

Optimization-based additive decomposition of weakly coercive problems with applications,^{☆☆}

Pavel Bochev, Denis Ridzal

Center for Computing Research, Sandia National Laboratories^{}, P.O. Box 5800, MS 1320, Albuquerque, NM 87185-1320*

Abstract

We present an abstract mathematical framework for an optimization-based additive decomposition of a large class of variational problems into a collection of concurrent subproblems. The framework replaces a given monolithic problem by an *equivalent* constrained optimization formulation in which the subproblems define the optimization constraints and the objective is to minimize the mismatch between their solutions. The significance of this reformulation stems from the fact that one can solve the resulting optimality system by an iterative process involving only solutions of the subproblems. Consequently, assuming that stable numerical methods and efficient solvers are available for every subproblem, our reformulation leads to robust and efficient numerical algorithms for a given monolithic problem by breaking it into subproblems that can be handled more easily. An application of the framework to the Oseen equations illustrates its potential.

Keywords: Optimization, additive decomposition, weakly coercive problems, Oseen's equations, finite elements.

1. Introduction

We present an abstract framework for the additive decomposition of a large class of variational problems into a collection of concurrent subproblems. The framework comprises three distinct stages. Given a weakly coercive variational equation, the first stage decomposes the corresponding bilinear form into a finite sum of weakly coercive bilinear subforms. At the second stage we use these

^{*}Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

^{☆☆}This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research.

Email addresses: pboche@sandia.gov (Pavel Bochev), dridzal@sandia.gov (Denis Ridzal)

subforms to define a collection of independent variational subproblems with undetermined right hand sides. The final, third stage reconnects the subproblems into a constrained optimization problem. The optimization objective is to minimize the mismatch between the solutions of the subproblems, the undetermined right hand sides provide the (distributed) controls, and the subproblems define the optimization constraints.

This paper continues our previous efforts [1, 2] on optimization-based additive operator decomposition. Its primary purpose is to enable an efficient numerical solution of a given monolithic problem by breaking it into smaller pieces that can be handled more easily. Specifically, optimization-based reformulation allows us to synthesize stable discretizations and efficient solvers for the monolithic problem from stable discretizations and efficient solvers for its subproblems. Such a capability can be useful in multiple contexts.

For instance, for concurrent multiphysics problems, i.e., problems where multiple physics operators act simultaneously on the same physical domain, the subproblems in our framework can correspond to the constituent physics components of the multiphysics problem, for which stable discretizations and efficient solvers can be obtained more easily than for the monolithic problem. Another possible scenario arises when a given problem can be written as a sum of subproblems with better computational properties. For example, one can write an advection-dominated operator as a sum of two diffusion-dominated suboperators [1]. In either case, the framework allows us to synthesize discretizations and solvers for the parent problem from discretizations and solvers for its subproblems.

This work improves upon the results in [1, 2] in several important ways. Specifically, we extend the abstract additive splitting framework [1] to weakly coercive problems, which effectively enables its application to virtually any Partial Differential Equation (PDE) problem. We also prove that the abstract optimization problem and its approximation are well-posed *without penalization of the objective by the control*. This significantly strengthens the result in [2] where well-posedness without penalization was shown only for the algebraic versions of the optimization problem. Most notably, the absence of a control penalty in the new formulation brings about important algorithmic advantages upon [1] and other operator-splitting strategies. First, it implies that the resulting optimization problem is an equivalent reformulation of the original equations. Second, it allows us to significantly simplify the associated optimality system and exploit it for the design of efficient iterative solvers. Third, the simplified optimality system prompts a simple yet efficient preconditioner that shows minimal mesh dependence.

In the recent years there has been a steady interest in exploiting optimization and control ideas for the efficient solution of PDEs. Two important examples are the optimization-based domain decomposition [3, 4, 5, 6] and heterogeneous domain decomposition [7, 8, 9, 10, 11, 12], which focus on merging operators acting in different parts of the computational domain. In the former case these operators are restrictions of the governing equations to overlapping or non-overlapping subdomains of the original domain and the primary purpose of the

decomposition is to enable a more efficient solution of the equations. In the latter case, optimization is used to merge two fundamentally different material descriptions separated by a physical interface, such as atomistic and continuum models [10], fluid and structure [12], or surface and subsurface flows [8].

The emphasis of this paper on the coupling of concurrent operators that act on the same physical domain is a key distinction between our work and the efforts cited above. From this point of view the decomposition of operators and energy spaces by replicas and virtual controls [13, 14, 15, 16] is the closest to our approach. However, there are some key differences in the manner in which we define the suboperators and effect their coupling. For instance, the virtual controls in [13, 14] coincide with the original PDE solution and the constraints are “replicas” of the original PDE. In contrast, the virtual controls in our approach *do not coincide* with the PDE solution and the constraints are not equivalent to the original PDE.

The paper is organized as follows. Section 2 reviews notation and the abstract setting for the framework. The core of the paper is Section 3, which introduces the optimization-based additive decomposition and analyzes the resulting optimization problem. This section also establishes the equivalence between the latter and the original equations. Section 4 discusses the discretization of the optimization problem, including well-posedness of the approximate optimization problem, while Section 5 focuses on the design of efficient solvers for the discrete optimality system. Finally, in Section 6 we apply the framework to synthesize an efficient iterative solver for the Oseen equations.

2. Abstract setting

We consider optimization-based additive operator decomposition of weakly coercive variational problems. The abstract setting for such problems involves two Hilbert spaces: a trial (solution) space U , and a test function space V . Let V^* denote the dual of V and $\langle \cdot, \cdot \rangle_{V^*, V}$ is the duality pairing between V^* and V . We assume that there exists a third, pivot Hilbert space H such that $\{V, H, V^*\}$ is a Gelfand triple [17, Definition 17.1, p.262]. This means that

$$\langle f, v \rangle_{V^*, V} = (f, v)_H \quad \forall f \in V^*; \quad \forall v \in V, \quad (1)$$

and $V \subset H \subset V^*$ with continuous embeddings.

The abstract variational problem is defined by a bilinear form $B(\cdot, \cdot) : U \times V \mapsto \mathbb{R}$ and a dual element $f \in V^*$. The objective is to find $u \in U$ such that

$$B(u, v) = (f, v)_H \quad \forall v \in V. \quad (2)$$

We assume that $B(\cdot, \cdot)$ is weakly coercive and continuous, i.e., there exist positive constants γ_{01} and γ_{02} such that

$$\begin{cases} \sup_{v \in V} \frac{B(u, v)}{\|v\|_V} \geq \gamma_{01} \|u\|_U & \text{and} & \sup_{u \in U} \frac{B(u, v)}{\|u\|_U} > 0 \\ B(u, v) \leq \gamma_{02} \|u\|_U \|v\|_V & \forall u \in U, \quad \forall v \in V. \end{cases} \quad (3)$$

Néčas theorem [18, 19] asserts that conditions (3) are sufficient for (2) to have a unique solution $u \in U$, which depends continuously on the data:

$$\|u\|_U \leq \frac{1}{\gamma_{01}} \|f\|_{V^*}. \quad (4)$$

3. Optimization-based additive operator splitting

We assume that $B(\cdot, \cdot)$ can be written as a finite sum of weakly coercive subproblems that are “easier” to solve¹ than the original equation (2). To explain the key ideas of the optimization-based additive operator splitting it suffices to consider the decomposition of $B(\cdot, \cdot)$ into a sum of two subproblems. Thus, we assume that there exist bilinear forms $B_i(\cdot, \cdot) : U \times V \mapsto \mathbb{R}$; and positive constants γ_{i1} and γ_{i2} , $i = 1, 2$, such that

$$B(\cdot, \cdot) = B_1(\cdot, \cdot) + B_2(\cdot, \cdot), \quad (5)$$

and

$$\left\{ \begin{array}{l} \sup_{v \in V} \frac{B_i(u, v)}{\|v\|_V} \geq \gamma_{i1} \|u\|_U \quad \text{and} \quad \sup_{u \in U} \frac{B_i(u, v)}{\|u\|_U} > 0; \\ B_i(u, v) \leq \gamma_{i2} \|u\|_U \|v\|_V \quad \forall u \in U, \quad \forall v \in V. \end{array} \right. \quad (6)$$

We propose to reformulate (2) as the constrained optimization problem

$$\left\{ \begin{array}{l} \min_{u_1, u_2 \in U; \theta \in V} \quad J(u_1, u_2) = \frac{1}{2} \|u_1 - u_2\|_H^2 \\ \text{subject to} \left\{ \begin{array}{ll} B_1(u_1, v_1) - (\theta, v_1)_V = (f, v_1)_H & \forall v_1 \in V \\ B_2(u_2, v_2) + (\theta, v_2)_V = 0 & \forall v_2 \in V, \end{array} \right. \end{array} \right. \quad (7)$$

where $u_1, u_2 \in U$ are state variables and $\theta \in V$ is a virtual distributed control.

Our strategy is to exploit the structure of (7) for the development of efficient solution algorithms for the original equation (2). For this strategy to work, the optimization problem (7) must be well posed and its optimal states u_1, u_2 must provide an accurate approximation to the solution of (2). In the next section we prove that both of these prerequisites hold for (7). In fact, we show that (7) is an equivalent reformulation of (2), i.e., for any solution u of the latter the pair $u_1 = u_2 = u$ is an optimal state of (7).

Remark 1. *Typically, PDE-constrained optimization problems fail to be well-posed without a control penalty term $\varepsilon \|\theta\|_V^2$ in the cost functional. This is not the case for (7), which turns out to be well-posed without a such a term. This result implies the equivalence of (2) and (7), and improves upon our previous work [1], where the analysis required a control penalty.*

¹The precise meaning of this statement depends on the context and could range from the availability of iterative solvers optimized for the component subproblems to discretizations that are better suited for the physics represented by these subproblems.

3.1. Lagrange multiplier solution

By using Lagrange multipliers $\{\lambda_1, \lambda_2\} \in V \times V$ to enforce the constraints one can replace the constrained minimization of (7) by the unconstrained optimization problem of finding the stationary points of the Lagrangian functional

$$L(\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\}) = J(u_1, u_2) + \left(B_1(u_1, \lambda_1) - (\theta, \lambda_1)_V - (f, \lambda_1)_H \right) + \left(B_2(u_2, \lambda_2) + (\theta, \lambda_2)_V \right). \quad (8)$$

Taking the first variations of L with respect to the states, controls and the Lagrange multipliers, one easily finds that the necessary optimality condition for the saddle point $\{\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\}\}$ of (8) is given by the following variational problem: *seek* $\{\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\}\} \in \{U \times U \times V\} \times \{V \times V\}$ such that

$$\begin{aligned} (u_1 - u_2, \hat{u}_1 - \hat{u}_2)_U + B_1(\hat{u}_1, \lambda_1) + B_2(\hat{u}_2, \lambda_2) &= 0 & \forall \hat{u}_1, \hat{u}_2 \in U \\ (\hat{\theta}, \lambda_2 - \lambda_1)_V &= 0 & \forall \hat{\theta} \in V \\ (\theta, \hat{\lambda}_2 - \hat{\lambda}_1)_V + B_1(u_1, \hat{\lambda}_1) + B_2(u_2, \hat{\lambda}_2) &= (f, \hat{\lambda}_1)_H & \forall \hat{\lambda}_1, \hat{\lambda}_2 \in V. \end{aligned} \quad (9)$$

Let $X = U \times U \times V$, $Y = V \times V$. Define a bilinear form $\mathcal{A} : X \times X \mapsto \mathbb{R}$

$$\mathcal{A}(\{u_1, u_2, \theta\}, \{v_1, v_2, \psi\}) = (u_1 - u_2, v_1 - v_2)_U,$$

and a bilinear form $\mathcal{B} : X \times Y \mapsto \mathbb{R}$,

$$\mathcal{B}(\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\}) = (\theta, \lambda_2 - \lambda_1)_V + B_1(u_1, \lambda_1) + B_2(u_2, \lambda_2).$$

In terms of \mathcal{A} and \mathcal{B} the optimality system (9) assumes a canonical mixed structure of [20]. To show that (9) is well-posed we check the conditions of the abstract saddle-point theory in [20]. Obviously, \mathcal{A} and \mathcal{B} are continuous and so, we focus on the verification of the coercivity on the nullspace condition for \mathcal{A} and the inf-sup condition for \mathcal{B} . The nullspace of \mathcal{B} is

$$Z = \{\{u_1, u_2, \theta\} \in X \mid \mathcal{B}(\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\}) = 0 \quad \forall \{\lambda_1, \lambda_2\} \in Y\}.$$

Our first result shows that control penalty is not necessary for \mathcal{A} to be coercive on this space.

Lemma 1. *There is a positive constant $\gamma_{\mathcal{A}}$ such that*

$$\gamma_{\mathcal{A}} (\|u_1\|_U^2 + \|u_2\|_U^2 + \|\theta\|_V^2) \leq \mathcal{A}(\{u_1, u_2, \theta\}, \{u_1, u_2, \theta\}) \quad (10)$$

for all $\{u_1, u_2, \theta\} \in Z$.

Proof. Any $\{u_1, u_2, \theta\} \in Z$ satisfies the identity

$$(\theta, \hat{\lambda}_2 - \hat{\lambda}_1)_V + B_1(u_1, \hat{\lambda}_1) + B_2(u_2, \hat{\lambda}_2) = 0 \quad \forall \hat{\lambda}_1, \hat{\lambda}_2 \in V.$$

Setting $\hat{\lambda}_1 = \hat{\lambda}_2 = \hat{\lambda}$ in this equation yields the identity

$$B_1(u_1, \hat{\lambda}) + B_2(u_2, \hat{\lambda}) = 0 \quad \forall \hat{\lambda} \in V,$$

while adding and subtracting $B_1(u_2, \hat{\lambda})$ and $B_2(u_1, \hat{\lambda})$ gives

$$B_1(u_1 - u_2, \hat{\lambda}) + B_1(u_2, \hat{\lambda}) + B_2(u_2, \hat{\lambda}) = 0 \quad \forall \hat{\lambda} \in V$$

$$B_2(u_2 - u_1, \hat{\lambda}) + B_1(u_1, \hat{\lambda}) + B_2(u_1, \hat{\lambda}) = 0 \quad \forall \hat{\lambda} \in V,$$

After accounting for $B(\cdot, \cdot) = B_1(\cdot, \cdot) + B_2(\cdot, \cdot)$,

$$B(u_2, \hat{\lambda}) = -B_1(u_1 - u_2, \hat{\lambda}) \quad \forall \hat{\lambda} \in V$$

$$B(u_1, \hat{\lambda}) = -B_2(u_2 - u_1, \hat{\lambda}) \quad \forall \hat{\lambda} \in V,$$

Using the inf-sup condition (3) and the continuity of B_i ,

$$\gamma_{01} \|u_2\|_U \leq \sup_{\hat{\lambda} \in V} \frac{B(u_2, \hat{\lambda})}{\|\hat{\lambda}\|_V} = \sup_{\hat{\lambda} \in V} \frac{-B_1(u_1 - u_2, \hat{\lambda})}{\|\hat{\lambda}\|_V} \leq \gamma_{12} \|u_1 - u_2\|_U,$$

$$\gamma_{01} \|u_1\|_U \leq \sup_{\hat{\lambda} \in V} \frac{B(u_1, \hat{\lambda})}{\|\hat{\lambda}\|_V} = \sup_{\hat{\lambda} \in V} \frac{-B_2(u_2 - u_1, \hat{\lambda})}{\|\hat{\lambda}\|_V} \leq \gamma_{22} \|u_1 - u_2\|_U.$$

Therefore, for any $\{u_1, u_2, \theta\} \in Z$ there holds

$$\|u_1 - u_2\|_U \geq \frac{\gamma_{01}}{2\gamma_{22}} \|u_1\|_U + \frac{\gamma_{01}}{2\gamma_{12}} \|u_2\|_U.$$

Furthermore, if $\{u_1, u_2, \theta\} \in Z$, we have that

$$(\theta, \hat{\lambda}_2)_V + B_2(u_2, \hat{\lambda}_2) = 0 \quad \forall \hat{\lambda}_2 \in V,$$

$$-(\theta, \hat{\lambda}_1)_V + B_1(u_1, \hat{\lambda}_1) = 0 \quad \forall \hat{\lambda}_1 \in V.$$

Setting $\hat{\lambda}_1 = \theta$ and $\hat{\lambda}_2 = -\theta$ above yields

$$\|\theta\|_V^2 = B_2(u_2, \theta) \leq \gamma_{22} \|u_2\|_U \|\theta\|_V,$$

$$\|\theta\|_V^2 = B_1(u_1, \theta) \leq \gamma_{12} \|u_1\|_U \|\theta\|_V$$

or, equivalently,

$$\|\theta\|_V \leq \gamma_{12} \|u_1\|_U \quad \text{and} \quad \|\theta\|_V \leq \gamma_{22} \|u_2\|_U.$$

Combining all inequalities yields the bound

$$\|u_1 - u_2\|_U \geq \frac{\gamma_{01}}{2\gamma_{22}} \|u_1\|_U + \frac{\gamma_{01}}{4\gamma_{12}} \|u_2\|_U + \frac{\gamma_{01}}{4\gamma_{22}\gamma_{12}} \|\theta\|_V$$

which implies (10) with

$$\gamma_{\mathcal{A}} = \gamma_{01}^2 \min \left\{ \left(\frac{1}{2\gamma_{22}} \right)^2, \left(\frac{1}{4\gamma_{12}} \right)^2, \left(\frac{1}{4\gamma_{22}\gamma_{12}} \right)^2 \right\}.$$

This completes the proof. \square

The following lemma verifies the inf-sup condition for \mathcal{B} .

Lemma 2. *There is a positive constant $\gamma_{\mathcal{B}}$ such that*

$$\sup_{\{u_1, u_2, \theta\} \in X} \frac{\mathcal{B}(\{u_1, u_2, \theta\}, \{\lambda_1, \lambda_2\})}{\|u_1\|_U + \|u_2\|_U + \|\theta\|_V} \geq \gamma_{\mathcal{B}} (\|\lambda_1\|_V + \|\lambda_2\|_V) \quad (11)$$

for all $\{\lambda_1, \lambda_2\} \in Y$.

Proof. Let $\{\lambda_1, \lambda_2\} \in Y$. The mapping

$$v \rightarrow (\lambda_i, v)_V, \quad i = 1, 2,$$

defines a continuous linear functional from V to \mathbb{R} . By assumption, the component forms $B_i(\cdot, \cdot)$ are weakly coercive and continuous. Therefore, for $i = 1, 2$, the variational equation: seek $u_i(\lambda_i) \in U$ such that

$$B_i(u_i(\lambda_i), v) = (\lambda_i, v)_V \quad \forall v \in V$$

has a unique solution $u_i(\lambda_i)$ and

$$\|u_i(\lambda_i)\|_U \leq \frac{1}{\gamma_{i1}} \|\lambda_i\|_V. \quad (12)$$

Setting $\{u_1, u_2, \theta\} = \{u_1(\lambda_1), u_2(\lambda_2), 0\}$ gives the identity

$$\begin{aligned} & \mathcal{B}(\{u_1(\lambda_1), u_2(\lambda_2), 0\}, \{\lambda_1, \lambda_2\}) \\ &= B_1(u_1(\lambda_1), \lambda_1) + B_2(u_2(\lambda_2), \lambda_2) = \|\lambda_1\|_V^2 + \|\lambda_2\|_V^2. \end{aligned}$$

The lemma follows from this identity and (12). \square

3.2. Equivalence of reformulated and original problems

From the abstract mixed theory of [20] it follows that the optimality system (9) has a unique solution $\{u_1, u_2, \theta\} \in X$, $\{\lambda_1, \lambda_2\} \in Y$, which depends continuously on the data. The component $\{u_1, u_2, \theta\}$ solves the constrained minimization problem (7). We establish equivalence of (2) and the optimization problem in two steps.

Proposition 1. *Let $u \in U$ denote a solution of (2). There exists a virtual control $\theta(u) \in V$ such that the triple $\{u, u, \theta(u)\}$ is a solution of (7).*

Proof. Let $u \in U$ be solution of (2). Owing to the continuity of component bilinear forms, $B_2(u, \cdot)$ is a continuous linear functional on V and the problem: seek $\theta \in V$ such that

$$(\theta, v)_V = -B_2(u, v) \quad \forall v \in V$$

has a unique solution $\theta(u)$. It follows that

$$B_2(u, v) + (\theta(u), v)_V = 0 \quad \forall v \in V,$$

that is the pair $\{u, \theta(u)\}$ satisfies the second constraint in (7). On the other hand, the identity $(f, v)_H = B(u, v) = B_1(u, v) + B_2(u, v)$ implies that

$$B_1(u, v) - (\theta(u), v)_V = B_1(u, v) + B_2(u, v) = (f, v)_H \quad \forall v \in V,$$

that is $\{u, \theta(u)\}$ also satisfies the first constraint. Therefore, the triple $\{u, u, \theta(u)\}$ is feasible. It is also a minimizer, as $J(u, u) = 0$. \square

Theorem 1. *Assume that $u \in U$ solves (2) and $\{u_1, u_2, \theta\}$ is a solution of (7). Then $u_1 = u_2 = u$.*

Proof. From Proposition 1 we know that there exist a virtual control $\theta(u)$ such that $\{u, u, \theta(u)\}$ is a solution of the optimization problem. Because the solution of (7) is unique it follows that $u_1 = u_2 = u$. \square

4. Approximation of the optimization problem

This section studies the discretization of the optimization reformulation problem (7). Assuming that weak coercivity (6) holds for the subproblems defining the split (5) we show that a stable discretization of the abstract problem (2) induces a stable discretization of (7). This implies well-posedness of the discrete optimization problem without a control penalty in the objective, optimal error estimates, and equivalence of the discrete solutions of (2) and (7).

Suppose that $U^h \subset U$ and $V^h \subset V$ form a stable pair for (2), i.e., there exists a positive constant γ_{01}^h , independent of h and such that²

$$\sup_{v^h \in V^h} \frac{B(u^h, v^h)}{\|v^h\|_V} \geq \gamma_{01}^h \|u^h\|_U \quad \text{and} \quad \sup_{u^h \in U^h} \frac{B(u^h, v^h)}{\|u^h\|_U} > 0. \quad (13)$$

A restriction of (2) to $\{U^h, V^h\}$ defines an approximation of the monolithic problem given by: seek $u^h \in U^h$ such that

$$B(u^h, v^h) = (f, v^h)_H \quad \forall v^h \in V^h. \quad (14)$$

Discrete weak coercivity conditions (13) imply that (14) is a well-posed problem.

²Continuity of $B(\cdot, \cdot)$ is inherited on any conforming approximations of U and V , i.e., $B(u^h, v^h) \leq \gamma_{02} \|u^h\|_U \|v^h\|_V$ for all $u^h \in U^h$ and $v^h \in V^h$.

A restriction of (7) to $\{U^h, V^h\}$ defines an approximation of this optimization problem given by

$$\left\{ \begin{array}{l} \min_{u_1^h, u_2^h \in U^h; \theta^h \in V^h} J(u_1^h, u_2^h) = \frac{1}{2} \|u_1^h - u_2^h\|_H^2 \\ \text{subject to } \left\{ \begin{array}{ll} B_1(u_1^h, v_1^h) - (\theta^h, v_1^h)_V = (f, v_1^h)_H & \forall v_1^h \in V^h \\ B_2(u_2^h, v_2^h) + (\theta^h, v_2^h)_V = 0 & \forall v_2^h \in V^h, \end{array} \right. \end{array} \right. \quad (15)$$

while a restriction of (8) to the spaces

$$X^h = U^h \times U^h \times V^h \subset X \quad \text{and} \quad Y^h = V^h \times V^h \subset Y, \quad (16)$$

defines a discrete Lagrangian functional

$$\begin{aligned} L(\{u_1^h, u_2^h, \theta^h\}, \{\lambda_1^h, \lambda_2^h\}) &= J(u_1^h, u_2^h) + \\ & (B_1(u_1^h, \lambda_1^h) - (\theta^h, \lambda_1^h)_V - (f, \lambda_1^h)_H) + (B_2(u_2^h, \lambda_2^h) + (\theta^h, \lambda_2^h)_V), \end{aligned} \quad (17)$$

associated with (15). Finally, it is straightforward to check that a restriction of the continuous optimality system (9) to $\{X^h, Y^h\}$, i.e., the discrete variational problem: *seek* $\{u_1^h, u_2^h, \theta^h\} \in X^h$ and $\{\lambda_1^h, \lambda_2^h\} \in Y^h$ such that

$$\begin{aligned} \mathcal{A}(\{u_1^h, u_2^h, \theta^h\}, \{\hat{u}_1^h, \hat{u}_2^h, \hat{\theta}^h\}) + \mathcal{B}(\{\hat{u}_1^h, \hat{u}_2^h, \hat{\theta}^h\}, \{\lambda_1^h, \lambda_2^h\}) &= 0 \\ \mathcal{B}(\{u_1^h, u_2^h, \theta^h\}, \{\hat{\lambda}_1^h, \hat{\lambda}_2^h\}) &= (f, \hat{\lambda}_1^h)_H \end{aligned} \quad (18)$$

for all $\{\hat{u}_1^h, \hat{u}_2^h, \hat{\theta}^h\} \in X^h$ and $\{\hat{\lambda}_1^h, \hat{\lambda}_2^h\} \in Y^h$, gives the necessary optimality conditions for the stationary points of the discrete Lagrangian (17).

The following lemma shows that (15), resp. (18), are well-posed by verifying that (16) is a stable pair for the continuous optimality system (9).

Lemma 3. *Assume that (6) holds for $B_1(\cdot, \cdot)$ and $B_2(\cdot, \cdot)$, $\{U^h, V^h\} \subset \{U, V\}$ is a stable pair for the composite form $B(\cdot, \cdot)$, and let*

$$Z^h = \{\{u_1^h, u_2^h, \theta^h\} \in X^h \mid \mathcal{B}(\{u_1^h, u_2^h, \theta^h\}, \{\lambda_1^h, \lambda_2^h\}) = 0 \quad \forall \{\lambda_1^h, \lambda_2^h\} \in Y^h\}.$$

Then, (16) defines a stable pair the pair $\{X^h, Y^h\}$ for (9): there exist positive constants $\gamma_{\mathcal{A}}^h$ and $\gamma_{\mathcal{B}}^h$, independent of h and such that

$$\gamma_{\mathcal{A}}^h (\|u_1^h\|_U^2 + \|u_2^h\|_U^2 + \|\theta^h\|_V^2) \leq \mathcal{A}(\{u_1^h, u_2^h, \theta^h\}, \{u_1^h, u_2^h, \theta^h\}) \quad (19)$$

for all $\{u_1^h, u_2^h, \theta^h\} \in Z^h$, and

$$\sup_{\{u_1^h, u_2^h, \theta^h\} \in X^h} \frac{\mathcal{B}(\{u_1^h, u_2^h, \theta^h\}, \{\lambda_1^h, \lambda_2^h\})}{\|u_1^h\|_U + \|u_2^h\|_U + \|\theta^h\|_V} \geq \gamma_{\mathcal{B}}^h (\|\lambda_1^h\|_V + \|\lambda_2^h\|_V) \quad (20)$$

for all $\{\lambda_1^h, \lambda_2^h\} \in Y^h$.

Proof. Assumption (6) implies that $B_1(\cdot, \cdot)$ and $B_2(\cdot, \cdot)$ are well-posed on the same pair of spaces as their parent form $B(\cdot, \cdot)$. As a result, any pair $\{U^h, V^h\} \subset \{U, V\}$ that is stable for the latter is also stable for its additive components. Thus, from (13) it follows the existence of positive constants γ_{i1}^h ; $i = 1, 2$, independent of h and such that

$$\sup_{v^h \in V^h} \frac{B_i(u^h, v^h)}{\|v^h\|_V} \geq \gamma_{i1}^h \|u^h\|_U \quad \text{and} \quad \sup_{u^h \in U^h} \frac{B_i(u^h, v^h)}{\|u^h\|_U} > 0, \quad (21)$$

respectively. The rest of the proof follows the steps in Lemmata 1–2. \square

Lemma 3 confirms that the discrete optimization problem (15), resp. the discrete optimality system (18), remain well-posed without a control penalty term in the cost functional. In particular, this lemma reveals that the well-posedness of the discrete subproblems defining the constraints is sufficient to ensure the well-posedness of (15), resp. (18). The following theorem provides information about the approximation error of (18).

Theorem 2. *Assume that (6) holds for $B_1(\cdot, \cdot)$ and $B_2(\cdot, \cdot)$, $\{U^h, V^h\} \subset \{U, V\}$ is a stable pair for the composite form $B(\cdot, \cdot)$ and $\{X^h, Y^h\}$ is the pair defined in (16). Let $\{u_1, u_2, \theta\} \in X$, $\{\lambda_1, \lambda_2\} \in Y$ denote a solution of (9). The solution $\{u_1^h, u_2^h, \theta^h\} \in X^h$, $\{\lambda_1^h, \lambda_2^h\} \in Y^h$ of (18) satisfies the optimal error estimate*

$$\begin{aligned} & \sum_{i=1}^2 \|u_i - u_i^h\|_U + \|\lambda_i - \lambda_i^h\|_V \\ & \leq C \left(\inf_{\hat{u}_1^h, \hat{u}_2^h \in U^h} \sum_{i=1}^2 \|u_i - \hat{u}_i^h\|_U + \inf_{\hat{\lambda}_1^h, \hat{\lambda}_2^h \in V^h} \sum_{i=1}^2 \|\lambda_i - \hat{\lambda}_i^h\|_V \right). \end{aligned} \quad (22)$$

The theorem follows directly from the abstract approximation result in [20]. \square

We conclude this section with an analogue of Theorem 1, which asserts that the discrete optimization problem (15) is an equivalent reformulation of the discrete monolithic problem (14).

Theorem 3. *Assume that (6) holds for $B_1(\cdot, \cdot)$ and $B_2(\cdot, \cdot)$, $\{U^h, V^h\} \subset \{U, V\}$ is a stable pair for the composite form $B(\cdot, \cdot)$. Assume that $u^h \in U^h$ solves (14) and $\{u_1^h, u_2^h, \theta^h\}$ is solution of (15). Then $u_1^h = u_2^h = u^h$.*

Proof. The proof follows verbatim the proof of Theorem 1. \square

5. Solution of the discrete optimization problem

Let $\{u_i^h\}_{i=1}^N$ and $\{v_i^h\}_{i=1}^N$ denote basis sets for U^h and V^h , respectively, and \vec{u}_i , $\vec{\theta}$, and $\vec{\lambda}_i$ be the coefficient vectors of u_i^h , θ^h and λ_i^h relative to these basis sets. Then, it is easy to see that the discrete optimality system (18) is equivalent

to a block linear system of algebraic equations

$$\begin{bmatrix} \mathbf{U} & -\mathbf{U} & \mathbf{0} & \mathbf{B}_1^T & \mathbf{0} \\ -\mathbf{U} & \mathbf{U} & \mathbf{0} & \mathbf{0} & \mathbf{B}_2^T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{V} & \mathbf{V} \\ \mathbf{B}_1 & \mathbf{0} & -\mathbf{V} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{V} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \vec{u}_1 \\ \vec{u}_2 \\ \vec{\theta} \\ \vec{\lambda}_1 \\ \vec{\lambda}_2 \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \vec{0} \\ \vec{0} \\ \vec{f} \\ \vec{0} \end{bmatrix} \quad (23)$$

with blocks given by

$\mathbf{U}_{rs} = (u_s^h, u_r^h)_U$, $\mathbf{B}_{i,rs} = B_i(u_s^h, v_r^h)$, and $\mathbf{V}_{rs} = (v_s^h, v_r^h)_V$, $r, s = 1, \dots, N$, respectively. The linear system (23) gives the Karush–Kuhn–Tucker (KKT) necessary optimality conditions for the discrete optimization problem (15). The discrete monolithic problem (14) is similarly equivalent to a block linear system

$$\mathbf{B}\vec{u} = (\mathbf{B}_1 + \mathbf{B}_2)\vec{u} = \vec{f} \quad (24)$$

where \vec{u} is the coefficient vector of the solution to (14).

Recall the assumption that the component forms $B_i(\cdot, \cdot)$ are “easier” to solve than the original problem involving $B(\cdot, \cdot)$. In the context of (24) this assumption implies that the blocks \mathbf{B}_1 , \mathbf{B}_2 are easier to invert than their sum $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$, i.e., that there are robust and efficient solvers for systems involving these matrices. We will take advantage of this fact and the structure of the KKT system (23), which exposes these blocks to synthesize a robust and efficient solver for (24) from the available solvers for \mathbf{B}_1 and \mathbf{B}_2 .

The KKT system (23) is similar to the one studied in [1, 2], with one subtle difference. Namely, the analysis carried out in the previous sections allows us to bypass control penalty, i.e., we may set the (3,3) block in [1, (3.17)] to zero, directly yielding (23). As the next theorem shows, this has important consequences for the design of iterative solvers for (24), and reveals a simpler approach to additive operator decomposition based on auxiliary variables.

Theorem 4. *Let $(\vec{u}_1^*, \vec{u}_2^*, \vec{\theta}^*, \vec{\lambda}_1^*, \vec{\lambda}_2^*)$ be the solution of the KKT system (23) and \vec{u}^* be the solution of (24). Then, $\vec{\lambda}_1^* = \vec{\lambda}_2^* = \vec{0}$, $\vec{u}_1^* = \vec{u}_2^* = \vec{u}^*$ and $(\vec{u}_1^*, \vec{u}_2^*, \vec{\theta}^*)$ solve the reduced KKT system*

$$\begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & -\mathbf{V} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{V} \\ \mathbf{0} & \mathbf{0} & (\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})\mathbf{V} \end{bmatrix} \begin{bmatrix} \vec{u}_1 \\ \vec{u}_2 \\ \vec{\theta} \end{bmatrix} = \begin{bmatrix} \vec{f} \\ \vec{0} \\ -\mathbf{B}_1^{-1}\vec{f} \end{bmatrix}. \quad (25)$$

Proof. Adding the first two rows of (23) and rearranging the matrix transforms (23) into the following 2×2 block upper triangular system of equations:

$$\begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & -\mathbf{V} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{V} & \mathbf{0} & \mathbf{0} \\ \mathbf{U} & -\mathbf{U} & \mathbf{0} & \mathbf{B}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{V} & \mathbf{V} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_1^T & \mathbf{B}_2^T \end{bmatrix} \begin{bmatrix} \vec{u}_1 \\ \vec{u}_2 \\ \vec{\theta} \\ \vec{\lambda}_1 \\ \vec{\lambda}_2 \end{bmatrix} = \begin{bmatrix} \vec{f} \\ \vec{0} \\ \vec{0} \\ \vec{0} \\ \vec{0} \end{bmatrix}. \quad (26)$$

The (2,2) block in (26) further reduces to

$$\begin{bmatrix} -\mathbf{V} & \mathbf{V} \\ \vec{\theta} & (\mathbf{B}_1 + \mathbf{B}_2)^T \end{bmatrix} \begin{bmatrix} \vec{\lambda}_1 \\ \vec{\lambda}_2 \end{bmatrix} = \begin{bmatrix} \vec{\theta} \\ \vec{\theta} \end{bmatrix}. \quad (27)$$

By assumption $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$ is nonsingular and so it follows that $\vec{\lambda}_1^* = \vec{\lambda}_2^* = \vec{\theta}$. As a result, we can neglect the (1,2) block in (26) to obtain a 3×3 block system for the states and the control only:

$$\begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & -\mathbf{V} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{V} \\ \mathbf{U} & -\mathbf{U} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \vec{u}_1 \\ \vec{u}_2 \\ \vec{\theta} \end{bmatrix} = \begin{bmatrix} \vec{f} \\ \vec{\theta} \\ \vec{\theta} \end{bmatrix}. \quad (28)$$

Performing two steps of block Gaussian elimination to remove blocks (3,1) and (3,2) in (28) results in the reduced KKT system (25).

To complete the proof it remains to verify that $\vec{u}_1^* = \vec{u}_2^* = \vec{u}^*$. Solving (28) by back substitution we find that

$$\vec{\theta}^* = -\mathbf{V}^{-1}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1}\mathbf{B}_1^{-1}\vec{f} \quad (29)$$

$$\vec{u}_1^* = \mathbf{B}_1^{-1}(\vec{f} + \mathbf{V}\vec{\theta}^*) \quad (30)$$

$$\vec{u}_2^* = -\mathbf{B}_2^{-1}\mathbf{V}\vec{\theta}^*. \quad (31)$$

It is easy to check that

$$\begin{aligned} (\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1} &= (\mathbf{B}_1^{-1}\mathbf{B}_2\mathbf{B}_2^{-1} + \mathbf{B}_1^{-1}\mathbf{B}_1\mathbf{B}_2^{-1})^{-1} \\ &= (\mathbf{B}_1^{-1}(\mathbf{B}_2 + \mathbf{B}_1)\mathbf{B}_2^{-1})^{-1} = \mathbf{B}_2(\mathbf{B}_1 + \mathbf{B}_2)^{-1}\mathbf{B}_1, \end{aligned}$$

or, equivalently,

$$(\mathbf{B}_1 + \mathbf{B}_2)^{-1} = \mathbf{B}_2^{-1}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1}\mathbf{B}_1^{-1}. \quad (32)$$

Using (32) and (29) we find that

$$\vec{u}_2^* = -\mathbf{B}_2^{-1}\mathbf{V}\vec{\theta}^* = \mathbf{B}_2^{-1}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1}\mathbf{B}_1^{-1}\vec{f} = \mathbf{B}_1^{-1}\vec{f} = \vec{u}^*.$$

Using again (32) and (29) together with $\vec{f} = \mathbf{B}\vec{u}$, and $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$ gives

$$\begin{aligned} \vec{u}_1^* &= \mathbf{B}_1^{-1}(\vec{f} + \mathbf{V}\vec{\theta}^*) \\ &= \mathbf{B}_1^{-1}\vec{f} - \mathbf{B}_1^{-1}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1}\mathbf{B}_1^{-1}\vec{f} \\ &= \mathbf{B}_1^{-1}\mathbf{B}\vec{u}^* - \mathbf{B}_1^{-1}\mathbf{B}_2(\mathbf{B}_2^{-1}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})^{-1}\mathbf{B}_1^{-1})\mathbf{B}\vec{u}^* \\ &= \mathbf{B}_1^{-1}(\mathbf{B}_1 + \mathbf{B}_2)\vec{u}^* - \mathbf{B}_1^{-1}\mathbf{B}_2\mathbf{B}_2^{-1}\mathbf{B}\vec{u}^* = \vec{u}^*. \end{aligned}$$

This completes the proof. \square .

The back substitution sequence (29)-(31) in Theorem 4 defines an iterative procedure for the solution of the KKT system (23). At the same time, the

related equation (32) reveals a more direct approach to deriving the same decomposition. Specifically, assuming that \mathbf{V} is the discrete identity operator, the splitting scheme defined by (32) is equivalent to the scheme defined by (29)-(31).

Furthermore, due to $\vec{\mathbf{u}}_1^* = \vec{\mathbf{u}}_2^* = \vec{\mathbf{u}}^*$, the system (28) can be reduced to the system

$$\begin{bmatrix} \mathbf{B}_1 & -\mathbf{V} \\ \mathbf{B}_2 & \mathbf{V} \end{bmatrix} \begin{bmatrix} \vec{\mathbf{u}} \\ \vec{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{f}} \\ \vec{\boldsymbol{\theta}} \end{bmatrix}, \quad (33)$$

illustrating a different application of auxiliary variables in deriving the additive decomposition, which bypasses the optimization formulation. We should note, however, that the derivation of the system (33) is not obvious at first. In contrast, the optimization framework automates and formalizes the discovery of operator decompositions. Additionally, optimization formulations may remain well posed even if \mathbf{B}_1 or \mathbf{B}_2 are singular.

We use the iterative procedure defined by the back substitution (29)-(31) for all numerical experiments in Section 6.4.

Remark 2. *In [1, p. 3947] we developed a solution procedure for the KKT system based on a reduced formulation in terms of the control vector $\vec{\boldsymbol{\theta}}$. In that context (using the current notation) the optimal control is given as the solution of the linear system*

$$\left(\mathbf{V}(\mathbf{B}_1^{-T} + \mathbf{B}_2^{-T})\mathbf{U}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})\mathbf{V} + \epsilon\mathbf{R} \right) \vec{\boldsymbol{\theta}} = -\mathbf{V}(\mathbf{B}_1^{-T} + \mathbf{B}_2^{-T})\mathbf{U}\mathbf{B}_1^{-1}\vec{\mathbf{f}},$$

where \mathbf{R} is the discretization of an inner product operator defined on the control space. In light of the new analysis, we can eliminate the term $\epsilon\mathbf{R}$, and obtain the equation

$$\mathbf{V}(\mathbf{B}_1^{-T} + \mathbf{B}_2^{-T})\mathbf{U}(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})\mathbf{V}\vec{\boldsymbol{\theta}} = -\mathbf{V}(\mathbf{B}_1^{-T} + \mathbf{B}_2^{-T})\mathbf{U}\mathbf{B}_1^{-1}\vec{\mathbf{f}},$$

which, after multiplication by $\mathbf{U}^{-1}(\mathbf{B}_1^{-T} + \mathbf{B}_2^{-T})^{-1}\mathbf{V}^{-1}$, reduces to

$$(\mathbf{B}_1^{-1} + \mathbf{B}_2^{-1})\mathbf{V}\vec{\boldsymbol{\theta}} = -\mathbf{B}_1^{-1}\vec{\mathbf{f}}.$$

Thus, in the absence of a control penalty and assuming \mathbf{U} is the identity, the approach in [1] amounts to solving the normal equations of (29).

6. Application to the Oseen equations

In this section we apply the optimization-based decomposition framework to the Oseen equations given by

$$\begin{cases} -\nu\Delta\mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega, \end{cases} \quad (34)$$

where ν is the kinematic viscosity, \mathbf{b} is a given advective vector, \mathbf{u} is the velocity, p is the pressure and \mathbf{f} is the body force. For simplicity, we complete the specification of the governing equations by augmenting (34) with the homogeneous velocity boundary condition

$$\mathbf{u} = 0 \quad \text{on } \Gamma$$

and the zero mean pressure constraint

$$\int_{\Omega} p dx = 0.$$

Our principal objective is to demonstrate the potential of the optimization-based decomposition approach for the design of robust and efficient iterative solvers for incompressible flows. Linearization of the Navier-Stokes equations by a fixed point method or a Newton-type method results in problems that are either identical or similar in structure to (34). Consequently, availability of robust solvers with optimal complexity for (34) is a prerequisite towards the robust and efficient solution of the nonlinear Navier-Stokes equations. The rate of convergence of these solvers should be independent of the mesh size and depend only mildly if at all on the kinematic viscosity ν .

Yet, formulation of such solvers remains a challenge, especially for small viscosity values; see e.g., [21, 22, 23] and the references therein. Numerical studies in these papers cover viscosities ranging from 0.1 to 0.001 and show deterioration of convergence rates at the lower end of this interval and as the mesh size is reduced.

Interestingly enough, some of the methods in these papers also utilize additive operator splittings but with two important distinctions. First, the splittings employed there follow either a physics-based decomposition of the equations, or a dimension-based decomposition along the coordinate directions. The Hermitian/skew-Hermitian splitting in [21] is an example of the first approach, which writes the Oseen operator as a sum of an elliptic vector Laplacian term and a hyperbolic convection term. The method in [23] writes the two dimensional Oseen operator as a sum of two scalar advection-diffusion equations acting in each one of the coordinate directions and is an example of the second kind of splittings. In either case, the design of the iterative algorithm is specifically tailored to the splitting employed. In contrast, our approach is agnostic to the kind of splitting used and only requires well-posedness of the subproblems. For instance, we could in principle apply the optimization-based decomposition in this paper with the same splittings as in [21, 23].

The second important distinction is in the type of iterative schemes resulting from the splittings. In [21, 23] the splittings are used in an alternating iteration scheme that switches back and forth between the subproblems. In contrast, our approach leads to an iteration process (29)-(31) involving an auxiliary variable (the virtual control), which is not equivalent to an alternating iteration scheme.

6.1. Variational formulation of the Oseen equations

We recall the space $L^2(\Omega)$ of all square integrable functions and its subspace $L_0^2(\Omega)$ of all square integrable functions with zero mean. We will also need the

Sobolev space $\mathbf{H}^1(\Omega)$ of all square integrable vector fields whose first derivatives are also square integrable, and its subspace $\mathbf{H}_0^1(\Omega)$ comprising all vector fields with zero trace on Γ .

To specialize the abstract framework in Section 3 to (34) we associate (2) with the weak variational form of the Oseen equations: *seek* $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ and $p \in L_0^2(\Omega)$ such that

$$\begin{cases} a(\mathbf{u}, \mathbf{v}) + b(p, \mathbf{v}) = (f, \mathbf{v}) & \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega) \\ -b(q, \mathbf{u}) = 0 & \forall q \in L_0^2(\Omega), \end{cases} \quad (35)$$

where

$$a(\mathbf{u}, \mathbf{v}) = \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + (\mathbf{b} \cdot \nabla \mathbf{u}, \mathbf{v}) \quad \text{and} \quad b(q, \mathbf{v}) = -(q, \nabla \cdot \mathbf{v}). \quad (36)$$

With the identifications $U = V = \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$, $H = L^2(\Omega)$, and

$$B(\mathbf{u}, p; \mathbf{v}, q) = a(\mathbf{u}, \mathbf{v}) + b(p, \mathbf{v}) - b(q, \mathbf{u}), \quad (37)$$

problem (35) assumes the abstract form (2). It is a standard result that (35) is well-posed [24] and that $B(\cdot, \cdot)$ satisfies the weak coercivity conditions (3), i.e., it fulfills the requirements of Section 2.

6.2. Optimization-based additive decomposition of the Oseen equations

Recall that the ability to decompose $B(\cdot, \cdot)$ into a sum of two subproblems that are “easier to solve” is a key assumption of the abstract framework. In the context of the Oseen equations this means that the subproblems should have better ratios between the viscosity and the velocity than the original problem. Such a split can be accomplished in many ways, including the decompositions described in [21, 23]. Here we choose to effect the splitting by adding and subtracting a Stokes operator with viscosity $\sigma > \nu$ to (37). In other words, we write

$$B(\mathbf{u}, p; \mathbf{v}, q) = B_1(\mathbf{u}, p; \mathbf{v}, q) + B_2(\mathbf{u}, p; \mathbf{v}, q) \quad (38)$$

where

$$B_1(\mathbf{u}, p; \mathbf{v}, q) = \sigma(\nabla \mathbf{u}, \nabla \mathbf{v}) + (\mathbf{b} \cdot \nabla \mathbf{u}, \mathbf{v}) - 2(p, \nabla \cdot \mathbf{v}) + 2(q, \nabla \cdot \mathbf{u})$$

is an Oseen operator with a larger viscosity coefficient and

$$B_2(\mathbf{u}, p; \mathbf{v}, q) = (\nu - \sigma)(\nabla \mathbf{u}, \nabla \mathbf{v}) + (p, \nabla \cdot \mathbf{v}) - (q, \nabla \cdot \mathbf{u}),$$

is a Stokes operator, respectively. Again, it is a standard result that both $B_1(\cdot, \cdot)$ and $B_2(\cdot, \cdot)$ are continuous and weakly coercive on $U = \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$, i.e., (6) holds for these forms. As a result, B_1 and B_2 fulfill the requirements of the abstract framework.

Remark 3. *To shed some more light on (38) it is instructive to examine the strong form of the splitting. It is easy to see that (38) corresponds to writing (34) as the following sum:*

$$\begin{bmatrix} -\sigma\Delta\mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + 2\nabla p \\ 2\nabla \cdot \mathbf{u} \end{bmatrix} + \begin{bmatrix} (\sigma - \nu)\Delta\mathbf{u} - \nabla p \\ -\nabla \cdot \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}. \quad (39)$$

By choosing a sufficiently large splitting parameter σ one can ensure that both subproblems are dominated by the Laplacian operator, which makes both of them easier to solve than the original equation. Numerical results in Section 6.4 reveal that optimization-based decomposition only mildly depends on the size of this parameter.

To complete the specialization of (7) to the Oseen equations we introduce the objective

$$J(\mathbf{u}_1, p_1; \mathbf{u}_2, p_2) = \frac{1}{2} (\|\mathbf{u}_1 - \mathbf{u}_2\|_1^2 + \|p_1 - p_2\|_0^2)$$

and the virtual controls $\theta = \{\boldsymbol{\xi}, r\} \in U = \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. With these specifications, (7) assumes the form

$$\left\{ \begin{array}{l} \text{minimize } J(\mathbf{u}_1, p_1; \mathbf{u}_2, p_2) \text{ subject to} \\ B_1(\mathbf{u}_1, p_1; \mathbf{v}_1, q_1) - (\boldsymbol{\xi}, \mathbf{v}_1)_1 - (r, q_1)_0 = (f, \mathbf{v}_1)_0 \quad \forall \{\mathbf{v}_1, q_1\} \in U \\ B_2(\mathbf{u}_2, p_2; \mathbf{v}_2, q_2) + (\boldsymbol{\xi}, \mathbf{v}_2)_1 + (r, q_2)_0 = 0 \quad \forall \{\mathbf{v}_2, q_2\} \in U. \end{array} \right. \quad (40)$$

Remark 4. *It is instructive to examine the alternative application of the auxiliary variables, which bypasses the optimization formulation to arrive directly at equations (33). In the present context, these equations can be derived by first writing (34) as the sum*

$$\begin{bmatrix} -\sigma\Delta\mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + 2\nabla p - \boldsymbol{\xi} \\ 2\nabla \cdot \mathbf{u} - r \end{bmatrix} + \begin{bmatrix} (\sigma - \nu)\Delta\mathbf{u} - \nabla p + \boldsymbol{\xi} \\ -\nabla \cdot \mathbf{u} + r \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}. \quad (41)$$

followed by formally separating the two equations and assigning the right hand side to the first operator. However, as pointed out, this derivation is not immediately obvious and depends on serendipity to discover the auxiliary variables.

6.3. Discretization

An attractive computational property of our framework is that any stable discretization of the original equations induces a stable discretization of the associated optimality system. As a result, in order to obtain stable and accurate discretization of the optimization reformulation (40) it suffices to take any finite element subspace of $U = \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ that is stable for the Oseen equations. Moreover, for any such space the optimal approximation result in Theorem 2

carries over to the approximation of (40). For brevity we omit the specialization of these results as they are straightforward.

Let $K(\Omega)$ denote a uniformly regular partition of the computational domain into triangular finite elements κ . In this paper we use the Taylor-Hood finite element, which is a standard stable velocity-pressure pair for incompressible flows [25], to discretize (40). In this element the velocity is approximated by C^0 piecewise quadratic polynomials, whereas the pressure is approximated by C^0 piecewise linear polynomials. Although there are several modifications of the Taylor-Hood element, e.g., aiming to improve its mass conservation, their consideration is beyond the scope of this paper.

6.4. Numerical examples

We present four numerical studies, the first verifying formulation equivalence, the second demonstrating mesh independence, the third examining the sensitivity to the splitting parameter σ , and the fourth examining the dependence on the viscosity parameter ν .

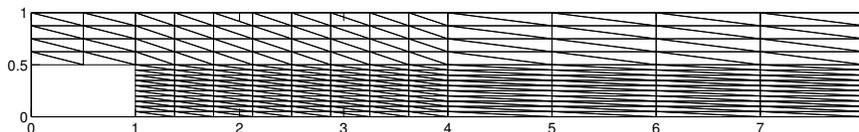


Figure 1: Computational mesh $K(\Omega)$; represents the ‘initial’ (Level-1) mesh.

The computational domain Ω is defined as the two-dimensional backward-facing step channel, a widely used benchmark. The geometry of the channel and the boundary conditions are described, e.g., in [26, p. 1767ff.]. In our setup, the advective vector \mathbf{b} is fixed, and given as the velocity solution of the Stokes equations. The magnitude of \mathbf{b} is independent of the viscosity ν , with the approximate value $\|\mathbf{b}\|_0 \approx 0.556$. As explained above, we discretize the Oseen equations and the two subproblems using the Taylor-Hood element.

We consider a sequence of computational meshes, based on a subdivision of the domain into rectangles, followed by a splitting of each rectangle into two triangles, as shown in Figure 1. We call the mesh depicted in Figure 1 the *Level-1 mesh*. Subsequent meshes are obtained via a uniform refinement of each rectangle in the Level-1 mesh into $n \times n$ rectangles, followed by a splitting into triangles. We will perform numerical studies on Level- n meshes with $n \in \{1, 2, 4, 8, 16\}$. Note that while the Level-1 mesh contains 352 triangles, the Level-16 mesh contains 90,112 triangles.

The first experiment verifies computationally the equivalence between the original problem and its optimization reformulation. In this experiment we compare finite element solutions of (35) and (40) computed on the Level-1 mesh. To solve equation (29) we use left-preconditioned GMRES with the preconditioner $\mathbf{V}^{-1}\mathbf{B}_1$, and recover the state variables by solving the equations (30) and (31). We note that the preconditioner is motivated by the form of equation (29). We

measure the ℓ_2 error in the obtained state variables, with respect to a direct solution of the Oseen equations, (35). For this experiment the GMRES stopping tolerance is set to 10^{-8} . The viscosity is set to $\nu = 5 \cdot 10^{-3}$, which is of interest because the flow separates (yet remains steady) and recirculation develops in a region around the point (1.3, 0.25). We set the splitting parameter to $\sigma = 1$. From Table 1 we observe that the ℓ_2 error approximately equals the GMRES stopping tolerance, for all solution components. Figure 2 and Figure 3 confirm this visually.

Solution component	ℓ_2 error
\mathbf{u}_1 horizontal	8.69e-07
\mathbf{u}_1 vertical	2.51e-06
p_1	6.21e-07
\mathbf{u}_2 horizontal	8.73e-07
\mathbf{u}_2 vertical	2.52e-06
p_2	8.29e-07

Table 1: The ℓ_2 error in the solution components of the optimization-based decomposition (40), with respect to the direct solution of (35). The error is roughly on the order of the GMRES stopping tolerance.

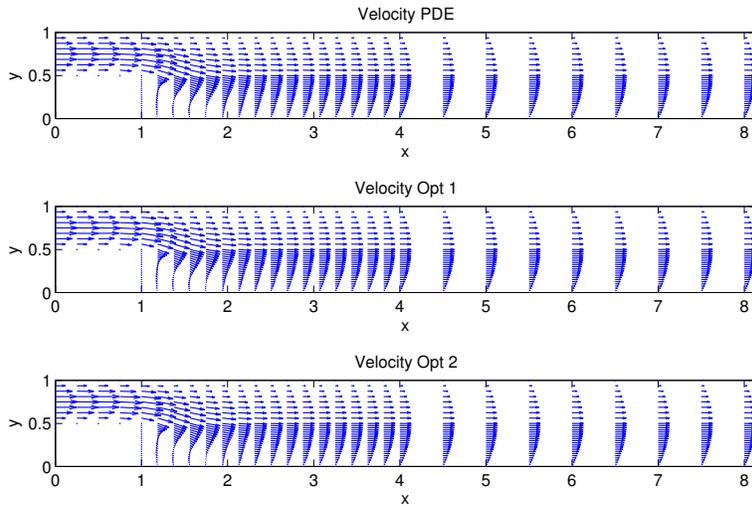


Figure 2: Velocity components of the solution of (35), labeled ‘PDE’, and the solution of (40), labeled ‘Opt 1’ and ‘Opt 2’.

The second experiment examines the mesh dependence of our optimization-

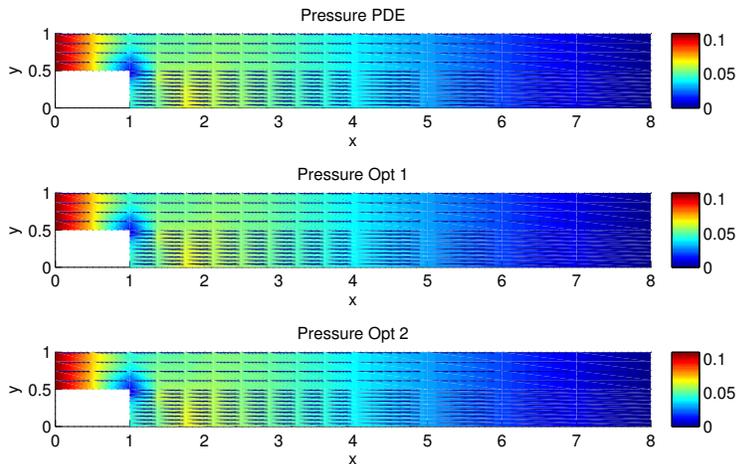


Figure 3: Pressure components of the solution of (35), labeled ‘PDE’, and the solution of (40), labeled ‘Opt 1’ and ‘Opt 2’.

based scheme. The viscosity is again set to $\nu = 5 \cdot 10^{-3}$. The GMRES stopping tolerance is set to 10^{-6} . We choose the splitting parameter to be $\sigma = 1$, resulting in the repeated solution of a Stokes problem and an Oseen problem with similar magnitudes of viscosity and advection. Table 2 clearly demonstrates that the performance of the iterative solver is independent of the mesh size. Additionally, the numbers of iterations required to solve the optimization problem are modest. We emphasize that each optimization iteration only involves the solution of “nice” Stokes and Oseen systems.

The third experiment tracks the performance of our algorithm for a wide range of splitting parameters σ . As before, the viscosity is set to $\nu = 5 \cdot 10^{-3}$, and the GMRES stopping tolerance is set to 10^{-6} . For this study we choose the Level-4 mesh. Table 3 shows that the performance of the optimization-based

Mesh Level	#Cells	#Degrees of Freedom	#Iterations
1	352	1,727	37
2	1,408	6,619	39
4	5,632	25,907	40
8	22,528	102,499	40
16	90,112	407,747	38

Table 2: The number of GMRES iterations (#Iterations) as the mesh is refined, for the viscosity $\nu = 5 \cdot 10^{-3}$ and the splitting parameter $\sigma = 1$. The performance of the optimization-based decomposition is independent of the mesh size. The total iteration numbers are modest.

Splitting Parameter σ	1e-2	1e-1	1	1e+1	1e+2	1e+3	1e+4
#Iterations	34	34	40	47	54	63	75

Table 3: The number of GMRES iterations (`#Iterations`) as the splitting parameter σ increases, for the viscosity $\nu = 5 \cdot 10^{-3}$ and the Level-4 mesh. The performance of the optimization-based decomposition is only mildly dependent on the size of the splitting parameter.

solver is only mildly dependent on the splitting parameter. The number of GMRES iterations merely doubles as σ increases six orders of magnitude. We note that for $\sigma = 10^4$ the Oseen system solved at every optimization iteration is dominated by an effective viscosity of 10^4 (recall the magnitude of the advection, ≈ 0.556). In other words, the optimization iteration comprises a Stokes solve and a near-Stokes solve.

Viscosity ν	1e+2	1e+1	1e-1	1e-2	5e-3
#Iterations	4	4	6	22	40

Table 4: The number of GMRES iterations (`#Iterations`) as the viscosity decreases, for the splitting parameter $\sigma = 1$ and the Level-4 mesh. The performance of the optimization-based decomposition is mildly dependent on the viscosity.

The fourth experiment studies our solver’s performance for a fixed splitting parameter, $\sigma = 1$, and viscosities ranging from $5 \cdot 10^{-3}$ to 100. For this study we use the Level-4 mesh and a GMRES stopping tolerance of 10^{-6} . Table 4 reveals a modest growth in GMRES iteration numbers as the viscosity decreases. Specifically, as ν decreases five orders of magnitude the iteration count goes up by a single order of magnitude. This experiment illustrates the “raw” performance of the optimization splitting without any attempts to optimize the parameter σ for each value of ν .

7. Conclusions

We formulated and analyzed an abstract optimization framework for the additive decomposition of weakly coercive variational problems. The purpose of the framework is to enable an efficient solution of the original equations by working with subproblems that are “easier” to solve.

A notable feature of the framework is the well-posedness of the reformulated problem without a control penalty term. It implies equivalence between the original and the reformulated problem, and enables a simplification of the attendant KKT optimality system to a form that lends itself to efficient iterative solvers.

An application of the framework to the Oseen equations, written as a sum of a Stokes problem and an Oseen problem with a better viscosity to velocity

ratio, results in iterative solvers that are mesh independent and have only mild dependence on the splitting parameter for a fixed viscosity value. Likewise, for a fixed splitting parameter the solvers exhibit a moderate dependence on the viscosity value. Tuning of the splitting parameter for each viscosity value is an interesting avenue of future research that may further improve the performance of the optimization-based solver.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research.

Key junctures of our approach have been influenced by the work and ideas of Max Gunzburger, who was among the first to explore the use of optimization and control for the numerical solution of PDEs. We dedicate this paper to the occasion of his 70th birthday.

References

- [1] P. Bochev, D. Ridzal, An optimization-based approach for the design of PDE solution algorithms, *SIAM Journal on Numerical Analysis* 47 (2009) 3938–3955.
- [2] P. Bochev, D. Ridzal, Additive operator decomposition and optimization-based reconnection with applications, in: I. Lirkov, S. Margenov, J. Wasniewski (Eds.), *Proceedings of LSSC 2009*, volume 5910 of *Springer Lecture Notes in Computer Science*.
- [3] M. D. Gunzburger, M. Heinkenschloss, H. K. Lee, Solution of elliptic partial differential equations by an optimization-based domain decomposition method, *Applied Mathematics and Computation* 113 (2000) 111 – 139.
- [4] M. D. Gunzburger, J. S. Peterson, H. Kwon, An optimization based domain decomposition method for partial differential equations, *Computers & Mathematics with Applications* 37 (1999) 77 – 93.
- [5] M. D. Gunzburger, H. K. Lee, An optimization-based domain decomposition method for the Navier-Stokes equations, *SIAM Journal on Numerical Analysis* 37 (2000) pp. 1455–1480.
- [6] Q. Du, M. D. Gunzburger, A gradient method approach to optimization-based multidisciplinary simulations and nonoverlapping domain decomposition algorithms, *SIAM Journal on Numerical Analysis* 37 (2000) pp. 1513–1541.
- [7] P. Gervasio, J.-L. Lions, A. Quarteroni, Heterogeneous coupling by virtual control methods, *Numerische Mathematik* 90 (2001) 241–264. 10.1007/s002110100303.

- [8] M. Discacciati, P. Gervasio, A. Quarteroni, Interface control domain decomposition methods for heterogeneous problems, *International Journal for Numerical Methods in Fluids* 76 (2014) 471–496.
- [9] M. Discacciati, P. Gervasio, A. Quarteroni, The interface control domain decomposition (icdd) method for elliptic problems, *SIAM Journal on Control and Optimization* 51 (2013) 3434–3458.
- [10] D. Olson, P. Bochev, M. Luskin, A. Shapeev, An optimization-based atomistic-to-continuum coupling method, *SIAM Journal on Numerical Analysis* 52 (2014) 2183–2204.
- [11] D. Olson, A. Shapeev, P. Bochev, M. Luskin, Analysis of an optimization-based atomistic-to-continuum coupling method for point defects., *Mathematical Modeling and Numerical Analysis (ESAIM) Accepted* (2015).
- [12] P. Kuberry, H. Lee, A decoupling algorithm for fluid-structure interaction problems based on optimization, *Computer Methods in Applied Mechanics and Engineering* (2013) –.
- [13] J. Lions, Virtual and effective control for distributed systems and decomposition of everything, *Journal d’Analyse Mathématique* 80 (2000) 257–297. 10.1007/BF02791538.
- [14] J.-L. Lions, O. Pironneau, Virtual control, replicas and decomposition of operators, *C. R. Acad. Sci. Paris* 330 (2000) 47–54.
- [15] J.-L. Lions, Decomposition of energy space and virtual control for parabolic systems, in: T. Chan, T. Kako, H. Kawarada, O. Pironneau (Eds.), *Proceedings of the 12th International Conference on Domain Decomposition Methods*, Chiba, Japan, pp. 41 – 53.
- [16] R. Glowinski, J.-L. Lions, O. Pironneau, Decomposition of energy spaces and applications, *Comptes Rendus de l’Académie des Sciences - Series I - Mathematics* 329 (1999) 445 – 452.
- [17] J. Wloka, *Partial differential equations*, Cambridge University Press, 1987.
- [18] A. Aziz (Ed.), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.
- [19] J. Nécas, *Les Methodes Directes en Theorie des Equations Elliptiques*, Masson, Paris, 1967.
- [20] F. Brezzi, On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers, *Model. Math. Anal. Numer.* 21 (1974).
- [21] S. Hamilton, M. Benzi, E. Haber, New multigrid smoothers for the Oseen problem, *Numerical Linear Algebra with Applications* 17 (2010) 557–576.

- [22] M. Benzi, M. A. Olshanskii, Z. Wang, Modified augmented lagrangian preconditioners for the incompressible navier–stokes equations, *International Journal for Numerical Methods in Fluids* 66 (2011) 486–508.
- [23] M. Benzi, X.-P. Guo, A dimensional split preconditioner for Stokes and linearized Navier-Stokes equations, *Applied Numerical Mathematics* 61 (2011) 66 – 76.
- [24] V. Girault, P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
- [25] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic, Boston, 1989.
- [26] L. S. Hou, S. S. Ravindran, Numerical approximation of optimal control problems by a penalty method: error estimates and numerical results, *SIAM Journal on Scientific Computing* 20 (1999) 1753–1777.