evaluations of the reduced functional $\hat{J}(m)$ and its derivative $D\hat{J}(m)$. Evaluating the reduced functional requires the numerical solution of the Navier-Stokes equations. This is described in section 3.3.1. The derivatives are computed by solving the adjoint equations, described in section 3.3.2.

3.3.1. Discretisation of the Navier-Stokes equations. The Navier-Stokes equations are discretised with a θ time-stepping scheme and the finite element method. The controlled Dirichlet boundary conditions are weakly enforced with the Nitsche method [34]. An advantage of the Nitsche approach is that the boundary values are explicitly included in the variational formulation, which simplifies the (automated) derivation of the adjoint equations. This is exploited in the implementation.

For the spatial discretisation, we consider conforming finite element spaces

$$\begin{aligned} V_h &\subset H^1_{0,\Gamma_{\text{walls}}}(\Omega) = \{u \in H^1(\Omega) : u|_{\Gamma_{\text{walls}}} = 0\} \\ Q_h &\subset L^2(\Omega). \end{aligned} \tag{10}$$

For the time discretisation we assume a partition of the interval $[0, T]$ with a constant timestep δt. Applying a standard θ time-stepping scheme to the Navier-Stokes equations (1), we obtain a sequence of nonlinear problems: For $k = 0, \ldots, N-1$, let $u^{k+\theta} = \theta u^{k+1} + (1-\theta)u^k$ and find $(u^{k+1}, p^{k+1}) \in V_h \times Q_h$ such that

$$\begin{aligned} \frac{u^{k+1} - u^k}{\delta t} - \nu \Delta u^{k+\theta} + (u^{k+\theta} \cdot \nabla) u^{k+\theta} - \nabla p^{k+1} &= 0, \\ \nabla \cdot u^{k+1} &= 0, \end{aligned} \tag{11}$$

subject to the boundary conditions (2). The equations (11) are integrated against test functions $v \in V_h$ and $q \in Q_h$ in order to obtain a nonlinear variational problem at each time t_k,

$$\begin{aligned} 0 = & \int_\Omega \left(\frac{u^{k+1} - u^k}{\delta t} \right) \cdot v + \nu \nabla u^{k+\theta} \colon \nabla v \, \mathrm{d}x \\ & + \int_\Omega (u^{k+\theta} \cdot \nabla) u^{k+\theta} \cdot v \, \mathrm{d}x \\ & + \int_\Omega q \nabla \cdot u^{k+1} + p^{k+1} \nabla \cdot v \, \mathrm{d}x \\ & - \int_{\Gamma_D} \left(\nu \frac{\partial u^{k+\theta}}{\partial n} - p^{k+1} n \right) \cdot v \, \mathrm{d}s \\ & - \int_{\Gamma_D} \left(\theta \nu \frac{\partial v}{\partial n} - q n \right) \cdot (u^{k+1} - g^{k+1}) \, \mathrm{d}s \\ & + \int_{\Gamma_D} \frac{\nu \sigma}{h} (u^{k+1} - g_D^{k+1}) \cdot v \, \mathrm{d}s. \end{aligned} \tag{12}$$

The nonlinear variational problem (12) consists of a volume integral and a boundary integral over Γ_D. The volume integral coincides with the "standard" variational form of (11) obtained when the boundary condition (2b) is strongly imposed. The second, boundary integral part of the variational problem arises from the weakly imposing the Dirichlet boundary condition (15) with Nitsche's method, and is discussed in detail below.

The discrete spaces for the state and control variables are

$$\begin{aligned} Y &= V_h^N \times Q_h^N \\ M &= V_h \times (T_\Gamma V_h)^N, \end{aligned}$$

and we introduce the notation

$$\begin{aligned} y &= (u, p) \in Y, \\ u &= (u_1, \ldots, v_N) \in V_h^N \\ p &= (p_1, \ldots, q_N) \in Q_h^N \\ m &= (u_0, g_1, \ldots, g_N) \in M \end{aligned}$$

The sequence of variational problems (12) is reformulated as an operator equation combining all the time steps,

$$\mathcal{F}(m, y) = \sum_{k=0}^{N-1} \left\{ \mathcal{F}_{k,\Omega}(m, y) + \mathcal{F}_{k,\Gamma_D}(m, y) \right\} = 0, \tag{13}$$

where $\mathcal{F}_{k,\Omega} : Y \to Y^*$ is the operator combining all the volume integrals in (12), i.e.

$$\begin{aligned}\mathcal{F}_{k,\Omega}(m, y; v, q) = &\int_\Omega \left(\frac{u^{k+1} - u^k}{\delta t}\right) \cdot v^{k+1} + \nu \nabla u^{k+\theta} \colon \nabla v^{k+1} \,\mathrm{d}x \\ &+ \int_\Omega (u^{k+\theta} \cdot \nabla) u^{k+\theta} \cdot v^{k+1} \,\mathrm{d}x \\ &+ \int_\Omega q^{k+1} \nabla \cdot u^{k+1} + p^{k+1} \nabla \cdot v^{k+1} \,\mathrm{d}x,\end{aligned} \tag{14}$$

for all $(v, q) = \{(v_k, q_k)\}_{k=1}^N \in Y^*$, for $k = 0, \ldots, N-1$. Note that that this part only involves the initial data u_0 from m. The operator $\mathcal{F}_{k,\Gamma_D} : Y \times M \to Y^*$ combines all the boundary integrals in (12) and reads

$$\begin{aligned}\mathcal{F}_{k,\Gamma_D}(m, y; v, q) = &-\int_{\Gamma_D} \left(\nu \frac{\partial u^{k+\theta}}{\partial n} - p^{k+1} n\right) \cdot v^{k+1} \,\mathrm{d}s \\ &- \int_{\Gamma_D} \left(\theta \nu \frac{\partial v^{k+1}}{\partial n} - q^{k+1} n\right) \cdot (u^{k+1} - g^{k+1}) \,\mathrm{d}s \\ &+ \int_{\Gamma_D} \frac{\nu \sigma}{h} (u^{k+1} - g^{k+1}) \cdot v^{k+1} \,\mathrm{d}s.\end{aligned} \tag{15}$$

The first integral in (15) arises when the integration by parts formula is applied to (11), and the integral would vanish if the Dirichlet boundary condition (2b) were strongly imposed on the space V_h. The remaining terms are added to obtain a variational problem that is consistent and stable, see e.g. [1, 3]. The form (15) is linear and symmetric, and positive definite provided that the parameter σ is sufficiently large. We must also require $\theta > 0$ to apply the Nitsche method.

The numerical examples in section 4 use two common finite element pairs: P2-P1 (Taylor-Hood) and P1-P1. The lowest order discretisation does not satisfy the LBB conditions, and hence requires stabilisation. We used the stabilisation $-\beta h^2 (\nabla p, \nabla q)_\Omega$ where h is the local mesh element size and $\beta = 10^{-3}$ is the stabilisation coefficient.

3.3.2. Adjoint equations. The adjoint equations are used to efficiently compute the functional derivative $\mathrm{d}J/\mathrm{d}m : Y \times M \to M^*$, at a cost of roughly one linearised Navier-Stokes solve.

To derive the adjoint equations consider the Navier-Stokes equations in the operator form $\mathcal{F}(m; y) = 0 \in Y^*$ and a functional $J(y, m) \in \mathbb{R}$. The total derivative of the functional in direction $\tilde{m}$ is

$$\left\langle \frac{\mathrm{d}J}{\mathrm{d}m}, \tilde{m} \right\rangle_{M^*,M} = \left\langle \frac{\partial J}{\partial y}, \frac{\mathrm{d}y}{\mathrm{d}m} \tilde{m} \right\rangle_{Y^*,Y} + \left\langle \frac{\partial J}{\partial m}, \tilde{m} \right\rangle_{M^*,M}. \tag{16}$$

Evaluating (16) directly is challenging because computing $\mathrm{d}y/\mathrm{d}m(m) \in \mathcal{L}(M, Y)$ is computationally expensive. The adjoint approach eliminates this term by taking the derivative of the PDE equation

$$\frac{\partial \mathcal{F}}{\partial y} \frac{\mathrm{d}y}{\mathrm{d}m} + \frac{\partial \mathcal{F}}{\partial m} = 0. \tag{17}$$

and substituting it into (16):

$$\left\langle \frac{\mathrm{d}J}{\mathrm{d}m}, \tilde{m} \right\rangle_{M^*,M} = -\left\langle \frac{\partial \mathcal{F}}{\partial m} \tilde{m}, \left(\frac{\partial \mathcal{F}}{\partial y}\right)^{-*} \frac{\partial J}{\partial y} \right\rangle_{Y^*,Y} + \left\langle \frac{\partial J}{\partial m}, \tilde{m} \right\rangle_{M^*,M}. \tag{18}$$

The functional derivative is then computed in two steps:

1. Compute the adjoint solution $\lambda \in Y$ by solving the adjoint PDE

$$\left(\frac{\partial \mathcal{F}}{\partial y}\right)^* \lambda = -\frac{\partial J}{\partial y} \tag{19}$$

2. Evaluate the derivative with

$$\frac{\mathrm{d}J}{\mathrm{d}m} = \left(\frac{\partial \mathcal{F}}{\partial m}\right)^* \lambda + \frac{\partial J}{\partial m} \tag{20}$$

The computational expensive part is the solution of (19), which involves the solution of a linear PDE.

The adjoint equations (19) can be derived before or after the discretisation of the Navier-Stokes equations [20]. Here, the discretise-then-optimise approach is chosen, which has the advantage that the discretised derivative is the exact derivative of the discretised system. The alternative approach does not guarantee this, and simple descent methods like algorithm 1 may fail and a more robust optimisation algorithm would need to be used.

The adjoint system (19) for the discretised Navier-Stokes operator (13) is

$$\left\langle \left(\frac{\partial \mathcal{F}}{\partial y}\right)^* \lambda, w \right\rangle = \left\langle \left(\frac{\partial \mathcal{F}}{\partial y}\right) w, \lambda \right\rangle = -\left\langle \frac{\partial J}{\partial u}, v \right\rangle, \tag{21}$$

for all $w = (v, q) \in Y$. Note that the derivative of the regularisation term in the functional vanishes because it does not depend on the state. Since the adjoint operator is linear, it can be written in matrix form:

$$\left(\frac{\partial \mathcal{F}}{\partial y}\right)^* = \begin{pmatrix} \frac{\partial \mathcal{F}_0}{\partial y^1} & 0 & 0 & \cdots \\ \frac{\partial \mathcal{F}_1}{\partial y^1} & \frac{\partial \mathcal{F}_1}{\partial y^2} & 0 & \ddots \\ 0 & \frac{\partial \mathcal{F}_2}{\partial y^2} & \ddots & \ddots \end{pmatrix}^* = \begin{pmatrix} \frac{\partial \mathcal{F}_0^*}{\partial y^1} & \frac{\partial \mathcal{F}_1^*}{\partial y^1} & 0 & \cdots \\ 0 & \frac{\partial \mathcal{F}_1^*}{\partial y^2} & \frac{\partial \mathcal{F}_2^*}{\partial y^2} & \ddots \\ 0 & 0 & \ddots & \ddots \end{pmatrix} \tag{22}$$

or more compactly,

$$\left(\frac{\partial \mathcal{F}}{\partial y^k}\right)^* \lambda = \begin{cases} \left(\frac{\partial \mathcal{F}_{k-1}}{\partial y^k}\right)^* \lambda_k + \left(\frac{\partial \mathcal{F}_k}{\partial y^k}\right)^* \lambda_{k+1} & \text{if } k < N \\ \left(\frac{\partial \mathcal{F}_{k-1}}{\partial y^k}\right)^* \lambda_k & \text{if } k = N. \end{cases}$$

The system (22) is upper-triangular, hence the adjoint (21) is solved by backwards substitution. Written explicitly, the volume integrals in the equation for λ_k, $k = 0, \ldots, N$, are

$$\begin{aligned} &\int_\Omega \left(\frac{\lambda^k - \lambda^{k+1}}{\delta t}\right) \cdot v^k \,\mathrm{d}x + \int_\Omega \nu \nabla \lambda^{k+\tilde{\theta}} \colon \nabla v^k \,\mathrm{d}x \\ &+\theta \int_\Omega (u^{k-1+\theta} \cdot \nabla) v^k \cdot \lambda_u^k \,\mathrm{d}x + \theta \int_\Omega (v^k \cdot \nabla) u^{k-1+\theta} \cdot \lambda_u^k \,\mathrm{d}x \\ &+\tilde{\theta} \int_\Omega (u^{k+\theta} \cdot \nabla) v^k \cdot \lambda_u^{k+1} \,\mathrm{d}x + \tilde{\theta} \int_\Omega (v^k \cdot \nabla) u^{k+\theta} \cdot \lambda_u^{k+1} \,\mathrm{d}x \\ &+ \int_\Omega \lambda_p^k \nabla \cdot v^k + q^k \nabla \cdot \lambda_u^k \,\mathrm{d}x, \end{aligned} \tag{23}$$

with $(\lambda_u^k, \lambda_p^k) = \lambda^k$, $\tilde{\theta} = 1 - \theta$ and setting $\lambda_u^{N+1} = 0$. Similarly, the boundary integrals are

$$\begin{aligned} &-\int_{\Gamma_D} \left(\nu \frac{\partial v^k}{\partial n}\right) \cdot \lambda_u^{k+\tilde{\theta}} \,\mathrm{d}s + \int_{\Gamma_D} \left(q^k n\right) \cdot \lambda_u^k \,\mathrm{d}s \\ &-\int_{\Gamma_D} \left(\theta\nu \frac{\partial \lambda_u^k}{\partial n} - \lambda_p^k n\right) \cdot v^k \,\mathrm{d}s + \int_{\Gamma_D} \frac{\nu\sigma}{h} v^k \cdot \lambda_u^k \,\mathrm{d}s. \end{aligned} \tag{24}$$

The adjoint equations are solved backwards in time, starting from a zero final condition. The timestepping scheme is the same θ-scheme as for the forward discretisation, but with a modified advective velocity. The homogeneous Dirichlet boundary conditions on the controlled surfaces are enforced with a Nitsche like approach.

3.4. Implementation and verification. The Navier-Stokes solver was implemented in the FEniCS finite element framework [31]. The adjoint solver was automatically derived via the algorithmic differentiation tool dolfin-adjoint [11]. The correctness of the adjoint equations, and the resulting derivatives of the goal functional, were verified using the Taylor remainder convergence test. This test checks that for a sufficiently smooth functional $\hat{J}$, a correct implementation should satisfy

$$\left| \hat{J}(m + h\delta m) - \hat{J}(m) - h \left\langle \frac{\mathrm{d}\hat{J}(m)}{\mathrm{d}m}, \delta m \right\rangle \right| = O(h^2), \tag{25}$$

where $\delta m = (\delta u_0, \delta g_D)$ is the perturbation direction and $h > 0$ the perturbation size. The Taylor remainder convergence test was applied to the 2D aneurysm example in section 4.1, for which second order convergence was consistently observed. This gives confidence that the adjoint solver is correctly implemented.

4. Experiments. The data assimilation is applied to two experiments. The first experiment uses an aneurysm-like domain in 2D with known exact solution (section 4.1). The second experiment aims to reconstruct the flow conditions in a real geometry in 3D with observations from an 4D MRI scan (section 4.2).

The implementation and data to reproduce the 2D results are available on https://bitbucket.org/biocomp/navier_stokes_data_assimilation. This website contains a Readme file with instructions for the installation and how to reproduce the paper results.

4.1. 2D Aneurysm. The first experiment tests the variational data assimilation under idealised conditions where the blood flow to be reconstructed is known. This is used to study the robustness of the reconstruction against incomplete data (both in space and time), noisy observations, the amount of regularisation. We also test two different types of observation operators.

The computational domain, shown in figure 3, resembles a blood vessel bifurcation with an aneurysm in 2D. The observations were generated with the same numerical model that was used in the data assimilation procedure. That is, the Navier-Stokes equations was first solved on an extended domain (including the gray area in figure 3) and the velocity solution used to generate the observations. For this setup, the initial velocity was set to zero. On the inlet and right outlet boundaries a parabolic velocity profiles was enforced with peak values of 1000 mm/s (inlet) and 870 mm/s (right outlet), multiplied by $\sin(\pi(1-t)^3)$ to obtain a pulse like flow pattern[1]. The simulation

[1] Note that real flow in cerebral arteries will never go to zero, but rather pulsate between around 0.5 m/s and one third of 0.5 m/s.

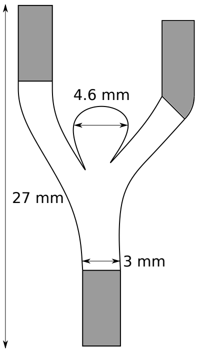


Fig. 3. *The computational domain for the 2D example. The grey area indicates the extended area used to generate the measurement data.*

Parameter	Symbol	Value	Unit
Viscosity	ν	3.5	m^2/s
Model timestep	Δt	0.004625	s
End time	T	0.555	s
Time discretisation	θ	0.5	
Spatial discretisation		$P1$-$P1$	
Mesh triangles		$20,989$	
Mesh element size		$0.03-0.20$	mm
Nitsche coefficient	σ	100.0	
Number of observations	N	16	
Regularisation	$\alpha=\gamma$	10^{-5}	

Table 1

The settings for the 2D aneurysm flow reconstruction. The first parameters specify the model setup, while the final two parameters configure the data assimilation.

started from a zero velocity at $t=0$ and was terminated after the peak velocity at $t=0.629$. The remaining model settings are listed in table 1.

The observation operator $\mathcal{T}$ was applied to the resulting velocity to obtain $N=16$ observations. We compared two observation operators: the *instantaneous observation operator*, which takes instantaneous measurements at evenly distributed times t_n:

$$\mathcal{T}^{\text{inst}}u := \mathcal{R}_{\Omega_{\text{obs}}}u(t_n), \tag{26}$$

where $\mathcal{R}_{\Omega_{\text{obs}}}$ restricts the velocity to the observation domain (the white area in figure 3). The restriction avoids the “inverse crime” and simulates the incompleteness of real measurement data. The *time-averaging observation operator* takes pointwise time-averaged observations over each observation time interval:

$$\mathcal{T}^{\text{avg}}u := \frac{1}{t_n - t_{n-1}}\int_{t_{n-1}}^{t_n} \mathcal{R}_{\Omega_{\text{obs}}}u(t)\mathrm{d}t. \tag{27}$$

The data assimilation was then applied to recover the original flow from the observations. The reconstructions were performed on the restricted domain Ω_{obs}, that is without any knowledge about the geometry of the extended domain. Furthermore, the outflow Dirichlet boundaries were swapped between the data generation and reconstruction to further avoid the "inverse crime". The optimisation was terminated when the relative change of the functional in one iteration dropped below $|J(x_k) - J(x_{k-1})|/|J(x_0) \leq 10^{-4}$ or if the number of iteration exceeded 100.

The results of the data assimilation with $\mathcal{T}^{\text{inst}}$ and $\mathcal{T}^{\text{avg}}$ are shown in figures 4 and 5 (left column), respectively. The first three plots show the observed and reconstructed velocities at $t = 0.296$ s. Visually, the observed and assimilated velocities agree well.

Since the true velocity is known from the initial simulation, the reconstruction error can also be quantified more rigorously. We define following two error measures: the first measures the relative error of the reconstructed velocity in the aneurysm

$$\mathcal{E}_{\Omega_{\text{ane}}} = \frac{\|u_{true} - u\|_{\Omega_{\text{ane}} \times (0,T]}}{\|u_{true}\|_{\Omega_{\text{ane}} \times (0,T]}}, \tag{28}$$

where u_{true} is the true velocity. The second measures the relative error of the reconstructed wall shear stress on the aneurysm wall, motivated by the fact that this an important diagnostic value in blood flow simulations:

$$\mathcal{E}_{\text{WSS}} = \frac{\|\text{WSS}(u_{true}) - \text{WSS}(u)\|_{\Gamma\text{ane} \times (0,T]}}{\|\text{WSS}(u_{true})\|_{\Gamma\text{ane} \times (0,T]}}, \tag{29}$$

with $\text{WSS}(u) = |\sigma n - (\sigma n \cdot n)n|$ and $\sigma = \rho\left(-pI + \nu(\nabla u + (\nabla u)^T)\right)$ with $\rho = 1060$ kg/m^3. The timeplots in figures 4 and 5 (left column) visualise these error measures over the simulation period. The results show a good agreement throughout the simulation period.

4.1.1. Sensitivity of reconstruction with respect to parameter changes. In this section, we investigate how the quality of the reconstruction depends on noise in the observations and the choice of the reconstruction parameters, such as the amount of regularisation. Since the exact solution is known for this example, we can visualise the reconstructed and the "true" velocities and compute the error measures (28) and (29). The following tests are based on the configuration listed in table 1, and in each test one parameter is varied and the quality of the reconstruction investigated.

Noisy observations (figures 4 and 5). Pointwise Gaussian white noise was added to the observations with zero mean and varying magnitude. This type of noise is not expected in real observations, in particular because it depends on the numerical mesh. Nevertheless, we consider it as a suitable benchmark setup. The results of the reconstruction for different signal to noise ratios are shown in figure 4, and for the instantaneous observation operator $\mathcal{T}^{\text{inst}}$ and figure 5 for the time-averaging observation operator $\mathcal{T}^{\text{avg}}$. With increasing level of noise, the optimised functional value $\mathcal{J} + \mathcal{R}$ increases, because of the increased the difference between reconstructed and observed velocity. Nevertheless, the error measures remain small, showing that the reconstruction works reliable even for high noise to signal ratios. Overall, the assimilated flows and metrics agree well for all noise levels, and one can conclude that the reconstruction is little affected by this type of noise. The data assimilation acts like a denoising procedure thanks to the mathematical model. This is consistent with the observations in [6].

Regularisation (figures 6 and 7). The regularisation terms (5) are enforcing smoothness on the control functions. Hence the choice of the regularisation coefficients α and γ could have a strong influence on the assimilation results. For the experiments, we varied α and γ coefficients simultaneously to retain the balance between the two regularisation terms. The results for different regularisation values are shown in figures 6 and 7 for $\mathcal{T}^{\text{inst}}$ and $\mathcal{T}^{\text{avg}}$, respectively. The reconstruction works well for values between 10^{-3} and 10^{-5}, but the quality starts to reduce visibly when $\alpha = \gamma > 10^{-2}$. For the case $\alpha = \gamma = 1$, the assimilated velocity is significantly lower than the true velocity, because the strong regularisation enforces spatially and temporally nearly constant controls. A possible approach is to use the discrepancy principle, that is to select the regularisation parameter such that the perturbation of the regularisation term affects the solution with the same order of the discrepancy induced by the noise. However, this approach is computationally demanding. An alternative, computationally cheaper approach has been proposed by [6].

Data sparsity (figures 8 and 9). Another important question is how many observations (N in (3)) are required to accurately reconstruct the blood flow. To address this question, the data assimilation was repeated with varying number of observations N. The base setup (left column in figures 4 and 5) used $N = 16$ and the results for $N = 4, 8$ and 32 are shown in figures 8 and 9 for $\mathcal{T}^{\text{inst}}$ and $\mathcal{T}^{\text{avg}}$, respectively.

With 4 observations the quality of the reconstruction suffers visibly, mostly at the beginning and the end of the simulation times. The time-averaging observation operator yields good results already with 8 observations, while the instantaneous observation operator requires 16 observations to yield an accurate reconstruction. The differences between 16 to 32 observations are minimal for both observation operators.

Choice of controlled outflow boundary (figure 10). In the problem definition (2b), we made a choice to control the outflow on Γ_{out_1}, and to enforce a no-stress condition on Γ_{out_2}. It is therefore natural to check if the reconstruction works well also if Γ_{out_2} is controlled and a non-stress condition is applied on Γ_{out_1}. The results for this setup are shown in figure 10. The reconstruction is similarly good as in the base setup, indicating that the assimilation is not impacted significantly by the choice of the controlled boundaries. Nevertheless, this choice might be more significant for other setups, in particular if one of the outflows is in close proximity to the aneurysm.

4.2. Flow reconstruction in an aneurysm from 4D MRA measurements. In this experiment, the variational data assimilation was applied to reconstruct the blood flow conditions in an artificially introduced aneurysm. The measurements were done using 4D PC-MRI, and are described in detail in [24]. The geometry was reconstructed from an image obtained by time-averaging the observations, using VMTK (www.vmtk.org). The observations were then linearly interpolated onto the resulting mesh nodes.

The numerical settings are listed in table 2. The assimilation terminated after 30 optimisation iterations, when the relative change in one optimisation iteration dropped below 0.09%.

For comparison, we additionally performed a high-resolution, low-viscosity flow simulation with commonly used strategies for the initial and boundary conditions, but without the data assimilation procedure described in this paper. To avoid spurious effects near the boundary for this simulation, the segmented geometry needed to be extended by artificial straight arteries on the in- and outlets. The high-resolution setup had 20 million DOFs with a Taylor-Hood pressure-corrector scheme and a timestep of $5.9 \cdot 10^{-4}$ s. Womersley boundary conditions were used on the inflow and outflows. The inflow flux $Q_{\Gamma_{\text{in}}}$ was interpolated from averaged observations as $(Q_{\Gamma_{\text{in}}} - Q_{\Gamma_{\text{out}_1}} - Q_{\Gamma_{\text{out}_2}})/3$, the outflow flux on Γ_{out_1} was averaged as $Q_{\Gamma_{\text{out}_1}}/(Q_{\Gamma_{\text{out}_1}} + Q_{\Gamma_{\text{out}_2}})$, and a traction free condition was applied on Γ_{out_2}.

The results for the data assimilation approach and the high-resolution solver are

Parameter	Symbol	Value	Unit
Viscosity	ν	7.5	m^2/s
Model timestep	Δt	0.004625	s
End time	T	0.629	s
Time discretisation	θ	1.0	
Spatial discretisation		$P1$-$P1$	
Dimension of spatial discretisation		184,464	
Nitsche coefficient	σ	100.0	
Number of observations	N	16	
Regularisation parameter	$\alpha = \gamma$	10^{-5}	

Table 2

The numerical settings for the reconstruction of blood flow from 4D MRA measurements. The first parameters specify the model setup, while the final two parameters configure the data assimilation.

shown in figure 11. The figure shows the data assimilation with the instantaneous observation operator - the results for the averaged observation operator look similar. The high-resolution solution has transient to turbulent behaviour in the aneurysm, while the assimilated solution is laminar. Visually, the assimilated solution fits better to the observed velocity, both in the vessel and the aneurysm areas.

5. Conclusion and future work. This paper presented the application of variational data assimilation to reconstruct transient blood flow from observations such as MRI images. This technique is well known in other scientific fields such as ocean science and meteorology, but has thus far not been applied to 3D transient blood flow reconstruction. Mathematically, the data assimilation problem is an optimisation problem constrained by the Navier-Stokes equations. We derived the reduced formulation and described the numerical solution with a focus on retaining the function spaces in the optimisation to obtain mesh-independent iteration numbers in the optimisation step.

The data assimilation was applied to two examples: first, the reconstruction of blood flow in an idealised blood vessel with known solution. This example was used to demonstrate that the proposed method is robust against user parameters and noisy observations. The second example was based on real 4D MRI measurements in a three-dimensional domain, and the result compared to a high-resolution blood flow simulation.

Even though the considered blood flow model and observation operators are simplified, the presented framework extends naturally to more complex setups. Possible extension is to take into account the movement of the vessel wall, non-Newtonian effects or a more realistic observation operator that reimplements an existing measurement device. Furthermore, the reconstruction controls could be extended, for example to also reconstruct the vessel geometry along with the initial and boundary conditions.

Future improvements of the model should include periodic boundary conditions and remove assumptions of traction free boundary on Γ_{out_2}.

The data assimilation procedure introduces an additional computational burden on the flow reconstruction process - for the discussed examples the data assimilation is typically around 50 times more computationally expensive than a single flow simulation. To keep the computational time feasible, the mesh and time resolutions had to be reduced compared to a single flow simulation study. A simple solution would be to first perform a data assimilation on a coarse setup, and then apply the reconstructed initial and boundary conditions on a high-resolution simulation.

Acknowledgments. This research was supported by The Research Council of Norway through a Centres of Excellence grant to the Center for Biomedical Computing at Simula Research Laboratory, project number 179578, and a FRIPRO grant, project number 251237. Computations were performed on the Abel supercomputing cluster at the University of Oslo via NOTUR projects NN9279K and NN9316K. In addition, the authors would like to thank Jingfeng Jiang and Charles Strother for providing the 4D phase-contrast MRA data, Lorenz John for comments on the regularisation and the reviewers for their constructive comments.

REFERENCES

[1] R. Becker, E. Burman, and P. Hansbo, *A Nitsche extended finite element method for incompressible elasticity with discontinuous modulus of elasticity*, Computer Methods in Applied Mechanics and Engineering, 198 (2009), pp. 3352–3360, doi:10.1016/j.cma.2009.06.017.

[2] L. Bertagna, M. D'Elia, M. Perego, and A. Veneziani, *Data assimilation in cardiovascular fluid–structure interaction problems: an introduction*, in Fluid-Structure Interaction and Biomedical Applications, Springer, 2014, pp. 395–481.

[3] E. Burman and P. Hansbo, *A unified stabilized method for Stokes' and Darcy's equations*, Journal of Computational and Applied Mathematics, 198 (2007), pp. 35–51, doi:http://dx.doi.org/10.1016/j.cam.2005.11.022.

[4] J. R. Cebral, M. A. Castro, S. Appanaboyina, C. M. Putman, D. Millan, and A. F. Frangi, *Efficient pipeline for image-based patient-specific analysis of cerebral aneurysm hemodynamics: technique and sensitivity*, IEEE transactions on medical imaging, 24 (2005), pp. 457–467.

[5] J. R. Cebral, F. Mut, J. Weir, and C. Putman, *Quantitative characterization of the hemodynamic environment in ruptured and unruptured brain aneurysms*, American Journal of Neuroradiology, 32 (2011), pp. 145–151, doi:10.3174/ajnr.A2419.

[6] M. D'Elia, L. Mirabella, T. Passerini, M. Perego, M. Piccinelli, C. Vergara, and A. Veneziani, *Applications of variational data assimilation in computational hemodynamics*, Springer Milan, 2012, pp. 363–394, doi:10.1007/978-88-470-1935-5_12.

[7] M. D'Elia, M. Perego, and A. Veneziani, *A variational data assimilation procedure for the incompressible Navier-Stokes equations in hemodynamics*, Journal of Scientific Computing, 52 (2012), pp. 340–359, doi:10.1007/s10915-011-9547-6.

[8] J. M. Dolan, J. Kolega, and H. Meng, *High wall shear stress and spatial gradients in vascular pathology: a review*, Annals of biomedical engineering, 41 (2013), pp. 1411–1427, doi:10.1007/s10439-012-0695-0.

[9] R. P. Dwight, *Bayesian inference for data assimilation using least-squares finite element methods*, in IOP Conference Series: Materials Science and Engineering, vol. 10, IOP Publishing, 2010, p. 012224.

[10] Ø. Evju, K. Valen-Sendstad, and K.-A. Mardal, *A study of wall shear stress in 12 aneurysms with respect to different viscosity models and flow conditions*, Journal of biomechanics, 46 (2013), pp. 2802–2808, doi:10.1016/j.jbiomech.2013.09.004.

[11] P. E. Farrell, D. A. Ham, S. W. Funke, and M. E. Rognes, *Automated derivation of the adjoint of high-level transient finite element programs*, SIAM Journal on Scientific Computing, 35 (2013), pp. C369–C393, doi:10.1137/120873558.

[12] L. Formaggia, A. Veneziani, and C. Vergara, *A new approach to numerical solution of defective boundary value problems in incompressible fluid dynamics*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2769–2794.

[13] K. Funamoto, Y. Suzuki, T. Hayase, T. Kosugi, and H. Isoda, *Numerical validation of MR-measurement-integrated simulation of blood flow in a cerebral aneurysm*, Annals of biomedical engineering, 37 (2009), pp. 1105–1116.

[14] A. V. Fursikov, M. D. Gunzburger, and L. S. Hou, *Boundary value problems and optimal boundary control for the Navier–Stokes system: the two-dimensional case*, SIAM Journal on Control and Optimization, 36 (1998), pp. 852–894, doi:10.1137/S0363012994273374.

[15] A. Gambaruto, D. Doorly, and T. Yamaguchi, *Wall shear stress and near-*

wall convective transport: Comparisons with vascular remodelling in a peripheral graft anastomosis, Journal of Computational Physics, 229 (2010), pp. 5339 – 5356, doi:http://dx.doi.org/10.1016/j.jcp.2010.03.029.

[16] A. Gambaruto, J. Janela, A. Moura, and A. Sequeira, *Shear-thinning effects of hemodynamics in patient-specific cerebral aneurysms*, Mathematical biosciences and engineering, 10 (2013), pp. 649–665.

[17] A. M. Gambaruto, J. Peiró, D. J. Doorly, and A. G. Radaelli, *Reconstruction of shape and its effect on flow in arterial conduits*, International Journal for Numerical Methods in Fluids, 57 (2008), pp. 495–517, doi:10.1002/fld.1642.

[18] T. Guerra, A. Sequeira, and J. Tiago, *Existence of optimal boundary control for the Navier-Stokes equations with mixed boundary conditions*, Portugaliae Mathematica, 72 (2015), pp. 267–283, doi:10.4171/pm/1968.

[19] T. Guerra, J. Tiago, and A. Sequeira, *Optimal control in blood flow simulations*, International Journal of Non-Linear Mechanics, 64 (2014), pp. 57–69, doi:10.1016/j.ijnonlinmec.2014.04.005.

[20] M. Gunzburger, *Perspectives in Flow Control and Optimization*, Society for Industrial and Applied Mathematics, 2002, doi:10.1137/1.9780898718720.

[21] R. Herzog and K. Kunisch, *Algorithms for pde-constrained optimization*, GAMM-Mitteilungen, 33 (2010), pp. 163–176, doi:10.1002/gamm.201010013.

[22] J. J. Heys, T. A. Manteuffel, S. F. McCormick, M. Milano, J. Westerdale, and M. Belohlavek, *Weighted least-squares finite elements based on particle imaging velocimetry data*, Journal of Computational Physics, 229 (2010), pp. 107–118.

[23] K. Jain, J. Jiang, C. Strother, and K.-A. Mardal, *Transitional hemodynamics in intracranial aneurysms–comparative velocity investigations with high resolution lattice Boltzmann simulations, normal resolution ANSYS simulations, and MR imaging*, Medical Physics, 43 (2016), pp. 6186–6198.

[24] J. Jiang, K. Johnson, K. Valen-Sendstad, K.-A. Mardal, O. Wieben, and C. Strother, *Flow characteristics in a canine aneurysm model: a comparison of 4D accelerated phase-contrast MR measurements and computational fluid dynamics simulations*, Medical physics, 38 (2011), pp. 6300–6312, doi:10.1118/1.3652917.

[25] L. J. John, *Optimal Boundary Control in Energy Spaces*, vol. 24, Verlag der Technischen Universität Graz, 2014, doi:10.3217/978-3-85125-373-3.

[26] B. M. Johnston, P. R. Johnston, S. Corney, and D. Kilpatrick, *Non-newtonian blood flow in human right coronary arteries: transient simulations*, Journal of biomechanics, 39 (2006), pp. 1116–1128.

[27] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM Journal on Control and Optimization, 25 (1987), pp. 1503–1516, doi:10.1137/0325083.

[28] Z. Kulcsár, A. Ugron, M. Marosfői, Z. Berentei, G. Paál, and I. Szikora, *Hemodynamics of cerebral aneurysm initiation: the role of wall shear stress and spatial wall shear stress gradient*, American Journal of Neuroradiology, 32 (2011), pp. 587–594, doi:10.3174/ajnr.A2339.

[29] T. Lassila, A. Manzoni, A. Quarteroni, and G. Rozza, *A reduced computational and geometrical framework for inverse problems in hemodynamics*, International journal for numerical methods in biomedical engineering, 29 (2013), pp. 741–776.

[30] H. Lee, *Optimal control for quasi-Newtonian flows with defective boundary conditions*, Computer Methods in Applied Mechanics and Engineering, 200 (2011), pp. 2498–2506, doi:10.1016/j.cma.2011.04.019.

[31] A. Logg, K. A. Mardal, and G. N. Wells, *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012, doi:doi:10.1007/978-3-642-23099-8.

[32] A. Manzoni, T. Lassila, A. Quarteroni, and G. Rozza, *A reduced-order strategy for solving inverse Bayesian shape identification problems in physiological flows*, in Modeling, Simulation and Optimization of Complex Processes-HPSC 2012, Springer, 2014, pp. 145–155.

[33] J. G. Myers, J. A. Moore, M. Ojha, K. W. Johnston, and C. R. Ethier, *Factors influencing blood flow patterns in the human right coronary artery*, Annals of Biomedical Engineering, 29 (2001), pp. 109–120, doi:10.1114/1.1349703.

[34] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, in Abhandlungen aus dem mathematischen Seminar der Universität Hamburg, vol. 36, Springer, 1971, pp. 9–15, doi:10.1007/BF02995904.

[35] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Science & Business Media,

2006, doi:10.1007/978-0-387-40065-5.

[36] M. Perego, A. Veneziani, and C. Vergara, *A variational approach for estimating the compliance of the cardiovascular tissue: An inverse fluid-structure interaction problem*, SIAM Journal on Scientific Computing, 33 (2011), pp. 1181–1211.

[37] A. Quarteroni and G. Rozza, *Optimal control and shape optimization of aorto-coronaric bypass anastomoses*, Mathematical Models and Methods in Applied Sciences, 13 (2003), pp. 1801–1823.

[38] S. Ramalho, A. Moura, A. M. Gambaruto, and A. Sequeira, *Sensitivity to outflow boundary conditions and level of geometry description for a cerebral aneurysm*, International Journal for Numerical Methods in Biomedical Engineering, 28 (2012), pp. 697–713, doi:10.1002/cnm.2461.

[39] S. Ramalho, A. Moura, A. M. Gambaruto, and A. Sequeira, *Influence of blood rheology and outflow boundary conditions in numerical simulation of cerebral aneursyms*, Springer, 2013, pp. 149–175, doi:10.1007/978-1-4614-4178-6_6.

[40] H. Samady, P. Eshtehardi, M. C. McDaniel, J. Suo, S. S. Dhawan, C. Maynard, L. H. Timmins, A. A. Quyyumi, and D. P. Giddens, *Coronary artery wall shear stress is associated with progression and transformation of atherosclerotic plaque and arterial remodeling in patients with coronary artery disease*, Circulation, 124 (2011), pp. 779–788, doi:10.1161/CIRCULATIONAHA.111.021824.

[41] T. Schwedes, S. W. Funke, and D. A. Ham, *An iteration count estimate for a mesh-dependent steepest descent method based on finite elements and Riesz inner product representation*, (2016), arXiv:1606.08069.

[42] D. A. Steinman, *Assumptions in modelling of large artery hemodynamics*, in Modeling of Physiological Flows, Springer, 2012, pp. 1–18, doi:10.1007/978-88-470-1935-5_1.

[43] H. Takao, Y. Murayama, S. Otsuka, Y. Qian, A. Mohamed, S. Masuda, M. Yamamoto, and T. Abe, *Hemodynamic differences between unruptured and ruptured intracranial aneurysms during observation*, Stroke, 43 (2012), pp. 1436–1439, doi:10.1161/STROKEAHA.111.640995.

[44] J. Tiago, A. Gambaruto, and A. Sequeira, *Patient-specific blood flow simulations: Setting Dirichlet boundary conditions for minimal error with respect to measured data*, Mathematical Modelling of Natural Phenomena, 9 (2014), pp. 98–116, doi:10.1051/mmnp/20149608.

[45] P. Triverro, L. Dede, A. Sequeira, S. Deparis, A. Robertso, and A. Quarteroni, *Fluid-structure interaction simulations of cerebral arteries by isotropic and anisotropic consitutive laws*, Computational Mechanics, (2015), doi:10.1007/s00466-014-1117-y.

[46] A. Veneziani and C. Vergara, *An approximate method for solving incompressible Navier–Stokes problems with flow rate conditions*, Computer methods in applied mechanics and engineering, 196 (2007), pp. 1685–1700.

[47] J. Xiang, S. K. Natarajan, M. Tremmel, D. Ma, J. Mocco, L. N. Hopkins, A. H. Siddiqui, E. I. Levy, and H. Meng, *Hemodynamic–morphologic discriminants for intracranial aneurysm rupture.*, Stroke, 42 (2011), pp. 144–52, doi:10.1161/STROKEAHA.110.592923.

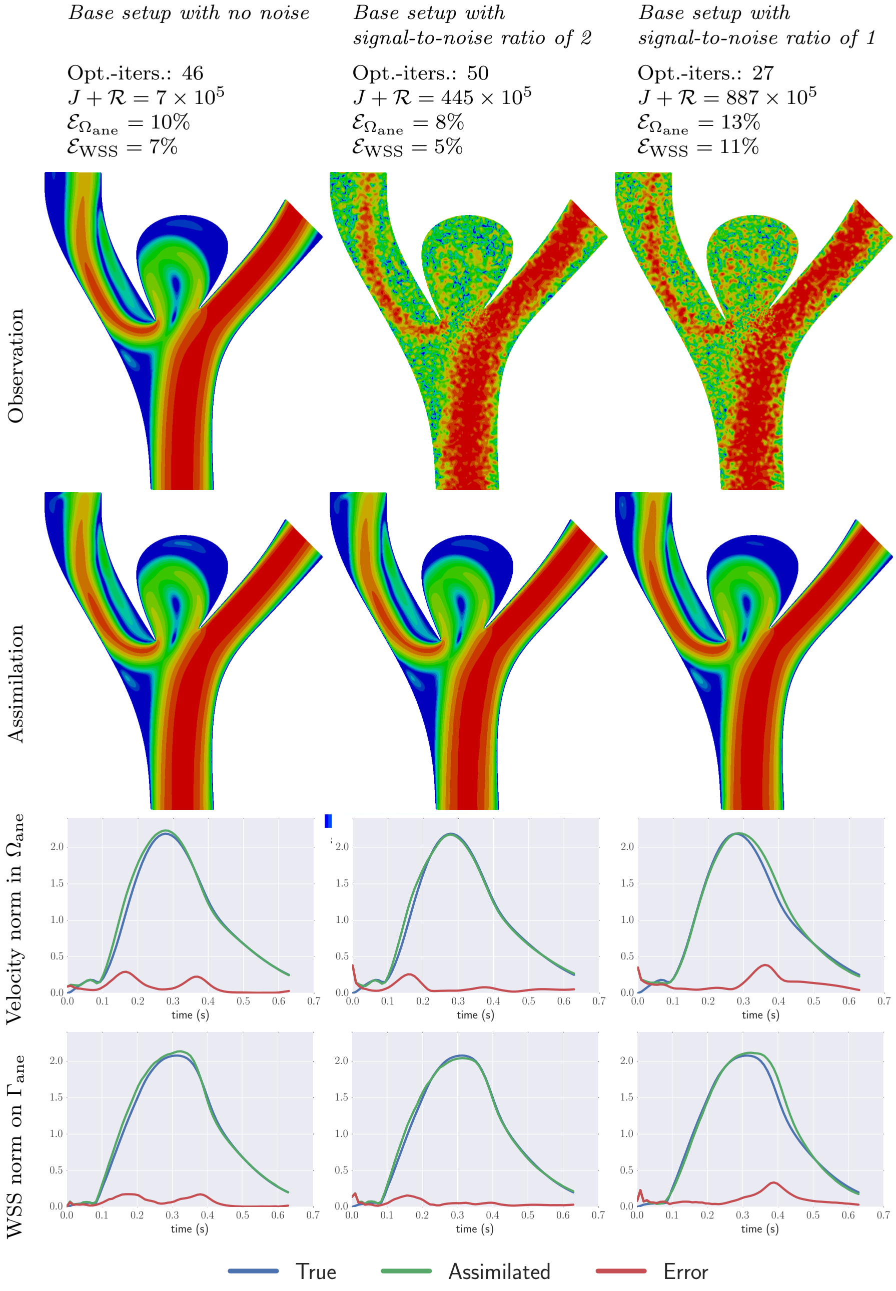


Fig. 4. *Results using the* ***instantaneous observation operator*** *with pointwise additive Gaussian white noise. The signal-to-noise-ratio was computed as* $\|\mathcal{T}^{inst}u_{true}\|^2/\|\mathcal{T}^{inst}u_{true} - d\|^2$, *where* d *is the noisy data. The snapshots on the top three rows are taken at* $t = 0.296s$.

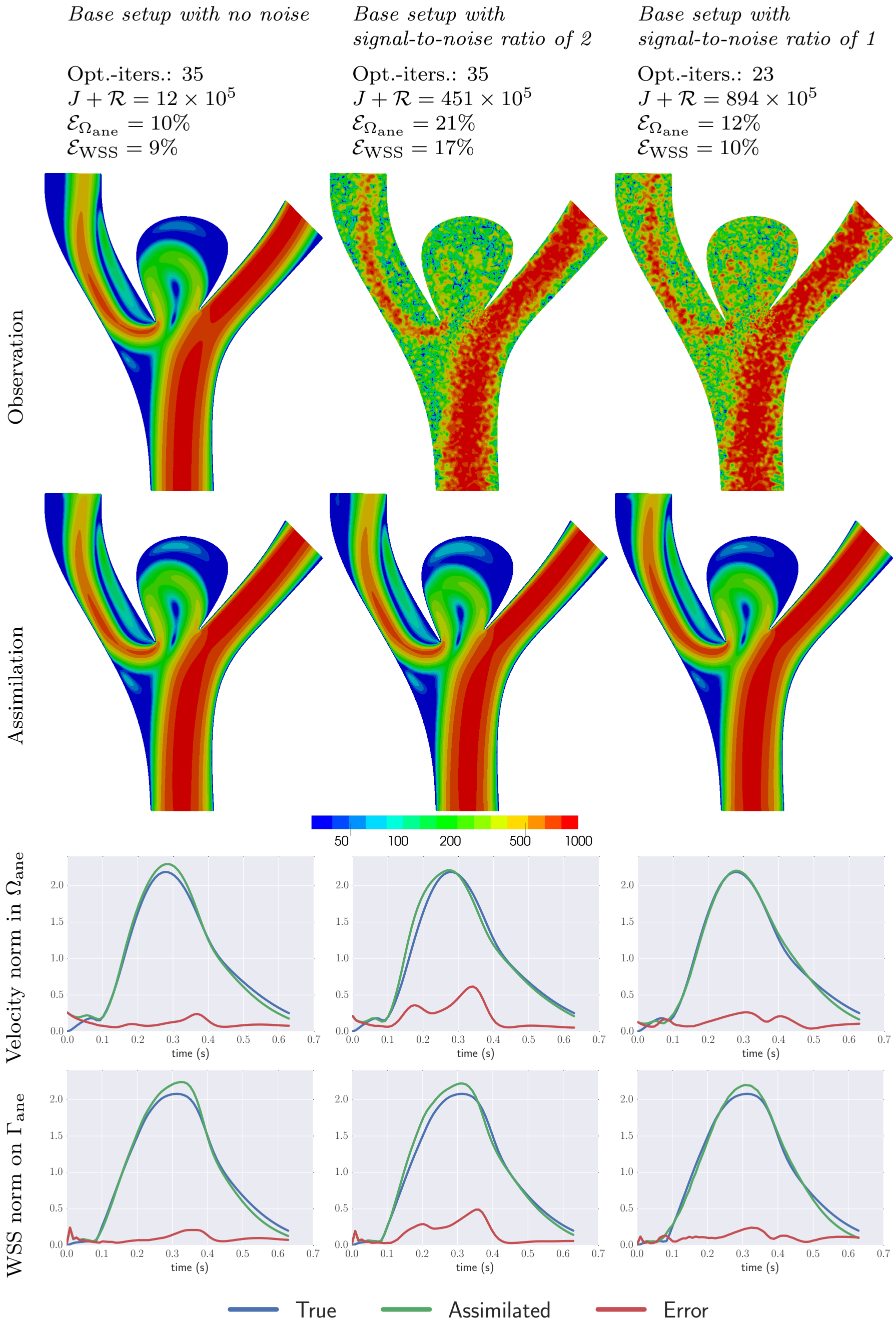


FIG. 5. *Results using the* ***time-averaging observation operator*** *with pointwise additive Gaussian white noise. The signal-to-noise ratio was computed as* $\|\mathcal{T}^{avg} u_{true}\|^2 / \|\mathcal{T}^{avg} u_{true} - d\|^2$*, where* d *is the noisy data. The snapshots on the top three rows are taken at* $t = 0.296s$*.*

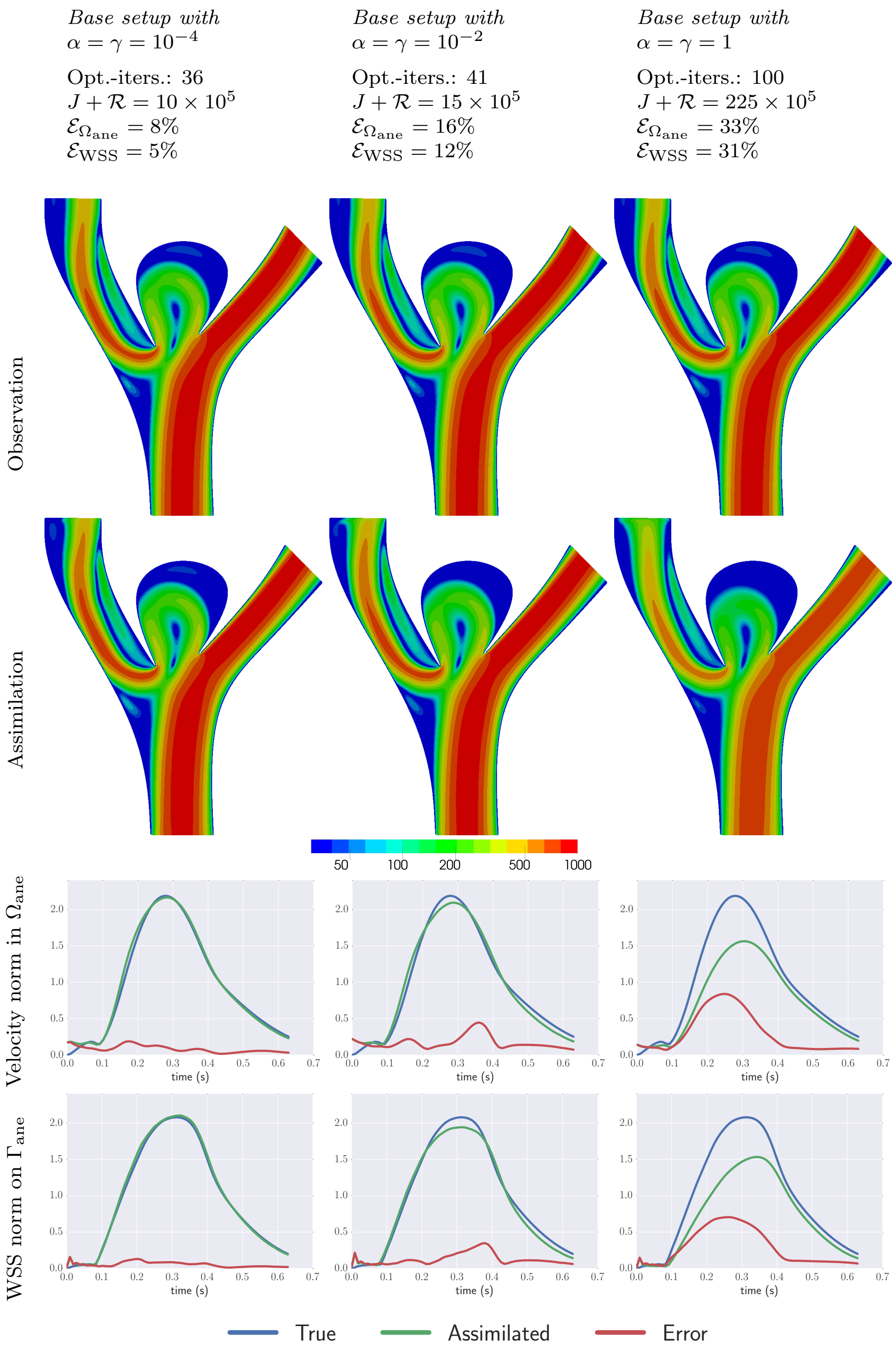
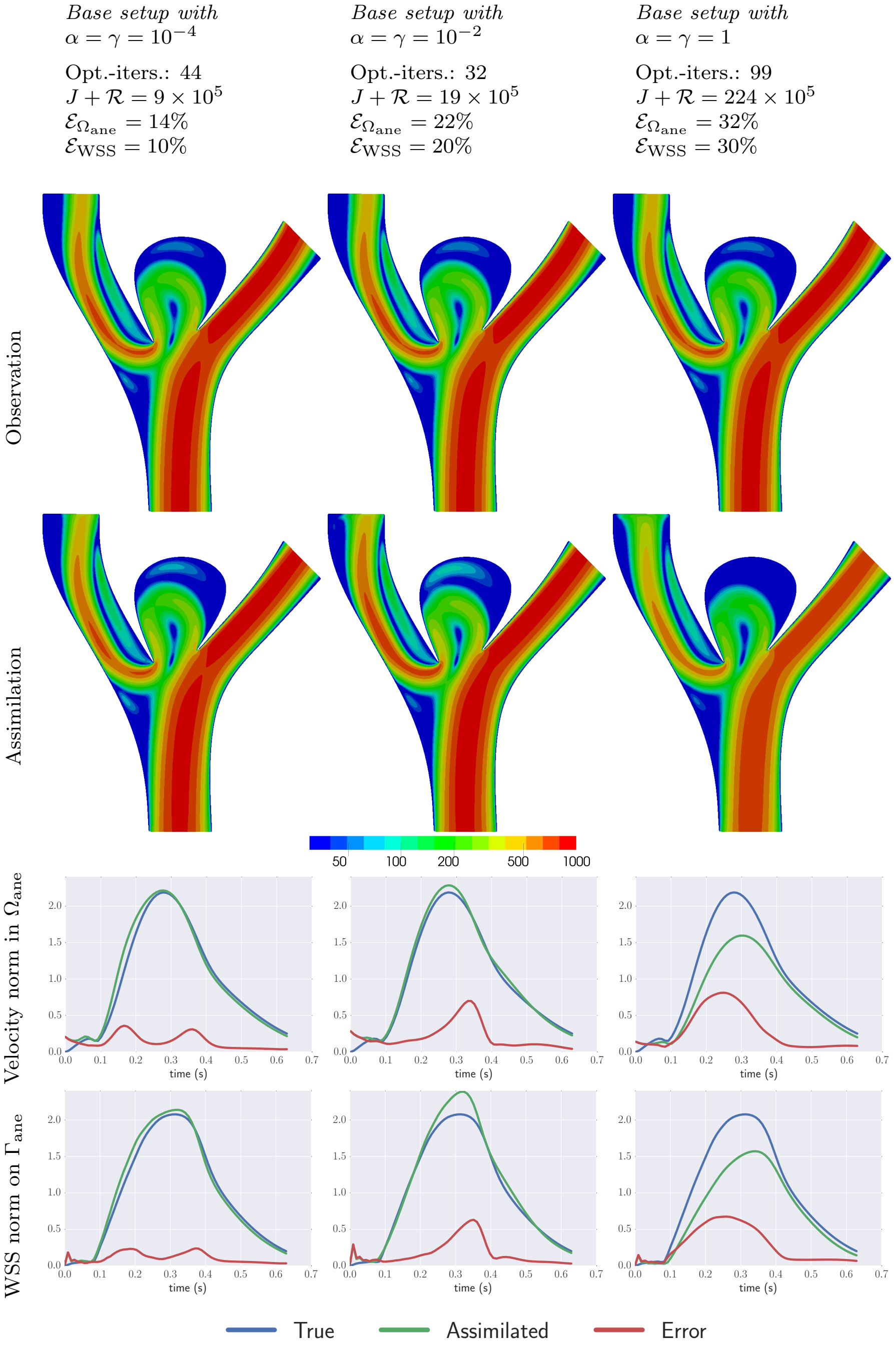


Fig. 6. *Results using the* ***instantaneous observation operator*** *with varying* α *and* γ *regularisation coefficients. The base setup (figure 4, left column) uses* $\alpha = \gamma = 10^{-5}$*. The snapshots on the top two rows are taken at* $t = 0.296s$*.*



Base setup with $\alpha = \gamma = 10^{-4}$

Opt.-iters.: 44
$J + \mathcal{R} = 9 \times 10^5$
$\mathcal{E}_{\Omega_{\text{ane}}} = 14\%$
$\mathcal{E}_{\text{WSS}} = 10\%$

Base setup with $\alpha = \gamma = 10^{-2}$

Opt.-iters.: 32
$J + \mathcal{R} = 19 \times 10^5$
$\mathcal{E}_{\Omega_{\text{ane}}} = 22\%$
$\mathcal{E}_{\text{WSS}} = 20\%$

Base setup with $\alpha = \gamma = 1$

Opt.-iters.: 99
$J + \mathcal{R} = 224 \times 10^5$
$\mathcal{E}_{\Omega_{\text{ane}}} = 32\%$
$\mathcal{E}_{\text{WSS}} = 30\%$

Fig. 7. *Results using the* ***time-averaging observation operator*** *with varying* α *regularisation parameters. The base setup (figure 5, left column) uses* $\alpha = \gamma = 10^{-5}$*. The snapshots on the top two rows are taken at* $t = 0.296s$*.*

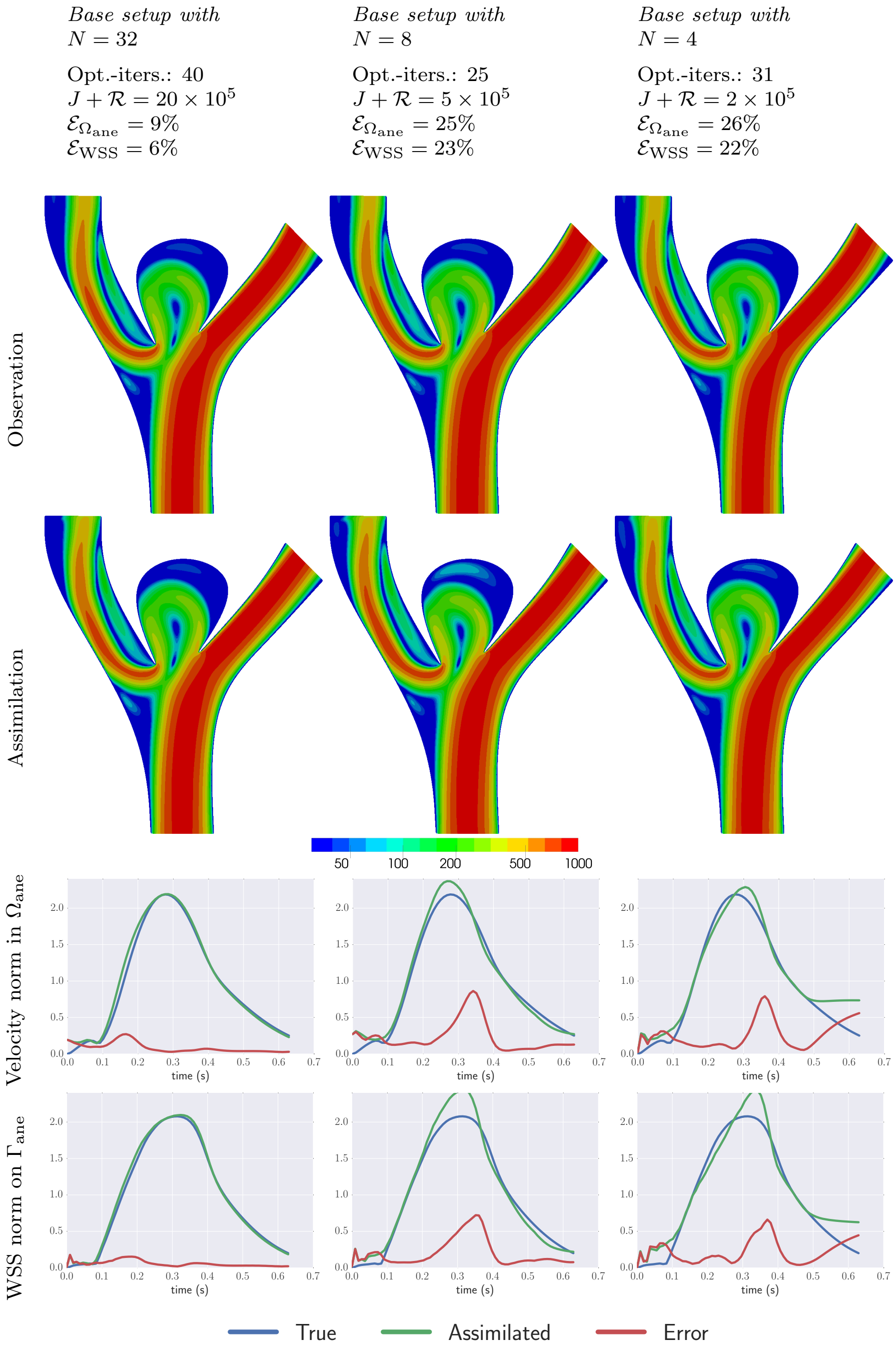


Fig. 8. *Results using the **instantaneous observation operator** with varying number of observations. The base setup (figure 4, left column) uses $N = 16$ observations. The snapshots on the top two rows are taken at $t = 0.296s$.*

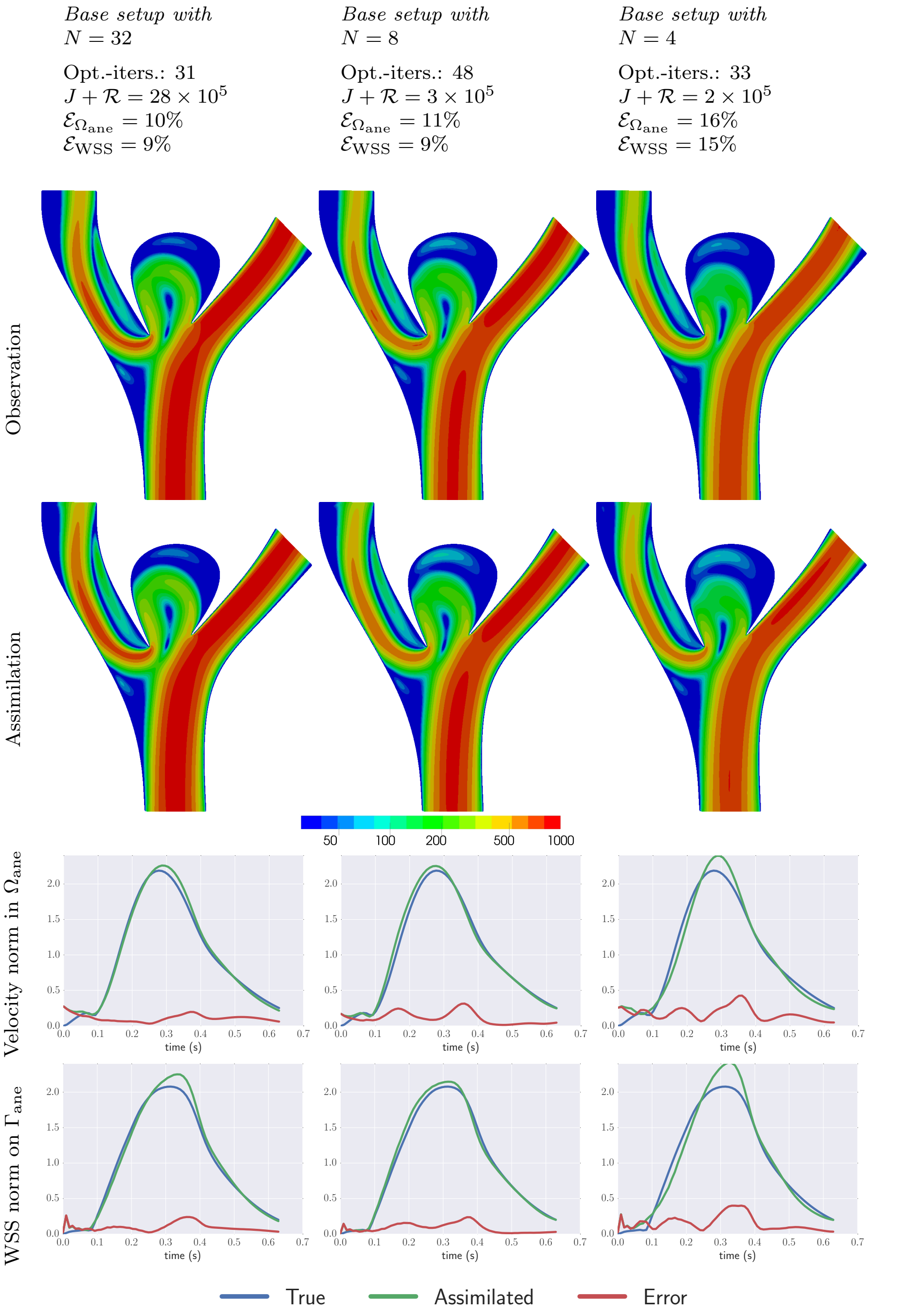
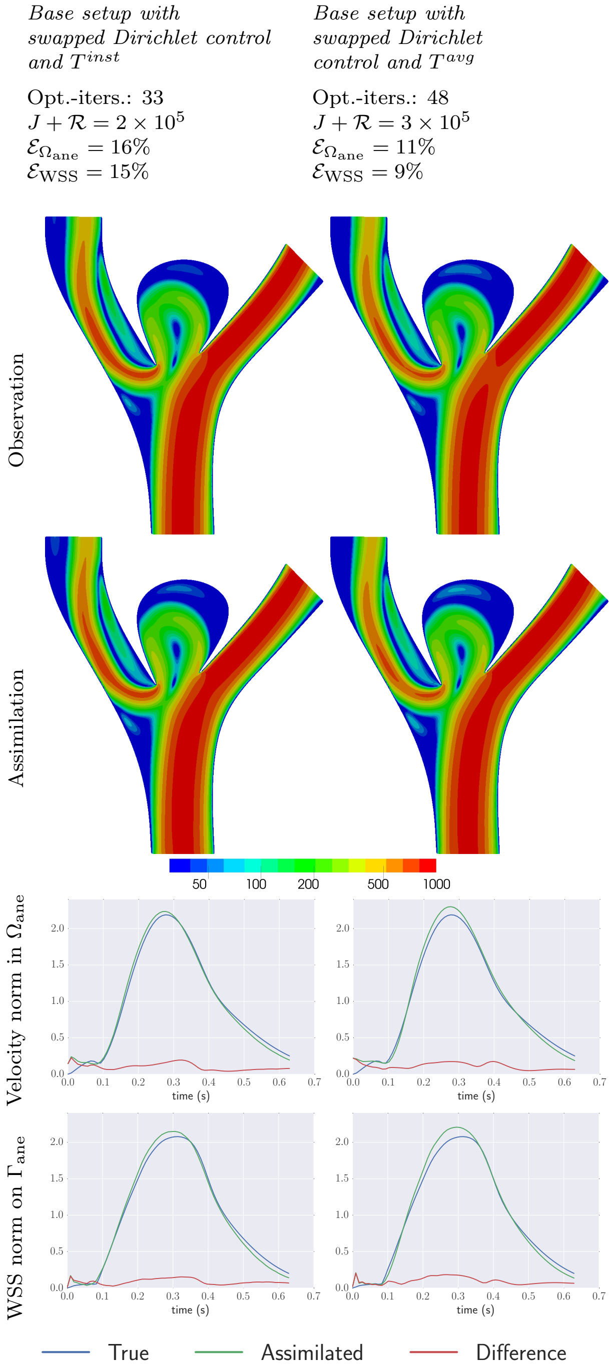


FIG. 9. *Results using the* ***time-averaging observation operator*** *with varying number of observations. The base setup (figure 5, left column) uses* $N = 16$ *observations. The snapshots on the top two rows are taken at* $t = 0.296s$.



Base setup with swapped Dirichlet control and T^{inst}

Opt.-iters.: 33
$J + \mathcal{R} = 2 \times 10^5$
$\mathcal{E}_{\Omega_{\text{ane}}} = 16\%$
$\mathcal{E}_{\text{WSS}} = 15\%$

Base setup with swapped Dirichlet control and T^{avg}

Opt.-iters.: 48
$J + \mathcal{R} = 3 \times 10^5$
$\mathcal{E}_{\Omega_{\text{ane}}} = 11\%$
$\mathcal{E}_{\text{WSS}} = 9\%$

Fig. 10. *Assimilation results where the controlled outlets are swapped from the base cases (figures 4 and 5, left columns). The snapshots on the top two rows are taken at $t = 0.296s$.*

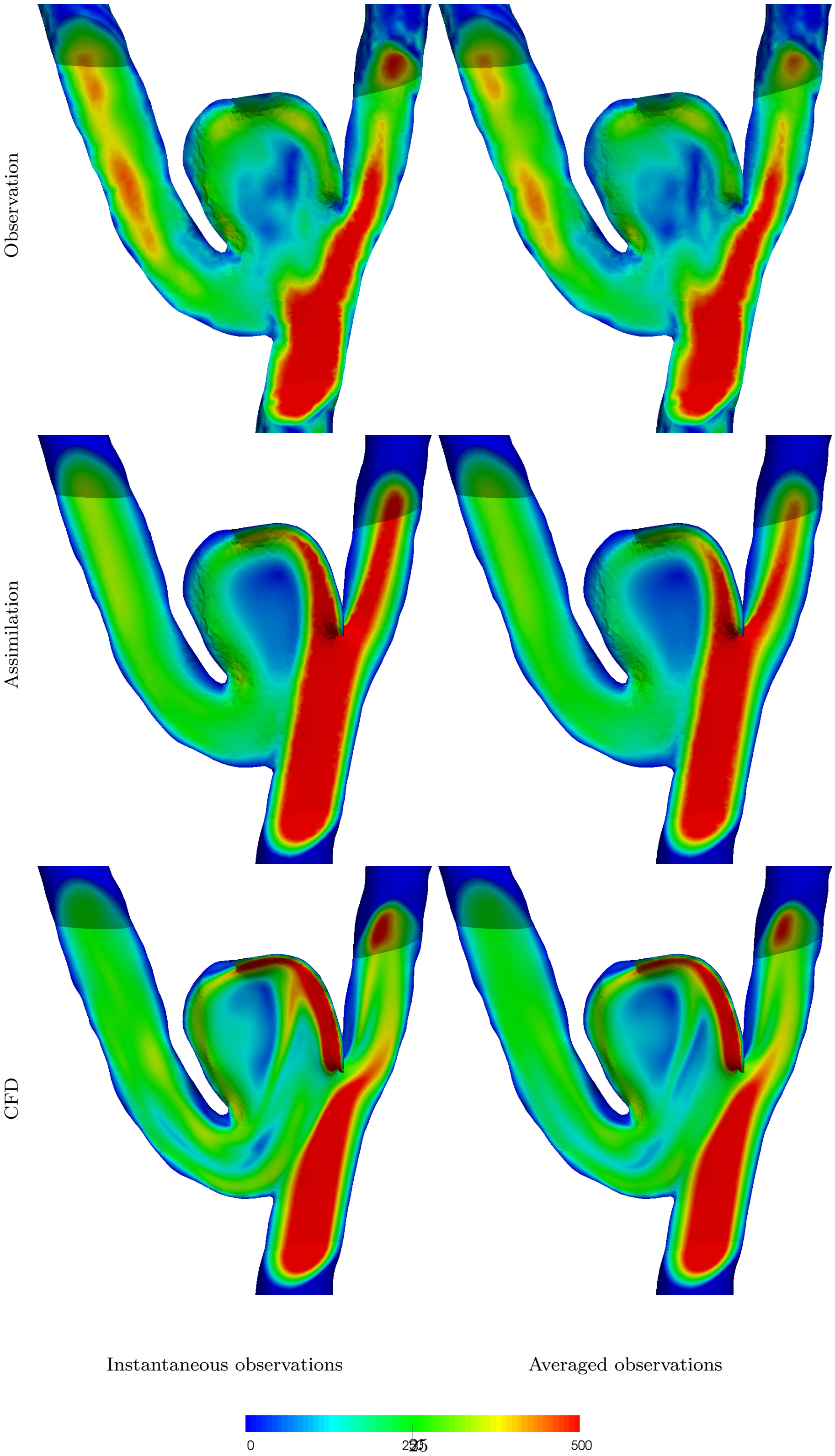


Fig. 11. *Flow speed visualised through a slice of the 3D dog vessel. The snapshots are taken at time $t = 0.296s$.*